ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ - ΤΜΗΜΑ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ

UNIVERSITY OF CRETE - DEPARTMENT OF APPLIED MATHEMATICS

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

# Μαθηματική και υπολογιστική μοντελοποίηση σύνθετων μοριακών συστημάτων με πολλαπλές κλίμακες

Ηράκλειο
Φεβρουάριος, 2017

**Αναστάσιος Τσούρτης**

Επιβλέπων Καθηγητής:
Ευάγγελος ΧΑΡΜΑΝΔΑΡΗΣ

**Επταμελής επιτροπή:**

Ε. Χαρμανδάρης<br>
Δ. Τσαγκαρογιάννης<br>
Χ. Μακριδάκης<br>
Μ. Κατσουλάκης<br>
Χ. Τσόγκα<br>
Γ. Μακράκης<br>
Γ. Ζουράρης

### Ευχαριστίες

Πάνω από όλους θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου Ευάγγελο Χαρμανδάρη για την αδιάλειπτη στήριξη, σε ακαδημαϊκό αλλά και προσωπικό επίπεδο, σε όλες τις δυσκολίες που αντιμετώπισα. Ήταν εκείνος που, με θετική σκέψη, με παρότρυνε να συνεχίζω να πιστεύω στον εαυτό μου τις στιγμές στις οποίες δεν μπορούσα να διακρινω έστω και ύπαρξη κάποιας λύσης.

Ο δεύτερος επιβλέπων καθηγητής μου, Δημήτρης Τσαγκαρογιάννης, αφιέρωσε αρκετό χρόνο και κόπο κατά τη διάρκεια των χρόνων που συνεργαστήκαμε. Τον ευχαριστώ και τον ετκιμώ ιδιαίτερα.

Ευχαριστώ τον εξαιρετικό άνθρωπο, καθηγητή Μάρκο Κατσουλάκη, ο οποίος έπαιξε καθοριστικό ρόλο στην πορεία μου από τις προπτυχιακές σπουδές μέχρι σήμερα, όντας πηγή αισιοδοξίας και έμπνευσης στόχων.

Ευχαριστώ τον πολύ καλό φίλο και συνεργάτη Γιάννη Πανταζή για τις αμέτρητες συζητήσεις από την αρχική περίοδο της διατριβής. Η συμβολή του ήταν ουσιαστικής σημασίας στην εκπαιδευτική διαδικασία του διδακτορικού.

Τέλος, ευχαριστώ θερμά την πολύ καλή φίλη και συνεργάτιδα Ευαγγελία Καλλιγιαννάκη, που αφιέρωσε αρκετό χρόνο για να με βοηθήσει όποτε είχα κάποιο πρόβλημα.

Αφιερωμένο στους γονείς και στον αδερφό μου, οι οποίοι με στήριξαν αλλά και με ανέχτηκαν σε αυτή την δύσκολη πορεία των τελευταιών ετών.

# Contents

## Abstract

The theoretical study of complex materials, through mathematical and computational modeling (computer simulations) of many-particle systems has been a vibrant field of study during the last decades. Such materials are used in technological applications, ranging from nanotechnology, to aerospace engineering materials up to biomedical applications, drug design etc. Recently, there has been significant progress both in the level of answering fundamental physical questions, and in the development of novel algorithms, numerical methods and simulations. At the same time there is a rapid increase in computer processing capabilities and architectures [1, 2]. Simulation techniques for such systems vary from quantum, to microscopic (atomistic) up to mesoscopic (Coarse-Grained, CG) and to the continuum level. In this thesis we develop novel mathematical and computational methodologies for studying molecular systems in multiple length scales. In more detail, we: (a) Develop/apply a sensitivity analysis methodology for the parameters of the model systems, by using ideas from information theory in a probabilistic framework. (b) Examine different parameterization schemes of CG effective potentials, and (c) Propose a new CG parameterization methodology based on rigorous cluster expansion techniques. The above numerical approaches can be used to a vast variety of different molecular systems.

Η θεωρητική μελέτη καθώς και οι υπολογιστικές προσομοιώσεις πολύπλοκων μοριακών υλικών/συστημάτων, αποτελούν ένα ενεργό διεπιστημονικό πεδίο κατά τη διάρκεια των τελευταίων δεκαετιών. Υλικά τέτοιου είδους χρησιμοποιούνται σε τομείς με εφαρμογές από τη νανοτεχνολογία, τη βιοτεχνολογία εώς την αεροδιαστημική μηχανική, την βιο-ιατρική (π.χ. ανάπτυξη φαρμάκων) κ.α. Οι τεχνικές προσομοιώσεων μοριακών συστημάτων αφορούν διαφορετικές μεθοδολογίες σε ένα πολύ μεγάλο εύρος, από το κβαντικό στο μικροσκοπικό (ατομιστικό) επίπεδο, το μεσοσκοπικό (αδροποιημένο) έως και το συνεχές επίπεδο. Στην παρούσα εργασία αναπτύξαμε μια σειρά καινοτόμων μαθηματικών και υπολογιστικών μεθοδολογιών για τη μελέτη πολύπλοκων μοριακών συστημάτων τόσο σε ατομιστικό όσο και σε αδροποιημένο επίεπεδο. Πιο συγκεκριμένα: (α) Αναπτύξαμε και εφαρμόσαμε μια ανάλυση ευαισθησίας στις παραμέτρους των μοριακών μοντέλων, χρησιμοποιώντας ιδέες από τη θεωρία πληροφορίας σε ένα πιθανοθεωρητικό πλαίσιο. (β) Εξετάσαμε διαφορετικές μεθόδους παραμετροποίησης αποτελεσματικών (δυναμικών) αδροποιημένων μοντέλων, για μια συστηματική μετάβαση από το μικροσκοπικό (ατομιστικό) στο μεσοσκοπικό επίπεδο. (γ) Παρουσιάσαμε μια νέα μεθοδολογία εξαγωγής των αδροποιημένων δυναμικών, η οποία βασίζεται σε τεχνικές αναπτυγμάτων ομάδων. Όλες οι παραπάνω τεχνικές δοκιμάστηκαν σε απλά μοριακά συστήματα, ενώ μπορούν περαιτέρω να

χρησιμοποιηθούν σε ένα μεγάλο εύρος πιθανών πολύπλοκων μοριακών συστημάτων.

# Chapter 1

# Introduction

The study of molecular systems, especially as complexity increases, e.g. proteins and other macromolecluar systems, is challenging for experiments as well as for numerical simulations, due to the broad range of time and length scales involved. Typically, the time scale in atomistic simulations if of the order of $\mathcal{O}(10^{-11} - 10^{-8})$sec (with time resolution $dt \in \mathcal{O}(10^{-9})$sec), the length scale for the non-bonded interactions are $\mathcal{O}(10^{-9})$m for system sizes of $\mathcal{O}(10^{4})$ particles. One can easily see that the differences in the corresponding scales are vast between real experiments. Phase transitions for instance, occur in the order of few seconds (depending on the system and temperature) and in many cases the computational cost is prohibitive.

Simulations can be used in order to **replace or complement** experiments, by providing detailed structural, conformational and dynamical properties of the system, down to a level of description where conventional macroscopic measurements cannot reach. The accuracy of simulations relies on the quality of the atomistic force-fields (potential) between particles. Typically, pair interaction potentials are utilized, where the shape is controlled by adjustable parameters (e.g. Lennard-Jones, Morse potentials). The parameters of such potentials can be derived either by fitting to experimental data, in order to reproduce properties of homogeneous bulk systems, or by performing detailed ab-initio calculations.

Having the above in mind, it is desirable to reduce the required computational cost, by describing the system through a *smaller number of degrees of freedom*, in comparison with the fully atomistic level. This is the main idea behind Coarse-Grained (CG) models, which have been proven very efficient means for increase of spatio-temporal scales accessible by simulations [3, 4, 5, 6, 7, 8, 9, 10]. Here we focus on particle-based approaches, in which groups of atoms are replaced by structureless interaction centres (beads or "superatoms" in literature) that interact through effective potentials. CG models produce correct results for the study of *equilibrium* and structural properties, however dynamical properties like diffusion (or other time-correlated properties) are not always recovered. The basic reason for this failure is that the CG procedure eliminates degrees of freedom that should appear in the CG dynamics in the form of dissipation and thermal noise, both connected through the fluctuation-dissipation theorem [11].

Another aspect is the degree of Coarse-Graining. In low CG degree models a small number of atoms (usually 5-10) are lumped together. These models can be used to predict properties at the monomeric level, while at the same time, atomistic detail can be re-introduced into the CG configurations, providing direct information in the all-atom level. Alternatively, in many cases, coarser models (large number of monomers, or even long molecules are represented as a single CG bead) are required in order to study more complex systems [7, 12].

We should note here, that from a mathematical point of view, Coarse-Graining is a subfield of *dimensionality reduction* [13, 14, 15] and some methods are principal component analysis, polynomial chaos and diffusion maps [15, 16]. By eliminating atomistic details that are considered "unnecessary", CG models may provide three or more orders of magnitude greater efficiency than atomically detailed models [17].

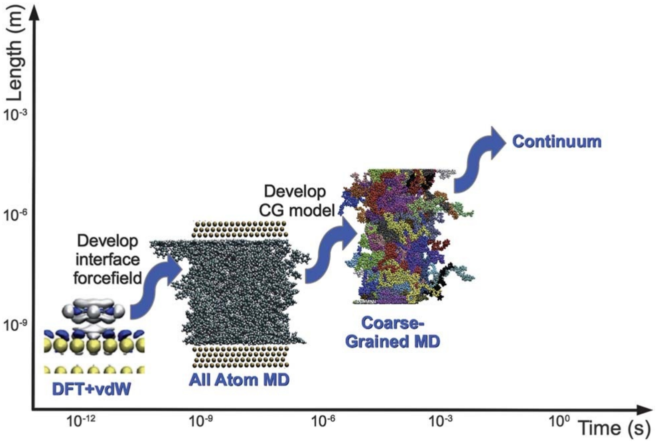

Figure 1.1: Multiscale modeling, from electronic structure level up to continuum for a polymer/solid interfacial system [18].

The accuracy of the CG models depends on the specific CG interaction potential (or force field) representation. There are several suggestions over the last decades [4, 19, 20]. On the one hand are more *qualitative bead-spring* type of models, in which molecules are represented as a series of beads interacting via Lennard-Jones like potentials, connected by (harmonic or FENE) springs to account for the intramolecular interactions of beads on the same molecule. On the other hand, are *systematic, more quantitative* CG models, in which atoms are lumped into beads or CG particles and they interact via effective CG potentials, which are derived from long atomistic simulations as an approximation of the (many-body) potential of mean force, PMF. PMF methods are further subdivided in categories, such as:

(a) The Force Matching [21, 19, 22], where a minimization technique is used for fitting a parameterized force functional according to mean forces in the atomistic level.

(b) Another PMF method [23] (discussed later in Chapters 4 and 5) is related to the reversible work theorem and the mapping $W(R) = -K_B T \ln Z(R)$, where $Z(R)$ is the mapped partition function and $W(R)$ is the PMF but it is a free energy function involving entropic contribution as well.

(c) An alternative approach to mean forcing, is to use structure based methods like the iterative Boltzmann Inversion (IBI) [5], in which one is interested in matching the radial distribution function (rdf) among the atomistic and CG levels, in an iterative fashion. All the above methods are rigorously valid in equilibrium.

As a general note, we stress that the use of CG models to describe (predict) the dynamics of complex systems is a subtle issue. The main reason is that the reduced degrees of freedom result to less friction between CG particles and therefore the time scale (and time-correlations) does not correspond to the atomistic one. To overcome this problem, in our description we used Langevin dynamics which involves friction forces in the equation of motion.

Uncertainty quantification (UQ) is the qualitative and quantitative estimation of uncertainties in complex physical, mechanical and engineering systems and is of paramount importance in multiscale modeling, where properties evaluated at the atomic-molecular scale are transferred to the macroscopic scale. UQ can determine how likely is the outcome of an experiment, if some of the system parameters are known up to a degree, meaning that they are described by a statistical distribution function. Sources of uncertainty can stem from i) numerical uncertainty ii) model uncertainty iii) parametric uncertainty. Numerical uncertainties are related to the finite time of the dynamic simulation, the number of particles, the time resolution (time-step) etc. Model uncertainty comes from the specific force field representation and its calibration to experimental properties. Parametric uncertainties stem from errors in parameter values due to noisy or insufficient measurements.

In the general context of UQ, Sensitivity Analysis (SA) is a powerful tool that gives insight into how small variations (uncertainty) in system parameters (input) can affect the results (output) if the system substantially. This means that this type of parameters have to be determined very accurately, as they are points of control in model dynamics. Such perturbations can occur form computational erros, uncertainty and errors resulting from experimental parameter estimation [24] (such as parameter fitting through ensemble averages of macroscopic thermodynamic quantities). Thus, parametric SA can provide critical insight into UQ. Depending on the magnitude of the perturbations, SA can be classifiend into local (infinitesimal, one-at-a-time parameter perturbation) and global (finit, multiple parameter perturbation).

Furthermore, SA in not restricted to UQ but it is of pivotal significance in several other applications. The notion of robustness of a system is the stability of its behavior under simultaneoud changes in model parameters or variations of orders of magnitude in insensitive parameters that insignificantly affect the dynamics and can be addressed by utilizing parametric SA approaches. SA can also be used in optimal experimental design [25], and identifiability analysis [26].

Stochastic modeling (models that use random forcing in the equations, according to probability density functions) is a powerful means of studying complex system. Such mathematical models involve slightly more calculations (to account for friction and collisions) and can accelerate the dynamics by increasing the time discretization. SA is suitable for this kind of systems, since it allows to extract information about the sensitivity and identifiability of parameters.

The main goals of this thesis are:

(a) To develop a novel sensitivity analysis methodology for complex molecular systems,

(b) To examine different parameterization of CG effective potentials, and

(c) To propose a new CG parameterization based on rigorous cluster expansion methodologies.

In more detail, we have developed a SA methodology suitable for complex stochastic systems, which relies on information loss due to parameter perturbations between *time-series* distributions, hence referred to as “pathwise”. This is achieved by employing the rigorously-derived Relative Entropy (RE) and the associated Relative Entropy Rate (RER), which can be thought of as “change of information per unit time”. RE or Kullback-Leibler divergence of two probability measures $\mu(dx) = \mu(x)dx$ and $\nu(dx) = \nu(x)dx$ is given by:

$$\mathcal{R}(\mu \backslash \nu) = \int \mu(x) \log \frac{\mu(x)}{\nu(x)} dx \tag{1.1}$$

and allows to define a psedo-distance between them. In the context of Chapter 3, $\mu(x)$ and $\nu(x)$ are probability distributions (or in weaker assumptions non-equilibrium steady-state measures) which come from perturbations in potential parameters. In the CG context of Chapter 4, the RE method employs the minimization of Relative Entropy between microscopic Gibbs measures $\mu(dx)$ and $\mu^{\theta}(dx)$ (or back-mapping [6]) representing approximations to the exact Coarse space Gibbs measure.

The Fisher Information Matrix (FIM) is associated with RER and constitutes a gradient-free approach to quantify parameter sensitivities. FIM is the Hessian of the RE and spectral analysis of this matrix provides further insight on sensitivities, parameter identifiability and dependencies (correlations) [27].

As mentioned earlier, we use CG methods based on comparing quantities between the atomistic and Coarse level, in order to construct an *effective* potential. The main issue is that although pair interactions are a good approximation for the microscopic level, after Coarse-Graining, a multi-body effective potential is derived, which for realistic system sizes cannot be calculated. Therefore, the **pair** effective CG potential is a (reasonable) approximation, which is made in an uncontrolled way, meaning that it is a solution to an "inverse problem". In Chapter 5, we suggest to explicitly compute the constrained configuration integral over all atomistic configurations that correspond to a given CG state and from that, suggest applications with a quantitative error.

Cluster expansions methods originate from the works of Mayer [28] and are valid in high temperature gas regime. The key idea is to form lumps of particles (clusters) and construct a CG potential from rigorous calculations of expanding the logarithm of the partition function in an absolutely convergent series. A hierarchy of CG Hamiltionians, based on the clustering, results to CG potentials of 2-Body, 3-Body and so on, in a *systematic way.* Constrained atomistic simulations of increasing sizes are performed in vacuum in order to construct controlled approximations of effective potentials and assess their validity, a posteriori, through the structural observable quantity rdf, in order to maintain correct thermodynamic properties. Finally, we assess the trade off between computational complexity-cost over accuracy between CG models.

We organize this thesis as follows. In Chapter 2, we briefly present the basic ideas behind molecular simulation, from Statistical Mechanics and derivation of the models, to common computational techniques. During this PhD thesis, three papers were written (two published, one submitted) and a fourth one is under development [29, 30, 31]. Chapters 3, 4, 5 correspond to the respective manuscripts, where there is some additional material. More specifically, Chapter 3 is on the sensitivity analysis of model parameters of stochastic $CH_4$ system. We use the information theoretic quantity of Relative Entropy and the associated Fisher Information Matrix on the equilibrium and non-equilibrium steady-state regimes. In Chapter 4, we reviewed, compared and contrasted various CG methodologies for molecular systems of different complexity, such as force-matching, Iterative Boltzmann Inversion and Relative Entropy. Chapter 5 is on systematic Coarse-Graining based on cluster expansions and mean force calculations. Construction of 2-Body and 3-Body effective potentials of different molecular systems is explained in detail in Chapter 6. It involves theoretical and especially the numerical/technical part of the 3-body computations and will appear as a future publication.

# Chapter 2

# Molecular Dynamics and Statistical Mechanics

## 2.1 Introduction

We carry out computer simulations in order to model molecular systems and to study their structural and dynamical properties, starting from the microscopic interactions between them. This can be used as a *computational design tool* that help us to predict structure-properties relations of complex molecular materials. At the same time serves as a complement to conventional experiments [32]. For example, a typical experiment measures the average of an observable quantity (property) over a "large" number of particles and over "large time scales". In contrast, molecular simulations (e.g. molecular dynamics) measure instantaneous positions and velocities of the particles where no real experiment can provide such detailed information. If we wish to use computer simulation as the numerical counterpart of experiments [33], we must know what kind of averages (with respect to pressure, internal energy etc.) we should aim to compute. In order to explain this, we need to introduce notions from statistical mechanics.

## 2.2 Statistical Mechanics

We start by reviewing basic ideas from the classical mechanics. The macroscopic thermodynamic state of a system is described by the following properties: the number of particles $N$. temperature $T$, pressure $P$ and volume $V$. Others can be derived by the equations of state and fundamental equations of thermodynamics. The instantaneous mechanical state is defined by the positions and momenta ($6N$ dimensions in total) and later on we will associate the aforementioned properties with them. Note, that in this description we neglect quantum effects.

The microscopic system (positions and momenta at a given time instant

$:= \{\mathbf{q(t)}, \mathbf{p(t)}\}$) evolves in time through Newton's equations of motion, so we define the (Hamiltionian) flow of a point in phase space ($6N$-dimensional sub-space) as $\Gamma(t)$. The macroscopic property $\mathcal{A}$ is a *time average* over a long time interval $T$:

$$\langle \mathcal{A} \rangle_t = \langle \mathcal{A}(\Gamma(t)) \rangle_t = \lim_{T \to \infty} \frac{1}{T} \int_0^T \mathcal{A}(\Gamma(t))dt \tag{2.1}$$

In computer simulations the integral becomes $\sum_{\tau=1}^{T} \mathcal{A}(\Gamma(\tau))$ and we require that the long trajectory explores the phase space satisfactorily, so that computed time averages correspond to averages of same parameters $(N, P, T)$ but different initial conditions, i.e. different starting phase point $\Gamma$.

Gibbs suggested replacing the time average by the **ensemble average**, (ergodic theorem) where ensemble is a collection of points $\Gamma$ in phase space. The points are described according to probability density $p(\Gamma) = p_{NVT}(\Gamma) = p_{NPT}(\Gamma) = p_{\text{ens}}(\Gamma)$. In this notation, phase point $\Gamma$ represents a typical system $(N, P, T = \textit{const})$ at any particular instant of time. Using the Liouville operator $L$ and Liouville's theorem, we may write:

$$\frac{\partial p_{\text{ens}}(\Gamma, t)}{\partial t} = -iLp_{\text{ens}}(\Gamma, t) \tag{2.2}$$

The rate of change of $p_{\text{ens}}$ at a fixed point $\Gamma$ in phase space, is related to the flows into and out of that point. If $p_{\text{ens}}(\Gamma)$ represents an equilibrium ensemble, then $\frac{\partial p_{\text{ens}}}{\partial t} = 0$. Such a system, is termed **ergodic**.

We now replace the time average in eq. (2.1) by an average taken over the ensemble phase space points

$$\langle \mathcal{A} \rangle_{\text{ens}} = \langle \mathcal{A} | p_{\text{ens}} \rangle = \sum_{\Gamma} \mathcal{A}(\Gamma) p_{\text{ens}}(\Gamma) \tag{2.3}$$

We define the sum over states $\Gamma$, the partition function $Q_{\text{ens}}$ as:

$$Q_{\text{ens}} = \sum_{\Gamma} W_{\text{ens}}(\Gamma), \quad p_{\text{ens}}(\Gamma) = \frac{W_{\text{ens}}(\Gamma)}{Q_{\text{ens}}} \tag{2.4}$$

where $W_{\text{ens}}(\Gamma)$ is the non-normalized form of $p_{\text{ens}}(\Gamma)$. $Q_{\text{ens}}$ is a function of the macroscopic properties defining the ensemble.

The above would suggest that the computation of thermodynamic quantities would be the evaluation of $Q_{\text{ens}}$, but this summation is not feasible. There are too many states (in the vast continuous description of phase space $\Gamma$) with very low weight $W_{\text{ens}}$due to non-physical overlaps between the repulsive cores of the molecules. In MD, $\langle \cdot \rangle_{\text{ens}}$ is replaced by a *trajectory average*, assuming that we are dealing with equilibrium (ergodic) systems, and since the Newton's equations generate a succession of states $\Gamma$ in accordance with the distribution function $p_{NVE}$; $E$ is constant energy.

We describe the energy of the system by the Hamiltonian at the phase point $\Gamma$ as $\mathcal{H}(\Gamma)$. The **Canonical Ensemble** has a probability density proportional to:

$$e^{-\mathcal{H}/(k_B T)} \tag{2.5}$$

(the constant is determined by the definition of zero entropy) $k_B$ is Boltzmann's constant, and the partition function is:

$$Q_{NVT} = \sum_{\Gamma} e^{-\mathcal{H}/(k_B T)} \tag{2.6}$$

At thermodynamic equilibrium, the function that is minimized in this case is the Helmholtz free energy

$$A_{\text{Helmholtz}} = -k_B T Q_{NVT} \tag{2.7}$$

(The corresponding representation for a quantum system of $\Omega(N, V, E)$ degenerate eigenstates is $S = k_B \ln(\Omega(N, V, E))$.) For a separable Hamiltionian (see next section) $\mathcal{H}(q,p) = \mathcal{K}(p) + U(q)$ the partition function $Q_{NVT}$ is rewritten as:

$$Z_{NVT} = \int e^{-U(q)/(K_B T)} dq \tag{2.8}$$

There are other ensembles, depending on the constant macroscopic quantities: the micro-Canonical ensemble $Q_{NVE}$, the isothermal-isobaric $Q_{NPT}$, the grand-Canonical $Q_{\mu VT}$ ($\mu$ is the chemical potential). In principle, all of them produce average properties which are consistent with one another, in the thermodynamic limit. Our MD simulations are under the $NVT$ and $NPT$ ensembles.

## 2.3 Classical Mechanics

We define the microscopic state of the system by specifying the generalized positions and momenta of the $N$ particles:

$$\begin{aligned} \mathbf{q} &= (q_1, q_2, \ldots, q_N) \\ \mathbf{p} &= (p_1, p_2, \ldots, p_N) \end{aligned} \tag{2.9}$$

In this classical approximation of the system, the particles interact with potential $U(\mathbf{q})$ and the energy of the system is the Hamiltonian $H(\mathbf{q}, \mathbf{p})$ which in turn is written as the sum of the kinetic and potential energy functions:

$$\mathcal{H}(\mathbf{p}, \mathbf{q}) = \mathcal{K}(\mathbf{p}) + U(\mathbf{q}) \tag{2.10}$$

where

$$\mathcal{K}(\mathbf{p}) = \sum_{i=1}^{N} \frac{\mathbf{p_i}^2}{2m_i} \tag{2.11}$$

The potential energy $U$ defines all interparticle interactions and it is possible to construct from $\mathcal{H}$, an equation of motion which governs the entire time-evolution of the system and all its mechanical properties. The Hamiltonian $\mathcal{H}$ defines the equilibrium distribution function for $\mathbf{p}$ and $\mathbf{q}$.

In principle, the potential energy of the system can be written as:

$$U(\mathbf{q}) = \sum_i \sum_{j>i} u^{(2)}(q_i, q_j) + \sum_i \sum_{j>i} \sum_{k>j>i} u^{(3)}(q_i, q_j, q_k) + \dots \tag{2.12}$$

where $u^{(2)}$ is the potential between pairs of atoms, $u^{(3)}$ between triplets and so on. We are interested in the relative distance between particles, so we can write $u^{(2)}(q_i, q_j) = u^{(2)}(|r_{ij}|)$ where $\mathbf{r_{ij}} = q_i - q_j$.

### 2.3.1 Equations of motion

The energy of the system is conserved in time. Let us denote the generalized coordinates $\mathbf{q}$ and $\dot{\mathbf{q}}$, the latter being the time derivatives of the positions. The classical equations of motion can be formulated in many ways [34]. Here, we use the Lagrangian formulation (equivalent to the Hamiltonian), in which eq's. of motion are derived though the Euler-Lagrange equations:

$$\frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \dot{\mathbf{q}}}\right) - \frac{\partial \mathcal{L}}{\partial \mathbf{q}} = 0$$
$$\mathcal{L}(\mathbf{q}, \dot{\mathbf{q}}) = \mathcal{K} - U$$

The generalized momentum is defined as

$$p_i = \frac{\partial \mathcal{L}}{\partial \dot{q}_i} \tag{2.13}$$

so the Hamiltonian form of the equations of motion is:

$$\begin{cases} \dot{q}_i = \frac{\partial \mathcal{H}}{\partial p_i} \\ \\ \dot{p}_i = -\frac{\partial \mathcal{H}}{\partial q_i} \end{cases} \tag{2.14}$$

$$\mathcal{H}(\mathbf{p}, \mathbf{q}) = \sum_i \dot{q}_i p_i - \mathcal{L}(q, \dot{q}) \tag{2.15}$$

for Cartesian coordinates, eq. (2.14) becomes

$$\begin{cases} \dot{r}_i = \frac{p_i}{m_i} \\ \\ \dot{p}_i = -\nabla_{r_i} U = f_i \end{cases} \tag{2.16}$$

$f_i$ is the force acting on atom $i$. Solving the equations of motion for the system involves the integration of first-order differential equations. Assuming

$V$ and $K$ do not depend explicitly on time ($\frac{\mathcal{H}}{dt} = 0$), the form of eq. (2.16) guarantees that the total derivative $\dot{\mathcal{H}} = \frac{\mathcal{H}}{dt}$ is zero i.e. the Hamiltonian is a constant of motion. The equations of motion are also reversible in time. We call *trajectory*, the evolution of the positions of a particle, in time:

$$\{\mathbf{q}^{(\mathbf{t})}, \mathbf{p}^{(\mathbf{t})}\}, \quad t \in [a, b] \subset \mathbb{Z} \tag{2.17}$$

## 2.4 Molecular Dynamics

Molecular Dynamics (MD) simulation consists of the numerical, step-by-step, solution of the classical equations of motion. There are many algorithms (numerical schemes) for solving systems of ODE's of the form (2.16). The key requirements of such an algorithm in our case should have the properties:

i) It must not involve a large number of force evaluations per time step *dt*. High order methods are more accurate but the computational cost is vast, in comparison to the gains of accuracy. Every MD code spends more than 80% of the computation time in force evaluation.

ii) It should satisfy the energy conservation law (or at least satisfy the energy on the average, instead of on every time-step. See BBK algorithm).

iii) It should permit the use of large time-step *dt*, in order to achieve longer time simulation. *stability*

iv) It should be efficient in terms of number of calculations **and** require as little memory as possible. This is because computations and storage, scale with the number of particles.

