Bayesian Calibration of Force-fields from Experimental Data: TIP4P Water

Ritabrata Dutta\textsuperscript{1,*}, Zacharias Faidon Brotzakis\textsuperscript{1,2},
Antonietta Mira\textsuperscript{1,3}

\textsuperscript{1}Institute of Computational Science, Università della Svizzera italiana, Switzerland
\textsuperscript{2}Department of Chemistry and Applied Bioscience, ETH Zürich
\textsuperscript{3}Department of Science and High Technology, Università degli Studi dell’Insubria, Italy

April 10, 2018

Molecular dynamics (MD) simulations give access to equilibrium and/or dynamic properties given an ergodic sampling and a force-field accuracy. These force-fields are indexed by many parameters that need to be calibrated given available experimental observations. The main contribution of this paper is a rigorous Bayesian framework for the calibration of the force-field parameters and for their uncertainty quantification. To this aim, since the likelihood function of the MD simulation models are intractable, we use a likelihood-free inference scheme known as approximate Bayesian computation (ABC). The second contribution is the adaption of ABC algorithms for High Performance computing to MD simulation within the Python ecosystem ABCpy. This adaptation includes a novel dynamic allocation scheme for MPI that we prove to be particularly efficient in the context of MD simulation. We illustrate the performance of the developed methodology to learn posterior distribution and Bayesian estimates of Lennard-Jones force-field parameters of TIP4P system of water implemented both for simulated and experimental datasets collected using Neutron and X-ray diffraction. For simulated data, the Bayesian estimate is in close agreement with the true parameter value used to generate the dataset. For experimental as well as for simulated data, the Bayesian posterior distribution shows a strong correlation pattern between the force-field parameters. Providing an estimate of the entire posterior distribution, our methodology allows to perform uncertainty quantification of the estimated parameter values. This research opens up the possibility to rigorously calibrate force-fields from available experimental datasets.

Keywords: Approximate Bayesian Computation, ABCpy, Bayesian inference, Force-field parameters, High performance computing, MPI, TIP4P, Uncertainty quantification.

\textsuperscript{*}Corresponding author: duttar@usi.ch
1 Introduction

In the last decades, molecular simulations have become a cornerstone for computing equilibrium and/or dynamic properties of classical many body systems, as well for bridging microscopic with macroscopic observables that could be of both experimental and basic importances [Allen and Tildesley, 1989, Frenkel and Smit, 2001, Karplus and Lavery, 2014]. Given a force-field formalism, in this paper we assume that the phase space can be ergodically explored by evolving the Newton’s equation of motion under a molecular mechanics force-field of interactions. For this purpose, we consider Molecular Dynamics (MD) simulations, which samples the phase space by integrating the deterministic Newton’s equations of motion, hence giving access to both dynamic and thermodynamic properties. The accuracy of the underlying molecular mechanics force-field used to solve the equations of motion defines the approximation in the phase space exploration.

Each force-field reproduces specific properties and is indexed by a set of parameters, whose values are unknown (e.g., bonded and non-bonded force-field parameters). If the parameters of the force-field could be rigorously learned with an automated data-driven methodology, we could reparametrize the force-field for different target properties experimentally obtained (e.g., radial distribution functions, self-diffusion coefficient, density etc.). There exist different force-field formalisms, constraining ourself only to water we have TIP3P, TIP4P [Jorgensen et al., 1983], TIP4P/2005 [Abascal and Vega, 2005] and TIP5P [Mahoney and Jorgensen, 2000] among others. This raises the need for calibrations and force-field comparisons in a data-driven manner. Given an experimentally observed dataset, the Bayesian inferential framework can address both questions, namely calibrating the parameters of a given force-field formalism and, given many formalism, choosing one. This is achieved, correspondingly, through Bayesian parameter inference and model selection [Marin et al., 2012].

One of the main difficulty in applying Bayesian inference for these purposes is the intractability of the likelihood functions of the force-field parameters for an observed dataset. Though this can be resolved by approximate Bayesian computation (ABC) [Lintusaari et al., 2017], a likelihood-free inference schemes recently developed in the field of Statistical science.

One of the ABC algorithms, ABCsubsim [Chiachio et al., 2014] has been recently used to calibrate parameters of Lennard-Jones force-field of Helium in Kulakova et al., 2016]. As illustrated there, ABC can be used for this purpose only under an optimal exploitation of High performance computing facility (HPC). HPC and parallelization are essential in the context of MD due to the long simulation time and chaotic and unstable behavior of more realistic and challenging force-field formalisms. This need is addressed thanks to our recently developed Python eco-system called ABCpy [Dutta et al., 2017a], which implements and parallelizes most of the existing ABC algorithms and adapts them optimally to HPC infrastructure.

