FROM ATOMISTIC TO SYSTEMATIC COARSE-GRAINED MODELS FOR MOLECULAR SYSTEMS

Vagelis Harmandaris[1,2], **Evangelia Kalligiannaki**[3], **Markos Katsoulakis**[3] **and Petr Plecháč**[4]

[1] Department of Mathematics and Applied Mathematics, University of Crete
Heraklion, GR-70013 Crete, Greece
e-mail: name@e-mail.address

[2] Institute of Applied and Computational Mathematics, Foundation for Research and Technology
Hellas, IACM/FORTH,
Heraklion, GR-70013 Crete, Greece

[3] Applied Mathematics and Computational Science, CEMSE, King Abdullah University of Science &
Technology
Thuwal 23955, Saudi Arabia
e-mail: evangelia,kalligiannaki@kaust.edu.sa

[4] Department of Mathematics and Statistics, University of Massachusetts at Amherst
Amherst, MA 01003, USA
e-mail: markos@math.umass.edu

[5] Department of Mathematical Sciences, University of Delaware
Newark, DE 19716, USA
e-mail: plechac@math.udel.edu

Keywords: coarse-graining, potential of mean force, relative entropy, force matching

Abstract. *The development of systematic (rigorous) coarse-grained mesoscopic models for complex molecular systems is an intense research area. Here we first give an overview of methods for obtaining optimal parametrized coarse-grained models, starting from detailed atomistic representation for high dimensional molecular systems. Different methods are described based on (a) structural properties (inverse Boltzmann approaches), (b) forces (force matching), and (c) path-space information (relative entropy). Next, we present a detailed investigation concerning the application of these methods in systems under equilibrium and non-equilibrium conditions. Finally, we present results from the application of these methods to model molecular systems.*

1 INTRODUCTION

Complex molecular systems characterize materials (e.g. plastics, rubbers, gels) that develop complex, multiphase morphologies through equilibrium self-assembly. Their properties, such as mixing behavior; phase diagrams; melting points, depend on the combination of molecular and macroscopic variables. On the microscopic level, all-atom simulations allow direct quantitative predictions of the properties of molecular systems over a range of length and time scales. However, due to the broad spectrum of characteristic lengths and times involved in complex molecular systems, it is not feasible to apply them to large realistic systems or molecules of complex structure. On the mesoscopic level, coarse-grained (CG) models have proven to be very efficient means in order to increase the length and time scales accessible by simulations.

Here we first give an overview of different algorithms for developing CG models of molecular systems. Methods such as inverse Monte Carlo [1], inverse Boltzmann [2], force matching [3], relative entropy [4], provide parameterizations of coarse-grained effective potentials at equilibrium by minimizing a fitting functional over a parameter space. Then, we further extend these studies using path-space methods (relative entropy rate) for coarse-graining and uncertainty quantification for non-equilibrium processes, [5, 6, 7, 8].

All the above methods mentioned in principle are employed to approximate a many-body potential, the (n-body) potential of mean force, describing the equilibrium distribution of coarse-grained sites observed in simulations of atomically detailed models. We present two main results of our latest studies, [9, 5]. Firstly, on optimizing coarse-grained models in equilibrium, we: (a) reveal the connection of the force matching method with thermodynamic integration. This connection provides us with information on how to construct a local mean force for equilibrium force matching implementations, to best approximate the potential of mean force. (b) We present in a mathematically consistent way the entropy and force matching methods and their equivalence, which we derive for general nonlinear coarse-graining maps, [9]. Secondly, we adopt the use of path-space methods, to define a dynamical analogue of the relative entropy minimization method. We provide a systematic derivation of Langevin type coarse-grained dynamics from fine-scale molecular simulations, based on the minimization of the relative entropy rate. It is shown that this minimization problem is equivalent to a weighted least squares problem, with weights that depend on the diffusion coefficient of the proposed stochastic dynamics for the coarse-grained system. At least for constant diffusion coefficient, it is nothing but the widely applied force matching method used in computational coarse-graining which, however, is restricted to equilibrium processes. Finally, we apply, and compare the above-described methodologies in several molecular systems: gas and fluid methane and water [10].

1.1 Atomistic and Coarse Grained Models

In the following we shortly describe the microscopic (atomistic) and mesoscopic (coarse-grained) representation of a prototypical molecular system. Assume the problem of N (classical) molecules in a box of volume V at temperature T. Let $\mathbf{q} = (q_1, \ldots, q_N) \in \mathbb{R}^{3N}$ describe the position of the N particles in the atomistic description, with potential energy $U(\mathbf{q})$. The probability that the system has a state $\mathbf{q}$ at the temperature T is given by the Gibbs canonical measure

$$\mu(d\mathbf{q}) = Z^{-1} \exp\{-\beta U(\mathbf{q})\} d\mathbf{q}, \tag{1}$$

where $Z = \int_{\mathbb{R}^{3N}} e^{-\beta U(\mathbf{q})} d\mathbf{q}$ is the partition function, $\beta = \frac{1}{k_B T}$ and k_B is the Boltzmann constant. We denote $f : \mathbb{R}^{3N} \to \mathbb{R}^{3N}$ the force corresponding to the potential $U(\mathbf{q})$, i.e., $f_j(\mathbf{q}) = -\nabla_{q_j} U(\mathbf{q}), \ j = 1, \ldots, N$, is the force exerted to the j-th particle.