Every numerical scheme is associated with numerical errors (discretization errors, round-off errors in computations and storage of variables etc) which means that after long times, we cannot actually have an exact solution. This does not pose a serious problem as, we demand energy conservation (with minor fluctuations) and we are interested in average quantities. Thermodynamic properties of the system are extracted through time averages of observable quantities and we don't need exact trajectories, just states sampled from the correct ensemble. Essentially we require exact solutions of equations for times comparable with the correlation times of interest, so that we may accurately calculate time correlation functions (little long-term energy drift).

### 2.4.1 Verlet algorithm

The most widely used method for integrating equations of motion is the Velocity Verlet algorithm and is obtained by using Taylor expansions at time instants $t - dt$ and $t + dt$. The trajectory at time-step $t + dt = n + 1$ is:

$$\begin{cases} p^{n+1/2} = p^n - \frac{dt}{2}\nabla U(q^n) \\ q^{n+1} = q^n + dtM^{-1}p^{n+1/2} \\ p^{n+1} = p^{n+1/2} - \frac{dt}{2}\nabla U(q^{n+1}) \end{cases} \tag{2.18}$$

where the equations are in vector form and $M$ is the matrix containing the masses. Velocity Verlet is a second order numerical scheme. Note, that we are not using higher order high-accuracy numerical schemes as they require more computations (and memory). In addition, such schemes are usually more accurate at short times, exhibiting larger errors fro very long times. On the other hand, Velocity Verlet (an Verlet methods in general) are reversible and conserve energy at long atomistic simulations, without energy drift.

### 2.4.2 Thermostats

The Verlet algorithm samples trajectories from the $NVE$ ensemble. In practice we are interested in performing MD simulations under constant pressure $P$ or temperature $T$ or both, in order to produce results (average properties) comparable to experimental observables. In this case, we reformulate the Lagrangian (**??**) to include constraints for these quantities. The most widely used thermostat to constraint $T$ is the *Nosé-Hoover* thermostat, which consists of additional degrees of freedom in the set of equations of motion (2.14). We add a potential term of the form

$$V_s = gK_BT\ln(s) \tag{2.19}$$

where $g = 3N_{\text{atoms}} - N_{\text{bonds}} - 3$, $s$ is the new degree of freedom. In conjunction with section 2.2, we are sampling from the $NVT$ ensemble. A good thermostat should be able to rapidly enforce the correct probability distribution function of the ensemble.

### 2.4.3 BBK algorithm

We mention another description of the system, which involves *stochastic* equations of motion. We will explain in the next chapters the need of this alternative description of the evolution of the system in time.

In general, we insert random and drift forces in the deterministic equations to account for collisions of "ghost" particles of the heat bath. These forces constitute the thermostat and model an infinite reservoir of energy. This model, Langevin dynamics, conserves energy on the average (not on every timestep as eq. (2.16)) and the corresponding stochastic equations of motion read:

$$\begin{cases} dq_t = M^{-1}p_t dt \\ \\ dp_t = -\nabla_q U(q_t)dt - \gamma(q_t)M^{-1}p_t d_t + \sigma(q_t)dW_t \end{cases} \tag{2.20}$$

where $W_t$ is a standard Brownian motion bringing energy to the systm, $\gamma(q_t)$ is the coupling coefficient (it is constant in our case) and $-\gamma(q_t)M^{-1}p_t dt$ is the viscous friction term [35].

$$\sigma\sigma^T = \frac{2\gamma}{\beta} \tag{2.21}$$

is the fluctuation-dissipation relation, which ensures that the Canonical measure at the correct temperature is sampled. The time discretization of the system (3.23) is given by:

$$\begin{cases} p^{n+1/2} = p^n - \frac{dt}{2}\nabla U(q^n) - \frac{dt}{2}\gamma M^{-1}p^n + \sqrt{\frac{dt}{2}}\sigma G^n \\ \\ q^{n+1/2} = q^n + dtM^{-1}p^{n+1/2} \\ \\ p^{n+1} = p^n - \frac{dt}{2}\nabla U(q^{n+1}) - \frac{dt}{2}\gamma M^{-1}p^{n+1} + \sqrt{\frac{dt}{2}}\sigma G^{n+1/2} \end{cases} \tag{2.22}$$

where Verlet is used for the Hamiltionian (deterministic part) and midpoint Euler for the thermostat part. $G^n$ are Gaussian random vectors $\sim \mathcal{N}(0,1)$.

### 2.4.4 Periodic boundary conditions

MD simulations aim to provide information about the properties of a macroscopic system, whereas we employ systems of a few hundred up to thousands of particles. This number is far away from the thermodynamic limit. In order to simulate bulk phases it is essential to choose boundary conditions that mimic the presence of an infinite bulk surrounding the $N$-particle system, otherwise we encounter finite volume effects. Periodic boundary conditions (PBC) is the solution and the volume containing the $N$ particles is replicated along 3-dimensions, having additional 26 cloned volumes. Every time a particle tries to contact the simulation volume walls, it enters from the wall opposite, instead of bouncing back. Moreover, this periodic image of the system affects interparticle non-bonded interactions. Schematically, figure (2.1) shows the periodic boundary conditions in 2 dimensions. Interactions are depicted by arrows towards cloned particles using PBC, and this setup is termed as “minimum image convention”.

### 2.4.5 Interaction potential

As mentioned in section 2.3, the force-field (or the interaction potential) contains the information regarding interaction between atoms and/or molecules.

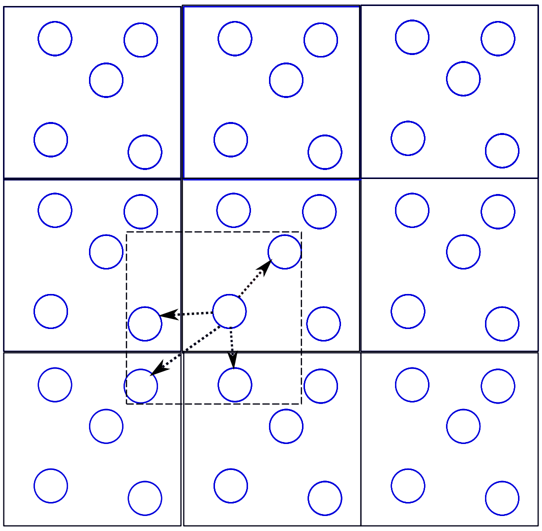

Figure 2.1: Minimum image convention. The dashed box includes interactions between the nearest clones.

The most widely used interaction potential for atoms belonging in different molecules (intermolecular interaction) is the Lennard-Jones potential (figure (2.2)) given by

$$u_{LJ}(r_{ij}) = \begin{cases} 4\epsilon_{LJ}\left[\left(\frac{\sigma_{LJ}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{LJ}}{r_{ij}}\right)^{6}\right] , & \text{if } r_{ij} < r_{cut} \\ 0 , & \text{otherwise,} \end{cases} \tag{2.23}$$

where $r_{cut}$ is the interaction cutoff radius we artificially set in order to reduce the computational cost while the total pairwise non-bonded potential energy of the system is

$$U_{\text{non-bond}}(\mathbf{r}) = \sum_{\substack{1\leq i,j\leq N \\ i<j}} u_{LJ}(r_{ij})$$

### 2.4.6 Bonds, Angles

Bonds and angles between atoms inside molecules are treated as (i) rigid or (ii) modeled with harmonic potentials. In case (i), bonds enter the Lagrangian in the form of constraints:

$$\mathcal{L}(\mathbf{r}, \dot{\mathbf{r}}) - \sum_{k} \lambda_k \sigma_k(r) \tag{2.24}$$

where $k$ is the set of constraints, $\lambda_k$ are the corresponding Lagrange multipliers and the constraints are of the form $\sigma_k = r_{ij}^2 - d_{ij}^2$. The constraining of rigid bonds requires extra degrees of freedom in the equations of motion which are handled by algorithms such as the SHAKE or the RATTLE [36]

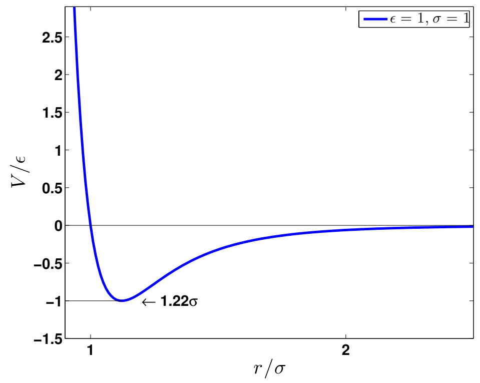

Figure 2.2: LJ potential for $\epsilon = 1, \sigma = 1$

When harmonic potentials are used, the bond and angle potentials are of the form:

$$u_{\text{bond}} = \frac{1}{2} K_{bond} (r_{ij} - r_0)^2, \quad u_{\text{angle}} = \frac{1}{2} K_{\theta} (\theta_{ijk} - \theta_0)^2 \tag{2.25}$$

between particles $i$ and $j$ and angle between particles $i, j, k$ respectively. Then, the bonded potential energies are defined as:

$$\begin{aligned} U_{\text{bond}} &= \sum_k u_{\text{bond}}(r^{(k)}) \\ U_{\text{angle}} &= \sum_k u_{\text{angle}}(\theta^{(k)}) \end{aligned} \tag{2.26}$$

$k$ being the id in the set of bonds and angles respectively. See Appendix C in chapter 3 for a more rigorous definition for the $CH_4$ model. The total potential energy of the system becomes:

$$U = U_{\text{non-bond}} + U_{\text{bond}} + U_{\text{angle}} \tag{2.27}$$

In the case of electrostatic charges, we add appropriate Coulomb potentials.

# Chapter 3

# Sensitivity Analysis

This chapter is primarily based on the published paper [29]. In the end, we have added the supplementary material of this publication.

In this work we present a parametric sensitivity analysis (SA) methodology for continuous time and continuous space Markov processes represented by stochastic differential equations. Particularly, we focus on stochastic molecular dynamics as described by the Langevin equation. The utilized SA method is based on the computation of the information-theoretic (and thermodynamic) quantity of relative entropy rate (RER) and the associated Fisher information matrix (FIM) between path distributions and it is an extension of the work proposed by Y. Pantazis and M. A. Katsoulakis [J. Chem. Phys. 138, 054115 (2013)].

A major advantage of the pathwise SA method is that both RER and pathwise FIM depend only on averages of the force field therefore they are tractable and computable as ergodic averages from a single run of the molecular dynamics simulation both in equilibrium and in non-equilibrium steady state regimes. We validate the performance of the extended SA method to two different molecular stochastic systems, a standard Lennard-Jones fluid and an all-atom methane liquid and compare the obtained parameter sensitivities with parameter sensitivities on three popular and well-studied observable functions, namely, the radial distribution function, the mean squared displacement and the pressure. Results show that the RER-based sensitivities are highly correlated with the observable-based sensitivities.

## 3.1 Introduction

Molecular simulation is the bridge between theoretically developed models and experimental approaches for the study of molecular systems in the atomistic level. [32, 33] Nowadays, molecular simulation methodologies are used extensively to predict structure-properties relations of complex systems. The importance of numerical simulations in material sciences, biology and chemistry has been recently acknowledged by the 2013 Nobel price in Chemistry "for the development of multiscale models in complex chemical systems". The computational modeling of realistic complex molecular systems at the molecular level requires long molecular simulations for an enormous distribution of length and time scales. [14, 37, 38, 18] The properties of the model systems depend on a large number of parameters, which are usually obtained utilizing optimization techniques matching specific data taken either from more detailed (e.g. ab-initio) simulations or from experiments. Furthermore, stochastic modeling is especially important for describing the inherent randomness of molecular dynamics in various scales.

All the above complexities imply the need of rigorous mathematical tools for the analysis of both deterministic and stochastic molecular systems. Uncertainty quantification (UQ) in computational chemistry is of paramount importance, especially in multiscale modeling, where properties evaluated at the atomic-molecular scale are transferred to the mesoscopic scale. [38, 39] Sources of epistemic uncertainty can stem from (i) numerical uncertainty, (ii) model uncertainty and (iii) parametric uncertainty. Numerical uncertainties are related to the finite time of the dynamic simulation, the number of particles, as well as values of the parameters related to the numerical method used (e.g. time-step), to name some. Model uncertainty comes from the specific force field representation and its calibration to experimental properties, and the usage of specific boundary conditions. Parametric uncertainties stem from errors in parameter values due to noisy or insufficient measurements. Of all the above, the uncertainty associated with the parameters of the potential is the least understood. [39, 40] Furthermore, intrinsic stochasticity of the system is added on top of epistemic uncertainty. This type of uncertainty is also called aleatoric.

There exists a diverse range of UQ approaches proposed in the literature. Variance-based methods such as analysis of variance (ANOVA), [41] Bayesian statistical analysis [42, 40] (applications to deal with large uncertainties) and polynomial chaos expansions [43] have been widely used. The first two methods are based on multiple and usually expensive Monte Carlo runs resulting in huge computational cost whereas the latter becomes intractable when the parameter space is large. An in depth study of this last method in MD has recently been presented by Rizzi et al. [43].

Sensitivity analysis (SA) is a powerful tool that gives insight of how small variations (uncertainty) in system parameters (input), can affect the output of the system substantially. Such perturbations occur from computational errors, uncertainty and errors resulting from experimental parameter estimation [24] (such as parameter fitting through ensemble averages of macroscopic thermodynamic quantities). Thus, parametric SA can provide critical insights in uncertainty quantification. Especially in the stochastic setting (e.g., Langevin dynamics in molecular systems), SA is performed by analysis of the system's mean behavior, i.e., several simulations starting from a configuration at the stationary (or steady states) regime. The stationary regime is crucial for complex molecular systems since it captures not only static quantities such as the radial distribution function but also dynamical quantities which includes transitions between metastable states in complex, high-dimensional energy landscapes and intermittency. [44] Depending on the magnitude of the perturbations, SA can be classified into local (infinitesimal, one-at-a-time parameter perturbation) and global (finite, multiple parameter perturbation).

Furthermore, the role of SA is not restricted to UQ but it is of pivotal significance in several other applications. First, robustness of a system meaning the stability of the behavior under simultaneous changes in model parameters or variations of orders of magnitude in insensitive parameters that insignificantly affect the dynamics can be addressed utilizing parametric SA approaches. Second, sensitivity analysis on experiment conditions under which information loss is minimized, establish optimal experimental design. [25] Furthermore, identifiability analysis employs SA to determine a priori whether certain parameters can be estimated from experimental data of a given type. The work in Ref. contains a general framework of SA in MD (proteins) using the observable helicity while cross-validation with experimental data is also displayed. Overall, SA plays a fundamental role in multiscale design and as it has been highlighted by Braatz et al.[45].

Typically in a stochastic setting, the most common local parametric SA method is based on partial derivatives on ensemble averages of quantities of interest around a nominal parameter value. [46] Large derivatives indicate strong sensitivity of the observable to the particular parameter while the opposite holds when the derivative values are small. There has been an increasing number of methods to compute the partial derivatives especially in discrete-event systems whose applications range from biochemical reaction networks to operations research and queuing theory. Finite-difference approaches based on common random numbers, [46] on common reaction path [47] which exploits positive correlations among coupled perturbed/unperturbed reaction paths as well as coupling methods [48, 49] have been proposed. There are certain issues associated with these finite-difference approaches; the estimator of the partial derivative has bias while the variance of the gradient estimator increases with the dimension of the parameter space. Instead of using the finite-difference approaches, one can utilize Girsanov measure transformation to directly compute the infinitesimal sensitivity. [50, 51, 52] In MD simulations where both time and states are typically continuous, Iordanov et al. [53] performed SA in three potentials of different functional forms (and number of parameters) in order to compare their influence in thermodynamic quantities. Instead of perturbing the potential parameters they scaled the potentials one-at-a-time (local SA) aiming to minimize the discrepancy from the experimental values of each observable separately.

Another class of sensitivity methods which is not focused on specific observable functions but on the overall properties of the stochastic process is based on information theory concepts. Application of information-theoretic SA methods to analysis of stochastic models uses quantities such as entropy, relative entropy (or Kullback-Leibler divergence), the corresponding Fisher information matrix as well as mutual information. Relative entropy (RE) measures the inefficiency of assuming a perturbed (or "wrong") distribution instead of assuming the unperturbed (or "true") one. RE have been used for the SA study of climate models [54] where the equilibrium probability density function (PDF) has been obtained through an entropy maximization subject to constraints induced by the measurements, while Fisher information matrix (FIM) is an indispensable tool for optimal experimental design. [25] In a typical SA approach based on RE, explicit knowledge of the equilibrium PDF is assumed. However, in systems with non-equilibrium steady states (NESS) (i.e., systems in which a steady state is reached but the detailed balance condition is violated), there are no explicit formulas for the stationary distribution and even when a Gibbs measure is available, it is usually computationally inefficient to sample from. Such non-equilibrium systems are common place in molecular systems with multiple mechanisms such as reaction-diffusion systems or driven molecular systems. [55]

In Ref. [44], the RE between path distributions (i.e., distributions of the particle trajectories) for discrete time or discrete event systems has been utilized as a measure of sensitivity. When the system is in stationarity the relative entropy of the two path distributions decomposes into two parts; (i) the relative entropy rate (RER) that scales linearly with time and (ii) a constant term related to the relative entropy of the initial distribution of the system. For long times, the first term dominates providing major insights on the sensitivity of the system with respect to parameter perturbations. In this work, we extend the SA method proposed in Ref. [44] to the case of stochastic differential equations (i.e., continuous-time and continuous state space) and particularly the Langevin equation. Furthermore, when perturbations are small, a Taylor expansion on RER is performed revealing the lower order of this expansion which is the pathwise FIM associated with the RER. Practically, RER is an observable of the stochastic process which can be computed numerically as an ergodic average in a straightforward manner as it only requires local dynamics (in our case the forces). Similarly, FIM computations are feasible in the same fashion with the advantage of being more informative since any perturbation direction can be explored. Both of these observables can be sampled on the fly, from a single MD run since only the reference process needs to be simulated. Finally, spectral methods for the calculation of RER were introduced for the over-damped Langevin case in Ref. [56].

The studied pathwise SA method has major advantages which can be listed as follows. First, it is a gradient-free method which does not require knowledge of the equilibrium PDF. By gradient-free we mean that the pathwise FIM does not depend on the extent of the perturbation, so when the computation for different perturbations is necessary (especially in high-dimensional problems) the extra cost is minimal in comparison to the straightforward RER calculation. Second, it is rigorously valid for long-time, stationary dynamics in path-space including metastable dynamics in a complex landscape. Third, it is suitable for non-equilibrium systems from statistical mechanics perspective; for example in NESS processes such as dissipative systems where the structure of the equilibrium PDF is unknown. Fourth, it is fast since it requires samples only from the unperturbed process which can be also obtained in a trivially parallel manner.

Overall, major novelties of the current work compared to previous work [44] are the following: i) *Validation* through observables of interest to molecular simulation of realistic complex fluids in comparison to the low dimensional systems studied previously. ii) Results are utilized from the continuous time. In Ref. [[44]], as well as in the appendix A, the current SA approach (RER and pathwise FIM) based on a discretized version of the numerical scheme is also presented, whereas in the following section we derived RER for the continuous time SDE. iii) The new approach is *independent* of the integration scheme one employs to integrate the equations numerically.

The work presented here is a part of a more general hierarchical simulation scheme that involves multiple simulation level and a broad range of length and time scales. [57, 58, 59] Here, we apply the above methodology on stochastic molecular systems as the RER and FIM methods are based on this setting i.e., non-deterministic with random noise. As test cases we examine: (a) a benchmark Lennard-Jones (LJ) fluid model, [33, 60] and, (b) a detailed all-atom methane ($CH_4$) model. [61] The LJ fluid system is the most widely used in molecular simulations model of a simple fluid, whereas the second one employs more complexity due to the intramolecular bond and angle potentials in addition to the LJ intermolecular potential. Methane has been also extensively studied over the years due to the fact that it is in abundance in nature and has environmental impacts as well as it can be used as fuel being the main component of natural gas. The method can be applied to a general SDE of the form $dY(t) = \beta(t,\omega)dt + \sigma(t,\omega)dB(t)$ where $B(t)$ is a (finite) dimensional Brownian motion provided that the diffusion term $\sigma(t,\omega)$ remains the same (Assumption II.1). This also holds for DPD or Brownian dynamics where the equations are almost identical.

The proposed pathwise SA method is validated through proper observable quantities upon perturbation of the potential parameters, which include structural, dynamical and thermodynamic properties of both LJ and methane model systems. We stress here that the utilized SA method based on RER and the pathwise FIM is independent of the observable quantities, which is not the case for derivative-based SA methods where they suffer from smoothness assumptions on the observable functionals. The partial derivative of an observable is related with the RE through the Pinsker inequality (see ineq. (3.16)). The Pinsker inequality asserts that small RER (or FIM) values result in small changes in observable expectation values under perturbation; thus RER and FIM can serve as a screening tool for specific observables. The present work provides a detailed quantitative study concerning the relation between the pathwise SA method (RER / FIM tools) and specific observables of molecular systems.

The organization of the paper is as follows. The following Section describes the path-wise sensitivity analysis method for Langevin dynamics in detail. In Section 3.3, the LJ fluid model, the methane model as well as various observable functions are presented followed by Section 3.4 where the validation of the proposed pathwise SA method is demonstrated. Finally, we conclude the paper in Section 3.5.

## 3.2 Pathwise Sensitivity Analysis for Langevin Dynamics

This Section describes and motivates the info-theoretic approach for sensitivity analysis of stochastic Molecular Dynamics. Particularly, the RER and the corresponding pathwise FIM are derived for the Langevin equation.

### 3.2.1 Stochastic equation of motion

Langevin dynamics models a Hamiltonian system which is coupled with a thermostat. [35] The thermostat serves as a reservoir of energy. In Langevin dynamics, the motion of particles is governed through a probabilistic framework by a system of stochastic differential equations given by

$$\begin{cases} dq_t = M^{-1}p_t dt \\ dp_t = F^\theta(q_t)dt - \gamma M^{-1}p_t dt + \sigma dW_t \ , \end{cases} \tag{3.1}$$

where $q_t \in \mathbb{R}^{dN}$ is the position vector of the $N$ particles in $d$-dimensions, $p_t \in \mathbb{R}^{dN}$ is the momentum vector of the particles, $M$ is the (diagonal) mass matrix, $F^\theta(\cdot) : \mathbb{R}^{dN} \to \mathbb{R}^{dN}$ is the driving (conservative) force which depends on a parameter vector $\theta \in \mathbb{R}^K$ (e.g. parameters of the specific atomistic force field), $\gamma$ is the friction matrix, $\sigma$ is the diffusion matrix and $W_t$ is a $dN$-dimensional Brownian motion. In the equilibrium regime, the forces are of gradient form, i.e., $F^\theta(q_t) = -\nabla V^\theta(q_t)$ where $V^\theta(\cdot)$ is the potential energy. Moreover, the fluctuation-dissipation theorem asserts that friction and diffusion terms are related with the inverse temperature $\beta \in \mathbb{R}$ of the system by

$$\sigma\sigma^T = 2\beta^{-1}\gamma \, .$$

Under gradient-type forces and the fluctuation dissipation theorem, the Langevin system has a Gibbs equilibrium (or invariant) distribution, $\mu^\theta(\cdot,\cdot)$, given by

$$\mu^\theta(dq, dp) = \frac{1}{Z} e^{-\beta(V^\theta(q)+\frac{1}{2}p^T M^{-1}p)} dqdp \ . \tag{3.2}$$

In non-equilibrium steady states, however, the stationary distribution, $\mu^\theta(\cdot,\cdot)$, is generally not known restricting the sensitivity analysis methods that rely on the explicit knowledge of the steady states. Though, as we show below, the proposed pathwise sensitivity methodology is not limited to equilibrium systems and it works equally well in the non-equilibrium steady states regime since it only necessitates the explicit knowledge of the driving forces (i.e., the local dynamics).

### 3.2.2 Relative Entropy Rate and Fisher Information Matrix for Langevin Processes

Let the path space $\mathcal{X}$ be the set of all trajectories $\{(q_t, p_t)\}_{t=0}^T$ generated by the Langevin equation in the time interval $[0, T]$. Let $Q_{[0,T]}^\theta$ denote the path space distribution, i.e., the probability to see a particular element of path space, $\mathcal{X}$, for a specific set of parameters $\theta$. Consider also a perturbation vector, $\epsilon_0 \in \mathbb{R}^K$, and denote by $Q_{[0,T]}^{\theta+\epsilon_0}$ the path space distribution of the perturbed process, $(\bar{q}_t, \bar{p}_t)$. The proposed sensitivity analysis approach is based on the quantification of the difference between the two path space probability distributions by computing the relative entropy (RE) between them. Thus, the pathwise RE of the unperturbed distribution, $Q_{[0,T]}^\theta$, with respect to the perturbed distribution, $Q_{[0,T]}^{\theta+\epsilon_0}$, assuming that they are absolutely continuous with respect to each other is defined as

$$\mathcal{R}(Q_{[0,T]}^\theta | Q_{[0,T]}^{\theta+\epsilon_0}) := \int \log\left(\frac{dQ_{[0,T]}^\theta}{dQ_{[0,T]}^{\theta+\epsilon_0}}\right) dQ_{[0,T]}^\theta \ , \tag{3.3}$$

where $\frac{dQ_{[0,T]}^\theta}{dQ_{[0,T]}^{\theta+\epsilon_0}}$ is the Radon-Nikodym derivative and it is well-defined due to the absolute continuity assumption. A key property of RE is that $\mathcal{R}(Q_{[0,T]}^\theta | Q_{[0,T]}^{\theta+\epsilon_0}) \geq 0$ with equality if and only if $Q_{[0,T]}^\theta = Q_{[0,T]}^{\theta+\epsilon_0}$, which allows us to view relative entropy as a "distance" (more precisely a semi-metric) between two probability measures capturing the relative importance of parameter vector changes. [41] Moreover, from an information theory perspective, the relative entropy measures *loss/change of information* when $Q_{[0,T]}^{\theta+\epsilon_0}$ is considered instead of $Q_{[0,T]}^\theta$. [62]

The necessary and sufficient conditions of the two path distributions (perturbed and unperturbed) to be absolutely continuous are provided next.

**Assumption 3.2.1.** *Assume that*

*(a) the diffusion matrix, $\sigma$, is invertible, and,*

*(b)* $\mathbb{E}_{Q^\theta_{[0,T]}}[\exp\big\{\int_0^T |u(q_t,p_t)|^2 dt\big\}] < \infty$, *where the function* $u(\cdot,\cdot) : \mathbb{R}^{2dN} \to \mathbb{R}^{2dN}$ *is defined such that for all pairs* $(q,p)$ *it should hold that*

$$\left[\begin{array}{cc} 0 & 0 \\ 0 & \sigma \end{array}\right] u(q,p) = \left[\begin{array}{c} M^{-1}p - M^{-1}p \\ F^\theta(q) - \gamma M^{-1}p - (F^{\theta+\epsilon_0}(q) - \gamma M^{-1}p) \end{array}\right] ,$$

*or, equivalently,*

$$\sigma u(q,p) = F^\theta(q) - F^{\theta+\epsilon_0}(q) \ .$$

Notice that such a function, $u(\cdot,\cdot)$, exists due to (a). Furthermore, (a) implies that the noise is non-degenerate for the momenta. In practice (b) means that the difference in the conservative forces is bounded over the path [0,T] as $F^\theta, F^{\theta+\epsilon_0}$ are bounded away from arbitrarily small intermolecular distances. Then, the RE of the path distribution defined in (3.3) is finite and an explicit formula can be estimated as the following proposition asserts.