In this paper we only address the issue of calibrating the parameters given a force-field formalism. For this purpose we consider a force-field formalism of water, specifically the TIP4P developed by Jorgensen et al., 1983]. Water is universal in science and life. From protein folding [Brotzakis et al., 2016] to ion mobility [Ohtaki and Radnai, 1993] and from
enzymatic activity [Abel et al., 2008] to anti-freeze activity [Brotzakis et al., 2018], local and/or global water structure and dynamics regulates biological and physicochemical processes. Hence, force-field parametrization by targeting structural or dynamical properties of water is of great impact in biology and science. We illustrate the performance of the inferential scheme for simulated dataset, generated using GROMACS [Pronk et al., 2013], and for experimental datasets, obtained under ambient conditions using Neutron diffraction [Soper, 2000] and X-ray diffraction [Skinner et al., 2013]. To infer the non-bonded force-field parameters we use the adaptive population Monte Carlo ABC (APMCABC), which is one of the most suitable algorithms to optimally exploit the HPC architecture. Further, to mitigate the imbalance in simulation times of TIP4P model using GROMACS for different values of non-bonded force-field parameters, in this paper we use a new dynamic allocation scheme for MPI.

In Section 2 we explain the TIP4P force-field formalism of water and describe the model used to forward-simulate dataset using GROMACS. A short introduction to Bayesian inference and approximate Bayesian computation used for inferring the non-bonded force-field parameters is given in Section 3. Finally in Section 4 we infer the posterior distribution of the parameters and their Bayes estimates given a dataset, for a simulated and two experimentally obtained datasets.

## 2 TIP4P force-field of water

TIP4P force-field [Jorgensen et al., 1983] of water is a four interaction site water force-field with an extra charge placed on a dummy atom, besides the charges existing on the oxygen and hydrogens. It has been parametrized to reproduce the enthalpy of vaporization [Vega and Abascal, 2011]. The waters formalism is shown in the system of Eqs. 1-3

\[
U = U_{cow} + U_{nocow}
\]

\[
U_{cow} = \sum_{i,j} \frac{1}{2} k_a^{ij} (R_{ij} - R_{ij}^0)^2 + \sum_{i,j,k} \frac{1}{2} k_b^{ijk} (\theta_{ijk} - \theta_{ijk}^0)^2
\]

\[
U_{nocow} = \sum_i \sum_j 4\epsilon \left[ \left( \frac{\sigma}{r_{ij}} \right)^{12} - \left( \frac{\sigma}{r_{ij}} \right)^{6} \right] + \sum_i \sum_j \sum_{\alpha} \sum_{\beta} \frac{q_{i\alpha} q_{j\beta}}{r_{ij}}
\]

where the potential energy function \(U\) is the sum of the covalent (\(U_{cow}\)) and non-covalent (\(U_{nocow}\)) interactions. The covalent interactions/bonds are described harmonically. In Equation 2 \(R_{ij}\) and \(\theta_{ijk}\) are, correspondingly, the bond length between atoms \(i\) and \(j\), and the angle between atoms \(i\), \(j\), \(k\). The equilibrium values for the bonds and angles are \(R_{ij}^0\) and \(\theta_{ijk}^0\), with force constants \(k_a^{ij}\) and \(k_b^{ijk}\) respectively. The non covalent interactions in Eqn. 3 consist of the 6-12 Lennard-Jones and the electrostatic interactions. The Lennard-Jones interaction term includes the \(\sigma\) and \(\epsilon\) parameters, respectively, the distance
at which the interparticle potential is zero and the value of the minimum energy. The Pauli repulsion is parametrized with a term $\propto r^{-12}$ and the van der Waals attraction with a term $\propto r^{-6}$. The $r_{ij}$ terms correspond to the intermolecular distance. Finally, the electrostatic term in Eqn. 3 involves the charges of atoms, where $i$ and $j$ correspond to different water molecules and $\alpha$ and $\beta$ are the indices of the partial charges of each molecule.

For the purpose of the present study, the TIP4P model of force-field, which we denote by $\mathcal{M}$, is parametrized in terms of the two non-bonded force-field parameters $\sigma$ and $\epsilon$, which we jointly name

$$\phi = (\sigma, \epsilon),$$

representing the repulsion and attraction of the Van der Waals forces. Although the proposed model contains additional parameters, (e.g., the charges of Eq. 3), we chose not to infer them and assume they are constant. However, we stress that the ABC method is not bound by the number of parameters to infer, and our choice is based on the illustrative purpose of this paper. If the initial values for coordinates of water molecules in an enclosed space are known, we can forward simulate the coordinates of the molecules over time using model $\mathcal{M}$ for given values of these parameters $\phi = \phi^*$:

$$\mathcal{M}[\phi = \phi^*] \rightarrow \{\text{coordinates}(t) \}, \ t = 0, \ldots, T\}.$$  \hspace{1cm} (4)

where coordinates($t$) are the position of the molecules at time $t$ in an enclosed space. The forward simulation is performed using MD (TIP4P implementation in GROMACS) for $t_0 = 0$, $T = 100$ ps and time steps of $2$ fs. We first construct a simulation box of $2.5 \times 2.5 \times 2.5$ cm. Then, after compiling the TIP4P force-field with $\phi^*$, we perform an energy minimization, followed by an NPT simulation. In the energy minimization step, we use the steepest descend algorithm for energy minimization [Jaidhan et al., 2014], the LINCS algorithm [Hess et al., 1997] for the bond constraints and the Particle Mesh Ewald (PME) [Darden et al., 1993] with a cut-off of $1$ nm to treat the electrostatics. The cut-off for the Van der Waals interactions is $1$ nm. In the NPT simulation we set the temperature and the pressure at 298 K and 1 atm correspondingly, using the v-rescale thermostat [Bussi et al., 2006] and the Parrinello-Rahman barostat [Parrinello and Rahman, 1981] respectively. The neighbor list is updated every 10 time steps. The electrostatic and Van der Walls parameters are treated the same as in the energy minimization.