Coarse-graining reduces the extended system's complexity by lumping together degrees of freedom into coarse-grained variables and investigates the mesoscopic range and the importance of atomistic detail across scales. Coarse-graining is considered as the application of a mapping (CG mapping) $\mathbf{\Pi} : \mathbb{R}^{3N} \to \mathbb{R}^{3M}$

$$\mathbf{q} \mapsto \mathbf{\Pi}(\mathbf{q}) \in \mathbb{R}^{3M} \tag{2}$$

on the microscopic state space, determining the $M(< N)$ CG particles as a function of the atomic configuration $\mathbf{q}$. We denote by $\mathbf{Q} = (Q_1, \ldots, Q_M)$ any point in the CG configuration space $\mathbb{R}^{3M}$ and use the bar " $^-$ " notation for quantities on the CG space. We call atoms the elements of the microscopic space with positions $q_j \in \mathbb{R}^3, j = 1, \ldots, N$ and 'CG particles' the elements of the coarse space with positions $Q_i \in \mathbb{R}^3$, $i = 1, \ldots, M$.

The mappings most commonly considered in coarse graining of molecular systems are linear mappings represented by a set of non-negative real constants $\{\zeta_{ij}\}_{\substack{i=1,\ldots,M \\ j=1,\ldots,N}}$, for which

$$\mathbf{\Pi}_i(\mathbf{q}) = \sum_{j=1}^N \zeta_{ij} q_j \in \mathbb{R}^3, \ i = 1, \ldots, M \,. \tag{3}$$

In this work we consider CG maps such that a CG particle is the center of mass of a group of atoms for which an atom contributes only to one CG particle, see for example figures 1a and b.
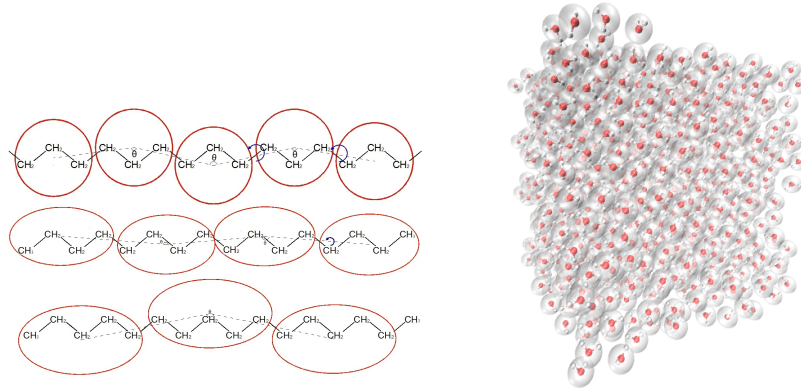

Figure 1: All-atom and CG representations of: (a) (left) a typical macromolecular chain, (b) (right) water bulk system.

The probability that the CG system has configuration $\mathbf{Q}$, is given by

$$\bar{\mu}(\mathbf{Q}) = \int_{\Omega(\mathbf{Q})} \mu(\mathbf{q}) d\mathbf{q}, \ \ \Omega(\mathbf{Q}) = \{\mathbf{q} \in \mathbb{R}^{3N} : \ \mathbf{\Pi}(\mathbf{q}) = \mathbf{Q}\} \,. \tag{4}$$

If we require that it is of the canonical Gibbs form then

$$\bar{\mu}(d\mathbf{Q}) = Z^{-1} \exp\{-\beta \bar{U}^{\mathrm{PMF}}(\mathbf{Q})\} d\mathbf{Q} \,,$$

and the corresponding free energy defines the $M-$body potential of the mean force (PMF),

$$\bar{U}^{\mathrm{PMF}}(\mathbf{Q}) = -\frac{1}{\beta} \log \int_{\Omega(\mathbf{Q})} e^{-\beta U(\mathbf{q})} d\mathbf{q} \,. \tag{5}$$

2 Parametrizations at equilibrium and Potential of Mean Force

The calculation of the PMF (5) is a task as difficult and costly as is calculating expectations of quantities of interest on the microscopic space. Therefore, one introduces parametric or non-parametric approximations

$$\bar{U}_{\text{eff}}(\mathbf{Q};\theta), \quad \theta \in \Theta \subseteq \mathbb{R}^k. \tag{6}$$

Methods such as inverse Monte Carlo (IMC), direct inverse Boltzmann (DBI) and iterative inverse Boltzmann (IBI) [11, 12, 1], force-matching [13, 14], and relative entropy minimization [15] provide optimal parameterizations of approximate coarse-grained models by considering a pre-selected set of observables ϕ and by then minimizing a cost functional over the parameter space,

$$\min_{\theta \in \Theta} \mathcal{L}(\phi;\theta). \tag{7}$$

The relative entropy (RE) minimization method is defined as the optimization problem

$$\min_{\theta} \mathcal{R}\left(\bar{\mu}|\bar{\mu}^\theta\right), \tag{8}$$

where

$$\mathcal{R}\left(\mu|\mu^\theta\right) = \mathbb{E}_\mu \left[\log \frac{\mu(\mathbf{q})}{\mu^\theta(\mathbf{q})}\right], \tag{9}$$