**Proposition 3.2.1.** *Let Assumption 3.2.1 holds. Assume also that* $(q_0,p_0) \sim \nu^\theta$ *and* $(\bar{q}_0,\bar{p}_0) \sim \nu^{\theta+\epsilon_0}$ *where* $\nu^\theta(\cdot,\cdot)$ *and* $\nu^{\theta+\epsilon_0}(\cdot,\cdot)$ *are two initial distributions which should be absolutely continuous with respect to each other. Then,*

$$\begin{aligned}\mathcal{R}(Q^\theta_{[0,T]}|Q^{\theta+\epsilon_0}_{[0,T]}) &= \mathcal{R}(\nu^\theta|\nu^{\theta+\epsilon_0}) \\ &\quad + \frac{1}{2}\mathbb{E}_{Q^\theta_{[0,T]}}\Big[\int_0^T |u(q_t,p_t)|^2 dt\Big]\end{aligned} \tag{3.4}$$

*Proof.* Under Assumption 3.2.1, the Girsanov theorem applies providing an explicit formula of the Radon-Nikodym derivative[63] which is given by

$$\begin{aligned}&\frac{dQ^\theta_{[0,T]}}{dQ^{\theta+\epsilon_0}_{[0,T]}}\Big(\big\{(q_t,p_t)\big\}_{t=0}^T\Big) = \frac{d\nu^\theta}{d\nu^{\theta+\epsilon_0}}(q_0,p_0)\times \\ &\exp\left\{-\int_0^T u(q_t,p_t)^T dW_t - \frac{1}{2}\int_0^T |u(q_t,p_t)|^2 dt\right\} \ .\end{aligned}$$

Moreover, $\hat{W}_t := \int_0^t u(q_s,p_s)dt + W_t$ is a Brownian motion with respect to the path distribution $Q^\theta_{[0,T]}$, meaning that, for any measurable function

$f(\cdot,\cdot)$, it holds $\mathbb{E}_{Q^\theta_{[0,T]}}\big[\int_0^T f(q_t,p_t)^T d\hat{W}_t\big] = 0$. Then,

$$\mathcal{R}(Q^\theta_{[0,T]}|Q^{\theta+\epsilon_0}_{[0,T]}) = \int \Bigg(\log \frac{d\nu^\theta}{d\nu^{\theta+\epsilon_0}}(q_0,p_0)$$
$$- \int_0^T u(q_t,p_t)^T dW_t - \frac{1}{2}\int_0^T |u(q_t,p_t)|^2 dt\Bigg) dQ^\theta_{[0,T]}$$
$$= \int \log \frac{d\nu^\theta}{d\nu^{\theta+\epsilon_0}}(q_0,p_0) dQ^\theta_{[0,T]} - \int\int_0^T u(q_t,p_t)^T d\hat{W}_t dQ^\theta_{[0,T]}$$
$$+ \frac{1}{2}\int\int_0^T |u(q_t,p_t)|^2 dt dQ^\theta_{[0,T]}$$
$$= \mathcal{R}(\nu^\theta|\nu^{\theta+\epsilon_0}) + \frac{1}{2}\int\int_0^T |u(q_t,p_t)|^2 dt dQ^\theta_{[0,T]}$$

□

We remark that this proposition is a result on the transient regime since the initial distributions can be anything as fas as they are absolutely continuous with respect to each other. In the stationary regime, a significant simplification of the pathwise RE occurs. As the following proposition asserts, pathwise RE is decomposed into a linear in time term plus a constant where the slope of the linear term is the relative entropy rate (RER).

**Proposition 3.2.2.** *Let Assumption 3.2.1 holds. Assume also that* $(q_0,p_0) \sim \mu^\theta$ *and* $(\bar{q}_0,\bar{p}_0) \sim \mu^{\theta+\epsilon_0}$ *where* $\mu^\theta(\cdot,\cdot)$ *and* $\mu^{\theta+\epsilon_0}(\cdot,\cdot)$ *are the stationary distributions for the unperturbed and the perturbed process, respectively, which should be absolutely continuous with respect to each other. Then, the pathwise RE equals to*

$$\mathcal{R}(Q^\theta_{[0,T]}|Q^{\theta+\epsilon_0}_{[0,T]}) = T\mathcal{H}(Q^\theta|Q^{\theta+\epsilon_0}) + \mathcal{R}(\mu^\theta|\mu^{\theta+\epsilon_0}) \tag{3.5}$$

*where*

$$\mathcal{H}(Q^\theta|Q^{\theta+\epsilon_0}) := \frac{1}{2}\mathbb{E}_{\mu^\theta}[(F^{\theta+\epsilon_0}(q) - F^\theta(q))^T(\sigma\sigma^T)^{-1}(F^{\theta+\epsilon_0}(q) - F^\theta(q))] \tag{3.6}$$

*is the Relative Entropy Rate.*

*Proof.* First notice that we drop the $T$ subscript from the definition of RER because RER is time-independent. Then, it is straightforward to show from the previous proposition that

$$
\begin{aligned}
&\mathcal{R}(Q^{\theta}_{[0,T]}|Q^{\theta+\epsilon_0}_{[0,T]})\\
&= \mathcal{R}(\mu^{\theta}|\mu^{\theta+\epsilon_0}) + \frac{1}{2}\int\int_0^T |u(q_t,p_t)|^2 dt dQ^{\theta}_{[0,T]}\\
&= \mathcal{R}(\mu^{\theta}|\mu^{\theta+\epsilon_0}) + \frac{1}{2}\int_0^T\int |u(q_t,p_t)|^2 dQ^{\theta}_{[0,T]} dt\\
&= \mathcal{R}(\mu^{\theta}|\mu^{\theta+\epsilon_0}) + \frac{1}{2}\int_0^T\int |u(q,p)|^2 \mu^{\theta}(dq,dp) dt\\
&= \mathcal{R}(\mu^{\theta}|\mu^{\theta+\epsilon_0}) + \frac{T}{2}\mathbb{E}_{\mu^{\theta}}[|u(q,p)|^2] \ .
\end{aligned}
$$

□

RER inherits all the properties of relative entropy (non-negativity, convexity, etc.) and it measures the change of information in path space per unit time. For large times, the term that involves RER is the significant term, since it scales linearly with time, while the constant one becomes less and less important. Moreover, the estimation of RER necessitates only the knowledge of the driving forces (i.e., the local dynamics) which is available since the driving forces are computed in any numerical scheme of the Langevin equation.

**Pathwise Fisher information matrix**: Generally, RE is locally a quadratic functional in a neighborhood of parameter vector, $\theta$. Under smoothness assumption in the parameter vector, the curvature of the RE around $\theta$, defined by its Hessian, is the FIM. Analogously, we define the Hessian of the RER to be the pathwise FIM denoted by $F_{\mathcal{H}}(Q^{\theta})$. The relation between the RER and the pathwise FIM is

$$
\mathcal{H}(Q^{\theta}|Q^{\theta+\epsilon_0}) = \frac{1}{2}\epsilon_0^T F_{\mathcal{H}}(Q^{\theta})\epsilon_0 + \mathcal{O}(|\epsilon_0|^3) \ . \tag{3.7}
$$

Under smoothness assumption of the force vector, $F^{\theta}(\cdot)$, with respect to the parameter vector, $\theta$, an explicit formula for the pathwise FIM for the Langevin process is straightforwardly obtained from (3.6) given by

$$
F_{\mathcal{H}}(Q^{\theta}) = \mathbb{E}_{\mu^{\theta}}[\nabla_{\theta}F^{\theta}(q)^T(\sigma\sigma^T)^{-1}\nabla_{\theta}F^{\theta}(q)] \ , \tag{3.8}
$$

where $\nabla_{\theta}F^{\theta}(\cdot)$ is a $dN \times K$ matrix containing all the first-order partial derivatives of the force vector (i.e., the Jacobian matrix). Observe that the pathwise FIM does not depend on the perturbation vector, $\epsilon_0$, making pathwise FIM an attractive "gradient-free" quantity for sensitivity analysis. Indeed, the RER for any perturbation can be recovered up to third-order utilizing only the pathwise FIM and (3.7). Moreover, the spectral analysis of $F_{\mathcal{H}}(Q)$ would allow to identify which parameter directions are most/least sensitive to perturbations.

**Example 1: Unknown stationary distribution**: In many molecular systems the steady state is not a Gibbs distribution and typically it is not known explicitly. This is commonplace in non-equilibrium molecular systems such as models with multiple mechanisms, e.g. reaction-diffusion systems, or driven molecular systems. [55, 64] Here we consider such a mathematically simple example, where we assume that the force field consists of two components; one conservative term given as minus the gradient of the potential energy and another term that is not the gradient of a potential function. Mathematically, the force field is given by

$$F^\theta(q) = -\nabla V^\theta(q) + G(q)$$

where we further assume for simplicity that only the conservative term depends on the parameter vector, $\theta$. Since, the resulting Langevin process is at the non-equilibrium regime, the steady states do not admit an explicit form. However, denoting by $\bar{\mu}^\theta$ the unknown stationary distribution of the Langevin process driven by the above forces, the RER is given by

$$\begin{aligned} &\mathcal{H}(Q^\theta|Q^{\theta+\epsilon_0}) = \\ &\frac{1}{2}\mathbb{E}_{\bar{\mu}^\theta}[(\nabla V^{\theta+\epsilon_0}(q) - \nabla V^\theta(q))^T(\sigma\sigma^T)^{-1}(\nabla V^{\theta+\epsilon_0}(q) - \nabla V^\theta(q))] \ . \end{aligned} \tag{3.9}$$

Notice that the expression in the expectation does not depend on the non-conservative forces and it is the same expression as in the equilibrium regime. However, the dependence on the non-conservative forces is evident through the (unknown) stationary distribution, $\bar{\mu}^\theta$.

**Example 2: Inverse temperature perturbation**: Using the fluctuation-dissipation relation, we can substitute the friction parameter $\gamma$ with the inverse temperature $\beta$ and compute the RER and the pathwise FIM for $\beta$ perturbations. Indeed, substituting in eq. (3.6) the relation $\gamma = \frac{1}{2}\beta\sigma\sigma^T$, we are looking for $u(\cdot,\cdot)$ such that

$$\begin{bmatrix} 0 & 0 \\ 0 & \sigma \end{bmatrix} u(q,p) = \begin{bmatrix} 0 \\ -\frac{1}{2}\beta\sigma\sigma^T M^{-1}p + \frac{1}{2}(\beta+\epsilon_\beta)\sigma\sigma^T M^{-1}p \end{bmatrix} ,$$

where $\epsilon_\beta$ is the perturbation of inverse temperature. Notice that the forces were cancelled out in this expression for $u$ because no perturbation is performed in the parameters of the forces. At the stationary regime, RER is then given by

$$\mathcal{H}(Q^\beta|Q^{\beta+\epsilon_\beta}) = \frac{\epsilon_\beta^2}{8}\mathbb{E}_{\mu^\beta}[p^T M^{-1}\sigma\sigma^T M^{-1}p] \ , \tag{3.10}$$

where $\mu^\beta(\cdot)$ is the stationary distribution of the process. It is evident that RER is a quadratic function of the perturbation of the inverse temperature and interestingly enough it depends only on the momenta, $p$. The above formula is valid for any force field and implies that the sensitivity of the (inverse) temperature as quantified by the relative entropy between path distributions is independent of the underlying system as it defined by the forces or by the potential function, $V^\theta(\cdot)$.

Furthermore, in the equilibrium regime where the stationary distribution is given by the Gibbs measure (eq. (3.2)), (3.10) can be further simplified because of the Gaussian nature of the momenta, $p$. Indeed, assuming for simplicity that $M = mI_{dN}$ and $\sigma = \sigma I_{dN}$ with $m, \sigma \in \mathbb{R}$, (3.10) is rewritten as

$$\mathcal{H}(Q^\beta | Q^{\beta+\epsilon_\beta}) = \frac{\epsilon_\beta^2 \sigma^2}{8\beta m} dN \ . \tag{3.11}$$

Consequently, the pathwise FIM in the logarithmic scale (see equation below) is given by

$$F_{\mathcal{H}}(Q^{\log\beta}) = \frac{\gamma}{2m} dN \ . \tag{3.12}$$

**SA in the logarithmic scale**: In many molecular systems, the model parameters may differ by orders of magnitude, thus, it is more appropriate to perform relative perturbations, i.e., the $i$-th element of the perturbation vector is $\theta_i \epsilon_{0,i}$. After straightforward algebra, the elements of the logarithmic-scale Fisher information matrix are given by

$$(F_{\mathcal{H}}(Q^{\log\theta}))_{i,j} = \theta_i \theta_j (F_{\mathcal{H}}(Q^\theta))_{i,j} \ , \quad i, j = 1, \ldots, K \ . \tag{3.13}$$

We refer to Ref. [44] for more details.

**Statistical estimators**: Even though the Langevin equation is degenerate since the noise applies only to the momenta, the process is hypo-elliptic and ergodic under mild conditions on the potential energy, $V(\cdot)$. Therefore, RER and the corresponding pathwise FIM can be computed as ergodic averages. Note though that in order to obtain samples from the Langevin process, a numerical scheme should be employed resulting in errors due to the discretization procedure. There exist several numerical integrators such as BBK and BAOAB for the Langevin equation. [35, 65] In Appendix 3.6.1, BBK integrator is briefly reviewed. The inserted bias is of order $O(\Delta t)$ where $\Delta t$ is the time-step as it has been shown for Langevin equation under compactness condition [66, 67, 68] (e.g., under bounded domain). Then, the statistical estimator for the RER is given by

$$\begin{aligned}\bar{\mathcal{H}}(Q^\theta | Q^{\theta+\epsilon_0}) =& \frac{1}{2n} \sum_{i=1}^{n} \left(F^{\theta+\epsilon_0}(q^{(i)}) - F^\theta(q^{(i)})\right)^T \\ & (\sigma\sigma^T)^{-1} \left(F^{\theta+\epsilon_0}(q^{(i)}) - F^\theta(q^{(i)})\right) \ ,\end{aligned} \tag{3.14}$$

where $n$ is the number of samples and, similarly, the statistical estimator for the pathwise FIM is given by

$$\bar{F}_{\mathcal{H}}(Q^\theta) = \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta F^\theta(q^{(i)})^T (\sigma\sigma^T)^{-1} \nabla_\theta F^\theta(q^{(i)}) \ . \tag{3.15}$$

**Sensitivity Bound**: Relative entropy provides a mathematically elegant and computationally tractable methodology for the parameter sensitivity analysis of Langevin systems. Such an approach focuses on the sensitivity of the entire probability distribution, either at equilibrium or at the path-space level, i.e., for the entire stationary time-series quantifying among others the transferability skills of the molecular models. However, in many situations in molecular simulations, the interest is focused on observables such as radial distribution function, pressure, mean square displacement, etc. Therefore, it is desirable to attempt to connect the parameter sensitivities of observables to the relative entropy methods proposed here. Indeed, relative entropy can provide an upper bound for a large family of observable functions, $g$, through the Pinsker (or Csiszar-Kullback-Pinsker) inequality, [62]

$$|\mathbb{E}_{Q^{\theta+\epsilon_0}_{[0,T]}}[g] - \mathbb{E}_{Q^{\theta}_{[0,T]}}[g]| \le ||g||_\infty \sqrt{2\mathcal{R}(Q^{\theta}_{[0,T]}|Q^{\theta+\epsilon_0}_{[0,T]})} \tag{3.16}$$

where $||\cdot||_\infty$ denotes the supremum (here, maximum) of $g$. In the context of sensitivity analysis, inequality (3.16) states that if the relative entropy is small, i.e., insensitive in a particular parameter direction, then, any bounded observable $g$ is also expected to be insensitive towards the same direction. In this sense, ineq. (3.16) can be viewed as a screening tool for parametric “insensitivity analysis” of observables. Sharper sensitivity bounds than inequality (3.16) were also developed recently. [69] Specifically, the authors showed that for path observables $g$ we have $|\mathbb{E}_{Q^{\theta+\epsilon_0}_{[0,T]}}[g] - \mathbb{E}_{Q^{\theta}_{[0,T]}}[g]| \le \sqrt{\frac{1}{T}Var_{Q^{\theta}_{[0,T]}}[Tg]}\sqrt{\frac{2}{T}\mathcal{R}(Q^{\theta+\epsilon_0}_{[0,T]}||Q^{\theta}_{[0,T]})} + \mathcal{O}(\frac{1}{T}\mathcal{R}(Q^{\theta+\epsilon_0}_{[0,T]}||Q^{\theta}_{[0,T]}))$ . In contrast to (3.16), the latter inequality provides bounds that involve the (time rescaled) variance of observables. We refer to Ref. [69] for other related bounds. Note that the inverse is not necessarily true.[69] If the relative entropy is large for a specific parameter direction, then an observable $g$ may, or may not, exhibit sensitivity with respect to the same parameter direction.

## 3.3 Models and Observables

This section describes the two molecular models discussed here and several observable functions on which the proposed sensitivity analysis method is validated. A prototypical Lennard-Jones fluid model with two force field parameters and a methane model with ten parameters are presented. Observables such as the radial distribution function, the mean square displacement and the pressure spanning from a wide range of model properties are also provided. All simulations are performed under constant number of atoms, volume and temperature (NVT ensemble).

### 3.3.1 LJ fluid model

In order to investigate the sensitivity analysis for a realistic system we examine the LJ fluid model. In this model, the atoms are identical, interacting with the Lennard Jones potential with reduced non-dimensional parameters

$\epsilon_{LJ} = 1, \sigma_{LJ} = 1$. One of the advantages of the LJ fluid is that there exists a phase diagram of the reduced density $\rho^*$ versus the reduced temperature $T^*$. [60] The popularity of this model relies on the generality of systems of molecular liquids that can be described as well as computational efficiency. We restrict the force field interactions in the vicinity of cutoff radius $r_{cut}$. Thus, the (truncated) LJ pair potential is given by

$$V_{LJ}(r_{ij}) = \begin{cases} 4\epsilon_{LJ}\left[\left(\frac{\sigma_{LJ}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{LJ}}{r_{ij}}\right)^{6}\right] \ , & \text{if } r_{ij} < r_{cut} \\ 0 \ , & \text{otherwise,} \end{cases} \tag{3.17}$$

while the total potential energy of the system is

$$V_{LJ}(q) = \sum_{\substack{1 \leq i,j \leq N \\ i<j}} V_{LJ}(r_{ij}) \ ,$$

with

$$r_{ij} = |q_i - q_j| = \sqrt{(q_i^x - q_j^x)^2 + (q_i^y - q_j^y)^2 + (q_i^z - q_j^z)^2} \ ,$$

being the Euclidean distance between the atoms.

Sensitivity analysis is performed on the LJ potential parameters $\epsilon_{LJ}$ and $\sigma_{LJ}$ and as we show later (see section 3.4), the most sensitive parameter is the latter. We consider a system of $N = 2048$ atoms in a cubic simulation box of side length $L = 14.3\sigma_{LJ}$ with periodic boundary conditions (PBC). The reduced temperature of the run is $T^* = 0.85\tau$ which means that the system is in liquid phase (number density $\rho^* = 0.7$). For the numerical scheme, the time-step is $\Delta t = 10^{-3}$ while the length of the run is $10^5$ time-steps. An equilibration period of $10^4$ steps is sufficient for the fcc lattice to melt and standard reduced units are used throughout the simulations.

### 3.3.2 $CH_4$ model

Methane is a more complicated molecule combined of two different types of atoms; carbon (C) and hydrogen (H). Active research is targeted on $CH_4$ because of its environmental impact and energy utilization. [70] Our sensitivity study is expanded and validated on this more complex molecular model which consists of different intermolecular potentials between the pairs of atoms (bonded and non-bonded) as well as additional parameters imposed by the geometry of the molecule (bonds and angles). We define $V(q)$ the total potential and $N$ the total number of atoms (both C's and H's).

$$V(q) = V_{bond}(q) + V_{angle}(q) + V_{LJ}(q) \ . \tag{3.18}$$

where $V_{bond}(q), V_{angle}(q)$ are quadratic intramolecular potential functions of the bonds and angles respectively. $V_{LJ}(q)$ is the non-bonded potential as

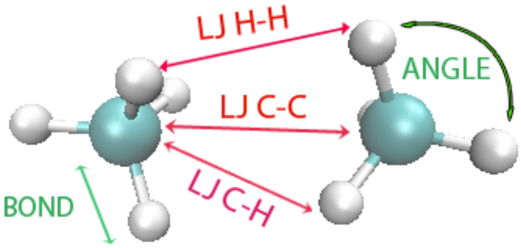

Figure 3.1: Visualization of the $CH_4$ interactions. The site to site non-bonded LJ interactions (intermolecular) are marked in red whereas the intramolecular potential interactions are marked in green.

defined in the previous subsection. For more details concerning the model see Appendix C.

The parameter vector, $\theta$, consists of six LJ parameters (three different LJ potentials depending on the atom type, see also Figure 3.1), two bond parameters and two angle parameters. The parameters values of $CH_4$ are summarized in Table 5.1 whereas the values of the simulation parameters are presented in Table 3.2.

| | $\epsilon_{LJ}[\frac{Kcal}{mol}]$ | $\sigma_{LJ}$ [Å] | $r_{cut}$ [Å] |
|---|---|---|---|
| $C-C$ | 0.0951 | 3.473 | 15.0 |
| $C-H$ | 0.0380 | 3.159 | 15.0 |
| $H-H$ | 0.0152 | 2.846 | 15.0 |

| $K_b$ $[\frac{Kcal}{mol\mathring{A}^2}]$ | $r_0$ [Å] | $K_\theta$ $[\frac{Kcal}{mol\cdot deg^2}]$ | $\theta_0$ [rad] |
|---|---|---|---|
| 700 | 1.1 | 100 | 1.909 |

Table 3.1: Non-bonded $LJ$ coefficients as well as bond and angle coefficients for methane. [61]

| N(molecules) | T [K] | L [Å] | $\rho$ $[\frac{moles}{\mathring{A}^3}]$ | $\gamma$ |
|---|---|---|---|---|
| 512 | 100 | 32.9 | 0.0143 | 0.5 |

Table 3.2: Simulation parameters for $CH_4$

### 3.3.3 Observables

To validate the proposed pathwise SA approach we have calculated various observables that are related to thermodynamical, structural and dynamical properties of the molecular stochastic models. These quantities are experimentally tractable and are related to the microscopic as well as the macroscopic level. In more detail, here, we focus on radial distribution function (RDF), mean square displacement (MSD) and pressure. Other studies [53] in the literature computed observables such as the Helmholtz free energy, density, enthalpy to name some. Despite the fact that the RDF as well as the pressure are equilibrium quantities, MSD is related to the dynamics (time-series averaging) making the proposed pathwise method suitable for such long-time quantities. Note also, that there are no closed analytic expressions for all the above observables with respect to the force field (model) parameters.

### Radial distribution function

The structure of liquids is characterized by the pair radial distribution function, $g(r)$, ($g^{(2)}(r)$ to be more precise) and it is the most important observable of molecular simulations due to the fact that the ensemble average of any pair function may be expressed by it. [32, 71] Furthermore, $g(r)$ can be calculated experimentally by X-ray diffraction. [71] The RDF is the pair distribution function that indicates the normalized distribution of a pair of identical atoms (or molecules) at a given distance. For long intermolecular distance $r$ in liquids, $g(r)$ fluctuates around unity. This static observable is based on the equilibrium structure of the system and it is constructed by histogram averages. For $N$ identical atoms let the two-atom distribution function be

$$P_N^{(2)}(q_1, q_2) := \frac{1}{(N-2)!} \int e^{-\beta V(\mathbf{q})} dq_3 \dots dq_N \ , \tag{3.19}$$

where $q_1, q_2$ are the positions of the first and second atoms kept fixed, irrespective of the configuration of the rest of the particles. For a (homogeneous) liquid, it holds that

$$P_N^{(2)} = \rho^2 g^{(2)}(|r_{1,2}|), \quad \rho = \frac{N}{\mathrm{Vol}} \tag{3.20}$$

where $\rho$ is the number density while $Vol$ is the volume of the simulation box. If the atoms were independent of each other, $P^{(2)}$ would equal $\rho^2$ so in practice $g(r)$ corrects for the spatial (density) correlation between atoms. For the $CH_4$ model, we consider the molecular $g(r)$ which is based on the center of mass of each individual molecules.

### Mean square displacement

The mean square displacement associates the diffusion coefficient, $D$, with the atom (or center of mass for molecules) coordinates and is a measure of the spatial extent of random motion of the Langevin dynamics. It is defined as

$$MSD = \langle (q_t - q_{t_0})^2 \rangle = \mathbb{E}_{Q_{[t_0,t]}} \left[ (q_t - q_{t_0})^2 \right] \tag{3.21}$$

where $q_t, q_{t_0}$ are vectors of particle positions at time $t$ and reference time instant $t_0$, while the brackets, $\langle\cdot,\cdot\rangle$, denote ensemble averaging over all configurations of all the atoms (or molecules). This quantity provides us with information about the dynamical properties of the system. The MSD and the diffusion coefficient, $D$, are related by Einstein's equation

$$2D = \frac{1}{d}\lim_{t\to\infty}\frac{\partial\langle(q_t - q_{t_0})^2\rangle}{\partial t} \tag{3.22}$$

Where $d$ is the dimension of the system (here $d = 3$).

#### Pressure

Temperature and pressure are macroscopic thermodynamic parameters defined in an experimental setup but they can also be defined microscopically. Pressure is given by the expression [32]

$$P = \frac{\rho}{\beta} + \frac{vir}{\mathrm{Vol}} \ ,$$

where the first term is the kinetic energy contribution while $vir$ is the atomic (or molecular) virial given by

$$vir = \frac{1}{3}\sum_{1\le i\le N}\sum_{j>i} F_{ij} r_{ij} \ .$$

Note that $F_{ij}$ is the total force (both non-bonded and bonded in the $CH_4$ case) between atoms (or molecules) $i$ and $j$.

## 3.4 Results

Every model at hand has a domain of applicability; i.e., the forcefield representation allows to calculate (usually thermodynamic) properties of interest in accordance to experimental values within a margin of error. This means that a force field might represent well one property, such as density, but may not be valid for others, or might represent all of them less accurately. In the following we perform simulations where the RER and FIM for each perturbed variable are computed. Discussion on the results as well as validation with respect to the observable quantities defined in section 3.3.3 supports our results.

### 3.4.1 LJ fluid

RER and FIM calculations for the LJ fluid are summarized in Figure 3.2. We compare the RER value using the continuous time statistical estimators, Eqs. 3.14, 3.15. The middle bar corresponds to the FIM-based RER whereas the left and right bars are the values of estimator 3.14 for a negative and positive perturbation by $\epsilon_0 = 5\%$ respectively. All the plots are normalized upon division with the number of particles. As the figure suggests $\sigma_{LJ}$ is the most sensitive parameter. Systems size effects have been thoroughly examined by performing test simulations of bigger systems under the same parameters, which produce similar results to those presented here. It has been shown for a similar model that uncertainty in thermodynamic and transport properties based on the potential parameters is larger than statistical simulation uncertainty. [42]

The corresponding results for the discrete time case using the BBK integrator are shown in the Appendix. There's minor discrepancy of order $O(\Delta t)$ as previously mentioned in section  3.2 due to the discretization error bias. We note here that the continuous time computations are faster since the RER and FIM formulas are less complex.