We use the term observed data, denoted by $x^0$, to refer to a dataset generated by some real-world process (e.g., experimental studies using X-ray and Neutron diffraction, that allow us to measure different properties of water molecules), and our goal is to learn values of the force-field parameters characterizing this process. Assuming that the first observation of the process occurs at time $t_0$ and the last observation at time $T$, the observed dataset is $x^0 \equiv \{\text{coordinates}(t), t_0, \ldots, T\}$. From experimental studies it is not possible to track the coordinates of water molecules, but we can learn their properties, e.g. different radial distribution functions, using different diffraction techniques.

In Jorgensen et al. [1983], the TIP4P water forcefield was parametrized for the physical properties: heat of vaporization, critical density, temperature and liquid density at 298K, using comparative trial and error method. We denote by $\phi^0 \equiv (\sigma^0, \epsilon^0) = (3.15e^{-1}, 6.48e^{-1})$ the value of the parameter which accurately reproduces the above-mentioned properties and
performs poorly in reproducing the radial distribution function and self-diffusion coefficient [Vega and Abascal, 2011] measured in different recent experimental studies [Soper, 2000; Skinner et al., 2013].

Here we address the question of calibrating force-fields for available structural and dynamical properties, such as the radial distribution function and self-diffusion coefficient. To this aim, we develop a rigorous Bayesian inference scheme that allows to quantify the uncertainty in the inferred model parameters, uncertainty which is inherent to the inferential process given the chaotic nature of the models described in Equations 4. We only calibrate force-fields targeting structural and dynamical properties, but we stress that this methodology can be used for any property which is available experimentally.

3 Inference framework

3.1 Bayesian inference

We can quantify the uncertainty of the inferred parameter $\phi = (\sigma, \epsilon)$ by its posterior distribution $p(\phi|x)$ given the observed dataset $x = x^0$. The posterior distribution is obtained by Bayes’ theorem as,

$$p(\phi|x^0) = \frac{\pi(\phi)p(x^0|\phi)}{m(x^0)},$$

where $\pi(\phi)$, $p(x^0|\phi)$ and $m(x^0) = \int \pi(\phi)p(x^0|\phi)d\phi$ are, correspondingly, the prior distribution on the parameter $\phi$, the likelihood function, and the marginal likelihood. The prior distribution $\pi(\phi)$ enables to incorporate, in the inferential process, prior knowledge on the parameter values. If the likelihood function could be evaluated, at least up to a normalizing constant, then the posterior distribution could be approximated by drawing a representative sample of parameter values from it using (Markov chain) Monte Carlo sampling schemes [Robert and Casella, 2005]. Unfortunately, the likelihood function induced by the TIP4P model of water is analytically intractable due to the complex nature of the model. In this setting, approximate Bayesian computation (ABC) [Lintusaari et al., 2017] offers a way to sample from the approximate posterior distribution and opens up the possibility of sound statistical inference on the parameter $\phi$. In this paper we only focus on parameter estimation/calibration and uncertainty quantification but we stress that ABC easily allows to also perform parameter hypothesis testing and model selection.

3.2 Approximate Bayesian computation (ABC)

Models that are easy to forward simulate, given values of the parameters, are called simulator-based models in the ABC literature and are used in a wide range of scientific disciplines to describe and understand different aspects of nature ranging from dynamics of sub-atomic particles [Martinez et al., 2016] to evolution of human societies [Turchin et al., 2013] and formation of universes [Schaye et al., 2013]. In the fundamental rejection ABC sampling scheme, we simulate a synthetic dataset $x^{\text{sim}}$ from the simulator-based model
\( \mathcal{M}(\phi) \) for a fixed parameter value, \( \phi \), and measure the closeness between \( \mathbf{x}^{\text{sim}} \) and \( \mathbf{x}^0 \) using a pre-defined discrepancy measure \( d(\mathbf{x}^{\text{sim}}, \mathbf{x}^0) \). Based on this discrepancy, ABC accepts the parameter value \( \phi \) when \( d(\mathbf{x}^{\text{sim}}, \mathbf{x}^0) \) is less than a pre-specified threshold value \( \epsilon \).

The intractable likelihood \( p(\mathbf{x}^0|\phi) \) is approximated by \( p_{d,\epsilon}(\mathbf{x}^0|\phi) \) for some choice of distance, \( d \), and threshold, \( \epsilon > 0 \), where

\[
p_{d,\epsilon}(\mathbf{x}^0|\phi) \propto P(d(\mathbf{x}^{\text{sim}}, \mathbf{x}^0) < \epsilon)
\]

and, as a consequence, the sampled parameters follow the posterior distribution of \( \phi \) conditional on \( d(\mathbf{x}^{\text{sim}}, \mathbf{x}^0) < \epsilon \):

\[
p_{d,\epsilon}(\phi|\mathbf{x}^0) \propto P(d(\mathbf{x}^{\text{sim}}, \mathbf{x}^0) < \epsilon)\pi(\phi).
\]