is the RE (or Kullback-Leibler divergence) between the microscopic Gibbs measure $\mu(\mathbf{q})$ and and a back-mapping $\mu^\theta(\mathbf{q})$ of the approximate CG measure $\bar{\mu}^\theta(\mathbf{Q}) \propto \exp(-\beta \bar{U}_{\text{eff}}(\mathbf{Q};\theta))$, [16, 4]. $\mathbb{E}_\mu[\cdot]$ denotes averages with respect to the probability measure $d\mu(\mathbf{q})$. The minimization of RE is thus equivalent to

$$\min_{\theta \in \Theta} \left\{\beta \mathbb{E}_\mu \left[\bar{U}_{\text{eff}}(\mathbf{\Pi}(\mathbf{q});\theta) - U(\mathbf{q})\right] - \left[\log Z^\theta - \log Z\right]\right\}, \tag{10}$$

where $Z^\theta = \int_{\mathbb{R}^{3M}} e^{-\beta \bar{U}_{\text{eff}}(\mathbf{Q};\theta)} d\mathbf{Q}$ and $Z = \int_{\mathbb{R}^{3N}} e^{-\beta U(\mathbf{q})}$.

The force-matching method determines a CG effective force $\bar{F}(\mathbf{Q};\theta)$ –and thus an effective potential– from atomistic force information as the solution of the mean least-square minimization problem

$$\min_{\theta \in \Theta} \mathbb{E}_\mu \left[\|h(\mathbf{q}) - \bar{F}(\mathbf{\Pi}(\mathbf{q});\theta)\|^2\right], \tag{11}$$

where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^{3M}$. The reference field $h(\mathbf{q}) \in \mathbb{R}^{3M}$ is the *local mean force* whose component $h_i(\mathbf{q})$, $i = 1, \ldots, M$ is the force exerted at the i-th CG particle that is a function of the microscopic forces. For example, if the CG mapping is the one that defines the CG particles as the center of mass of a group of atoms then $h_i(\mathbf{q}) = \sum_{j \in \{\text{group } i\}} f_j(\mathbf{q})$, $i = 1, \ldots M$. In work [9], we presented a rigorous probabilistic formulation and a generalization of the traditional force matching approach that applies to more complex and non-linear coarse-graining maps. For example, for

$$h(\mathbf{q}) = \mathbf{J}_{\mathbf{\Pi}}^{-1}(\mathbf{q})\mathbf{D}_{\mathbf{\Pi}}(\mathbf{q})f(\mathbf{q}) + \frac{1}{\beta}\nabla_\mathbf{q} \cdot \mathbf{J}_{\mathbf{\Pi}}^{-1}(\mathbf{q})\mathbf{D}_{\mathbf{\Pi}}(\mathbf{q}),$$

we prove that the solution of (11) is a best approximation of the PMF, which holds both for linear and non-linear mappings $\mathbf{\Pi}(\mathbf{q})$. Here $\mathbf{J}_{\mathbf{\Pi}}(\mathbf{q}) = \mathbf{D}_{\mathbf{\Pi}}(\mathbf{q})\mathbf{D}_{\mathbf{\Pi}}^t(\mathbf{q})$, $\mathbf{D}_{\mathbf{\Pi}} \in \mathbb{R}^{m \times 3N}$ $(\mathbf{D}_{\mathbf{\Pi}})_{ij}(\mathbf{q}) =$

$\nabla_{q_j} \mathbf{\Pi}_i(\mathbf{q})$. A more general result is available in [9]. These results are based on the probabilistic reformulation of the FM method and the connection to known results of the thermodynamic integration theory [?].

Additionally, the RE and FM methods are asymptotically equivalent, for small discrepancies from the PMF, see section VI in [9].

The DBI, IBI and IMC methods use the pair correlation function $g^{(2)}(\mathbf{Q})$ and the assumption that the interactions depend only on the distance R between particles, that is $g^{(2)}(\mathbf{Q}) =: \bar{g}(R)$. Thus the CG effective interaction is given by

$$\bar{U}_{\text{eff}}(R) = -\frac{1}{\beta} \log \bar{g}(R) \,. \tag{12}$$

DBI employs directly relation (12) to infer the interaction potential $\bar{U}_{\text{eff}}(R)$ from a reference CG (pair) distribution function $\bar{g}^{(\text{ref})}(R)$ obtained from the analysis of the all-atom configurations. In IBI methods, [2], an iterative numerical minimization problem is introduced based on $\bar{g}(R)$. The (pair) CG potential is refined at the iteration $(i+1)$ according to the following scheme:

$$\bar{U}_{\text{eff}}^{(i+1)}(R) = \bar{U}_{\text{eff}}^{(i)}(R) + c k_B T \log \frac{\bar{g}^{(i)}(R)}{\bar{g}^{(\text{ref})}(R)} \,, \tag{13}$$

where c is a constant to ensure stability of the iterative process.

3 Parametrizations away from equilibrium

We present an extension of the RE minimization approach to systems with *non-equilibrium steady states* as well as dynamics in *finite times*. The presented method also allows for approximation of *dynamical* observables, i.e., quantities that are averaged over the path distribution instead of over a distribution at a terminal time.