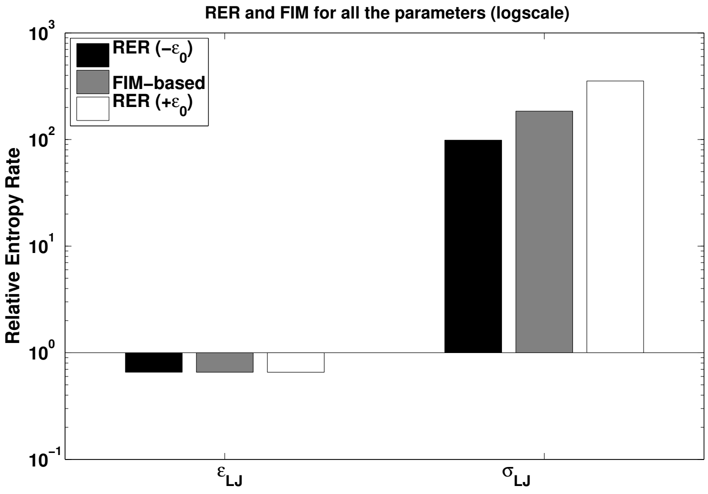

Figure 3.2: RER and FIM of continuous time estimators (3.14), (3.15). Comparing with the discrete time case (supplementary material), the values are almost identical. $\sigma_{LJ}$ is the most sensitive parameter.

Validation on the stronger sensitivity on $\sigma_{LJ}$ compared to $\epsilon_{LJ}$, is demonstrated by the RDF ($g(r)$) plots shown in Figure 3.3. Note that the gradient of the potential, i.e. the interatomic force, depends linearly with respect to $\epsilon_{LJ}$. An increase in this parameter leads to a deeper potential well and stronger attraction between the atoms at the same distance. Thus, as expected, positive perturbation in $\epsilon_{LJ}$ leads to an increase of the first peak in the RDF graph. In addition, only the first peak of the $g(r)$ is affected, the rest of the curve remains the same. Positive (negative) perturbations on the $\sigma_{LJ}$ parameter shift the whole RDF graph due to the fact that the atoms

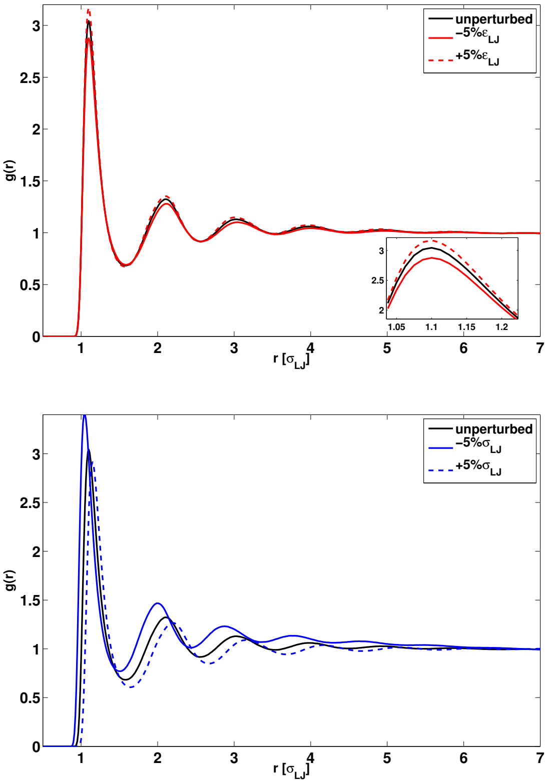

Figure 3.3: Effect of perturbation of $\epsilon_{LJ}$ parameter by ±5% (upper panel) and $\sigma_{LJ}$ parameter by ±5% (lower panel) on RDF. The first peak is shifted vertically for fluctuations around $\epsilon_{LJ}$ whereas it is shifted to the right or left when the fluctuations concern the $\sigma_{LJ}$ parameter. It is clear that $\sigma_{LJ}$ (lower pannel) is more sensitive as the plots differ substantially, which is in agreement with the RE method.

sense greater (weaker) repulsion forces. Hence, the distribution maximum is transferred to a longer (shorter) distance. We also notice that the peak of the curve has increased at the new maximum which can be explained by the finite volume of the same simulation box (NVT ensemble) of the unperturbed system.

In order to get a more detailed insight on the RDFs shown in Figure 3.3 we have also computed the $L_2$ norm, shown in Table 3.3. The $L_2$ norm is suitable for a comparison of the unperturbed versus the perturbed plots $g^{\theta}(r)$ and $g^{\theta+\epsilon_0}(r)$ respectively. As the RER/FIM computations have shown that there is a relative entropy difference of about 2 orders of magnitude with respect to the two potential parameters (Figure 3.2), we now observe a consistent difference, of about 5 times in $L_2$, for the RDF observable. Table 3.3 and Figure 3.3 suggest that the positively and negatively perturbed RDF plots exhibit a more symmetric behavior on the $\epsilon_{LJ}$ parameter than the $\sigma_{LJ}$. Moreover $-5\%\sigma_{LJ}$ changes the packing of the LJ fluid completely; all the density distribution peaks are moved to a shorter distance. This result is consistent with Pinsker's inequality ( refPinsker) as the more sensitive direction allows for greater differences in the expected values of the observables.

Opposite perturbation directions yield different RER values whereas this is not the case for the FIM based RER which is a second-order (quadratic) approximation. In one of the realizations in our example, RER for $+(-)5\%\sigma_{LJ}$ is 360.7 (101.1) and FIM is 196.6 meaning that $\mathcal{H}(Q^{\theta}|Q^{\theta \pm \epsilon_0})$ is not symmetric FIM and the negative direction being more sensitive.

There is no analytic formula that relates the MSD to the potential parameters but we expect that a larger deviation will result upon perturbation of a more sensitive parameter. As we can see in Figure 3.4, the line for the insensitive perturbed parameter $\epsilon_{LJ}$ slightly differs from the black one, both for positive and negative $\epsilon_0$. On the contrary, the line that corresponds to the increased $\sigma_{LJ}$ is further away and under the unperturbed one. Based on the aforementioned discussion on $g(r)$ this is reasonable, as an increase in the $\sigma_{LJ}$ values leads to stronger repulsive forces at the same distance, hence more atom collisions and consequantely to a larger friction coefficient, i.e. lower mobility of the LJ atoms. A decrease in $\sigma_{LJ}$ lowers the interatomic repulsive forces and there's no significant effect at this density because the random forcing dominates the dynamics. This result is consistent with Pinsker's inequality as it provides an upper bound only, meaning that although this parameter is indicated as more sensitive (bigger RER value on the r.h.s.) the expected value with respect to this observable slightly changes upon perturbation. Additional runs (realizations of the Markov chain starting from different configurations) for the same negative $\sigma_{LJ}$ direction have shown that the errorbars are within 2.5%. The linear dependence of the interatomic forces with respect to $\epsilon_{LJ}$ accounts for increased (decreased) interatomic interaction strength when this parameter is changed upwards (downwards).

| perturbation | $\|\|g^{\theta}(r) - g^{\theta+\epsilon_0}(r)\|\|_{L_2}$ | $\frac{\|g^{\theta}\|-\|g^{\theta+\epsilon_0}\|}{\|g^{\theta}\|}$ | RER |
|---|---|---|---|
| $+5\%\epsilon_{LJ}$ | 0.049 | 0.8 % | 0.79 |
| $-5\%\epsilon_{LJ}$ | 0.066 | -1.17% | 0.79 |
| $+5\%\sigma_{LJ}$ | 0.47 | -3.83 % | 409 |
| $-5\%\sigma_{LJ}$ | 0.59 | 7.4 % | 115 |
| $r_{cut} = 1.6\sigma_{LJ}$ | 0.189 | -3.44 % | 0.71 |
| $r_{cut} = 7\sigma_{LJ}$ | 0.01 | 0.19% | $1.6 \times 10^{-4}$ |

Table 3.3: $L_2$ norm of the difference of the unperturbed minus the perturbed g(r) and normalized area difference.

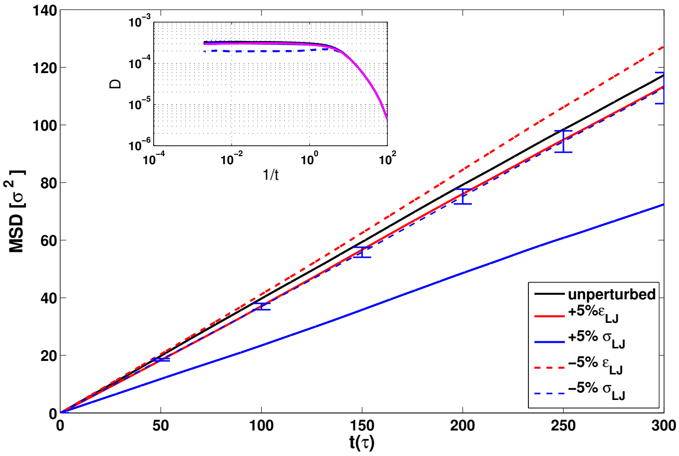

Figure 3.4: MSD for different perturbed directions by ±5%. The $\epsilon_{LJ}$ parameter has a small impact in comparison with the more sensitive $\sigma_{LJ}$. The MSD plot for the positive perturbation of $\sigma_{LJ}$ stands out as the increased collisions dominate the random forcing. Errorbars indicate the standard deviation for $-5\%\sigma_{LJ}$ and the deviation propagates with time. The inset illustrates the diffusion coefficient difference in logscale.

Table 3.4 contains the diffusion coefficient $D$ (from eq. 3.22) related to Figure 3.4 and depicts a quantitative aspect. The perturbation direction of $+5\%\sigma_{LJ}$ is dominant and clearly results in slowing down the diffusion of the LJ fluid particles.

As mentioned above for simulation of the LJ fluid standard non-dimensional (reduced units) are used. The reduced pressure is denoted by $P^*$ and Table 3.4 contains the simulation results. Once more we observe a greater influence in perturbation of the parameter $\sigma_{LJ}$ especially for a positive increase. This is consistent with the fact that the volume remains unchanged and the repulsive forces increase as discussed earlier, giving a pressure rise of fourteen times more for a +5% perturbation. We observe the opposite fact for a reduction in $\sigma_{LJ}$. The $\epsilon_{LJ}$ parameter has a more symmetric influence and this is explained by the linear increase in the forces (derivative of the potential formula) between the atoms and consequently via the virial coefficient it is depicted at the pressure. The third column of Table 3.4 compares the relative pressure change with respect to the unperturbed run and the pressure standard deviation is on the column that follows. Parameter $\sigma_{LJ}$ alters the pressure by an order of magnitude, a result which is consistent with the RER/FIM calculations in the previous subsection.

### Discontinuous model parameter: cutoff radius

The $r_{cut}$ is a parameter of the model but the potential is not differentiable with respect to it. Hence we can compute the relative entropy rate but we cannot have an estimate of the Fisher Information matrix because the computation of FIM involves products of partial derivatives (see also Eq. 3.15). Figure 3.5 summarizes the quantity $\mathcal{R}({r_{cut}}^{ref}|r_{cut})$ per particle, where in this notation we mean that the RER integral differential is with respect to the path space measure corresponding to the model's $r_{cut}$ as reference. The potential tends to zero at distance $r_{cut} = 2.5\sigma_{LJ}$, that is a typical value also used in the literature, so information lost upon trimming the potential tail is small in comparison to that when $r_{cut}$ is shifted to the left. We expected that the RER should be higher for a negative perturbation of $r_{cut}$ as validated in Figure 3.5 and the asymmetry (exponential form for negative perturbation) comes from the formula (plot) of the potential; more information regarding the attractive part is lost rapidly for a $-10\%$ reduction step from the reference $r_{cut} = 4\sigma_{LJ}$. Indeed the trick of the $r_{cut}$ convention has been used in molecular simulations in order to reduce the computations at the expense of minimal information loss, so our results using this pseudo-metric indicate that our choice of $r_{cut}$ is suitable. Additional runs for an increase in $r_{cut}$ suggest a trivial gain of information based on the $\mathcal{H}(Q^{r_{cut}^{ref}}|Q^{r_{cut}})$ value as well as the RDF (see next).

The RDF plot changes with a change in $r_{cut}$ as shown in Figure 3.6. When the potential tail is restricted up to $r_{cut}$, the long-range attractive part is zero after that distance. This results to weaker long-range attractive forces (loss of cohesive energy) hence the first peak in the RDF graph is lower and the mass is distributed to the right. We have included the plot of a 60% decrease to illustrate the higher dependence on a "premature" truncation and a plot of 75% increase for comparison. The empirical value of $2.5\sigma_{LJ}$ is adequate for simulations, but a further reduction to 1.6 $\sigma_{LJ}$ results to huge

| perturbation | $P^*$ | $\frac{P^*_{\theta+\epsilon_0}-P^*_{\theta}}{\lvert P^*_{\theta}\rvert}$ | $\sigma_{STD}$ | $D[\frac{\sigma_{LJ}^2}{\tau}]$ |
|---|---|---|---|---|
| unperturbed | 0.11 | - | 0.25 | $3.2 \times 10^{-4}$ |
| $+5\%\epsilon_{LJ}$ | -0.10 | -1.92 | 0.26 | $2.9 \times 10^{-4}$ |
| $-5\%\epsilon_{LJ}$ | 0.28 | 1.56 | 0.23 | $3.4 \times 10^{-4}$ |
| $+5\%\sigma_{LJ}$ | **1.71** | **14.34** | 0.6 | $\mathbf{1.8 \times 10^{-4}}$ |
| $-5\%\sigma_{LJ}$ | -0.47 | -5.24 | 0.12 | $3.04 \times 10^{-4}$ |

Table 3.4: (left)Pressure change with respect to different perturbation directions of the LJ fluid parameters. $\sigma_{LJ}$ is the most sensitive direction. (right)Diffusion coefficient of the MSD plots. The errorbars are within ±2.5%.

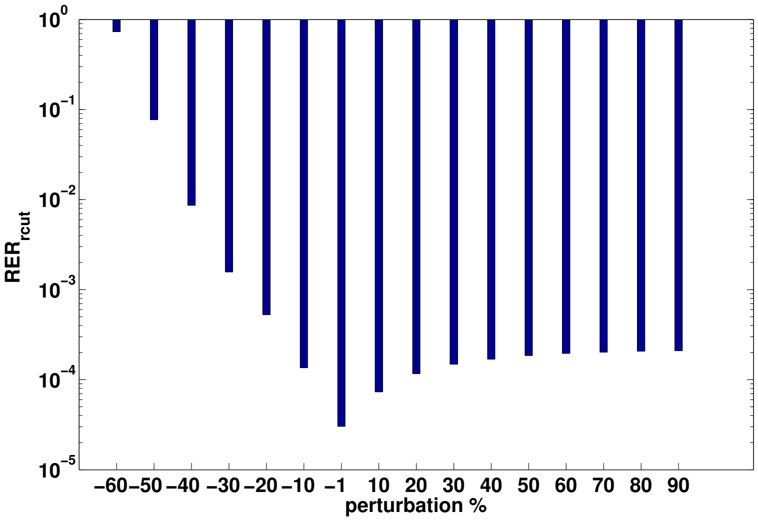

Figure 3.5: RER per particle for different $r_{cut}$ values in logscale. Perturbation of -10% corresponds to the 90% of the reference $r_{cut} = 4\sigma_{LJ}$. There is significant loss of information when we restrict the potential tail ($r_{cut}$) to less than one half, as that value is near the minimum of the potential well and a fraction of the attractive forces is lost.

loss of information, especially for the attractive part. On the contrary, if we almost double the reference value of $4\sigma_{LJ}$ to $7\sigma_{LJ}$, the gain is minimal and this can be seen in Figures 3.5 and 3.6.

We have seen here that the influence of this parameter is minimal in comparison with the potential parameters in Figure 3.2 for this reference value in terms of RER. RER per parcticle for a $-5\%\epsilon_{LJ}$ pertrubation is similar to a $-60\%$ redution in $r_{cut}$. The $L_2$ norm of the g(r) difference for different $r_{cut}$ values (Table 3.3 and Figure 3.6 ) illustrate the same behavior too. At this point we should stress that the sensitivity of the observables on $r_{cut}$ changes if we choose another reference value; however in practice usually $r_{cut}$ is not one of the parameters tuned during the force field development/optimization.

### Non-equilibrium regime LJ fluid

Finally, we have also studied a non-reversible LJ fluid. In more detail, we have checked the effect of an additional non-gradient term in the force in the y-direction i.e. $F^{\theta}(q) = -\nabla V^{\theta}(q) - G(q), G(q) = [0, \alpha, 0, 0, \alpha, 0, ..., 0, \alpha, 0]^T$. $\alpha = 1$ for the irreversible case and the term $G(q)$ is divergence-free. The eigenvalues and dominant eigenvectors are summarized in Table 3.5 and the corresponding RDF plot is given for comparison in Figure 3.7. We

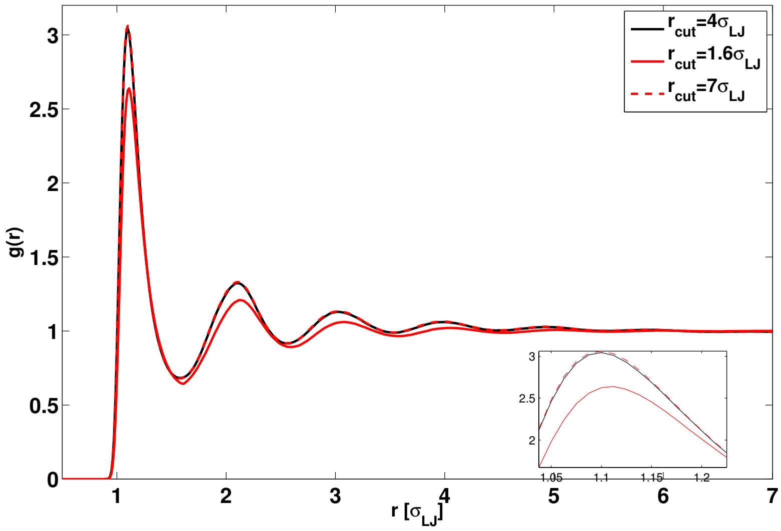

Figure 3.6: LJ fluid $g(r)$ for different $r_{cut}$ values. Bigger $r_{cut}$ results to longer range attractive forces binding the atoms closer (higher peak). As the $L_2$ norm quantifies, the influence of almost double $r_{cut}$ value ($4\sigma_{LJ}$ is considered as reference), slightly affects the plot and Pinsker inequality validates this fact. On the other hand, a decrease of this parameter leads to loss of information and the corresponding $g(r)$ describes a completely different model. Note that for this plot we increased the system size as the simulation box dimensions restrict the maximum value of $r_{cut}$.

expected that despite the fact that this process has a different measure close to the stationary measure of the reversible one, the extra non-gradient term cancels out in eq. (3.6). Hence our results as expected are similar but we have demonstrated that the method is general and can be used for a process equipped with a steady state measure. We aim to the study of more complex systems in non-equilibrium [64] in future work.

| $\alpha = 0$<br>eigenvalues | eigenvector | $\alpha = 1$<br>eigenvalues | eigenvector |
|---|---|---|---|
| $9.434 \times 10^4$ | 0.062 | $1.012 \times 10^5$ | 0.0621 |
| $4.33 \times 10^{11}$ | 0.998 | $4.48 \times 10^{11}$ | 0.998 |

Table 3.5: eigenvalues and eigenvectors for the non-reversible case

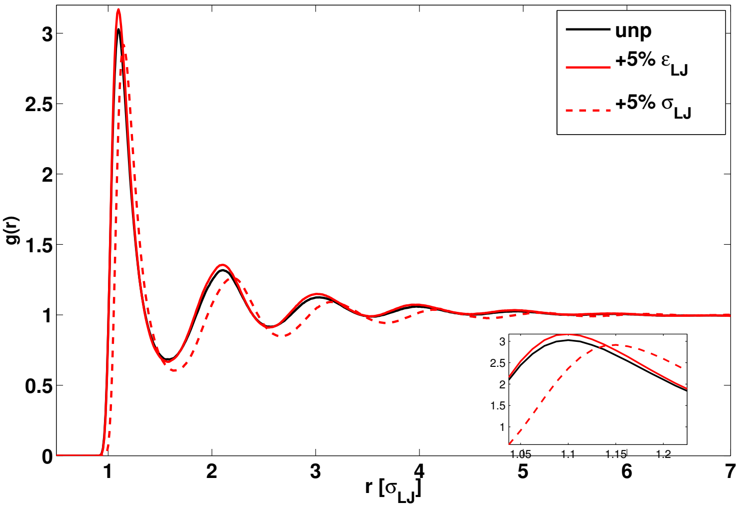

Figure 3.7: RDF plot for the irreversible case. The norm values are: $||g^{\theta+5\%\epsilon_{LJ}}(r) - g^{\theta}(r)||_{L^2}$=0.058, $||g^{\theta+5\%\sigma_{LJ}}(r) - g^{\theta}(r)||_{L^2}$=0.473

### 3.4.2 $CH_4$

In the following we discuss calculations of RER-FIM as well as various observables for the all-atom methane liquid. FIM and RER calculations are summarized in Figures 3.8 and 3.9. In more detail, Figure 3.8 shows that the RER values vary orders of magnitude for the various parameter perturbations, hence we grouped them in four panels a)-d). In Figure 3.9 the FIM-based RER data is plotted in logscale for comparison. We note here that due to the uneven number of the different pairs of $C-C$, $C-H$, $H-H$ we have divided with 8 and 16 the quantities corresponding to the second and third type of pairs in order to obtain comparable plots. All RER values are normalized with the number of corresponding interactions. Furthermore, bigger systems consisting of 4000 molecules conclude with identical results. As in the LJ paradigm, we can see a greater sensitivity on the $\sigma_{LJ}$ parameters instead of the corresponding $\epsilon_{LJ}$ ones. The errorbars indicate that the variance of the estimators were small and that a positive perturbation increased the value of the RER with respect to the FIM-based RER estimate. Clearly the most sensitive parameter is the $C-H$ bond length $r_0$ followed by the bending angle $\theta_0$.

The fact that $r_0$ and $\theta_0$ are more sensitive is not surprising if we consider that the type of all harmonic potentials is very steep. $K_b, K_\theta$ constants are of the order $\mathcal{O}(10^2 - 10^{-3})$ as obtained from more detailed (ab initio) calculations or from fittings of experimental data (see Table 3.2). These constants are part of the $\nabla V_{bond}, \nabla V_{angle}$ which is contained in the estimators. The asymmetry in the $\sigma_{LJ}$ RER values in comparison to the FIM values (panel b in Fig 3.8) is explained by the third order term contribution in the expansion of RER. A rigorous calculation in Appendix B shows that this term includes the Hessian of the gradient of the potential with respect to the parameters and is non-zero for $\sigma_{LJ}$.

### Observables

We perform the same observable computations as with the LJ fluid model in order to validate the predicted sensitivity of the parameters provided by the RER and pathwise FIM methods. Although we have performed simulations for various values of the parameters, we chose 5% as a suitable value for better representation of our results. Note that in principe parameter sensitivities change as we change phase space point; in higher temperatures or low densities each observable is affected differently and our proposed RE method incorporates this behavior through the force differences (eq. 3.6). Here we have performed simulations in the temperature range from 80 to 180 K and qualitatively similar results were observed. A more detailed study of SA over various temperatures of more complex (macromolecular) systems will be the subject of a future work.

As in the case of the RDF of the LJ fluid, an increase in the $\sigma_{LJ}$ parameters shifts the graphs to the right (Figure 3.10) due to the repulsive forces. All the differences with respect to the $L_2$ norm are summarized in Table 3.6 for clarity.

In addition, from the set of RDF data presented in Figure 3.10, an increase in $\sigma_{LJ}^{C-H}$ values results to larger deviations. As we keep the volume fixed, the contribution of the $C-H$ interactions in the packing is larger than that of the $C-C$ pairs because of the larger number of $C-H$ pairs. Following this graph is the one involving $\sigma_{LJ}^{H-H}$ increase because of the even smaller numerical value in comparison to the other $\sigma_{LJ}$'s. At this point the smaller mass of the hydrogens is the reason although the number of pairs (hence interactions) is the largest.

The MSD plots indicate the $\sigma_{LJ}^{CH}, \sigma_{LJ}^{HH}$ as the most sensitive parameters. An increase in $\sigma_{LJ}$ results to increased collisions and smaller diffusion coefficient (smaller MSD) as can be seen in Figure 3.11. As in the LJ case, positive $\sigma_{LJ}$ perturbations (for all three types) result to greater repulsive forces, hence reduced diffusivity. $\epsilon_{LJ}$ variations slightly affect the MSD with respect to the other parameters and the same holds for $K_b$ and $K_\theta$ too (we have omitted the plots for brevity). Under this dynamic observable the intramolecular interactions are less relevant than the intermolecular ones, for the specific state point (temperature and density) studied here.

Pressure calculations for different perturbation directions are summa-

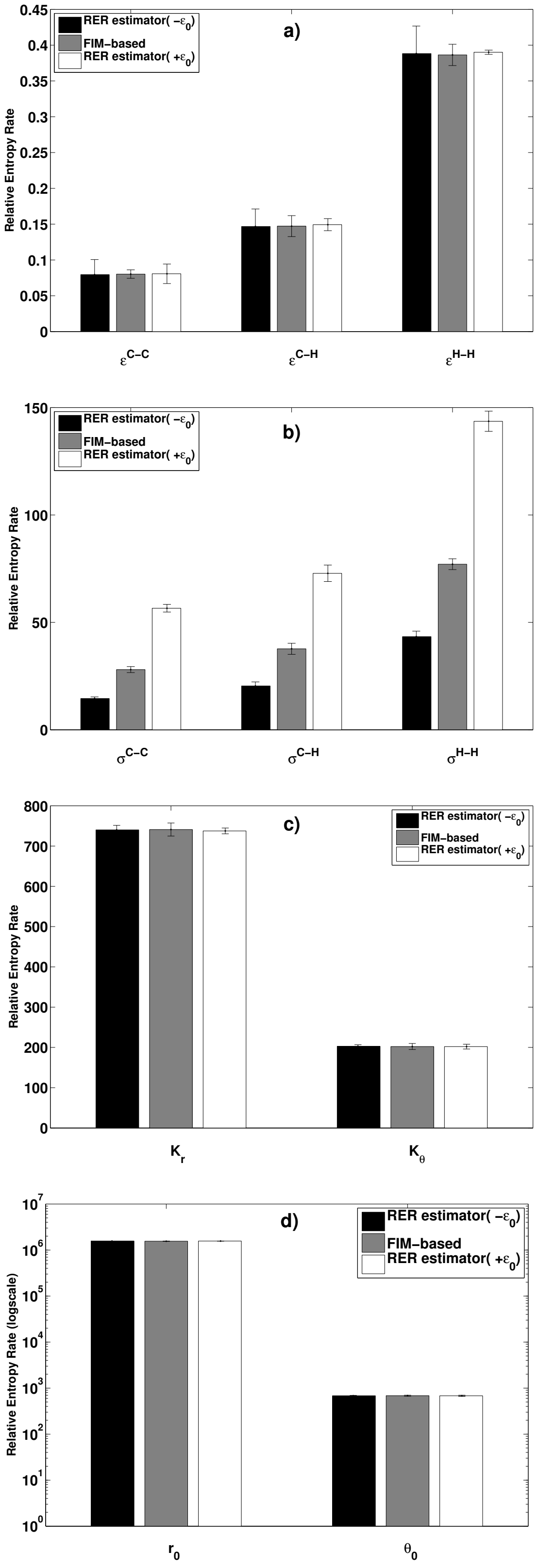

Figure 3.8: $CH_4$ per molecule RER-FIM comparison with error bars using the two different estimators for ±5% perturbations in all the parameters. Non-bonded (a and b) and bonded (c and d) potential parameters are shown. The parameters are grouped according to their order of magnitude. The most sensitive one is $r_0$ followed by $\theta_0$ and there has been a minor scaling according to the number of atom-atom pairs.