For a better approximation of the likelihood function, computationally efficient sequential ABC algorithms [Marin et al., 2012] decrease the value of the threshold \( \epsilon \) adaptively while exploring the parameter space. Here, we use adaptive population Monte Carlo ABC (APMCABC) [Lenormand et al., 2013] implemented in our ABCpy Python package [Dutta et al., 2017a] for optimal exploitation of an HPC environment. We note that the parallelization schemes in ABCpy were primarily meant for inferring parameters from models which take almost equal time to simulate dataset for any values of \( \phi \). Though, due to the chaotic nature of the MD systems, we observe that, for different values of \( \phi \), the simulation-time, for a fixed \( T \) value, is quite variable (from 10 minutes to an hour while using a single node with 36-core of Piz Daint Cray architecture (Intel Broadwell + NVidia TESLA P100)). To solve this imbalance, we use a new dynamic allocation scheme for MPI (see Subsection 3.4 and Dutta et al. [2017b]).

Using APMCABC, we approximate the posterior distribution \( p(\phi|\mathbf{x}^0) \) by drawing samples from it. All the tuning parameters for the APMCABC algorithm are fixed at the default values in the ABCpy package with the exception of the number of samples, the number of steps and the acceptance rate cut-off, which are set to 100, 10 and 0.03, respectively. Below we provide the summary statistics extracted from the dataset, the discrepancy measure, the parameters prior distribution, and the perturbation kernel used to explore the parameter space in the APMCABC algorithm.

One of the crucial aspects for a good ABC approximation is the choice of the summary statistics, as we define the discrepancy measure between \( \mathbf{x}^{\text{sim}} \) and \( \mathbf{x}^0 \) through a distance between summary statistics computed on \( \mathbf{x}^{\text{sim}} \) and \( \mathbf{x}^0 \). The summary statistics are usually chosen to minimize the loss of information on \( \phi \) contained in the data. Here we use intuitive discrepancy measures between interpretable and domain-driven summary statistics. Subjectivity of these decisions can be removed through automatic summary selection for ABC, described in Fearnhead and Prangle [2012], Pudlo et al. [2015], Jiang et al. [2015], Gutmann et al. [2018], where an informative linear or non-linear combination of the summaries is chosen.
Summary statistics

Given a dataset \( \mathbf{x} \equiv \{ \text{coordinates}(t) : t = t_0, \ldots, T \} \) for TIP4P model of water simulated using GROMACS, we compute an array of summary statistics,

\[
\mathcal{F} : \mathbf{x} \rightarrow (S_1, S_2, S_3, S_4, S_5, S_6, S_7, S_8)
\]
defined as follows:

- \( S_1 \): Estimate of the number of hydrogen bonds per water molecule - The area under the curve \( r_{OH} \) vs \( g_{OH} \) until the first minimum;
- \( S_2 \): Estimate of the donor acceptor hydrogen bond distance - Value of \( r_{OH} \) at the first minimum of the radial distribution function \( g_{OH} \);
- \( S_3 \): Mean of \( g_{OH} \);
- \( S_4 \): Estimate of number of water molecules in the first hydration shell - The area under the curve \( r_{OO} \) vs \( g_{OO} \) until the first minimum;
- \( S_5 \): Estimate of the maximum distance of the first hydration shell - Value of \( r_{OO} \) at the first minimum of the radial distribution function \( g_{OO} \);
- \( S_6 \): Mean of \( g_{OO} \);
- \( S_7 \): The height of \( g_{OO} \) at the first maximum of \( g_{OO} \);
- \( S_8 \): Slope of the line, fitted to mean square displacement (msd), which is an estimate of \( 6 \times \) self-diffusion coefficient.

To compute the above summary statistics, we first compute the radial distribution functions for the \( O-H \) and \( O-O \) atoms and the mean square displacement (msd) from the simulated coordinates of the dynamical system. We then compute the above mentioned summary statistics from the radial distribution functions and the msd as explained in Figure 1. The intuition behind choosing the above mentioned quantities is that they are broadly used characteristic quantities of the structure and dynamics of liquids [Allen and Tildesley, 1989, Frenkel and Smit, 2001].

Discrepancy measure

The discrepancy measure between two datasets \( \mathbf{x}^{(1)} \) and \( \mathbf{x}^{(2)} \) is constructed by considering the distance functions between the summary statistics extracted from them.

\[
d(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) := d \left( \mathcal{F}(\mathbf{x}^{(1)}), \mathcal{F}(\mathbf{x}^{(2)}) \right) = \frac{1}{8} \sum_{i=1}^{8} |S_i^{(1)} - S_i^{(2)}|
\]
Figure 1: **Summary statistics:** \((S_1, S_2, S_3, S_4, S_5, S_6, S_7, S_8)\) are computed from the radial distribution functions for \(O-H\) (a) and \(O-O\) (b) atoms and the mean square displacement (c), generated from the simulated coordinates of the dynamical system.

**Prior distributions**

We use independent continuous uniform prior distributions on the range \([2.81e^{-1}, 5.3e^{-1}]\) and \([2e^{-1}, 9e^{-1}]\) correspondingly for \(\sigma\) and \(\epsilon\). Outside this parameter range the TIP4P model of water in GROMACS becomes extremely chaotic and simulated data set can not be obtained in a reasonable time span.