We consider the evolution of the N particles described by a diffusion process $\{X_t\}_{t \geq 0}$, a continuous time Markov process satisfying the stochastic differential equation (SDE)

$$\begin{cases} dX_t = b(X_t)dt + \sigma(X_t)dB_t \,, & t > 0 \,, \\ X_0 \sim \mu_0 \,, \end{cases} \tag{14}$$

where $b(x) \in \mathbb{R}^{6N}$ and $\sigma(x) \in \mathbb{R}^{6N \times k}$, $k \leq 6N$ are the drift and diffusion coefficients and B_t denotes the standard k-dimensional Brownian motion.

The *proposed coarse space dynamics* are described by a Markov process $\{\bar{X}_t\}_{t \geq 0}$ in $\mathbb{R}^m$ approximating the process $\{\mathbf{\Pi}X_t\}_{t \geq 0}$ which is, in principle, non-Markovian. The Markov process $\{\bar{X}_t\}_{t \geq 0}$ is given as the solution of the *parametrized* stochastic differential equations

$$\begin{cases} d\bar{X}_t = \bar{b}(\bar{X}_t; \theta)dt + \bar{\sigma}(\bar{X}_t)d\bar{B}_t, & t > 0, \\ \bar{X}_0 \sim \bar{\mu}_0 \,, \end{cases} \tag{15}$$

where $\bar{\sigma}(x) \in \mathbb{R}^{m \times l}$, $l \leq m$ is the diffusion and $\bar{b}(x; \theta) \in \mathbb{R}^m$ the drift coefficient is parametrized with $\theta \in \Theta$. $\bar{B}_t$ is an l-dimensional standard Brownian motion. The goal is to find the most effective among the proposed CG models such that $\{\bar{X}_t\}_{t \geq 0}$ "best approximates" the process $\{\mathbf{\Pi}X_t\}_{t \geq 0}$, that is to find optimal $\bar{b}(x; \theta)$ and $\bar{\sigma}(x; \theta)$.

The best approximation is fitted using entropy based criteria in order to find the best Markovian approximation of the coarse-grained process. We consider the optimization principle

$$\min_{\theta \in \Theta} \mathcal{R}\left(P_{[0,T]} | \mathbf{\Pi}_*^\dagger \bar{Q}_{[0,T]}^\theta\right) \,, \tag{16}$$

where $P_{[0,T]}$ is the path distribution of the original microscopic process and $\mathbf{\Pi}_*^\dagger \bar{Q}_{[0,T]}^\theta$ is the parametrized path-space coarse-grained distribution back-mapped to the microscopic space.

Then, the variational inference problem (16) is equivalent to the following *path-space force matching* problem

$$\text{argmin}_{\theta \in \Theta} \mathcal{R}\left(P_{[0,T]} | \mathbf{\Pi}_*^\dagger \bar{Q}_{[0,T]}^\theta\right) = \text{argmin}_{\theta \in \Theta} \mathbb{E}_{P_{[0,T]}} \left[\frac{1}{2} \int_0^T \|\mathbf{\Pi} b(X_s) - \bar{b}(\mathbf{\Pi} X_s; \theta)\|_{\mathbf{\Pi}^\sharp \Xi}^2 \, ds\right] .$$

where

$$\|z\|_{\mathbf{\Pi}^\sharp \Xi}^2 = z^{tr} \mathbf{\Pi}^{\sharp,tr} \Xi^{tr} \Xi \mathbf{\Pi}^\sharp z, \; z \in \mathbb{R}^m \text{ and } \Xi = \left[\sigma^{tr}(x)\sigma(x)\right]^{-1} \sigma^{tr}(x) .$$

under the assumption that the auxiliary reconstructed diffusion process with path-space distribution $\mathbf{\Pi}_*^\dagger \bar{Q}_{[0,T]}^\theta$ has diffusion coefficient $\sigma(x)$. Here $\mathbf{\Pi}^\sharp$ is defined such that $\mathbf{\Pi}\mathbf{\Pi}^\sharp = \mathbb{I}$. If moreover $\{X_t\}_{t \geq 0}$ is stationary with the invariant measure μ, then

$$\mathcal{R}\left(P_{[0,T]} | \mathbf{\Pi}_*^\dagger \bar{Q}_{[0,T]}^\theta\right) = T\mathcal{H}(P \,|\, \mathbf{\Pi}_*^\dagger \bar{Q}^\theta) + \mathcal{R}\left(\mu | \mathbf{\Pi}_*^\dagger \bar{\mu}_0\right) ,$$

and

$$\text{argmin}_{\theta \in \Theta} \mathcal{H}(P \,|\, \mathbf{\Pi}_*^\dagger \bar{Q}^\theta) = \text{argmin}_{\theta \in \Theta} \mathbb{E}_\mu \left[\frac{1}{2}\|\mathbf{\Pi} b(X) - \bar{b}(\mathbf{\Pi} X; \theta)\|_{\mathbf{\Pi}^\sharp \Xi}^2\right] .$$