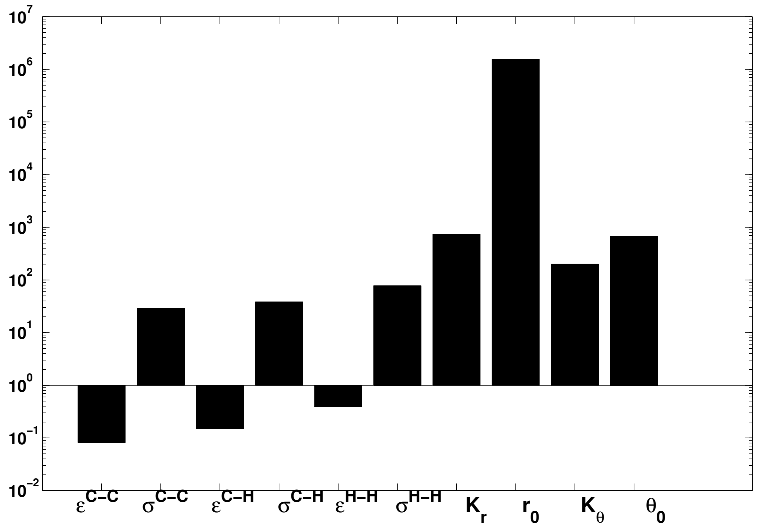

Figure 3.9: $CH_4$ FIM-based RER comparison for ±5% perturbations in logscale.

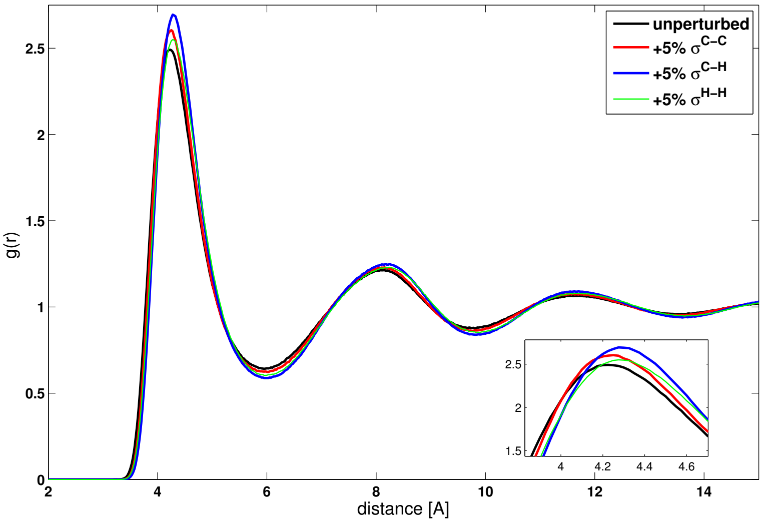

Figure 3.10: $CH_4$ molecular g(r) for +5% perturbations on $\sigma_{LJ}$. The tail of the plot varies slightly hence the zoomed region differs more. As in the LJ fluid case, the $\sigma_{LJ}$ defines the shift of the curve horizontally.

| perturbation | $\|\|g^{\theta}(r) - g^{\theta+\epsilon_0}(r)\|\|_{L_2}$ | $\|\|g^{\theta}(r) - g^{\theta-\epsilon_0}(r)\|\|_{L_2}$ |
|---|---|---|
| $\epsilon^{C-C}$ | $1.0 \times 10^{-2}$ | $1.2 \times 10^{-2}$ |
| $\sigma^{C-C}$ | $\mathbf{1.1 \times 10^{-1}}$ | $5.7\times10^{-2}$ |
| $\epsilon^{C-H}$ | $1.7 \times 10^{-2}$ | $9.6 \times 10^{-3}$ |
| $\sigma^{C-H}$ | $\mathbf{2.8 \times 10^{-1}}$ | $\mathbf{1.7 \times 10^{-1}}$ |
| $\epsilon^{H-H}$ | $1.4 \times 10^{-2}$ | $1.15 \times 10^{-2}$ |
| $\sigma^{H-H}$ | $\mathbf{2.05 \times 10^{-1}}$ | $\mathbf{1.6 \times 10^{-1}}$ |
| $K_b$ | $1.1 \times 10^{-2}$ | $9.7 \times 10^{-3}$ |
| $r_0$ | $\mathbf{1.01 \times 10^{-1}}$ | $\mathbf{1.1 \times 10^{-1}}$ |
| $K_\theta$ | $8.7 \times 10^{-3}$ | $8.2 \times 10^{-3}$ |
| $\theta_0$ | $9 \times 10^{-3}$ | $9 \times 10^{-3}$ |

Table 3.6: $L_2$ norm of the difference of the unperturbed minus the perturbed g(r) for ±5% perturbation.

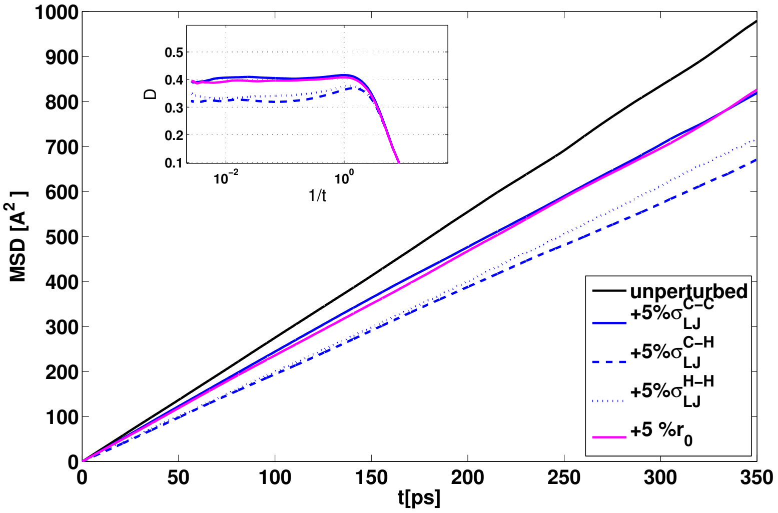

Figure 3.11: CH4 MSD for 5% perturbations. We have summarized the most important directions. With respect to this observable, the most sensitive parameter is $\sigma^{C-H}$ followed by $\sigma^{H-H}$. This is in accordance with the RER in Figure 3.9. The inset illustrates the diffusion coefficient differences in logscale.

| perturbation | $P^{\theta+\epsilon_0}$[atm] | $\frac{P^{\theta+\epsilon_0}-P^{\theta}}{\|P^{\theta}\|}$ | $\sigma_{STD}$ | $P^{\theta-\epsilon_0}$[atm] | $\frac{P^{\theta-\epsilon_0}-P^{\theta}}{\|P^{\theta}\|}$ |
|---|---|---|---|---|---|
| unperturbed | 19.7 | - | 58.4 | - | - |
| $\epsilon^{C-C}$ | -3.9 | -1.2 | 51.7 | 33.1 | +0.7 |
| $\sigma^{C-C}$ | -2.3 | -1.1 | 53.1 | 87.4 | +3.4 |
| $\epsilon^{C-H}$ | -31.3 | -2.6 | 52.2 | 63.6 | +2.2 |
| $\sigma^{C-H}$ | **177.2** | **+8** | 56.7 | 44.3 | +1.2 |
| $\epsilon^{H-H}$ | 8.27 | -0.6 | 49.1 | 23.7 | +0.2 |
| $\sigma^{H-H}$ | **437** | **+21.2** | 56.3 | **195** | **+10.9** |
| $K_b$ | 15.3 | -0.2 | 49.6 | 14.3 | -0.3 |
| $r_0$ | **281** | **+13.3** | 56.5 | **217** | **+12** |
| $K_\theta$ | 18.6 | -0.05 | 52.9 | 14.9 | -0.2 |
| $\theta_0$ | 13.7 | -0.3 | 52.45 | 13.3 | -0.3 |

Table 3.7: Pressure for ±5% perturbation of different directions and the corresponding standard deviation. The most sensitive parameters $r_0$ and $\sigma_{LJ}^{H-H}$ increase the pressure substantially.

rized in Table 3.7. According to this observable quantity $\sigma_{LJ}^{H-H}$ and $r_0$ are the most sensitive parameters, which are also indicated by the RE methods. As in the case of the LJ fluid, a change in $\epsilon_{LJ}$ (in all pair types) affects the pressure less than a change in $\sigma_{LJ}$. Pressure rises through an increase in $\sigma_{LJ}$ due to more atom collisions. Additionally, stronger forces account for a higher pressure virial. The presented results are in accordance with earlier work [72] on an LJ model of water, in which sensitivity analysis using partial derivatives of observables with respect to the parameters were used. That study also demonstrated that pressure is greatly affected by variations in $\sigma_{LJ}$ and classified the bond length, $\sigma_{LJ}$ and the bond constant as the most sensitive ones.

A change in the bending angle $\theta_0$ does not affect the pressure [73] as well as the impact of the constants $K_b, K_\theta$ on the pressure is minimal. We note that the unperturbed system pressure is higher than 1atm because the model we chose (forcefield and integrator) does not reproduce the whole $CH_4$ phase diagram precisely. Such small deviations from the equations of state and experiments are expected. We refer to the supplementary materials for more figures and results which were omitted here for brevity.

From the computational efficiency point of view, the main computational load is another force calculation, for each perturbed parameter, on every step. This cost is of the order of $k$ parallel simulations, where $k$ is the number of parameters to be perturbed. The FIM requires another $k$ force evaluations (partial derivative with respect to $\theta_k$ of the force) per step and one cheap matrix multiplication with the perturbation vector $\epsilon_0$. The computational advantage of the FIM at this point is that it is independent of the values of $\epsilon_0$, meaning that if we wanted to change the nominal parameter values

$\theta$ by a new $\epsilon_0$, we just have to multiply FIM by that vector. For the $CH_4$ system, a trajectory of 20000 timesteps (sample every 40 timesteps) was sufficient for our RER/FIM calculations whereas the $g(r)$ required circa 100000 timesteps (50 ps). Moreover, the RE methodology can be seen as a screening tool for "insensitivity" analysis meaning that if the system is insensitive towards a perturbation direction, we do not have to estimate observable quantities through expensive long runs. Unlike the widely used finite differences method which requires observable estimations of two (or even worse, the average of two) perturbed directions, our method requires only two short *observable independent* runs.

## 3.5 Conclusion

In this paper we present a parametric SA approach for complex stochastic molecular systems. The focus was set particularly to the Langevin equation, however, it is applicable to any molecular system that can be described by a system of stochastic differential equations. The presented SA approach is an extension of the work in Ref. [44] and is based on the relative entropy per unit time of the path distribution at a reference parameter point with respect to the path distribution at a perturbed parameter point.

Major advantages of this method are that: i) it is capable of handling non-equilibrium steady state systems, ii) it is independent of the numerical scheme, iii) it is computationally tractable through the expansion of the RER which results in the pathwise FIM. Pathwise FIM provides a fast "gradient-free" method for parametric SA since it provides an estimate –up to third-order accuracy– of the RER for different perturbation directions through a simple matrix multiplication. iv) it is based on the continuous time SDE.

We examined two systems; the well-known prototypical LJ fluid and a more complex one: methane ($CH_4$). SA on the LJ fluid system was based on the potential parameters $\epsilon_{LJ}, \sigma_{LJ}$ with the latter being the more sensitive to perturbations whereas $CH_4$ involved 6 intermolecular and 4 intramolecular potential parameters with the intramolecular parameters being the most sensitive in terms of RER.

For the validation of our proposed SA approach various observables have been monitored: static and dynamic observable quantities such as the radial distribution function, the mean square displacement and the pressure. Theoretical justification of the SA approach is also provided through the Pinsker inequality. We also investigated the effect of the potential cutoff radius, $r_{cut}$, by numerically computing the RER showing first that RER can be used as an information criterion for assigning appropriate values to parameters of the system and second that 5% perturbation of $\sigma_{LJ}$ produce greater impact than changing $r_{cut}$ from $4\sigma_{LJ}$ to $1.6\sigma_{LJ}$.

We highlight physics-driven limitations of the FIM method; i) higher order corrections (with respect to the expansion of RE in $\epsilon_0$) may be needed depending on the physical system at study (density, temperature, phase etc). ii) no FIM estimates for discontinuous parameters with respect to the derivatives are available (see $r_{cut}$ in sec. IV.A.1) iii) FIM provides information on the sensitivity of a parameter but cannot indicate if a positive or a negative perturbation is more substantial.

As far as computational efficiency is concerned, the RER estimation requires only one additional force evaluation per perturbation direction and it is observable independent unlike other SA methods. In addition, less steps are required for an accurate estimate than the typical number of steps needed for the calculation of an observable. Moreover the partial derivatives of the conservative forces with respect to the parameter vector $\theta$ are needed for the FIM estimation where the tradeoff is the independence of the nominal value of the perturbation vector $\epsilon_0$.

Finally, RE for high-dimensional systems was used as a measure of loss of information in coarse-graining. [74, 75, 25] Coarse-graining (CG) methods of stochastic systems allow for constructing optimal parametrized Markovian coarse-grained dynamics within a parametric family, by minimizing the information loss (i.e., the relative entropy) on the path space. Application of RE to the error analysis of coarse-graining of stochastic particle systems have been pioneered in these papers. [76, 77, 78] Recent ongoing work on application of the RE framework for CG in the non-equilibrium regime where there's no Gibbs structure can be found in Ref. [79]. We aim to utilize the current SA method to tackle with more complex hybrid macromolecular materials or biomolecular systems in and out-of equilibrium conditions. [80, 59, 58] Another goal is to adapt the RE method to quantify and indicate the most efficient CG interaction potential of mesoscale simulations. [81]

## 3.6 Appendix

### 3.6.1 Pathwise SA at the discrete-time level

In Section 3.2, we perform SA by first deriving RER and the corresponding pathwise FIM for the continuous-time stochastic Langevin process and then discretizing the process to get numerical estimates for these quantities.[82] We can reverse the order of SA and first discretize the Langevin process and then derive the RER and the pathwise FIM. Here, the latter approach is presented using the BBK algorithm as a numerical integrator of the Langevin process which defines a discrete-time Markov chain. A preliminary example of this approach can be found in Ref. [44]. In the BBK integrator, the Hamiltonian part of the Langevin equation (3.1) is integrated with the Verlet propagator whereas the thermostat is an Ornstein-Uhlenbeck process and the explicit/implicit propagator is used.

The BBK algorithm [35] reads

$$\begin{cases} p_{i+\frac{1}{2}} = p_i - \nabla V(q_i)\frac{\Delta t}{2} - \gamma M^{-1} p_i \frac{\Delta t}{2} + \sigma \Delta W_i \\ q_{i+1} = q_i + \Delta t M^{-1} p_{i+\frac{1}{2}} \\ p_{i+1} = p_{i+\frac{1}{2}} - \nabla V(q_{i+1})\frac{\Delta t}{2} - \gamma M^{-1} p_{i+1} \frac{\Delta t}{2} + \sigma \Delta W_{i+\frac{1}{2}} \end{cases} \tag{3.23}$$

$\Delta W_i, \Delta W_{i+\frac{1}{2}}$ are iid Gaussian random vectors with zero mean and covariance matrix $\frac{\Delta t}{2} I_{dN}$ while $\Delta t$ is the time step of the numerical scheme. Notice that other choices of numerical integrators can be utilized such as the ones proposed by Leimkuhler et al. [83, 65] which introduce a relatively weak perturbative effect on the physical dynamics.

We define the state of the discrete-time system at time-step $i$ as $z_i = (q_i, p_i) \in \mathbb{R}^{2dN}$. The process $\{z_i\}_{i=0}^M$ for the BBK integrator is a Markov chain with transition probability $P^\theta(z_i, z_{i+1})$ where $\theta \in \mathbb{R}^K$ is the vector of the system's parameters. Notice that the length of the discrete-time process is related with the time window of the continuous-time process through $T = M\Delta t$. The path space probability density, $\bar{Q}^\theta_{0:M}(\cdot)$, is defined as

$$\bar{Q}^\theta_{0:M}(\{z_i\}_{i=0}^M) = \bar{\mu}^\theta(z^0) \prod_{i=0}^{M-1} P^\theta(z_i, z_{i+1}) \ , \tag{3.24}$$

where $\bar{\mu}^\theta(\cdot)$ denotes the stationary distribution of the discrete-time. As in the continuous-time case, we perturb the parameter vector, $\theta$, by adding a perturbation vector $\epsilon_0 \in \mathbb{R}^K$. At the stationary regime, the pathwise relative entropy of $\bar{Q}^\theta_{0:M}$ with respect to $\bar{Q}^{\theta+\epsilon_0}_{0:M}$ admits also a decomposition into a linear in time term plus a constant. [44] Indeed, it holds that

$$\mathcal{R}(\bar{Q}^\theta_{0:M} | \bar{Q}^{\theta+\epsilon_0}_{0:M}) = M\mathcal{H}(\bar{Q}^\theta | \bar{Q}^{\theta+\epsilon_0}) + \mathcal{R}(\bar{\mu}^\theta | \bar{\mu}^{\theta+\epsilon_0}) \ , \tag{3.25}$$

where $\mathcal{R}(\bar{\mu}^\theta | \bar{\mu}^{\theta+\epsilon_0})$ is the relative entropy between the stationary distributions while $\mathcal{H}(\bar{Q}^\theta | \bar{Q}^{\theta+\epsilon_0})$ is the RER of the discrete-time Markov chain given by

$$\mathcal{H}(\bar{Q}^\theta | \bar{Q}^{\theta+\epsilon_0}) = \mathbb{E}_{\bar{\mu}^\theta}\left[ \int_{\mathbb{R}^{2dN}} P^\theta(z, z') \log \frac{P^\theta(z, z')}{P^{\theta+\epsilon_0}(z, z')} dz' \right] . \tag{3.26}$$

The discrete-time RER is related with the continuous-time RER through [81]

$$\mathcal{H}(Q^\theta | Q^{\theta+\epsilon_0}) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathcal{H}(\bar{Q}^\theta | \bar{Q}^{\theta+\epsilon_0}) \ . \tag{3.27}$$

As expected, discrete-time RER is locally a quadratic functional in a neighborhood of $\theta$ hence its curvature around $\theta$, defined by the Hessian, is the pathwise FIM which is given by [44]

$$\begin{aligned} &F_{\mathcal{H}}(\bar{Q}^\theta) = \\ &\mathbb{E}_{\bar{\mu}^\theta}\left[ \int_{\mathbb{R}^{2dN}} P^\theta(z, z') \nabla_\theta \log P^\theta(z, z') \nabla_\theta \log P^\theta(z, z')^T dz' \right] . \end{aligned} \tag{3.28}$$

We refer to ([44]) for statistical estimators of the discrete-time RER and the corresponding pathwise FIM while in Supplementary Materials we provide detailed formulas for the numerical calculation of (3.26) and (3.28) for the BBK integrator.

### 3.6.2 Expansion of the continuous-time RER

We now expand the RER in eq. (3.6) through Taylor series expansion around the point $\theta$. We start with expanding the $m$-th component of the force, $F^{\theta+\epsilon_0}(q)$, around $\theta$

$$F_m^{\theta+\epsilon_0}(q) = F_m^{\theta}(q) + \nabla_\theta F_m^{\theta}(q)\epsilon_0 + \frac{1}{2}\epsilon_0^T \nabla_\theta^2 F_m^{\theta}(q)\epsilon_0 + \mathcal{O}(|\epsilon_0|^3) \tag{3.29}$$

where $\nabla$ denotes the $1 \times K$ gradient vector while $\nabla^2$ denotes the $K \times K$ Hessian matrix. Then, the RER is written as

$$\begin{aligned}
&\mathcal{H}(Q^\theta|Q^{\theta+\epsilon_0}) \\
&= \frac{1}{2}\mathbb{E}_{\mu^\theta}[(F^{\theta+\epsilon_0}(q) - F^\theta(q))^T(\sigma\sigma^T)^{-1}(F^{\theta+\epsilon_0}(q) - F^\theta(q))] \\
&= \frac{1}{2}\sum_{m,n=1}^{dN} \mathbb{E}_{\mu^\theta}[(F_m^{\theta+\epsilon_0}(q) - F_m^\theta(q))((\sigma\sigma^T)^{-1})_{m,n}(F_n^{\theta+\epsilon_0}(q) - F_n^\theta(q))] \\
&= \frac{1}{2}\sum_{m,n=1}^{dN} ((\sigma\sigma^T)^{-1})_{m,n}\mathbb{E}_{\mu^\theta}[\nabla_\theta F_m^\theta(q)\epsilon_0 \nabla_\theta F_n^\theta(q)\epsilon_0] \\
&+ \frac{1}{2}\sum_{m,n=1}^{dN} ((\sigma\sigma^T)^{-1})_{m,n}\mathbb{E}_{\mu^\theta}[\nabla_\theta F_m^\theta(q)\epsilon_0\epsilon_0^T \nabla_\theta^2 F_m^\theta(q)\epsilon_0] + O(|\epsilon_0|^4)\ .
\end{aligned}$$

The pathwise FIM comes from the second-order term while the third-order term defines a tensor matrix.

For the LJ non-bonded potential, the leading term of the second-order term (i.e., the pathwise FIM) in the RER expansion when $\sigma_{LJ}$ is perturbed is of order $O\big(\big(\frac{\sigma_{LJ}}{r}\big)^{\mathbf{10}}\big)$ while the leading term of the third-order term of RER is of order $O\big(\big(\frac{\sigma_{LJ}}{r}\big)^{\mathbf{9}}\big)$ with (typically) $\sigma_{LJ} < r$. The fact that the leading term of the third-order term has smaller exponent compared to the second-order term, makes the contribution of the third-order term to the value of RER significant on average. Therefore, the asymmetry between $\mathcal{H}(Q^\theta|Q^{\theta+\epsilon_0})$ and $\mathcal{H}(Q^\theta|Q^{\theta-\epsilon_0})$ observed both in the LJ fluid (Figure 3.2) and the methane (Figure 3.8) stems exactly from the significance of the third-order term. Notice that asymmetries between positive and negative perturbations are not rare and have been observed in biological reaction models and one method that is employed for assessing parameter identifiability in non-linear models is the profile likelihood method. [84]

### 3.6.3 Potential energy terms of $CH_4$

In this section, the details of the total potential $V(q) = V_{bond}(q) + V_{angle}(q) + V_{LJ}(q)$ for the methane model are presented. The total bond potential equals to

$$V_{bond}(q) = \sum_{\mathcal{A}} V_{bond}(|q_j - q_i|) \tag{3.30}$$

where

$$\mathcal{A} = \{q_i\text{=C},\ q_j\text{=H}\ q_i, q_j \in \text{same } CH_4, \\ \text{4 bonds per } CH_4\}$$

while the local bond potential is

$$V_{bond}(|q_j - q_i|) = V_{bond}(r_{ij}) = \frac{1}{2}K_b(r_0 - q_{ij})^2\ . \tag{3.31}$$

The two constants $r_0$ and $K_b$ determine the distance and the strength of the bond between the two atoms, respectively.

The angle defined for each triplet $H - C - H$ on the same $CH_4$ molecule is denoted by $\theta_{jik}$. Then, the total angular potential is

$$V_{angle}(q) = \sum_{\mathcal{B}} V_{angle}(\angle q_j q_i q_k) \tag{3.32}$$

where

$$\mathcal{B} = \{q_i\text{=C},\ q_j, q_k\text{=H},\ q_i, q_j, q_k \in \text{same } CH_4, \\ \text{6 angles per } CH_4\}\ ,$$

while the local angular potential is given by

$$V_{angle}(\angle q_j q_i q_k) = V_{angle}(\theta_{ijk}) = \frac{1}{2}K_\theta(\theta_0 - \theta_{ijk})^2\ . \tag{3.33}$$

The two constants $\theta_0$ and $K_\theta$ determine the degree and the strength of the angle, respectively.

Moreover, the non-bonded term of the potential energy, $V_{LJ}(q)$, is given by

$$V_{LJ}(q) = \sum_{\mathcal{C}} V_{LJ}(|q_j - q_i|) \tag{3.34}$$

where

$$\mathcal{C} = \{q_i, q_j\text{=H or C},\ q_i, q_j \in \text{different } CH_4\}$$

while the functional form of the LJ potential, $V_{LJ}(r_{ij})$, is given by (3.17).

Since the LJ potential is the non-bonded term, the sum in (3.34) is over all the atoms of the other methanes. It is convenient furthermore to divide this sum into three sums, each one corresponding on a different class of interactions between $C-C, C-H, H-H$. Thus, we can rewrite

$$V_{LJ}(q) = \sum_{\mathcal{C}_1} V_{LJ}^{C-C}(r_{ij}) + \sum_{\mathcal{C}_2} V_{LJ}^{H-H}(r_{ij}) + \sum_{\mathcal{C}_3} V_{LJ}^{H-C}(r_{ij}) \ , \tag{3.35}$$

where

$$\begin{aligned} &\mathcal{C}_1 = \{q_i, q_j = \text{C}\} \\ &\mathcal{C}_2 = \{q_i = \text{C},\ q_j = \text{H},\ q_i, q_j \in \text{different } CH_4\} \\ &\mathcal{C}_3 = \{q_i, q_j = \text{H},\ q_i, q_j \in \text{different } CH_4\} \ . \end{aligned}$$

Each LJ potential has its own parameter values.

## 3.7 Supplementary Material

## Details on discrete time case

In this section we illustrate the derivation of the numerical implementation of our proposed method based on the BBK integrator. We calculate the RE estimators based on the derived state transition probabilities.