**Perturbation kernel**

To explore the parameter space of \(\phi = (\sigma, \epsilon) \in [2.81e^{-1}, 5.3e^{-1}] \times [2e^{-1}, 9e^{-1}]\), we consider a truncated two-dimensional multivariate Gaussian distribution as the perturbation kernel. APMCABC inference scheme centers the perturbation kernel at the particle it is perturbing and updates the variance-covariance matrix of the perturbation kernel based on the particles learned from the previous step.
3.3 Parameter estimation

Our main goal is to estimate $\phi$, given $x^0$. In decision theory, the Bayes estimator minimizes the posterior expected loss, $E_{p(\phi|x^0)}(L(\phi, \cdot)|x^0)$. Here we consider the following loss function,

$$L(\phi_1, \phi_2) := d_E(\phi_1, \phi_2)$$

were $d_E$ is the Euclidean distance. If we have $Z$ samples $(\phi_i)_{i=1}^Z$ from the posterior distribution $p(\phi|x^0)$, the Bayes estimator can be approximated by

$$\hat{\phi} = \arg\min_{\phi} \frac{1}{Z} \sum_{i=1}^Z L(\phi_i, \phi).$$

(7)

which is also the estimated posterior mean $\hat{\phi} = \frac{1}{Z} \sum_{i=1}^Z \phi_i$.

3.4 Dynamic Allocation for MPI

In this section, we discuss the importance of a dynamic work allocation strategy for map-reduce and provide an empirical comparison of a straightforward allocation approach versus a online greedy approach.

In the straightforward approach, the allocation scheme initially distributes $m$ tasks to $n$ executors, sends the map function to each executor, which in turn applies the map function, one after the other, to its $m/n$ map tasks. This approach is visualized in Figure 2, where a chunk represents the set of $m/n$ map tasks. For example, if we want to draw 10,000 samples from the posterior distribution and we have $n = 100$ cores available, at each step of APMCABC we create groups of 100 parameters and each group is assigned to one individual core.

On the other hand, the dynamic allocation scheme initially distributes $k < m$ tasks to the $k$ executors, sends the map function to each executor, which in turn applies it to the single task available. In contrast to the straightforward allocation, the executor requests a new map task as soon as the old one is terminated. This clearly results in a better balance.
The unbalanced behavior is apparent if we visualize the run time of the individual map tasks on each executor. In Figure 3, the individual map tasks processing time is shown for an ABC algorithm. Each row corresponds to an executor (or rank) and each bar corresponds to the total time spent on all tasks assigned to the respective rank (row) for one map call. For the straightforward allocation strategy (a), one can easily verify that most of the ranks finish their map tasks in half the time of the slowest rank. This clearly leads to large inefficiencies. Conversely, using the dynamic allocation strategy (b), the work is more evenly distributed across the ranks. The unbalancedness is not a problem that can be overcome easily by adding resources, rather speed-up and efficiency can drop drastically compared to the dynamic allocation strategy with increasing number of executors. For a detailed description and comparison, we direct readers to Dutta et al. [2017b].

4 Results

In this section, we illustrate how Bayesian inference scheme introduced in Section 3.1 can be used to infer the posterior distribution and Bayes estimates of parameter $\phi = (\sigma, \epsilon)$ given an observed dataset $x^0$ both in simulated and experimental settings.
Simulated Data.

We first consider a simulated setting where the observed data has been generated, using GROMACS, from the TIP4P model of water reported in Equation 4. To generate the simulated observed data, we fix the non-bonded force-field parameter values at \( \phi^0 \equiv (\sigma^0, \epsilon^0) = (3.15e^{-1}, 6.48e^{-1}) \), which has been found to accurately reproduce the heat of vaporization, critical density, temperature and liquid density at 298K [Jorgensen et al., 1983]. As the true parameter value \( \phi^0 \) is known, we can assess the performance of the posterior distribution and the Bayes estimate correspondingly by their concentration and closeness to \( \phi^0 \). In Figure 4, the inferred posterior distribution, Bayes estimate \( \hat{\phi} \) and the true parameter value \( \phi^0 \) clearly show a very good performance of our inference scheme in estimating the underlying parametrization and quantifying the uncertainty in the inference. Additionally, to point the strength of having a posterior distribution for the parameters, we compute the posterior correlation between \( \sigma \) and \( \epsilon \), highlighting a strong negative correlation of \(-9.9e^{-1}\).

![Figure 4: Inference on Simulated Data: Contour plot of posterior distribution \( p(\phi|x^0) \) (red), Bayes estimate \( \hat{\phi} \equiv (\hat{\sigma}, \hat{\epsilon}) = (3.15e^{-1}, 6.18e^{-1}) \) (black star) and true value \( \phi^0 \equiv (\sigma^0, \epsilon^0) = (3.15e^{-1}6.48e^{-1}) \) (blue star) used to generate the simulated dataset. A strong negative posterior correlation of \(-9.9e^{-1}\) between the parameters is present. The posterior distribution is obtained using a Gaussian kernel density estimator with bandwidth 7.0e^{-1}.](image)

Experimental Data

We now illustrate the performance of the inference scheme for experimental dataset of water molecules under ambient conditions (temperature and pressure being fixed at 298K and 1 atm), assuming TIP4P force-field formalism of water. Though the exact coordinates over time of water molecules in an enclosed space can not be observed, the radial distribution functions of different molecular bonds and self-diffusion coefficient can be learned by experimental studies [Soper, 2000, Skinner et al., 2013]. As our inference scheme only depends on the summary statistics extracted form the radial distribution functions of \( O-O, O-H \) and self-diffusion coefficient, provided we have access to the experimentally obtained radial...
distribution functions and self-diffusion coefficient, we can compute the summary statistics and thus infer the non-bonded force-field parameters.