Let us now consider Langevin dynamics for the N-particle molecular system, described by the process $\{(\mathbf{q}_t, \mathbf{p}_t)\}_{t \geq 0}$, with positions $\mathbf{q} \in \mathbb{R}^{3N}$ and momenta $\mathbf{p} \in \mathbb{R}^{3N}$

$$\begin{cases} d\mathbf{q}_t = \mathbb{M}^{-1}\mathbf{p}_t dt \,, \\ d\mathbf{p}_t = F(\mathbf{q}_t)dt - \gamma\mathbb{M}^{-1}\mathbf{p}_t dt + \sigma dB_t \,, \end{cases} \tag{17}$$

a Hamiltonian system coupled with a thermostat, where $F(\mathbf{q})$ is the force field that is not necessarily a gradient. $\mathbb{M} = \text{diag}\left(m_1 I_3, \ldots, m_N I_3\right) \in \mathbb{R}^{3N \times 3N}$ is the mass matrix, $\gamma \in \mathbb{R}^{3N \times 3N}$ is the friction and $\sigma \in \mathbb{R}^{3N \times 3N}$ the diffusion coefficients respectively, and B_t is the $3N-$dimensional Brownian motion. The diffusion and friction coefficients satisfy the fluctuation-dissipation relation $\sigma\sigma^{tr} = 2\beta^{-1}\gamma$.

The proposed dynamics for the coarse variables $\bar{X} = (\mathbf{Q}, \mathbf{P}) \in \mathbb{R}^{6M}$ are given by the Langevin system

$$\begin{cases} d\mathbf{Q}_t = \overline{\mathbb{M}}^{-1}\mathbf{P}_t dt \,, \\ d\mathbf{Q}_t = \bar{F}(\mathbf{Q}_t; \theta)dt - \bar{\gamma}\overline{\mathbb{M}}^{-1}\mathbf{P}_t dt + \bar{\sigma} d\bar{B}_t \,, \end{cases} \tag{18}$$

where $\bar{B}_t$ is a $3M$-dimensional Brownian motion. The diffusion coefficient $\bar{\sigma}$ is defined by

$$\bar{\sigma}\bar{\sigma}^{tr} = \mathbf{\Pi}\sigma\sigma^{tr}\mathbf{\Pi}_p^{tr} .$$

The optimal parameter set for which the process $\{(\mathbf{Q}_t, \mathbf{P}_t)\}_{t \geq 0}$ best approximates $\{(\mathbf{q}_t, \mathbf{p}_t)\}_{t \geq 0}$ at the time interval $[0, T]$ is given by, if $\bar{\gamma}\mathbf{\Pi}_q = \mathbf{\Pi}_p\gamma$,

$$\theta^*(T) = \text{argmin}_\theta \mathbb{E}_{P_X^T} \left[\frac{1}{2} \int_0^T \|\mathbf{\Pi} F(\mathbf{q}_s) - \bar{F}(\mathbf{Q}_s; \theta)\|_{\mathbf{\Pi}^\sharp \Xi}^2 ds\right] , \tag{19}$$

Note that the norm $\|\cdot\|_{\mathbf{\Pi}^\sharp \Xi}$ is a weighted Euclidean norm with weights $\mathbf{\Pi}^\sharp \Xi = \mathbf{\Pi}^\sharp \left[\sigma^{tr}(x)\sigma(x)\right]^{-1} \sigma^{tr}(x)$. This retrieves the relation of the path-space force matching approach to the diffusion coefficient of the atomistic dynamics. Moreover, if we assume e equilibrium dynamics and constant diffusion $\sigma(x) = \sigma$ the optimization problem is reduced to the known FM method, see also Remark 6.3 in [5],

$$\theta^= \text{argmin}_\theta \mathbb{E}_\mu \left[\frac{1}{2}\|\mathbf{\Pi} F(\mathbf{q}) - \bar{F}(\mathbf{Q}; \theta)\|^2\right] PERIOD \tag{20}$$

4 Models and Simulations

We test the above described methods on different molecular systems at equilibrium:

(a) *Methane*. Methane liquid was simulated at constant temperature (NVT conditions) at T = 100 K cfor several ns. 512 CH_4 molecules were modeled, whereas the density was calculated after equilibrating the system in the NPT ensemble for 5 ns ($\rho = 0.38 \ g/cm^3$). The time step was 0.5 fs and a cut-off distance of $10\mathring{A}$ was used.

For the coarse-grained representation of CH_4, we have used a one-site representation with a pair potential.

(b) *Water*. One of the most well-studied liquids both through atomistic and coarse-grained models in the literature is water [17]. Here we have simulated all-atom water, using one of the most typical atomistic force fields, the SPC/E [18]. The model system consists of 1192 molecules at ambient conditions (T = 300 K, P = 1 atm). The time step was 1 fs. A cut-off distance of $10\mathring{A}$ was used, while electrostatic interactions were calculated using PME. We first equilibrate the system under NPT conditions for about 50 ns. Then, NVT simulations, in the average density, were performed for 20 ns. All-atom configurations were recorded every 10 ps.

For the coarse-grained representation of H_2O, we have also used a one-site representation with a pair potential. In the CG representation of water electrostatic interactions were not required to be introduced.

5 Results

5.1 Methane

In the following we present results from the atomistic and CG simulations of the bulk methane liquid. We approximate the many-body PMF between the CG particles through the different approaches discussed above.