## Calculation of transition probabilities

After reordering the equations of the BBK integrator we get (matrix form) the expressions for the new timestep $i+1$

$$\left\{\begin{array}{l} \boldsymbol{q}_{i+1} = \boldsymbol{q}_i + M^{-1}\Delta t(I - \gamma M^{-1}\frac{\Delta t}{2})\boldsymbol{p}_i - M^{-1}\frac{\Delta t^2}{2}\nabla V(\boldsymbol{q}_i) + M^{-1}\Delta t\sqrt{2\gamma\beta^{-1}}\Delta W_i \\ (I + \gamma M^{-1}\frac{\Delta t}{2})\boldsymbol{p}_{i+1} = (\frac{M}{\Delta t})\Delta\boldsymbol{q}_i - \nabla(V(\boldsymbol{q}_{i+1}))\frac{\Delta t}{2} + \sqrt{2\gamma\beta^{-1}}\Delta W_{i+\frac{1}{2}} \end{array}\right. \tag{3.36}$$

From the above set of normal distributions we define the transition probability as a product of two independent normal ones:

$$P(\boldsymbol{q}_i, \boldsymbol{p}_i \rightarrow \boldsymbol{q}_{i+1}, \boldsymbol{p}_{i+1}) = P(\boldsymbol{q}_{i+1}|\boldsymbol{q}_i, \boldsymbol{p}_i)P(\boldsymbol{p}_{i+1}|\boldsymbol{q}_i, \boldsymbol{p}_i, \boldsymbol{q}_{i+1}) \tag{3.37}$$

This splitting of P is feasible because the numerical scheme of BBK is non-degenerate. The corresponding formulas after reordering eq. (3.36) are:

$$\begin{aligned} P(\boldsymbol{q}_{i+1}|\boldsymbol{p}_i, \boldsymbol{q}_i) = & \frac{1}{((2\pi)^{dN} det(\Delta t^3\gamma\beta^{-1}M^{-2}))^{1/2}} \times \exp\Big\{ - \frac{\beta}{2\Delta t^3\gamma} \\ & ||\Delta\boldsymbol{q}_i - M^{-1}\Delta t(I - \frac{\gamma\Delta t M^{-1}}{2})\boldsymbol{p}_i + M^{-1}\frac{\Delta t^2}{2}\nabla V(\boldsymbol{q}_i)||^2_{M^2}\Big\} \end{aligned}$$

$$\begin{aligned} P(\boldsymbol{p}_{i+1}|\boldsymbol{p}_i, \boldsymbol{q}_i, \boldsymbol{q}_{i+1}) = & \frac{1}{((2\pi)^{dN} det(\Delta t\gamma\beta^{-1}Id)^{1/2}} \times \exp\Big\{ - \frac{\beta}{2\Delta t\gamma} \\ & ||(I + \frac{\gamma\Delta t M^{-1}}{2})\boldsymbol{p}_{i+1} - \frac{M}{\Delta t}\Delta\boldsymbol{q}_i + \nabla V(\boldsymbol{q}_{i+1})\frac{\Delta t}{2}||^2\Big\} \end{aligned}$$

where

$$||x||_M = x^T M x, \quad M \in \mathbb{R}^{dN\times dN}, x \in \mathbb{R}^{dN}$$

$$M^{-1} = diag\Big(\underbrace{\frac{1}{m_1}, \dots, \frac{1}{m_1}}_{\text{d-times}}, \dots, \frac{1}{m_N}\Big)$$

hence

$$det\Big(\Delta t^3\gamma\beta^{-1}M^{-2}\Big) = \Big(\Delta t^3\gamma\beta^{-1}\Big)^{dN}\prod_{i=1}^{N} m_i^{-2d}$$

## Detailed Calculation of RER and path-wise FIM for BBK

The statistical estimator $\bar{H}_1$ (obtained from the Radon-Nikodym derivative) is utilized (see Ref. [1] in main text)

$$\bar{H_1}^{(n)} = \frac{1}{n\Delta t}\sum_{i=0}^{n-1}\log\frac{P^{\theta}(\boldsymbol{z}_i,\boldsymbol{z}_{i+1})}{P^{\theta+\epsilon_0}(\boldsymbol{z}_i,\boldsymbol{z}_{i+1})} \tag{3.38}$$

The corresponding estimator for FIM derived in the same fashion is:

$$\bar{F_1}^{(n)} = \frac{1}{n\Delta t}\sum_{i=0}^{n-1}\nabla_\theta \log P^{\theta}(\boldsymbol{z}_i,\boldsymbol{z}_{i+1})\nabla_\theta \log P^{\theta}(\boldsymbol{z}_i,\boldsymbol{z}_{i+1})^T \tag{3.39}$$

From eq.'s (3.37) and (3.38) the RER is given by:

$$\begin{aligned}\bar{H}_1(\bar{Q}^{\theta}|\bar{Q}^{\theta+\epsilon_0}) &= \frac{1}{n\Delta t}\sum_{i=0}^{n-1} log\frac{P^{\theta}(\boldsymbol{q}_i,\boldsymbol{p}_i\rightarrow\boldsymbol{q}_{i+1},\boldsymbol{p}_{i+1})}{P^{\theta+\epsilon_0}(\boldsymbol{q}_i,\boldsymbol{p}_i\rightarrow\boldsymbol{q}_{i+1},\boldsymbol{p}_{i+1})} \qquad (3.40)\\ &= \frac{1}{n\Delta t}\sum_{i=0}^{n-1}\Big[log\frac{P^{\theta}(\boldsymbol{q}_{i+1}|\boldsymbol{q}_i,\boldsymbol{p}_i)}{P^{\theta+\epsilon_0}(\boldsymbol{q}_{i+1}|\boldsymbol{q}_i,\boldsymbol{p}_i)}\\ &+ log\frac{P^{\theta}(\boldsymbol{p}_{i+1}|\boldsymbol{q}_i,\boldsymbol{p}_i,\boldsymbol{q}_{i+1})}{P^{\theta+\epsilon_0}(\boldsymbol{p}_{i+1}|\boldsymbol{q}_i,\boldsymbol{p}_i,\boldsymbol{q}_{i+1})}\Big]\end{aligned}$$

where

$$\begin{aligned}\log P^{\theta}(\boldsymbol{q}_{i+1}|\boldsymbol{p}_i,\boldsymbol{q}_i) &= -\frac{1}{2}\log((2\pi)^{dN}(\Delta t^3\gamma\beta^{-1})^{dN}\prod_{j=1}^{N}m_j^{2d})\\ -\frac{\beta}{2\Delta t^3\gamma}\sum_{j=1}^{dN}m_j^2\Big[(\Delta\boldsymbol{q}_i)_j &- \frac{\Delta t}{m_j}(1-\frac{\gamma\Delta t}{2m_j})(\boldsymbol{p}_i)_j + \frac{\Delta t^2}{2m_j}(\nabla_j V^{\theta}(\boldsymbol{q}_i))\Big]^2\end{aligned}$$

$$\begin{aligned}\log P^{\theta}(\boldsymbol{p}_{i+1}|\boldsymbol{p}_i,\boldsymbol{q}_i,\boldsymbol{q}_{i+1}) &= -\frac{dN}{2}\log((2\pi\Delta t\gamma\beta^{-1})\\ -\frac{\beta}{2\gamma\Delta t}\sum_{j=1}^{dN}\Big[(1+\frac{\gamma\Delta t}{2m_j})(\boldsymbol{p}_{i+1})_j &- \frac{m_j}{\Delta t}(\Delta\boldsymbol{q}_i)_j + \frac{\Delta t}{2}(\nabla_j V^{\theta}(\boldsymbol{q}_{i+1}))\Big]^2\\ ,\quad (\Delta\boldsymbol{q}_i)_j = (\boldsymbol{q}_{i+1})_j - (\boldsymbol{q}_i)_j,\quad \nabla_j &= \frac{\partial}{\partial q_j}\end{aligned}$$

$(\Delta\boldsymbol{q}_i)_j$ is the momentum difference of atom $j$ in time. The Fisher information matrix (FIM) is $k\times k$ in dimension, (for the $CH_4$ model studied here k = 10, which include the LJ $\epsilon_{LJ}, \sigma_{LJ}$, bond and angle coefficients) and the $(l, m)$-th element at the $i$-th timestep is given by the partial derivatives of (3.37) with respect to the potential coefficients:

$$
\begin{aligned}
&(\nabla_\theta \log P^\theta(\boldsymbol{z}_i, \boldsymbol{z}_{i+1}) \nabla_\theta \log P^\theta(\boldsymbol{z}_i, \boldsymbol{z}_{i+1})^T)_{l,m} = \\
&\Big[\frac{\partial}{\partial \theta_l}(\log P^\theta(\boldsymbol{q}_{i+1}|\boldsymbol{q}_i, \boldsymbol{p}_i)) + \frac{\partial}{\partial \theta_l}(\log P^\theta(\boldsymbol{p}_{i+1}|\boldsymbol{q}_i, \boldsymbol{p}_i, \boldsymbol{q}_{i+1}))\Big] \times \\
&\Big[\frac{\partial}{\partial \theta_m}(\log P^\theta(\boldsymbol{q}_{i+1}|\boldsymbol{q}_i, \boldsymbol{p}_i)) + \frac{\partial}{\partial \theta_m}(\log P^\theta(\boldsymbol{p}_{i+1}|\boldsymbol{q}_i, \boldsymbol{p}_i, \boldsymbol{q}_{i+1}))\Big]
\end{aligned}
\tag{3.41}
$$

where

$$
\begin{aligned}
&\frac{\partial}{\partial \theta_m}(\log P^\theta(\boldsymbol{q}_{i+1}|\boldsymbol{q}_i, \boldsymbol{p}_i)) = \\
&\qquad -\frac{1}{\sigma^2 \Delta t} \sum_{j=1}^{dN} m_j \Big[(\Delta \boldsymbol{q}_i)_j - \frac{\Delta t}{m_j}(1 - \frac{\Delta t \gamma}{2 m_j})(\boldsymbol{p}_i)_j + \frac{\Delta t^2}{2 m_j}(\nabla_j V^\theta(\boldsymbol{q}_i))\Big] \\
&\qquad \times \frac{\partial}{\partial \theta_m}(\nabla_j V^\theta(\boldsymbol{q}_i)) \\
&\frac{\partial}{\partial \theta_m}(\log P^\theta(\boldsymbol{p}_{i+1}|\boldsymbol{q}_i, \boldsymbol{p}_i, \boldsymbol{q}_{i+1})) = \\
&\qquad -\frac{1}{\sigma^2} \sum_{j=1}^{dN} \Big[(1 + \frac{\gamma \Delta t}{2 m_j})(\boldsymbol{p}_{i+1})_j - \frac{m_j}{\Delta t}(\Delta \boldsymbol{q}_i)_j + \frac{\Delta t}{2}(\nabla_j V^\theta(\boldsymbol{q}_{i+1}))\Big] \\
&\qquad \times \frac{\partial}{\partial \theta_m} \nabla_j V^\theta(\boldsymbol{q}_{i+1})
\end{aligned}
$$

RER and FIM calculations for the LJ fluid are summarized in Figure 3.12. We compare the RER value using the discrete time estimators (3.38), (3.39) and the middle bar corresponds to the FIM-based RER (eq.(3.5) in paper, when $T \to \infty$) whereas the left and right bars are the values of estimator (3.38) for a negative and positive perturbation by $\epsilon_0 = 5\%$ respectively. The perturbation in the figures is in logscale. The errorbars (variance) of the RER estimator is larger than the one corresponding to FIM, necessitating more samples for accurate estimation. All the plots are normalized upon division with the number of particles and simulations of bigger systems under the same parameters produce the same results. As the figure suggests $\sigma_{LJ}$ is the most sensitive parameter.

We conclude that the discrete time version is in very good agreement, $O(\Delta t)$, with the continuous time version presented in the main text.

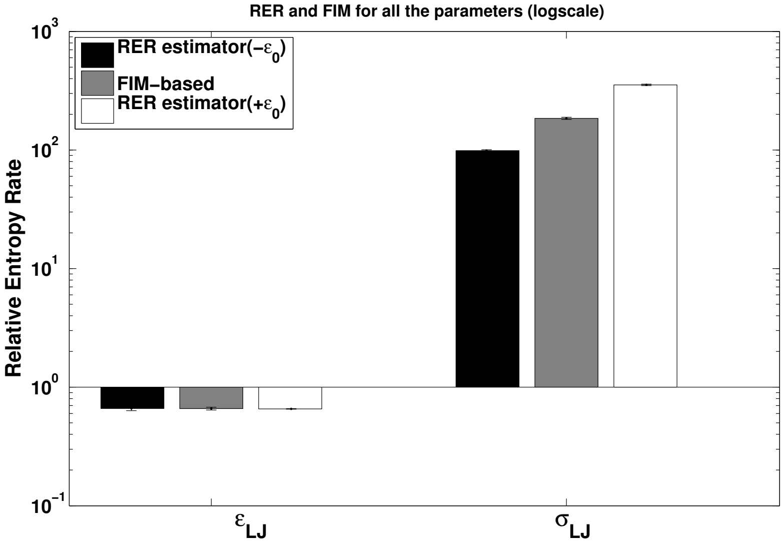

Figure 3.12: FIM based RER (per particle) for various directions using the two different estimators (3.38), (3.39). LJ parameters perturbed by ±5%. The plot is in logscale and the most sensitive variable is $\sigma_{LJ}$ (the errorbars are indecipherable). The variance of the RER estimator is larger than the one corresponding to FIM thus more samples are needed.

# Chapter 4

# Relative Entropy

This chapter is primarily based on the published paper [30].

Hierarchical coarse graining of atomistic molecular systems at equilibrium is an intensive research topic during the last few decades. In this work we discuss theoretical and numerical aspects of different parametrization methods (structural-based, force matching and relative entropy) to derive the effective interaction potential between coarse-grained particles. All methods approximate the many body potential of mean force; resulting, however, in different optimization problems. We apply and compare these methods to: (a) a benchmark system of two isolated methane molecules; (b) methane liquid; (c) water; and (d) an alkane fluid. Differences between the effective interactions, derived from the various methods, are found that depend on the actual system under study. The results further reveal the relation of the various methods and the sensitivities that may arise in the implementation of the numerical methods used in each case.

## 4.1 Introduction

Soft matter fluids involve a very broad range of materials, from polymers, to colloids up to biomolecular systems. The theoretical and computational modeling of such systems is a very intense research area due to both basic scientific aspects and technological applications [33, 14]. Indeed, a direct quantitative link between chemical structure at the molecular level and measurable macroscopic quantities over a broad range of length and time scales is still missing. Such a knowledge would be especially important for the tailored design of materials with the desired properties, over an enormous range of possible applications in nano-, bio-technology, food science, drug industry, cosmetics etc.

A common characteristic of all above systems is that they exhibit multiple characteristic lengths and times, that cannot be described by a single simulation technique. Therefore a quantitative study of specific molecular systems, over a broad range of spatio-temporal scales, require the application of hierarchical simulation methodologies that rigorously connect different levels of description. On the most detailed (classical) level simulation methods (such as molecular dynamics, MD, and Monte Carlo, MC) using all-atom models allow direct quantitative predictions of the properties of molecular systems [14, 32, 8]. However, it is desirable to reduce the required computational cost by describing the system through a smaller, compared to the full atomistic detail, number of degrees of freedom. This is the main idea behind the coarse-grained (CG) models, which have been proven very efficient means in order to increase the length and time scales accessible by simulations, in continuum as well as in lattice molecular systems, [33, 3, 4, 5, 6, 7, 8, 9, 10, 85, 86, 3, 19, 87, 88, 23, 89, 20, 15, 90, 91, 92, 93, 76, 94, 95, 96, 97].

Among several different types of CG models here we focus on systematic - hierarchical ones, which are developed by lumping groups of atoms into CG particles (beads or ”superatoms” in the literature), and deriving the effective CG interaction potentials *directly* from more detailed (microscopic) simulations. Such models are capable of predicting *quantitatively* the properties of *specific systems* and have been applied with great success to a very broad range of molecular systems (see for example refs. [3, 4, 5, 6, 10, 85, 86, 59, 87, 23, 89, 20, 98, 99, 100, 101, 102] and references therein). We also restrict our discussion on CG models with a low degree of coarse-graining, in which a small number of atoms (usually 5-10, up to 1-2 monomers) are lumped together. These models can be used to predict properties at the monomeric level, while at the same time atomistic detail can be re-introduced into the CG configurations, providing direct information in the all-atom level. Altertnatively, in many cases coarser models, in which a large number of monomers, or even long molecules are represented as a single CG bead, are required in order to study more complex systems [7, 103, 11, 93, 12].

The most important part in all such CG models is to develop rigorous all-atom to CG methodologies that allow, as accurate as possible, an approximation of the “exact” CG effective interaction. With such approaches the hierarchical combination of atomistic and CG models could be used in order to study specific molecular complex systems *without adjustable parameters*, and by that become truly predictive.

We should also note here that from a mathematical point of view coarse-graining is a sub-field of the dimensionality reduction [13, 14, 15], a very active research subject in scientific computing, applied statistics and numerical analysis. Indeed, several statistical methods have been developed for the reduction of the degrees of freedom under consideration, in a deterministic or stochastic model, such as principal component analysis, polynomial chaos and diffusion maps, [15, 16].

In this work we examine in detail numerical parameterizing methods to construct a reduced CG model that approximates the properties of reference (microscopic) molecular systems, based on statistical mechanics, and which have been used extensively the last two-three decades in the theoretical modeling of molecular systems across a very broad range of disciplines, from physics to chemistry and biology as well as in engineering sciences. Such methods usually consider the optimization of proposed parametric models using different minimization principles, that is considering a pre-selected observable $\phi$ and then minimizing (average) values over a parameter set $\Theta$

$$\min_{\theta\in\Theta} \mathcal{L}_{cost}(\phi;\theta)\,,$$

where $\mathcal{L}_{cost}$ is a cost function properly defined on the different observables. Different methods consider different sets of observables. For example:
(a) In structural, or correlation, based methods the observable is the *pair radial distribution function* $g(r)$, related to the two-body potential of mean force (see section 4.2), for the intermolecular interaction potential, and distribution functions of bonded degrees of freedom (e.g. bonds, angles, dihedrals) for CG systems with intramolecular interaction potential, [90, 92, 4, 5, 9, 7].
(b) Force matching (FM) or multi-scale CG (MSCG) methods [21, 3, 19, 22, 88] is a mean least squares problem that considers as observable function the total force acting on a coarse bead.
(c) The relative entropy (RE) [6, 89, 79] method employs the minimization of the relative entropy, or Kullback-Leibler divergence, between the microscopic Gibbs measure $\mu$ and $\mu^{\theta}$ representing approximations to the exact coarse space Gibbs measure. In this case, the microscopic probability distribution can be thought as the observable.Information inequalities, such as the Csiszár-Kullback-Pinsker inequality [62] and generalizations [69] suggest that the models obtained by the RE minimization method apply to many reasonable observables.

These methods, in principle, are employed to approximate a many body potential (PMF) describing the *equilibrium* distribution of CG particles observed in simulations of atomically detailed models. To achieve this, several numerical approaches have been used for the different observables. For example: (a) correlation based methods use the direct Boltzmann inversion [4, 9, 38] or iterative techniques such as iterative Boltzmann inversion, IBI, [104, 5] and inverse Monte Carlo, IMC, [92, 91]. (b) Force matching approaches solve a typical least squares problem [105, 19], whereas (c) minimization of the relative entropy is performed through standard Newton-Raphson approaches and stochastic optimization [106, 20].

Note that besides the above numerical parametrization schemes, more analytical approaches for the approximation of the CG effective interaction, based on traditional liquid state theory and on pair correlation functions have been also developed [107, 108, 109, 110, 93, 111, 12].

The main goal of this work is to examine different parameterization methods for obtaining rigorous CG model (i.e. effective CG interaction) through numerical approximations of the PMF. In more detail, we employ iterative inverse Boltzmann, force matching and relative entropy approaches for various molecular systems and numerical algorithms.

First, we discuss the theoretical background of the methods and their ability to approximate the PMF. We present the FM using the probabilistic language of conditional expectation, reformulating it as a projection onto spaces of coarse observables. Second, we provide a critical comparison/discussion of all above methods, applied on the same molecular systems. We also examine numerical aspects (e.g. different basis set functions, error analysis, convergence issues) related to the implementation of the different techniques for specific test cases. Furthermore, we compare the predictions of the different CG models by comparing the structural and dynamical behavior of the CG molecular systems, compared to the detailed all-atom ones.

The structure of this work is as follows. In the next Section, we introduce the atomistic molecular system and its coarse graining through the definition of the CG map, the n-body distribution function and its corresponding n-body potential of mean force. The different approximation methods for the CG effective interaction (potential of mean force) are presented in detail in Section 4.3. The molecular models and details about the atomistic and CG simulations are given in Section 4.4. Results for different molecular systems examined here are given in Section 5.5. Finally, we close with Section 4.6 summarizing and discussing the results of this work.

## 4.2 Theoretical Aspects

### 4.2.1 Atomistic and Coarse Grained Description

Assume a prototypical 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 (microscopic) description, with potential energy $U(\mathbf{q})$. The probability of a state $\mathbf{q}$ at temperature $T$ is given by the Gibbs canonical measure

$$\mu(d\mathbf{q}) = Z^{-1} \exp\{-\beta U(\mathbf{q})\} d\mathbf{q}\,, \tag{4.1}$$

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 $f(\mathbf{q})$ the force corresponding to the potential $U(\mathbf{q})$,

$$\begin{aligned} f: \quad & \mathbb{R}^{3N} \to \mathbb{R}^{3N} \\ & f_j(\mathbf{q}) = -\nabla_{q_j} U(\mathbf{q}), \quad j = 1, \ldots, N\,, \end{aligned} \tag{4.2}$$

i.e. $f_j(\mathbf{q})$ is the force exerted to the $j$-th particle.

Coarse-graining is considered as the application of a mapping (CG mapping)

$$\begin{aligned}\mathbf{\Pi}: \quad & \mathbb{R}^{3N} \to \mathbb{R}^{3M} \\ & \mathbf{q} \mapsto \mathbf{\Pi}(\mathbf{q}) \in \mathbb{R}^{3M}\end{aligned} \tag{4.3}$$

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 $\{\zeta_{ij}, i = 1, \ldots, M,\ j = 1, \ldots, N\}$, for which

$$\mathbf{\Pi}_i(\mathbf{q}) = \sum_j \zeta_{ij} q_j \in \mathbb{R}^3,\ i = 1, \ldots, M\,. \tag{4.4}$$

A nice discussion about the choice of CG mapping is given in [23, 81]. In this work we consider CG maps such that: A CG particle is the center of mass of a group of microscopic particles for which a particle contributes only to one CG particle, that is, if $\zeta_{ij} \neq 0$ for some $i = 1, \ldots, M$ and $j = 1, \ldots, N$, then $\zeta_{kj} = 0$ for all $k \neq i\ k = 1, \ldots, M$.

### 4.2.2 The many body Potential of Mean Force and approximations

Having defined the CG mapping $\mathbf{\Pi}$, (4.3), the probability that the CG system has configuration $\mathbf{Q}$, is given by

$$\bar{\mu}(\mathbf{Q}) = \int_{\Omega(\mathbf{Q})} \mu(\mathbf{q}) d\mathbf{q},\ \ \Omega(\mathbf{Q}) = \{\mathbf{q} \in \mathbb{R}^{3N} :\ \mathbf{\Pi}(\mathbf{q}) = \mathbf{Q}\}\,, \tag{4.5}$$

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 mean force (PMF),

$$\bar{U}^{\mathrm{PMF}}(\mathbf{Q}) = -\frac{1}{\beta} \log \int_{\Omega(\mathbf{Q})} e^{-\beta U(\mathbf{q})} d\mathbf{q}\,. \tag{4.6}$$

We denote the mean force $F^{\mathrm{PMF}} : \mathbb{R}^{3M} \to \mathbb{R}^{3M}$ corresponding to the PMF defined by (4.6), assuming it exists, by

$$F_i^{\mathrm{PMF}}(\mathbf{Q}) = -\nabla_{Q_i} \bar{U}^{\mathrm{PMF}}(\mathbf{Q}),\ i = 1, \ldots, M\,. \tag{4.7}$$

### Approximate forms for the Potential of Mean Force

The calculation of the PMF is a task as difficult and costly as is calculating expectations on the microscopic space. Instead, one seeks for an effective potential function $\bar{U}^*(\mathbf{Q})$ that 'best' approximates the PMF which is easy to formulate and calculate. This is the ultimate goal of all numerical methods discussed here (structural-based methods, force matching, relative entropy) for molecular systems at equilibrium. In all these methods one proposes a family of interaction potential functions $\bar{U}(\mathbf{Q})$ in a parametrized, or a functional, form, $\bar{U}(\mathbf{Q};\theta)$, $\theta \in \Theta$, and seeks for the optimal $\bar{U}^*(\mathbf{Q})$, $(\bar{U}^*(\mathbf{Q}) = \bar{U}(\mathbf{Q};\theta^*))$, that 'best approximates' the PMF. We denote by

$$\bar{\mu}_{\bar{U}}(d\mathbf{Q}) = \bar{Z}^{-1}\exp\{-\beta\bar{U}(\mathbf{Q})\}d\mathbf{Q}\,, \tag{4.8}$$

the equilibrium probability measure at the coarse grained configurational space for the given CG potential function $\bar{U}(\mathbf{Q})$, where $\bar{Z} = \int e^{-\beta\bar{U}(\mathbf{Q})}d\mathbf{Q}$ is the corresponding partition function.

In general the many body PMF can be described as being composed by two-body, three-body, e.t.c., interactions; thus, if we define the pairwise distance $R_{ij} = \|Q_i - Q_j\|,\ i,j = 1,\dots,M$, we may write

$$\bar{U}^{\mathrm{PMF}}(\mathbf{Q}) = \sum_{i,j} u_2(R_{ij}) + \sum_{i,j,k} u_3(R_{ij}, R_{ik}, R_{jk}) + \dots$$

Usually a two-body effective pair potential is assumed to approximate the PMF

$$\bar{U}(\mathbf{Q}) = \sum_{i,j} u(R_{ij}) \approx \bar{U}^{\mathrm{PMF}}(\mathbf{Q})\,.$$

Note that if the CG particle interactions are only two-body then we have exactly that

$$\bar{U}^{\mathrm{PMF}}(\mathbf{Q}) = \sum_{i,j} u_2(R_{ij})\,.$$

This is true for an "ideal" system of two isolated molecules, such as the example of two $CH_4$'s presented in section 4.4.1.

Next, we present different possible functional representations of the coarse interaction potential and force field used in the numerical studies of the current work. These are separated to non-bonded and bonded CG particles interactions.

*Non-bonded pair potentials.* We assume dependence of $\bar{U}(\mathbf{Q};\theta)$ on the pairwise distance $R_{ij}$ $i,j = 1,\dots,M$,

$$\bar{U}(\mathbf{Q};\theta) = \sum_{j}\sum_{i\neq j} u(R_{ij};\theta)\,, \tag{4.9}$$

then

$$\bar{F}_j(\mathbf{Q};\theta) = -\nabla_j \bar{U}(\mathbf{Q};\theta) = \sum_{i\neq j} \frac{\partial u(R_{ij};\theta)}{\partial R_{ij}} \frac{dR_{ij}}{dQ_j} \tag{4.10}$$

If a contribution to the pair interaction potential $u(R;\theta)$ is of the form $\sum_n \theta_k \phi_k(R)$, the corresponding contribution to the force $\bar{F}_j(\mathbf{Q};\theta)$, $j = 1,\ldots,M$ would depend on the derivatives of $\phi'_k(R)$. Then the choice of either representing $\phi_k(R)$ or $\phi'_k(R)$ depends on the problem studied.

Possible representations of the pair CG interaction potential $u(R;\theta)$ used in the literature are of the following form:

(a) Tabulated potentials: In this case values of the potentials (and possibly forces) at specific distances are used. Such potentials are typically used in DBI and IBI type of methods where the numerical scheme is performed directly on the effective potential.

(b) Polynomial basis approximation: Usually linear and cubic splines [112] are considered. The linear splines $\{\phi_k^{(l)}(R)\}_{k=0}^n$ and cubic splines $\{\phi_k^{(c)}(R)\}_{k=0}^n$ are determined for a given set of knots $I_n$, [112]. Note that for the chosen set of knots $I_n$ in $(R_0, R_c)$ the number of the parameters for cubic splines is $2n$.

(c) Lennard-Jones (LJ) potential: Here we have also incorporated LJ type of potentials that depend on two parameters $\theta_1$ and $\theta_2$,

$$u(R) = \theta_1 \frac{1}{R^{12}} - \theta_2 \frac{1}{R^6}, \quad R_0 < R < R_c\,. \tag{4.11}$$

Polynomial basis and Lehnard-Jones representations are linear in the parameters, that is any $w(R)$ is written as

$$w(R) = \sum_{k=1}^{n} \theta_k \phi_k(R), \quad R_0 < R < R_c\,, \tag{4.12}$$

where $n$ is the dimension of the representation and $R_0$ and $R_c$ are cut-off distances that are fixed and not parameters of the representation.