**Neutron Diffraction Dataset.** First we consider the Neutron diffraction derived radial distribution function of water from [Soper 2000](#). The experimentally obtained radial distribution function of $O-O$ and $O-H$ are shown in Figure 5a & 5b for further details we point readers to [Soper 2000](#). Additionally, we use the value of the self-diffusion coefficient, $1.3e^{-5} \text{cm}^2 \text{s}^{-1}$, reported in [Vega and Abascal 2011](#). The experimental values of the other summary statistics calculated are: $S_1^e = 9.628e^{-1}$, $S_2^e = 2.432e^{-1}$, $S_3^e = 0.59e^{-1}$, $S_4^e = 8.634e^{-1}$, $S_5^e = 3.3e^{-1}$, $S_6^e = 1.30e^{-1}$ and $S_7^e = 2.69$.

![Neutron Diffraction Dataset](#)

**Figure 5:** Neutron diffraction dataset: (a) & (b) The radial distribution functions of water obtained by Neutron diffraction and reported in [Soper 2000](#). (c) Contour plot of posterior distribution $p(\phi|x^0)$ (red), Bayes estimate $\hat{\phi} \equiv (\hat{\sigma}, \hat{\epsilon}) = (3.18e^{-1}, 5.87e^{-1})$ (black star). A strong negative posterior correlation of $-9.48e^{-1}$ between the parameters is present. The posterior distribution is obtained using a Gaussian kernel density estimator with bandwidth $7.0e^{-1}$. 

12
In Figure 5c, we plot the posterior distribution of $\phi$ and the Bayes estimates $\hat{\phi} \equiv (\hat{\sigma}, \hat{\epsilon}) = (3.18e^{-1}, 5.87e^{-1})$ obtained using the proposed Bayesian inference scheme. The posterior correlation between $\sigma$ and $\epsilon$, has a negative correlation of $-9.48e^{-1}$ as in the case of simulated dataset.

**X-ray Diffraction Dataset.** Next we consider an experimental dataset where only radial distribution function of $O-O$ of water using the X-ray diffraction is obtained [Skinner et al., 2013]. The experimentally obtained radial distribution function of $O-O$ is shown in Figure 6a, for further details we point readers to [Skinner et al., 2013]. Additionally, we use the value of the self-diffusion coefficient, $1.3e^{-5} cm^2 s^{-1}$, reported in [Vega and Abascal, 2011]. The experimental values of the other summary statistics calculated are: $S^e_4 = 9.8e^{-1}$, $S^e_5 = 3.47e^{-1}$, $S^e_6 = 1.41e^{-1}$ and $S^e_7 = 2.57$. In the absence of radial distribution function of $O-H$, we only consider these 5 summary statistics and the following distance function between them as the discrepancy measure:

$$d(x^{(1)}, x^{(2)}) := d\left(\mathcal{F}(x^{(1)}), \mathcal{F}(x^{(2)})\right) = \frac{1}{5} \sum_{i=4}^{8} |S^i_{(1)} - S^i_{(2)}|$$

![Figure 6](image_url)

(a) radial distribution function of $O-O$  
(b) Inference on X-ray diffraction dataset

Figure 6: **X-ray diffraction dataset:** (a) The radial distribution function of $O-O$ of water obtained by Neutron diffraction and reported in [Skinner et al., 2013]. (c) Contour plot of posterior distribution $p(\phi|x^0)$ (red), Bayes estimate $\hat{\phi} \equiv (\hat{\sigma}, \hat{\epsilon}) = (3.36e^{-1}, 4.11e^{-1})$ (black star). A strong negative posterior correlation of $-9.95e^{-1}$ between the parameters is present. The posterior distribution is obtained using a Gaussian kernel density estimator with bandwidth $7.3e^{-1}$.

In Figure 6b, we plot the posterior distribution of $\phi$ and the Bayes estimates $\hat{\phi} \equiv (\hat{\sigma}, \hat{\epsilon}) = (3.35e^{-1}, 4.31e^{-1})$ obtained using the proposed Bayesian inference scheme. Also in this case, we see a strong negative correlation of $-9.95e^{-1}$ between $\sigma$ and $\epsilon$. 
Figure 7: Comparison of radial distribution function: Radial distribution function of $O - O$ obtained using X-ray diffraction (red), simulated in GROMACS using $\phi^0$ (blue) and $\hat{\phi}$ (green). We notice a close similarity between experimentally obtained and simulated data using $\hat{\phi}$ with reference to the properties concerning the first hydration shell.

Further to illustrate the performance of the estimated parameter values in reproducing the dynamical properties, we simulate radial distribution function of $O - O$ in GROMACS using both $\phi^0$ and $\hat{\phi}$. In Figure 7 we compare these simulated radial distribution functions with the experimentally obtained one. We notice that the estimated force-field is able to reproduce the dynamical properties concerning the first hydration shell better than $\phi^0$. The discrepancy measure defined in Equation 8 also gives a significantly higher value for $\phi^0$ as the discrepancy values computed for $\phi^0$ and $\hat{\phi}$ are correspondingly 0.13 and 0.02.