First, we apply the IBI method for this system using the all-atom data. IBI converges for this system (tolerance is 10^{-4}) after 14 iterations. Data for the CG pair correlation function, $g(R)$, and the resulting potential for various iterations are presented elsewhere [10].

Next, we examine the FM method for the CH_4 fluid, by analyzing the reference data from the all-atom simulations. In order to solve the minimization problem we have tested different basis function sets: linear splines, cubic splines, LJ and Morse. Linear splines, cubic splines and a Morse type basis give the same results, within the numerical accuracy. Only the results using the LJ basis slightly deviate. Results concerning the CG effective interaction were also found to be very close to the data obtained from the two isolated CH_4 in the vacuum [10].

We have also examined the application of the relative entropy minimization problem for the CH_4 liquid. Here we have used a typical Newton-Raphson scheme. Convergence is achieved after about 20 iterations [10].

Data about the pair PMF, that is an approximation of the many-body PMF, for the bulk methane fluid derived from the different approaches (IBI, FM and RE) are shown in Figure 2. It is clear that different methods give slightly different approximations of the PMF. However, the differences between the various sets of data are rather small, less than 5% in overall.

5.2 Water

The next example considered here is water. First, we apply the IBI method for this system using the all-atom data. Convergence of IBI for water is more sensitive than for the methane fluid discussed before. Indeed more than 100 iterations are required for the CG radial distribution
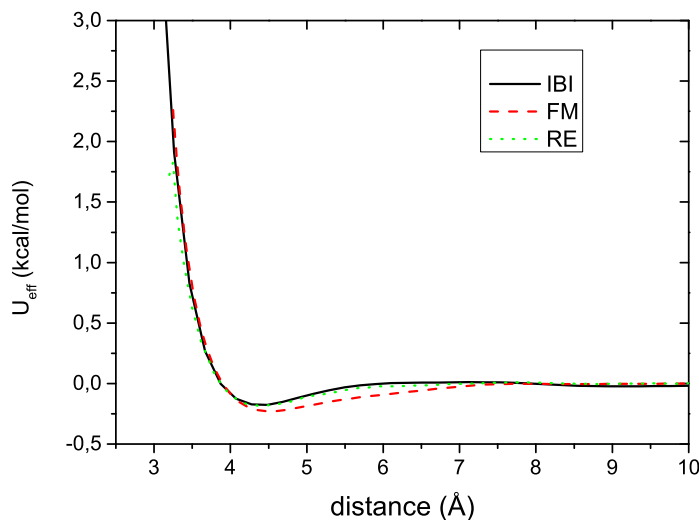
Figure 2: Bulk CH_4 CG effective potentials approximated from different methods ($T = 100K$).

function in order to match the atomistic data.

Then, we apply the RE and the FM methods for water. Numerical implementations for these methods are very sensitive to poor sampling. In FM the matrix the canonical system becomes singular, while in IBI and RE the iterative procedure fails. Thus very careful extrapolation methods must be considered at small distances as well as smoothing approaches to reduce the inherent noise of the statistical sampling. Specifically, the speed of convergence for the Newton-Raphson iterative scheme is based on the χ parameter, whereas its stability primarily depends on the condition number of the Hessian. The latter depends on several parameters: the trajectory length (at what extent does the sample size evenly cover the chosen basis function); the chosen basis set; and the correlation of the above to the model parameters, such as the number of atoms and the complexity of the coarse graining mapping [10].

Another important issue is poor sampling towards the minimum distance of the pair potential R_{min}. If this is the case, in the Newton-Raphson method for the RE minimization the Jacobian may involve negative values while the Hessian matrix becomes singular and the iterative scheme either stops or produces enormous fluctuations. A way to overcome this issues is the enrichment of nodes towards R_{min}, together with extrapolation of the potential on the first couple of nodes. Another good practice is smoothing out the potential after every iteration to reduce the noise in the updated forces.

In Figure 3 results for the effective CG potential from the IBI and the FM method (using cubic spline basis set) are presented. Note, that results from RE are similar to those of IBI. Although both IBI and FM potentials have a very similar structure with two minima, the actual values of the potential are considerably different, in contrast to the CH_4 fluid discussed in the previous subsection. Possible reasons for these discrepancies are related to the fact that FM and RE are only asymptotically equivalent, meaning that finite size basis sets effects might be important during the numerical optimization procedure. Clearly more work is required in order to clarify such differences [19, 9, 10].

Finally, in Figure 4 we show the CG $\bar{g}(R)$ obtained from RE minimization problem together with the reference curve, obtained from the analysis of the all-atom data. The curves are very
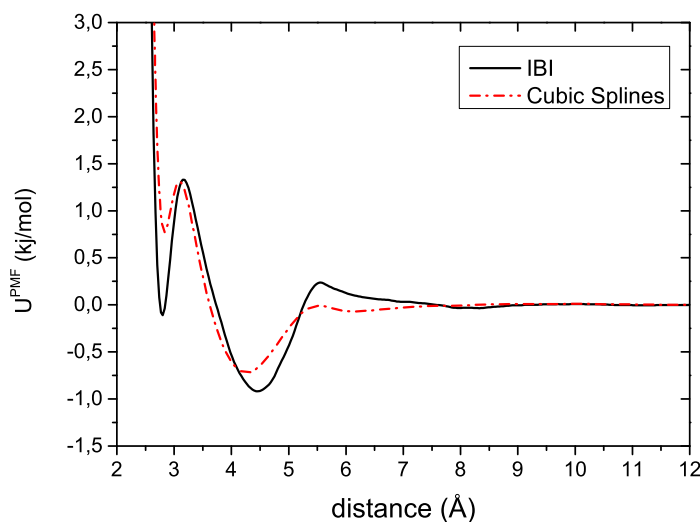
Figure 3: CG effective interactions of bulk water derived through IBI and FM methods.