(d) Morse potential: Finally we have employed Morse type representation for the CG effective interaction, which depends in a non-linear form on three parameters,

$$u(R) = \theta_1 \left(1 - e^{-\theta_2(R-\theta_3)}\right)^2 - \theta_1\,, \quad R_0 < R < R_c\,. \tag{4.13}$$

Note that the choice of the proper basis set is particularly important, [113]. Besides the above functional forms other possible representations of the interatomic potentials are also considered in the literature, such as (a) Parametrized wavelets [114] and (b) Gaussian basis representation [115].

*Bonded pair potentials.* For many coarse graining procedures it is necessary to consider intra-molecular bonded interactions, i.e. bonds, angles, dihedrals between CG particles. This is the case in the example of the alkane liquid we present in section 4.4.3. The interactions for bonds and angles are represented either again by splines or by harmonic potentials of the type:

$$u_b(r) = \theta_1^b (r - \theta_2^b)^2 \,, \tag{4.14}$$

where $b$ refers to bonds or to angles.

**Remark.** *Beyond pair interaction potentials: Note that in all above cases for the non-bonded potential pair interaction potentials are considered. The incorporation of many-body non-bonded potentials is also an active research field [116, 23].*

## 4.3 Parametrization methods for CG models at equilibrium

### 4.3.1 Boltzmann Inversion

The first family of numerical methods for obtaining the CG interaction potential that we examine are the structural, or correlation, based ones. Typical such methods are: the direct inverse Boltzmann, DBI [4]; the iterative inverse Boltzmann, IBI [90, 5, 104]; and the inverse Monte Carlo, IMC [92, 91]. The relation between these methods and the potential of mean force is straightforward, since the n-body PMF is defined through the n-body distribution (correlation) function $g^{(n)}(R)$ [71],

$$U^{(n),PMF}(R) = -\frac{1}{\beta} \log g^{(n)}(R) \,.$$

Actually, in all these methods the CG effective interaction is calculated through a conditional (pair) distribution function $\bar{g}(R)$, defined over all atomistic configurations that correspond to a specific CG one, through

$$\bar{U}(R) = -\frac{1}{\beta} \log \bar{g}(R) \,, \tag{4.15}$$

where $R$ is the pairwise distance between two CG particles. In more detail:

(a) DBI employs directly relation (4.15) to infer the interaction potential $\bar{U}(R)$ from a reference CG (pair) distribution function $\bar{g}^{(ref)}(R)$ obtained from the analysis of the all-atom configurations.

In DBI type of methods the CG effective interaction is decomposed in independent bonded and non-bonded parts. The former are derived from bonded (usually bonds, angles, dihedrals) distributions obtained from simulations of a single isolated molecule, whereas the latter are either fully repulsive [4, 9], or they are obtained from two isolated molecules in vacuum [117, 10]. Such approaches are computationally efficient since a full atomistic sampling of the reference system is not required; at the same time they are expected to be exact in the gas phase, since they neglect higher order correlations. However, for homogeneous systems, such as simple liquids and bulk polymers [4, 9, 10, 118, 119, 86, 57], they can provide an accurate prediction of the structure and in some cases of thermodynamic properties.

(b) In IBI methods [104] an iterative numerical minimization problem is introduced based on $\bar{g}(R)$. In more detail the (pair) CG potential is refined at the iteration $(i+1)$ according to the following scheme:

$$\bar{U}^{(i+1)}(R) = \bar{U}^{(i)}(R) + ck_BT \log \frac{\bar{g}^{(i)}(R)}{\bar{g}^{(ref)}(R)} \,. \tag{4.16}$$

where c is a constant to ensure stability of the iterative process. Convergence is checked in each iteration by examining whether the CG non-bonded distribution function matches the reference (derived from the atomistic run) one, within the numerical accuracy. Thus, the two-body potential of mean force, also converges to the (two-body) reference PMF. An analogous scheme exist also for the bonded part of the potential, based on bonded distribution functions.

(c) In IMC an alternative to the above numerical problem is introduced also by "matching" a set of reference distribution functions. The difference with the IBI methods is that the CG effective interaction update scheme is expressed in a thermodynamically consistent way in terms of the number of particle pairs with a specific inter-particle distance, which correspond to the tabulated value of the potential. The latter can be directly related to the radial distribution function; therefore IMC type of potentials are expected to match perfectly those of the IBI ones. For more details, concerning implementation issues of IMC see [91, 92, 120].

The latter two approaches should result exactly into the same potential since both are based on the same distribution functions. A comparison of these techniques can be found in [121]. Here we focus on DBI and IBI methods.

### 4.3.2 Force Matching

The force-matching method determines a CG effective force $\bar{F}(\mathbf{Q};\theta)$, and thus an effective potential in view of (4.10), from *atomistic force information* through the mean least-square minimization

$$\min_{\theta\in\Theta} \mathbb{E}_\mu \left[\|h(\mathbf{q}) - \bar{F}(\mathbf{\Pi}(\mathbf{q});\theta)\|^2\right] \,, \tag{4.17}$$

where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^{3M}$ and $\mathbb{E}_\mu[\cdot]$ averages with respect to the probability measure $d\mu(\mathbf{q})$. $h(\mathbf{q}) \in \mathbb{R}^{3M}$ is the *local mean force*, whose component $h_i(\mathbf{q})$, $i = 1, \ldots, M$ is the force exerted at $i-th$ 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 $h_i(\mathbf{q}) = \sum_{j\in\{\text{group } i\}} f_j(\mathbf{q}), \quad i = 1, \dots M$.

In work [81], we presented a rigorous probalistic formulation and a generalization of the traditional force matching approach that applies to more complex and *nonlinear coarse-graining maps*. The formulation of the force matching method in the probabilistic language of conditional expectations allowed us, moreover, to prove that the relative entropy and force matching are equivalent when the PMF

$$F^{\text{PMF}}(\mathbf{Q}) \in \mathcal{E} := \{\bar{F}(\mathbf{Q};\theta), \theta \in \Theta\} \,. \tag{4.18}$$

This probabilistic formulation gives in addition a geometric representation of the force matching method (see also Figure 1 of ref [81]). Within this formulation the $F^{\text{PMF}}(\mathbf{Q})$ is the conditional expectation (a projection) of a local mean force $h(\mathbf{q})$ onto the coarse forces space $L^2(\mu;\mathbf{\Pi}) = \{F \in L^2(\mu)|$ there exists $\bar{F} : R^{3M} \to R^{3M}$ s.t. $F(\mathbf{q}) = \bar{F}(\mathbf{\Pi}(\mathbf{q}))\}$. Here $L^2(\mu)$ denotes square integrable functions with respect to the probability measure $\mu$. This property guarantees that the solution of the mean least squares problem (4.17) "best approximates" the PMF. The generalized force matching approach is based on the observation that $h(\mathbf{q})$ is not uniquely defined, one choice is

$$h(\mathbf{q}) = \mathbf{J}^{-1}(\mathbf{q})\mathbf{D}(\mathbf{q})f(\mathbf{q}) + \frac{1}{\beta}\nabla_{\mathbf{q}} \cdot \mathbf{J}^{-1}(\mathbf{q})\mathbf{D}(\mathbf{q}) \,, \tag{4.19}$$

where $\mathbf{J}(\mathbf{q}) = \mathbf{D}(\mathbf{q})\mathbf{D}^t(\mathbf{q})$ is the Jacobian matrix of the CG map and $\mathbf{D} \in \mathbb{R}^{3M\times 3N}$ with elements $\mathbf{D}_{ij}(\mathbf{q}) = \nabla_{\mathbf{q}_j}\mathbf{\Pi}_i(\mathbf{q})$, $i = 1, \dots, M, j = 1, \dots, N$, and $\cdot^t$ denotes matrix transpose. The second term in (4.19) depends on the curvature $\nabla_{\mathbf{q}} \cdot \mathbf{J}^{-1}(\mathbf{q})\mathbf{D}(\mathbf{q})$ that contributes only when the CG map in nonlinear. Also, its dependence on the inverse temperature suggests that it will be significant at low temperatures. In [81] we present in detail examples of CG mapping and corresponding local mean forces $h(\mathbf{q})$. For the linear CG map (4.4) the local mean force

$$h(\mathbf{q}) = (\mathbf{W}\mathbf{\Pi}^t)^{-1}\mathbf{W}f(\mathbf{q}) \,, \tag{4.20}$$

for any $\mathbf{W} : \mathbb{R}^{3M} \to \mathbb{R}^{3N}$ such that $\mathbf{W}\mathbf{\Pi}^t$ is invertible, ensures that the least squares problem for if $\mathcal{E} = L^2(\mu;\mathbf{\Pi})$ has optimal solution the PMF. This representation agrees with the findings in work [19], where the local mean force is given by $h_i(\mathbf{q}) = \sum_{\{\text{group } i\}} \frac{d_{ij}}{\zeta_{ij}} f_j(\mathbf{q})$ for the case of CG mapping where particles that contribute to a single CG particle.

**Numerical approaches.** According to (4.10) we write the force in the form

$$\bar{F}_j(\mathbf{Q};\theta) = \sum_{k=1}^{n} \theta_k \psi_k(\mathbf{Q}) \,, \tag{4.21}$$

where $\psi_k(\mathbf{Q}) = \sum_{i \neq j} \theta_k \phi_k(\|Q_i - Q_j\|)$. The set $\{\phi_k(r)\}_k$ may be chosen as described in section 4.2.2 or corresponding to the potential parametrization are given from relation (4.10). A set of i.i.d samples $\{\mathbf{q}_\ell,\ \ell = 1, \dots, n_s\}$ with distribution $\mu(\mathbf{q})$ is generated once from atomistic simulations. Thus the estimator for the mean least squares problem is

$$\frac{1}{n_s} \sum_{\ell=1}^{n_s} \sum_{i=1}^{M} \left( h_i(\mathbf{q}_\ell) - \sum_{k=1}^{n} \theta_k \psi_k(\mathbf{\Pi}(\mathbf{q}_\ell)) \right)^2 .$$

Recall that $h_i(\mathbf{q})$ is the local mean force given by (4.19) or (4.20) in the generalized FM formulation. For the CG map to the center of mass of groups $h_i(\mathbf{q}) = \sum_{j \in \text{ group}i} f_j(\mathbf{q})$ is chosen.

There are two approaches to the solution of this minimization problem. Either, directly solving the system

$$\mathbf{F}\boldsymbol{\theta} = \mathbf{h} \tag{4.22}$$

where $\boldsymbol{\theta} = [\theta_1, \dots, \theta_n]$, $\mathbf{F}$ is an $3Mn_l \times n$ matrix with entries $\psi_k(\mathbf{\Pi}(\mathbf{q}_\ell))\psi_\nu(\mathbf{\Pi}(\mathbf{q}_\ell))$, and $\mathbf{h}$ is a $3Mn_l$ vector with entries $h_j(\mathbf{\Pi}(\mathbf{q}_\ell))$. Or, solving the system of canonical equations

$$\mathbf{G}\theta = \tilde{\mathbf{h}} \tag{4.23}$$

where $\mathbf{G} = \mathbf{F}^t\mathbf{F}$ is an $M \times n$ and $\tilde{\mathbf{h}} = \mathbf{F}^t\mathbf{h}$.

The advantage of the later is its low dimensionality compared to the first. On the other hand the condition number of matrix $\mathbf{G}$ may be very large for some systems. In this case one should consider to solve the system (4.22). To avoid the direct solution of the very high-dimensional linear system (4.22) block-averaging is introduced [3, 87]. An iterative approach has been introduced to solve the FM problem in [105], employing CG simulations for updating the parameters similar to the IBI, IMC and some methods for the RE minimization. In the numerical results we present in section 4.4 we solve the system of canonical equations. Tests using the block averaging with SVD for the CH4 example 4.4.1 resulted similar results.

### 4.3.3 Relative Entropy

This method considers the minimization of the relative entropy (RE),

$$\mathcal{R}\left(\mu|\mu^\theta\right) = E_\mu \left[\log \frac{\mu(\mathbf{q})}{\mu^\theta(\mathbf{q})}\right], \tag{4.24}$$

between the microscopic Gibbs measure $\mu(\mathbf{q})$ and and a back-mapping $\mu^\theta(\mathbf{q})$ of the approximate CG measure $\bar{\mu}(\mathbf{Q})$ (4.8), [6, 89]. Its success on approximating the PMF is based on (a) the properties of RE, that $\mathcal{R}(\mu|\pi) \geq 0$ for all probability measures $\mu, \pi$ and $\mathcal{R}(\mu|\pi) = 0$ if and only $\mu \equiv \pi$ and (b) the definition of the PMF $\bar{\mu}(\mathbf{Q})$ based on which we can write $d\mu(\mathbf{q}) =$

$d\bar{\mu}(\mathbf{Q})d\nu(\mathbf{q}|\mathbf{Q})$, where $\nu(\mathbf{q}|\mathbf{Q})$ is a back-mapping probability. The minimization of RE is thus equivalent to

$$\operatorname*{argmin}_{\theta\in\Theta}\left\{\beta\mathbb{E}_{\mu}\left[\bar{U}(\mathbf{\Pi}(\mathbf{q});\theta)-U(\mathbf{q})\right]-\Big[\log Z^{\theta}-\log Z\Big]\right\}\,, \tag{4.25}$$

where $Z^{\theta}=\int_{\mathbb{R}^{3M}}e^{-\beta\bar{U}(\mathbf{Q};\theta)}d\mathbf{Q},\ Z=\int_{\mathbb{R}^{3N}}e^{-\beta U(\mathbf{q})}$. Note that as the relative is positive the an optimal solution(s) always exist.

We should also state here that, as shown in the above relations RE minimization problem can be defined either directly on the path or on the equilibrium ensemble. The former is expected to be more general since it is valid for systems out of equilibrium too [96]. In addition, such an approach can be used to employ relative entropy as a tool for sensitivity analysis [79, 29]. Preliminary numerical data show good agreement between the two approaches, for the systems at equilibrium we examine here.

**Numerical approaches.** Different optimization algorithms have been used for the RE minimization problem, Newton-Raphson, Robins-Monro and modifications, [6, 89, 20].

In the standard Newton-Raphson algorithm the parameters $\theta$ are updated by the following iterative scheme. Given an initial guess $\theta^{(0)}$, $\theta^{(k+1)}$, is updated by

$$\theta^{(k+1)}=\theta^{(k)}-\chi\left(H_k^{(n)}\right)^{-1}(\theta^{(k)})J_k^{(n)}(\theta^{(k)}),\quad k=0,\dots\,,$$

where $\chi>0$, $J_k^{(n)}(\theta)$ and $H_k^{(n)}(\theta)$ are estimators of the Jacobian

$$J(\theta)=\beta E_{\mu}[\nabla_{\theta}\bar{U}(\mathbf{\Pi}(\mathbf{q});\theta)]-\beta E_{\bar{\mu}}[\nabla_{\theta}\bar{U}(Q;\theta)]\,,$$

and the Hessian matrix $H(\theta)$ with entries

$$\begin{aligned}H_{ij}(\theta)\;&=\;\beta E_{\mu}\left[\frac{\partial^2\bar{U}(\mathbf{\Pi}(\mathbf{q});\theta)}{\partial\theta_i\partial\theta_j}\right]-\beta E_{\bar{\mu}}\left[\frac{\partial^2\bar{U}(\mathbf{Q};\theta)}{\partial\theta_i\partial\theta_j}\right]\\&\quad+\beta^2E_{\bar{\mu}}\left[\frac{\partial\bar{U}(\mathbf{Q};\theta)}{\partial\theta_i}\frac{\partial\bar{U}(\mathbf{Q};\theta)}{\partial\theta_j}\right]-\beta^2E_{\bar{\mu}}\left[\frac{\partial\bar{U}(\mathbf{Q};\theta)}{\partial\theta_i}\right]E_{\bar{\mu}}\left[\frac{\partial\bar{U}(\mathbf{Q};\theta)}{\partial\theta_j}\right]\,.\end{aligned}$$

A set of i.i.d samples $\{\mathbf{q}_{\ell},\ \ell=1,\dots,n_s\}$ from $\mu(\mathbf{q})$ from atomistic simulations is generated once. While, for each updated $\theta^{(k)}$ i.i.d samples $\{\mathbf{Q}_{\ell}^{(k)},\ \ell=1,\dots,m_s\}$ from $\bar{\mu}(\mathbf{Q};\theta^{(k)})$ are generated from coarse grained simulations for each iteration $k$. Thus the estimator of the Jacobian is

$$J_k^{(n)}(\theta)=\beta\frac{1}{n_s}\sum_{\ell=1}^{n_s}\nabla_{\theta}\bar{U}(\mathbf{\Pi}(\mathbf{q}_{\ell});\theta)-\beta\frac{1}{m_s}\sum_{\ell=1}^{m_s}\nabla_{\theta}\bar{U}(\mathbf{Q}_{\ell}^{(k)};\theta)\,, \tag{4.26}$$

and similarly for the Hessian. Additionally, the Hessian provides information on the asymptotic convergence of the solution. In works [106, 122] authors introduce a deterministic iterative scheme based on importance sampling with respect to a fixed parameter $\theta_0$ for computing averages on the CG space. With the importance sampling the CG sampling at each update is avoided though the algorithm becomes very sensitive to small differences of the updated $\theta$ to $\theta_0$. This drawback can be improved with (a) a good $\theta_0$ close to the optimal and (b) update periodically the CG sample set used.

As the minimization of RE involves the minimization of expectations stochastic optimization methods may be more efficient, [123, 124, 125], than direct Newton-Raphson or steepest descent. In [20] the authors propose modified Robbins-Monro algorithm that is essentially a stochastic optimization version of the Newton-Raphson algorithm. The modified Robbins-Monro algorithm reduces the instabilities that may appear in the Newton-Raphson due to singular Hessian estimators as result of poor sampling. The one step update of the iterative scheme consists of two parts:
(1) The first $t > 0$ updates of $\theta$ are given by

$$\theta^{(k+1)} = \theta^{(k)} - \chi p^{(k)} ,$$

where $H_j^{(n)}(\theta^{(k)})p^{(k)} = J_k^{(n)}(\theta^{(k)})$, that are Newton-Raphson steps.
Then (2) for $k > t$ a Robbins-Monro step is employed

$$\theta^{(k+1)} = \theta^{(k)} - \alpha_j J_k^{(n)}(\theta^{(k)}) ,$$

and $\alpha_j$ are rates introduced to ensure convergence of the algorithm, with properties $\alpha_j > 0, \alpha_j \to 0,\ \sum_{j=0}^{\infty} \alpha_j = \infty, \sum_{j=0}^{\infty} \alpha_j^2 < \infty$.

At this point, we would like to shortly discuss the relation of the above methods. As it has been shown before in [106] RE and IBI methods are directly related at equilibrium, both should convergence in the same potential, for the same basis set. In addition, we have recently discussed the relation between RE and FM methods, which are in principle asymptotically equivalent both for the case of linear and nonlinear coarse-graining maps [81].

Note also that in all above methods the information required from the more detailed all-atom level concerns canonical sampling of all-atom systems; i.e. we do not use results related to their dynamics. Thus, any sampling technique, such as Monte Carlo, molecular dynamics or Langevin (stochastic) dynamics, is appropriate for the atomistic simulations.

## 4.4 Molecular models and Simulations

### 4.4.1 Methane

**Two isolated methane molecules**

The simplest system to begin with is the one with only two interacting methane, $CH_4$, molecules in vacuum. This is a reference system for which the many-body PMF is exactly equal to the two-body one.

In order to compute the effective non-bonded two-body potential for this simple system we have used the following methods:

(a) *Constraint runs*: A series of Langevin dynamics (LD) runs are performed in which the distance between the two $CH_4$ molecules, $R = R_{1,2} := ||\mathbf{Q}_1 - \mathbf{Q}_2||$, is constant by keeping both centres of mass (COM) fixed in space. Essentially, on every step throughout the trajectory, we subtract the total force acting on each COM, allowing the atoms to move, resulting in rotations but not translations of the $CH_4$ COM. $R_{1,2}$ ranges from $R_{min}$ to $R_{max}$ ($R_{max} = R_{cutoff}$), with $R_{min} = 3\AA$ and $R_{max} = 12\AA$. Each run is about $1ns$, whereas the time step is $1fs$. During these runs the constraint forces are recorded. Note that for such constraint runs a stochastic numerical scheme, such as LD, is required since standard MD simulations might be trapped in configurations of minimum energy. The effective potential $\bar{U}_{CR}(R)$ is calculated by numerical integration of the constraint force $\langle f \rangle_{R_{12}=R}$ from $R_{min}$ up to $R_{max}$.

(b) *Geometric direct calculation*: In addition, for this simple system, we have calculated directly the two-body PMF (constraint partition function), through "full sampling" of all possible configurations using a geometrical method proper for rigid bodies. In more detail, the geometric averaged constrained two-body effective potential $\bar{U}_{geom}(R)$, is obtained by rotating the two $CH_4$ molecules around their COM's, through their Eulerian angles and taking account of all the possible (up to a degree of angle discretization) orientations. In this case the molecules are treated as rigid bodies; i.e. bond lengths and bond angles are kept fixed, essentially it is assumed that intra-molecular degrees of freedom do not affect the intermolecular (non-bonded potential) ones. The advantage of this method is that we avoid long (and more expensive) molecular simulations of the canonical ensemble, which might also get trapped in local minima and inadequately sample the phase space. This method is very similar to the one used by McCoy and Curro in order to develop a $CH_4$ united-atom model from all-atom configurations [117].

(c) *DBI method*: We have also performed free LD runs for the two isolated methane molecules and calculated the pair distribution function. By direct inversion we estimate the two-body potential of mean force, $\bar{U}_{DBI}(R)$.

All above calculations have been performed using the all-atom Dreiding force field [61]. We have also checked different temperatures, i.e. $T = 80K$,

$T = 100K$, $T = 120K$ and $T = 300K$.

### Bulk methane

Methane liquid was also simulated at constant temperature (NVT conditions) at $T = 80K$, $100K$ and $120K$ for several $ns$. 512 $CH_4$ molecules were modeled, whereas the density was calculated after equilibrating the system in the NPT ensemble for $5ns$. The time step was $0.5fs$ 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.

## 4.4.2 Water

One of the most well-studied liquids both through atomistic and coarse-grained models in the literature is water [126]. Here we have simulated all-atom water, using one of the most typical atomistic force fields, the SPC/E [127]. The model system consists of 1192 molecules at ambient conditions ($T = 300K$, $P = 1atm$). The time step was $1fs$. 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 $50ns$. Then, NVT simulations, in the average density, were performed for $20ns$. All-atom configurations were recorded every $10ps$.

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.

In Figure 4.1 we show a snapshot from the bulk water simulations. Both the all-atom and the CG representations are shown.

## 4.4.3 Alkane liquid

The above systems (methane and water) illustrate examples with only non-bonded CG degrees of freedom; i.e. the whole molecule is represented as a single CG bead. To further examine CG models with bonded degrees of freedom we have also examined liquid hexane.

All-atom simulations of hexane were performed using the OPLS all-atom force field [128]. The system consists of 512 hexane molecules at T=$300K$. The time step was $1fs$. A cut-off distance of $10\mathring{A}$ was used, while electrostatic interactions were calculated using PME. The model system was first equilibrated in the NPT ensemble ($T = 300K$, $P = 1atm$) for about $10ns$. Then, NVT simulations, in the average density, were performed for $10ns$. All-atom configurations were recorded every $1ps$. In all above systems the radial distribution function was calculated using a 0.01 $nm$ grid spacing. For the coarse-grained representation of hexane we use two CG beads, i.e.

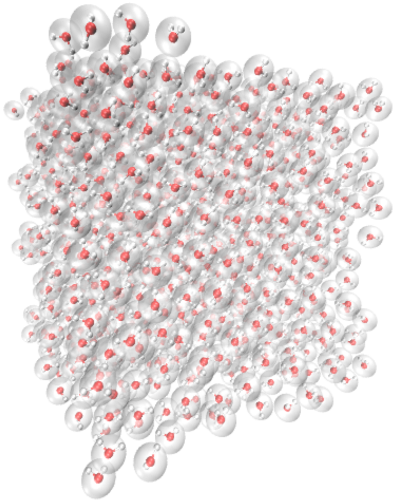

Figure 4.1: All-atom and CG representation of water.

a 3:1 (3 monomers correspond to 1 CG particle) mapping scheme.

We should also state here that all atomistic and coarse-grained simulations have been performed using a home-made parallel simulation package, whereas all analysis has been executed performed through home-made codes in Matlab, Python and C++ (all codes are available upon request).

## 4.5 Results

### 4.5.1 Two Methane

First, we examine all the different approaches discussed in the previous sections for a reference system of two isolated molecules (methanes here) in vacuum. In such an "ideal" system the n-body PMF is a two-body one; i.e. the pair approximation in the calculation of PMF is exact.

In Figure 4.2 we present data from the geometric and the structural-based methods for this system at a specific temperature ($T = 100K$). In more detail, we estimate $U_{nb}(R)$ through:

(a) A "direct" method ($U^{\mathrm{PMF}}_{geom}$), using the geometrical approach described in section 4.2 that involves the direct calculation of the constraint partition function, treating the two molecules as rigid bodies. Note that in this case in the all-=atom description bond lengths and bond angles are kept fixed.

(b) The constraint force approach. In this case the constraint force required to keep fixed two methane molecules at a specific distance is computed. Then through a numerical integration the effective potential between the two molecules (CG particles), $U_{\mathrm{CF}}^{\mathrm{PMF}}$, is computed. This is a method that has been used extensively in the literature to estimate effective pair CG interaction between two molecules, as well as differences in the free energy between two states.

(c) The DBI method: CG effective potential, $U_{\mathrm{DBI}}^{\mathrm{PMF}}$, is obtained by inverting the pair (radial) correlation function, g(R), computed through a stochastic LD run with only two methane molecules in the simulation box. The g(R) of the two methane molecules is also shown in Figure 4.2.

It is clear from Figure 4.2 that all above methods give the same estimate for the CG effective interaction (PMF). There are only slight differences between the various sets of data in the regions of high potential (short distances). This is not surprising if we consider that high energy data from any simulation technique that samples the canonical ensemble, exhibit large error bars, due to difficulties in sampling. The only method that provides a "full", within the numerical discretization, sampling at any distance is the geometric one; however as discussed before (see section 4.4) such a method is possible to be applied only in relatively simple molecules (such as methane) and assumes rigid bond lengths and bond angles.

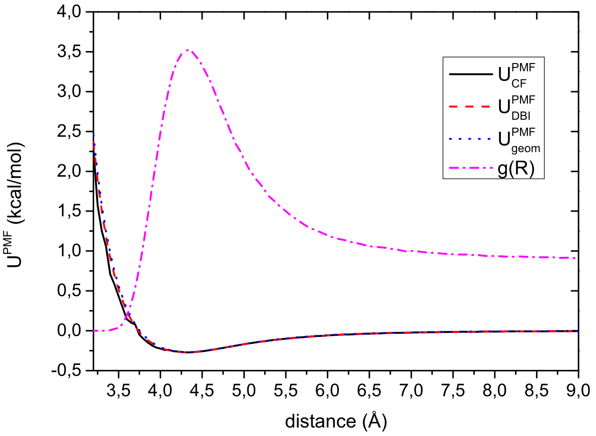

Figure 4.2: "Exact" PMF for two isolated $CH_4$ molecules, calculated through a direct geometric calculation, constraint force approach and DBI method ($T = 100K$). The $g(r)$ from a MD run is also shown.