5 Discussion

We propose a Bayesian inference framework to calibrate the force-field parameters of Molecular systems. This is achieved using approximate Bayesian computation and High performance computing. The methodology is illustrated by learning the force-field parameters of TIP4P system of water, both for simulated and experimental datasets. Bayesian inference is based on the posterior distribution, we are thus not only able to compute the Bayes estimate of the parameters in a data-driven manner but also the posterior correlation between them, which turns out to be strongly negative.

The estimated force-fields better reproduce the experimentally obtained structural and dynamical properties of water (e.g. radial distribution functions and self-diffusion coefficients), compared to the previous force-field parametrization which targets heat of vaporization, critical density, temperature and liquid density at 298K. Further, the negative correlation structure between non-bonded force-fields has been observed across two different experimental datasets, obtained using Neutron and X-ray diffraction. This points towards
a strong underlying mechanism. Though further investigation is necessary to understand this mechanism, we hypothesize that: At smaller values of $\sigma$, the repulsion is present at shorter distances and oxygen atoms can come closer. For oxygen atoms to come closer, while maintaining the same summary statistics (i.e. the same number of first neighbors), without creating vacuum vacancies, there should be more attraction between each other. This scenario is equivalent to a bigger $\epsilon$. The opposite holds for higher values of $\sigma$.

We also observe a strong imbalance in the simulation time of the TIP4P system using GROMACS for different parameter values. This creates problems in the performance of ABC which are mitigated using a dynamic-allocation scheme for MPI. We foresee that, for more challenging force-fields, more sophisticated ABC algorithms might be needed, possibly using surrogate and less computationally intensive models.

Finally, ABC model selection can be used to learn the most suitable force-field formalism for an experimental dataset. This will be a line of future investigation.

6 Acknowledgment

We thank Alessandro Laio and Michelle Parrinello for suggesting the specific force-field of TIP4P, for insightful comments on the choice of the summary statistics and for discussing earlier drafts of this manuscript. This research was supported by Swiss National Science Foundation Grant No. 105218_163196 (Statistical Inference on Large-Scale Mechanistic Network Models). We also thank Dr. Marcel Schoengens, CSCS, ETH Zurich for help regarding HPC and the Swiss National Super Computing Center for providing computing resources.

References

Jose LF Abascal and Carlos Vega. A general purpose model for the condensed phases of water: Tip4p/2005. *The Journal of chemical physics*, 123(23):234505, 2005.

Robert Abel, Tom Young, Ramy Farid, Bruce J Berne, and Richard A Friesner. Role of the active-site solvent in the thermodynamics of factor xa ligand binding. *Journal of the American Chemical Society*, 130(9):2817–2831, 2008.

M P Allen and D J Tildesley. *Computer Simulation of Liquids*. Clarendon Press, New York, NY, USA, 1989.

Z Faidon Brotzakis, Ilja K Voets, Huib J Bakker, and Peter G Bolhuis. Water structure and dynamics in the hydration layer of a type iii anti-freeze protein. *Physical Chemistry Chemical Physics*, 20(10):6996–7006, 2018.

ZF Brotzakis, CCM Groot, W Homsi Brandeburgo, HJ Bakker, and PG Bolhuis. Dynamics of hydration water around native and misfolded $\alpha$-lactalbumin. *The Journal of Physical Chemistry B*, 120(21):4756–4766, 2016.
Giovanni Bussi, Francesco Luigi Gervasio, Alessandro Laio, and Michele Parrinello. Free-energy landscape for $\beta$ hairpin folding from combined parallel tempering and metadynamics. *Journal of the American Chemical Society*, 128(41):13435–13441, 2006.

Manuel Chiachio, James L. Beck, Juan Chiachio, and Guillermo Rus. Approximate Bayesian computation by subset simulation. *SIAM J. Sci. Comput.*, 36(3):A1339–A1358, 2014.

Tom Darden, Darrin York, and Lee Pedersen. Particle mesh Ewald: An $N \log (N)$ method for Ewald sums in large systems. *The Journal of chemical physics*, 98(12):10089–10092, 1993.

Ritabrata Dutta, Marcel Schoengens, Jukka-Pekka Onnela, and Antonietta Mira. ABCpy: A user-friendly, extensible, and parallel library for approximate Bayesian computation. In *Proceedings of the Platform for Advanced Scientific Computing Conference*, page 8. ACM, 2017a.

Ritabrata Dutta, Marcel Schoengens, Avinash Ummadisingu, Jukka-Pekka Onnela, and Antonietta Mira. Abcppy: A high-performance computing perspective to approximate bayesian computation. *arXiv preprint arXiv:1711.04694*, 2017b.

Paul Fearnhead and Dennis Prangle. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474, 2012.

Daan Frenkel and Berend Smit. *Understanding Molecular Simulation*. Academic Press, Inc., Orlando, FL, USA, 2nd edition, 2001.

Ronald L Graham. Bounds for certain multiprocessing anomalies. *Bell Labs Technical Journal*, 45(9):1563–1581, 1966.