close to each other; however there are small differences, in particular in small distances, close to the first maximum. Note that theoretically it is expected that the RE outcome, at the level of $g(R)$, should agree with the IBI one [15]. We should report here that we have calculated the CG potential derivatives appearing in the Jacobian and Hessian in the Newton-Raphson scheme by direct sampling during the corresponding CG run.

6 CONCLUSIONS

The development of systematic high fidelity coarse-grained models for molecular systems is a very challenging research area. In this aspect, finding the optimum effective interaction potential between CG particles for a given model involves crucial theoretical and numerical issues related to numerical parametrization of the (many-body) potential of mean force.

Here we have discussed different parametrization methods: (a) structural/correlation- based methods (direct Boltzmann inversion and iterative Boltzmann inversion), (b) Force matching, and (c) Relative entropy methods. We further examine the implementation of these approaches to molecular systems at equilibrium and non-equilibrium conditions in various molecular systems. In more detail, we have seen that CG methods based on relative entropy and force matching are in principle asymptotically equivalent. If we consider that RE methods are expected to give the same solution as the IBI methods, for a given CG mapping and a specific basis set, then we see the direct theoretical relation of all methods discussed here.

Furthermore, we applied the above numerical schemes (IBI, RE and FM) on the same atomistic systems. Despite the fact that all the above methods are approximations of the same (many-body) potential of mean force, it is not clear that their numerical implementation will converge into the same solution, since there are differences in the derived numerical schemes. We first consider a simple liquid (methane fluid) the CG effective potentials derived from the different methods are very similar. Slight differences of the order of 5-10% are found that are within the numerical accuracy. CG simulations with the derived force field also show structural properties in very good agreement with the reference (all-atom) data for all models. Different is the case of
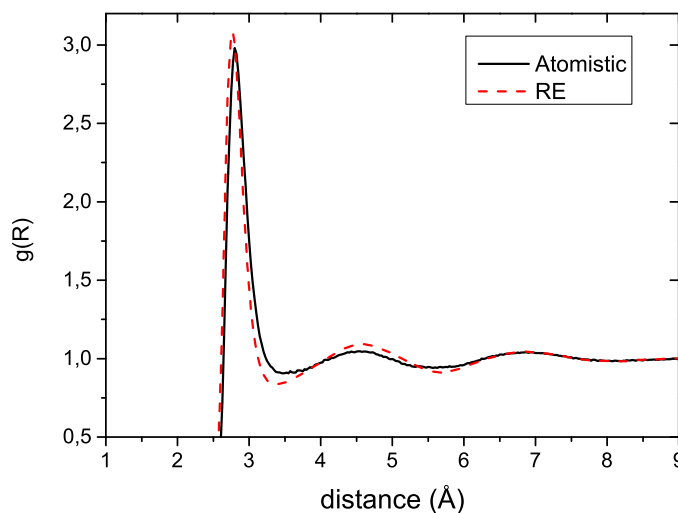
Figure 4: The RE optimizal CG model captures the atomistic RDF details for the bulk water.

the dynamic properties; friction in the CG models is clearly more sensitive to slight differences in the CG potential used. Second, we apply the methods on bulk water, one of the most well studied systems in the literature. Larger, compared to the methane, differences in the derived CG potential from the various methods are observed for water.

A general comment valid for all methods, is related to the actual numerical problems for parts of the phase space, where the energy is very high; i.e. areas with rare sampling. Practically, numerical extrapolation schemes using proper functional forms, as well as smoothing approaches, for the CG potentials should be used, in order to prevent sampling from such regimes.

Application of all above methods requires a very good sampling of the reference all-atom system. Such a sampling could be even more problematic for complex (e.g. polymeric) molecules. On the contrary, the DBI correlation-based approach, that is based on the decomposition of the CG potential in bonded and non-bonded components, can be a computationally efficient alternative. Such a methodology neglects many body terms; however, for several systems can provide an accurate prediction of the structural and thermodynamic properties [20, 21, 22, 23].

Several issues remain to be examined in order to systematically compare different numerical parametrization schemes for realistic molecular systems. For example, all systems studied here concern pair non-bonded CG effective potential; the use of many-body, or density dependent, CG potentials would expect to be important, in particular in systems of high density. In addition, non-linear CG maps could be also relevant especially when free energy differences, such as in thermodynamic integration, are to be computed. Parametrization of the dynamics of CG models is also one of the most challenging issues, in particular for non-equilibrium molecular systems [3, 24, 5, 25].

REFERENCES

[1] A. P. Lyubartsev and A. Laaksonen. On the Reduction of Molecular Degrees of Freedom in Computer Simulations. In M. Karttunen, A. Lukkarinen, and I. Vattulainen, editors,

Novel Methods in Soft Matter Simulations, volume 640 of *Lecture Notes in Physics, Berlin Springer Verlag*, pages 219–244, 2004.

[2] D. Reith, M. Pütz, and F. MüllerPlathe. Deriving effective mesoscale potentials from atomistic simulations. *Journal of computational chemistry*, 24(13):1624–1636, 2003.

[3] S. Izvekov and G.A. Voth. Modeling real dynamics in the coarse-grained representation of condensed phase systems. *J. Chem. Phys.*, 125(15), 2006.

[4] A. Chaimovich and M. S. Shell. Anomalous waterlike behavior in spherically-symmetric water models optimized with the relative entropy. *Phys. Chem. Chem. Phys.*, 11:1901–1915, 2009.

[5] V. Harmandaris, E. Kalligiannaki, M.A. Katsoulakis, and P. Plecháč. Path-space variational inference for non-equilibrium coarse-grained systems. *Journal of Computational Physics*, 314:355383, 2016.

[6] M. A. Katsoulakis and P. Plechac. Information-theoretic tools for parametrized coarse-graining of non-equilibrium extended systems. *J. Chem. Phys.*, 139:4852–4863, 2013.

[7] P. Dupuis, M. A. Katsoulakis, Y. Pantazis, and P. Plecháč. Path-space information bounds for uncertainty quantification and sensitivity analysis of stochastic dynamics. *SIAM J. Uncert. Quant.*, (4(1)):80–111, 2016.

[8] Georgios Arampatzis, Markos A. Katsoulakis, and Yannis Pantazis. *Pathwise Sensitivity Analysis in Transient Regimes*, pages 105–124. Springer International Publishing, Cham, 2015.

[9] E. Kalligiannaki, V. Harmandaris, M.A. Katsoulakis, and P. Plecháč. The geometry of generalized force matching and related information metrics in coarse-graining of molecular systems. *The Journal of Chemical Physics*, 143(8), 2015.

[10] E. Kalligiannaki, A. Chazirakis, A. Tsourtis, M.A. Katsoulakis, P. Plecháč, and V. Harmandaris. Parametrizing coarse grained models for molecular systems at equilibrium. *The European Physical Journal Special Topics*, 225(8):1347–1372, 2016.

[11] A.P. Lyubartsev and A. Laaksonen. Calculation of effective interaction potentials from radial distribution functions: A reverse Monte Carlo approach. *Phys. Rev. E*, 52:3730–3737, 1995.

[12] A.K. Soper. Empirical potential monte carlo simulation of fluid structure. *Chemical Physics*, 202(23):295 – 306, 1996.

[13] S Izvekov and GA Voth. Effective force field for liquid hydrogen fluoride from ab initio molecular dynamics simulation using the force-matching method. *The Journal of Physical Chemistry. B*, 109(14):6573–6586, 04 2005.

[14] S. Izvekov and G.A. Voth. Multiscale coarse graining of liquid-state systems. *The Journal of Chemical Physics*, 123(13):134105, 2005.

[15] A. Chaimovich and M.S. Shell. Coarse-graining errors and numerical optimization using a relative entropy framework. *The Journal of Chemical Physics*, 134(9):094112, 2011.

[16] M.S. Shell. The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *The Journal of Chemical Physics*, 129(14):–, 2008.

[17] H. Wang, C. Junghans, and K. Kremer. Comparative atomistic and coarse-grained study of water: What do we lose by coarse-graining? *Eur. Phys. J. E*, 28:221229, 2009.

[18] J. R.; Straatsma T. P. Berendsen, H. J. C.; Grigera. The missing term in effective pair potentials. *J. Phys. Chem.*, 91(-):62696271, 1987.

[19] L. Lu, J. F. Dama, and G. A. Voth. Fitting coarse-grained distribution functions through an iterative force-matching method. *J. Chem. Phys.*, 139:121906, 2013.

[20] W. Tschöp, K. Kremer, O. Hahn, J. Batoulis, and T. Bürger. Simulation of polymer melts. I. coarse-graining procedure for polycarbonates. *Acta Polym.*, 49:61, 1998.

[21] V. A. Harmandaris, N. P. Adhikari, N. F. A. van der Vegt, and K. Kremer. Hierarchical modeling of polystyrene: From atomistic to coarse-grained simulations. *Macromolecules*, 39:6708, 2006.

[22] V. Harmandaris and K. Kremer. Dynamics of polystyrene melts through hierarchical multiscale simulations. *Macromolecules*, 42:791, 2009.

[23] E. Brini, E.A. Algaer, P. Ganguly, C. Li, F. R. Ropero, and N.F.A. van der Vegt. Systematic coarse-graining methods for soft matter simulations: A review. *Soft Matter*, 9:21082119, 2013.

[24] P. Espanol and I. Zuniga. Obtaining fully dynamic coarse-grained models from md. *Phys. Chem. Chem. Phys.*, 13:10538–10545, 2011.

[25] C. Baig and V. Harmandaris. Quantitative Analysis on the Validity of a Coarse-Grained Model for Nonequilibrium Polymeric Liquids under Flow. *Macromolecules*, 43:3156–3160, 2010.