Michael U Gutmann, Ritabrata Dutta, Samuel Kaski, and Jukka Corander. Likelihood-free inference via classification. *Statistics and Computing*, 28(2):411–425, 2018.

Berk Hess, Henk Bekker, Herman JC Berendsen, Johannes GEM Fraaije, et al. LINCS: a linear constraint solver for molecular simulations. *Journal of computational chemistry*, 18(12):1463–1472, 1997.

BJ Jaidhan, P Srinivasa Rao, and Allam Apparao. Energy minimization and conformation analysis of molecules using steepest descent method. *International Journal of Computer Science and Information Technologies*, 5(3):3525–3528, 2014.

Bai Jiang, Tung-yu Wu, Charles Zheng, and Wing H Wong. Learning summary statistic for approximate Bayesian computation via deep neural network. *arXiv preprint arXiv:1510.02175*, 2015.

William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics*, 79(2):926–935, 1983.
Martin Karplus and Richard Lavery. Significance of Molecular Dynamics Simulations for Life Sciences. *Israel Journal of Chemistry*, 54(8-9):1042–1051, August 2014.

Lina Kulakova, Panagiotis Angelikopoulos, Panagiotis E Hadjidoukas, Costas Papadimitriou, and Petros Koumoutsakos. Approximate Bayesian computation for granular and molecular dynamics simulations. In *Proceedings of the Platform for Advanced Scientific Computing Conference*, page 4. ACM, 2016.

Maxime Lenormand, Franck Jabot, and Guillaume Deffuant. Adaptive approximate Bayesian computation for complex models. *Computational Statistics*, 28(6):2777–2796, 2013.

Jarno Lintusaari, Michael U Gutmann, Ritabrata Dutta, Samuel Kaski, and Jukka Corander. Fundamentals and recent developments in approximate Bayesian computation. *Systematic Biology*, 66(1):e66–e82, 2017. doi: 10.1093/sysbio/syw077. URL https://doi.org/10.1093/sysbio/syw077

Michael W Mahoney and William L Jorgensen. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *The Journal of Chemical Physics*, 112(20):8910–8922, 2000.

Jean-Michel Marin, Pierre Pudlo, Christian P. Robert, and Robin J. Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012. ISSN 0960-3174. doi: 10.1007/s11222-011-9288-2. URL http://dx.doi.org/10.1007/s11222-011-9288-2

Esteban A. Martinez, Christine A. Muschik, Philipp Schindler, Daniel Nigg, Alexander Erhard, Markus Heyl, Philipp Hauke, Marcello Dalmonte, Thomas Monz, Peter Zoller, and Rainer Blatt. Real-time dynamics of lattice gauge theories with a few-qubit quantum computer. *Nature*, 534(7608):516–519, 2016. doi: 10.1038/nature18318.

Hitoshi Ohtaki and Tamás Radnai. Structure and dynamics of hydrated ions. *Chemical Reviews*, 93(3):1157–1204, 1993.

Michele Parrinello and Aneesur Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics*, 52(12):7182–7190, 1981.

Sander Pronk, Szilárd Páll, Roland Schulz, Per Larsson, Pär Bjelkmar, Rossen Apostolov, Michael R Shirts, Jeremy C Smith, Peter M Kasson, David Van Der Spoel, et al. Gromacs 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29(7):845–854, 2013.

Pierre Pudlo, Jean-Michel Marin, Arnaud Estoup, Jean-Marie Cornuet, Mathieu Gautier, and Christian P Robert. Reliable ABC model choice via random forests. *Bioinformatics*, 32(6):859–866, 2015.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer-Verlag New York, Inc., 2005.
Joop Schaye, Robert A. Crain, Richard G. Bower, Michelle Furlong, Matthieu Schaller, Tom Theuns, Claudio Dalla Vecchia, Carlos S. Frenk, I. G. McCarthy, John C. Helly, Adrian Jenkins, Y. M. Rosas-Guevara, Simon D. M. White, Maarten Baes, C. M. Booth, Peter Camps, Julio F. Navarro, Yan Qu, Alireza Rahmati, Till Sawala, Peter A. Thomas, and James Trayford. The EAGLE project: simulating the evolution and assembly of galaxies and their environments. *Monthly Notices of the Royal Astronomical Society*, 446 (1):521–554, 2015. doi: 10.1093/mnras/stu2058.

Lawrie B Skinner, Congcong Huang, Daniel Schlesinger, Lars GM Pettersson, Anders Nilsson, and Chris J Benmore. Benchmark oxygen-oxygen pair-distribution function of ambient water from x-ray diffraction measurements with a wide q-range. *The Journal of chemical physics*, 138(7):074506, 2013.

AK Soper. The radial distribution functions of water and ice from 220 to 673 k and at pressures up to 400 mpa. *Chemical Physics*, 258(2-3):121–137, 2000.

Peter Turchin, Thomas E. Currie, Edward A. L. Turner, and Sergey Gavrilets. War, space, and the evolution of old world complex societies. *Proceedings of the National Academy of Sciences*, 110(41):16384–16389, 2013. doi: 10.1073/pnas.1308825110.

Carlos Vega and Jose LF Abascal. Simulating water with rigid non-polarizable models: a general perspective. *Physical Chemistry Chemical Physics*, 13(44):19663–19688, 2011.