Next, we examine the application of the force matching procedure (least squares minimization problem on forces) for this simple system. In Figure 4.3a we present the forces derived by solving the minimization, using different basis function sets: linear (linear splines, cubic splines, LJ) and non-linear ones (Morse). Here we have used in the FM minimization problem the total force on each CG bead. It is clear that linear splines, cubic splines and a Morse type basis give the same results. Only the results using the LJ basis slightly deviate. In the inset of Figure 4.3 the derived potential, through numerical integration of the forces are shown. All basis functions discussed here provide again the same solution, but the LJ basis one.

Note that theoretically it is not required to use this specific force for the local forces in the FM minimization problem, but rather a more general form (see also section 4.2) and references [81, 19]). Here we have also used this form with $W = D_{\mathbf{\Pi}}$ the in relation (4.20), that is actually a weighted average of the atomistic forces. Results are shown in Figure 4.3b. It is clear that data from such a minimization problem are different. Possible reasons for such discrepancies are related to the range of $\mathcal{E}$ (4.18) of possible CG effective interactions captured with the specific basis sets.

The above data further emphasize the importance of the functional form used for the parametrization of the CG potentials. Even for such a simple case as two isolated $CH_4$ molecules the standard LJ types potential, that are used extensively in the united-atom models are not capable to accurately describe the all-atom system, in contrast to more flexible functional forms, such as the Morse one.

For completeness we also examine the same system at three different temperatures, i.e. $T = 80K$, $T = 120K$ and $T = 300K$. Note that in all these temperatures but the highest one bulk methane is in fluid state. Our goals was to examine the behavior of the different minimization methods at various temperatures. All methods provide again the same estimate for the CG effective interaction. Data for $U^{\text{PMF}}(T)$ are shown in Figure 4.4. We observed slight differences in the CG effective interactions (free energies) for the different temperatures that become larger for the highest temperature.

Finally, note the role of the molecular structure. Methane has a rather simple molecule structure, with almost perfect spherical symmetry. This allows a rather accurate estimation of the PMF between two molecules even at relative short distances. This is not the case for other molecules. For example, even for two ethane molecules, sampling at short distances might become very difficult (data not shown here). The latter, as well as the effect of temperature, over a broader range from gas to fluid state, will be discussed elsewhere [31].

### 4.5.2 Bulk methane

#### Approximation of the many-body CG PMF

Here we examine a more realistic system, of a 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. Data for the CG pair correlation function, $g(R)$, and the resulting potential

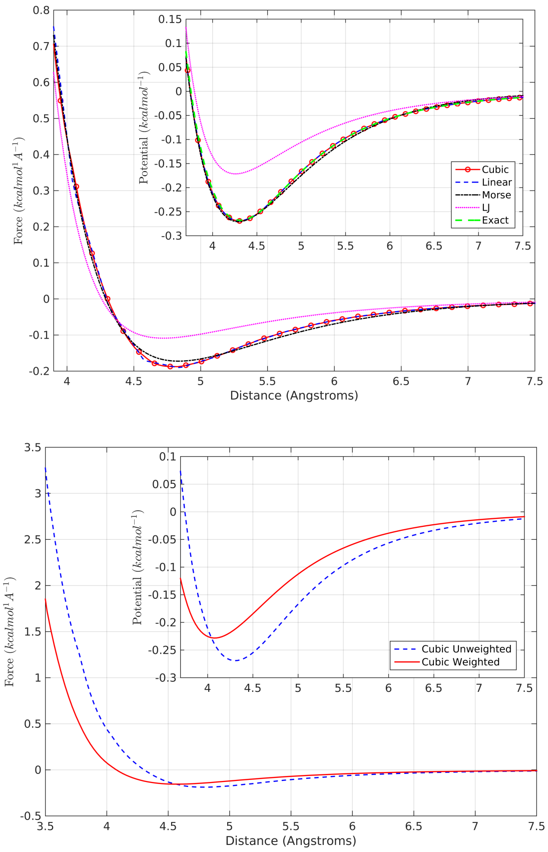

Figure 4.3: (a) Forces for two isolated $CH_4$ molecules through FM using different basis sets. In the inset the derived PMF is shown. (b) Forces using the weighted version ($T = 100K$).

for various iterations are presented in Figure 4.5. In the same plot the reference CG $g(R)$ data from the all-atom simulations are also shown. IBI converges for this system (tolerance is $10^{-4}$) after 14 iterations.

Next, we examine the FM method for the $CH_4$ fluid, by analyzing the reference data from the all-atom simulations. In Figure 4.6 we present the forces derived form the numerical solution of the FM least squares minimization problem for the bulk methane system. As before, we have solved the problem using 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. In the inset of Figure 4.6 the derived potential, through numerical integration of the forces are shown. All basis functions discussed here provide again the same solution, but the LJ basis one. Results concerning the CG effective interaction are also very close to the data

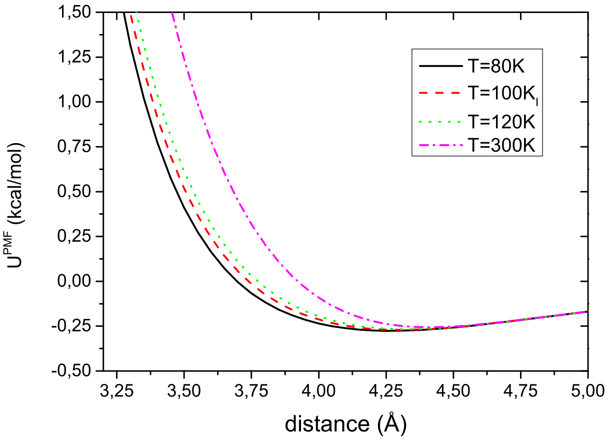

Figure 4.4: PMF for two $CH_4$ molecules in vacuum through different methods for different temperatures: $T = 80K$, $100K$, $120K$ and $300K$.

obtained form the two isolated $CH_4$ in the vacuum.

We have also examined the application of the relative entropy minimization problem for the $CH_4$ liquid. Here we have used the typical Newton-Raphson scheme presented in section 4.3. Convergence is achieved after about 20 iterations.

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 4.7. In the same graph the PMF of the two isolated methane molecules discussed before are shown. 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.

### Bulk CG methane runs

Here we use the different CG models (approximated pair CG interaction potentials) derived above, to predict the properties of the bulk CG methane fluid. In all cases we compare with the reference all-atom bulk system.

First, in Figure 4.8 we examine the structure of the model CG methane liquid by presenting the resulting CG pair correlation function, $g(R)$, from the different models and from the all-atom data for a system. As expected the CG model derived from the IBI method gives a $g(R)$ very close to the one derived from the analysis of the all-atom data. Interestingly the CG model derived from the FM model gives also in good agreement to the reference one, despite the small differences in the CG interaction potential (see Figure 4.7). This is not surprising if we consider that for most molecular systems small

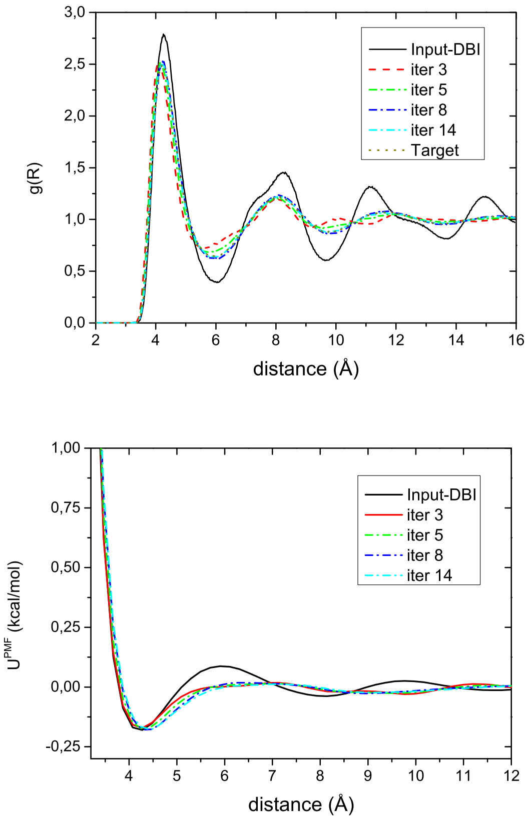

Figure 4.5: Iterative Boltzmann inversion for bulk $CH_4$ liquid ($T = 100K$): (a) $g(r)$ and (b) pair potential for different iterations.

differences in the interaction potential lead to even smaller differences in the obtained pair correlation function. Overall, differences between the different sets of data are less than 5% using square differences defined on the level of $g(R)$. Similar is the case also for the other temperatures ($T = 80K$) studied here (data not shown).

Second, in Figure 4.9 we shortly discuss the dynamics of the model CG methane liquid by presenting the mean square displacements (msd's) of the CG particles derived from the different models, as well as from the analysis of the all-atom methane simulations. Note that in all CG simulations used here dynamics is not expected to follow the all-atom one, since the intrinsic time scale of the CG model is not the same as that of the underlying chemical system. The reason is that due to the reduced degrees of freedom in the CG description, the friction between the CG beads is significantly reduced compared to what it would be if the monomers were represented in full atomistic detail [10, 100, 11]. Data for the methane liquid at $T = 100K$ are

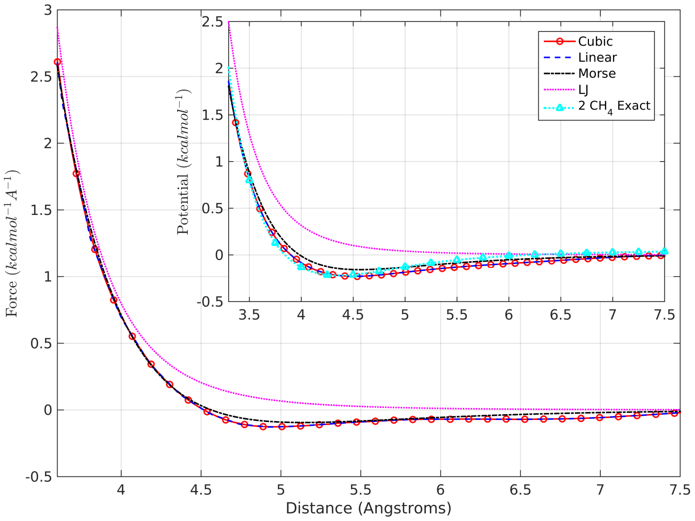

Figure 4.6: Forces obtained through the FM scheme for bulk $CH_4$ liquid. In the inset is the derived potential ($T = 100K$).

shown in Figure 4.9a. As we expect dynamics from the atomistic system is slower than of the CG models. At the same time there are clear differences between the CG models. This show us again that despite the fact that structure obtained from the different CG models is very similar, monomeric friction is much more sensitive into small differences in pair non-bonded potential, resulting into considerable differences in macroscopic quantities, such as mean square displacements and diffusion coefficient. A detailed examination of the dynamical behavior of the different CG models will be a subject of a future work.

### 4.5.3 Water model

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 RDF in order to match the atomistic data. Data for the CG pair correlation function, $g(R)$, are shown in Figure 4.10a. As we can see the final $g(R)$ curve obtained form the IBI method matches almost exactly the reference curve.

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 $\mathbf{F}$ in (4.22) becomes singular, while in IBI and RE the iterative procedure fails. Specifically, the speed of convergence for the NR iterative scheme is based on the $\chi$ parameter, whereas its stability primarily depends

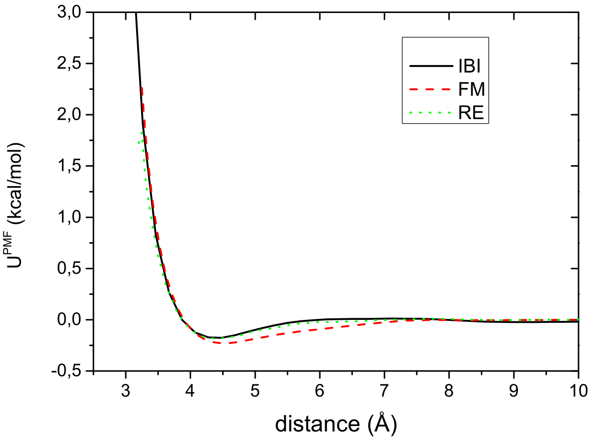

Figure 4.7: Bulk $CH_4$ PMF approximated from different methods ($T = 100K$).

on the condition number of the Hessian. The condition number depends on a number of 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 with the model parameters: number of atoms, complexity of the coarse graining mapping and other.

An important issue is poor sampling towards the minimum distance of the pair potential $R_{min}$. If this is the case, in RE 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 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 4.10b we show the CG $g(R)$ obtained from RE minimization problem together with the reference curve, obtained from the analysis of the all-atom data. The curves are very 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, in the level of $g(R)$, should agree to the IBI one [106]. 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.

In Figure 4.11 results for the effective CG potential from the RE and the FM method are presented. Although both RE and FM potentials have a very similar structure with two minima, the actual values of the potential

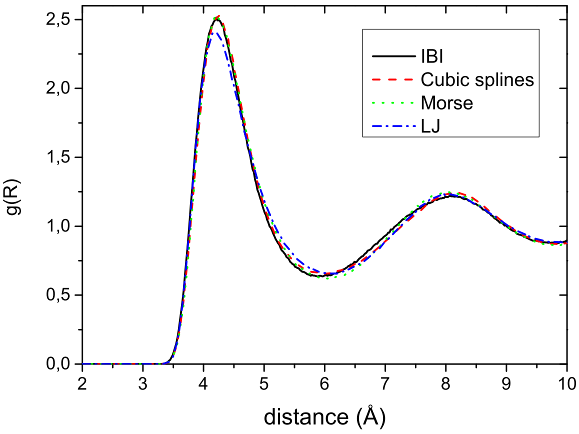

Figure 4.8: CG pair correlation function $g(r)$ for the different CG models and from the reference all-atom simulations ($T = 100K$), for the bulk $CH_4$ model.

are considerably different, in contrast to the $CH_4$ fluid discussed in the previous section. the target one. Possible reasons for these discrepancies are related, as also discussed in section 4.3, 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 requires to clarify such differences [105, 81].

### 4.5.4 Alkane

The last example we examine here concerns a short alkane, hexane. In the case of IBI method, first the bonded potential was calculated, convergence occurs after a few, 2-3, iterations. Then the non-bonded the bonded potentials were iteratively refined. The run length for each iteration was 100 ps with snapshots written every 1.0 ps. A grid spacing of $0.01nm$ and a cut-off distance of $1.2nm$ were used.

Concerning the FM method, an explicit separation of bonded and non-bonded interactions is not required (see also section 4.3). Therefore, both bonded and non-bonded CG potentials were obtained at the same time using the same type of basis set.

Both bonded and non-bonded potentials between FM and IBI were found to be in good agreement with each other. Data for the obtained CG effective potentials for the hexane liquid are shown in Figure 4.12. First, in Figure 4.12a the bonded effective potential derived form the DBI simulations are shown. To examine the difference between the DBI approach (first iteration in the IBI scheme) and the FM one, in Figure 4.12b the correspond-

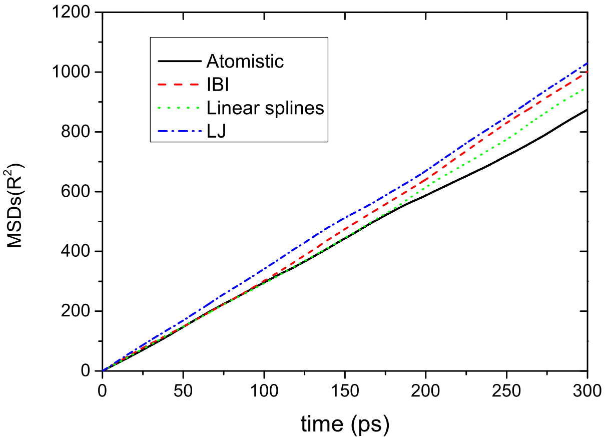

Figure 4.9: Mean square displacement of $CH_4$ molecules obtained from the different CG models and from the reference all-atom simulations, (a) $T = 100K$, (b) $T = 80K$.

ing non-bonded potentials are shown. Clear that small differences between the DBI and the FM predictions were founded.

Overall, the liquid hexane studied here is an example were IBI and FM type of methods found to give rather similar CG effective potentials. Note, however, that this is not always the case. Indeed strong differences betwee the predictions of the FM and the IBI type of methods might occur if the basis set used is not complete enough to represent the CG potential in the well sampled region of CG space. At the same time FM CG potential should be large and positive in the un-sampled regions of CG space. If there is a strong dependence between the CG degreess of freedom this might not be the case. For more details see refs [121, 113]. The detailed numerical investigation of such correlations will be the subject of a future work.

## 4.6 Discussion and Conclusions

Finding the optimum effective interaction potential between CG particles for a given model (i.e. CG mapping scheme) is a very challenging problem in molecular simulations of complex systems. Most of the more rigorous approaches for deriving such a CG potential (force field) are based on numerical parametrization of the (many-body) potential of mean force. In this work we have discussed different parametrization methods: (a) correlation-based methods (direct Boltzmann inversion and iterative Boltzmann inversion), (b) Force matching, and (c) Relative entropy methods. All methods were presented in the probabilistic language of conditional expectation.

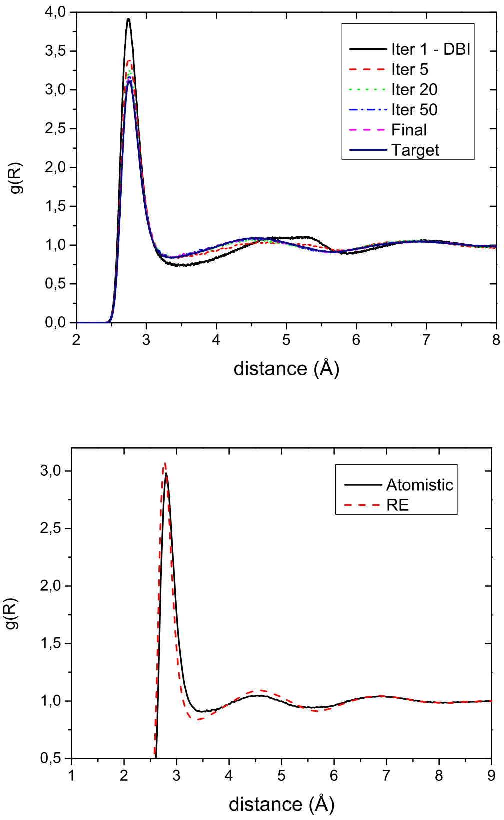

Figure 4.10: (a) $g(R)$ of CG water model through IBI for various iterations and (b) $g(R)$ of CG water model through RE method, and the reference curve. $(T = 300K)$

Below we summarize our findings.

(a) The probabilistic formalism, discussed shortly here, provides a generalized force matching formula, as a CG minimization problem both for linear and nonlinear CG maps (see also ref. [81]). In addition, it proves 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.

(b) 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. To further examine this issue we apply IBI, FM and RE on the same reference (atomistic) systems.

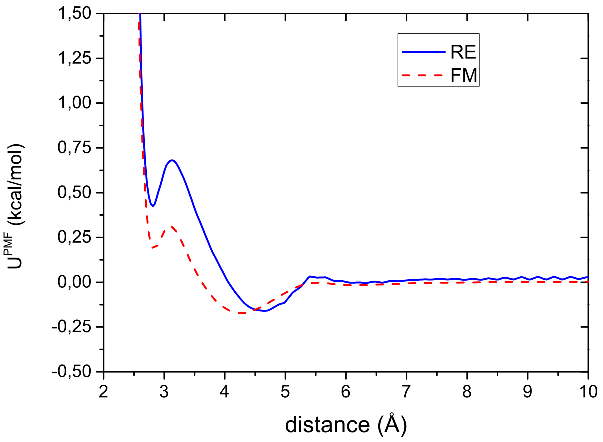

Figure 4.11: CG effective interactions for CG water molecules by analyzing the all-atom data, using force matching and relative entropy techniques. $(T = 300K)$

In all cases we have used the same (linear) mapping scheme, and we have approximated the effective CG interaction using pair potentials. Concerning the different systems: (1) First, for the simple system of two isolated methanes all methods give, within the numerical accuracy, the same CG effective potential, i.e. all of them approximate accurately the pair, exact in this case, CG PMF. For such a simple system a further geometric method, that treats molecules as solid objects was also presented that provides the "full" sampling of the phase space. (2) Second, for 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 for the dynamic properties; friction in the CG models is clearly more sensitive to slight differences in the CG potential used. (3) On the contrary, larger differences in the derived CG potential from the various methods are observed for water. (4) Finally, we also present and shorty discuss an example with bonded potential, the hexane fluid. Data were in good agreement between the different methods.

(c) The various methods discussed here show also different numerical difficulties: (1) IBI is a straight forward technique, however a large number of iterations might be required for convergence. Such a method is capable to describe pair distribution functions by construction, but higher order distributions, as well as cross-correlations, can not be described. (2) The applicability of FM depends strongly on the basis functional set, which should

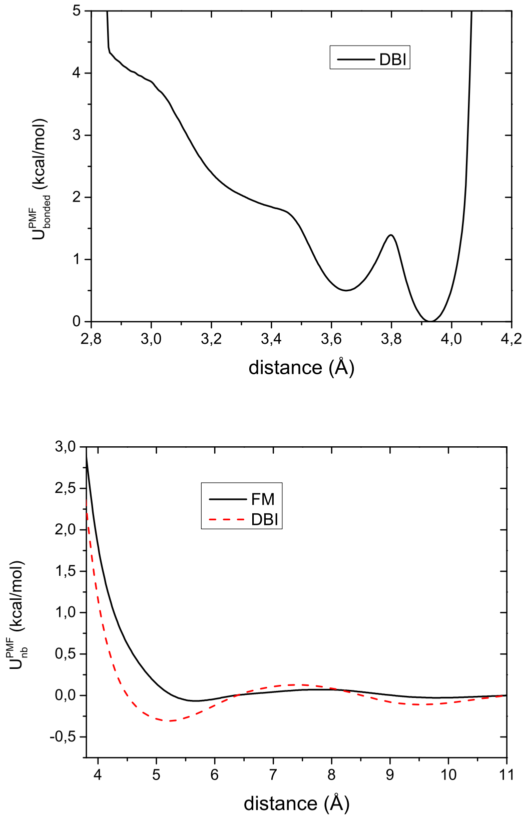

Figure 4.12: CG effective interactions for hexane molecules by analyzing the all-atom data, obtained through force-matching and DBI approaches: (a) bonded potential, and (b) non-bonded potential ($T = 300K$).

be capable to describe accurately the sampling regime. In addition, despite the fact that theoretically it is possible to construct a local mean force in order to best approximate the PMF with FM, using a coarsening transformation of the microscopic forces (local mean force), numerical test performed here resulted in different solution than the original numerical problem. (3) Completeness of basis set is also a main issue for RE.

(d) 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. The optimization problem in such areas can be problematic, therefore special techniques are required. Practically, first a detailed analysis of the all-atom configurations should be used to reveal these regions and then proper extrapolation functional forms, as well as smoothing approaches, for the CG potentials should be used, in order to prevent sampling from such regimes.

(e) Application of all above methods requires a very good sampling of the reference all-atom system. Such a sampling might be problematic for complex (e.g. polymeric) molecules. On the contrary, 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 such and can provide an accurate prediction of the structural and thermodynamic properties [4, 9, 10, 99].

Finally, we should state that the work presented is a first step towards a systematic comparison of different numerical parametrization schemes for realistic molecular systems. Several issues remain to be examined: 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, nonlinear 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 [129, 98, 96].

# Chapter 5

# Cluster Expansions

This chapter is primarily based on a submitted paper [31].

Here, we present a systematic coarse-graining (CG) strategy for many particle molecular systems based on cluster expansion techniques. We construct a hierarchy of coarse-grained Hamiltonians with interaction potentials consisting of two, three and higher body interactions. The accuracy of the derived cluster expansion based on interatomic potentials is examined over a range of various temperatures and densities and compared to direct computation of pair potential of mean force. The comparison of the coarse-grained simulations is done on the basis of the structural properties, against detailed all-atom data. We give specific examples for methane and ethane molecules in which the coarse-grained variable is the center of mass of the molecule. We investigate different temperature and density regimes, and we examine differences between the methane and ethane systems. Results show that the cluster expansion formalism can be used in order to provide accurate effective pair and three-body CG potentials at high $T$ and low $\rho$ regimes. In the liquid regime the three-body effective CG potentials give a small improvement, over the typical pair CG ones; however in order to get significantly better results one needs to consider even higher order terms.

## 5.1 Introduction

The theoretical study of complex molecular systems is a very intense research area due to both basic scientific questions and technological applications. [33] A main challenge in this field is to provide a direct quantitative link between chemical structure at the molecular level and measurable macroscopic quantities over a broad range of length and time scales. Such knowledge would be especially important for the tailored design of materials with the desired properties, over an enormous range of possible applications in nano-, bio-technology, food science, drug industry, cosmetics etc.

A common characteristic of all complex fluids is that they exhibit multiple length and time scales. Therefore, simulation methods across scales are required in order to study such systems. On the all-atom level description, classical atomistic models have successfully been used in order to quantitatively predict the properties of molecular systems over a considerable range of length and time scales. [32, 33, 8, 14] However, due to the broad spectrum of characteristic lengths and times involved in complex molecular systems it is desirable to reduce the required computational cost by describing the system through a small number of degrees of freedom. Thus, coarse-grained (CG) models have been used in order to increase the length and time scales accessible by simulations. [33, 3, 4, 5, 6, 7, 8, 9, 10, 85, 86, 3, 19, 87, 88, 23, 89, 20, 15, 90, 92]

From a mathematical point of view, coarse-graining is a sub-field of dimensionality reduction; there are several statistical methods for the reduction of the degrees of freedom under consideration in a deterministic or stochastic model, such as principal component analysis, polynomial chaos and diffusion maps.[14, 15] Here we focus our discussion on CG methods based on a combination of recent computational methods and old theoretical tools from statistical mechanics. Such CG models, which are developed by lumping groups of atoms into CG particles and deriving the effective CG interaction potentials directly from more detailed (microscopic) simulations, are capable of predicting *quantitatively* the properties of *specific molecular systems* (see for example refs. [3, 4, 5, 6, 7, 10, 85, 86, 59, 87, 23, 89, 20, 98, 103] and references therein).

The most important part in all systematic CG models, based on detailed atomistic data, is to develop rigorous all-atom to CG methodologies that allow, as accurate as possible, estimation of the CG effective interaction. With such approaches the combination of atomistic and hierarchical CG models could allow the study of a very broad range of length and time scales of *specific* molecular systems without adjustable parameters, and by that become truly predictive. [19, 10, 87] There exists a variety of methods that construct a reduced CG model that approximates the properties of molecular systems based on statistical mechanics. For example:
(a) In structural, or correlation-based, methods the main goal is to find effective CG potentials that reproduce the *pair radial distribution function* $g(r)$, and the *distribution functions* of bonded degrees of freedom (e.g. bonds, angles, dihedrals) for CG systems with intramolecular interaction potential. [90, 92, 4, 5, 9, 7] The CG effective interactions in such methods are obtained using the direct Boltzmann inversion, or reversible work, method [119, 9, 38, 99] or iterative techniques, such as the iterative Boltzmann inversion, IBI [104, 5], and the inverse Monte Carlo, IMC, (or inverse Newton) [92, 91] approach.
(b) Force matching (FM) or multi-scale CG (MSCG) methods [21, 3, 19, 22, 88, 105] is a mean least squares problem that considers as observable

function the total force acting on a coarse bead.
(c) The relative entropy (RE) [6, 89, 79] method employs the minimization of the relative entropy, or Kullback-Leibler divergence, between the microscopic Gibbs measure $\mu$ and $\mu^{\theta}$, representing approximations to the exact coarse space Gibbs measure. In this case, the microscopic probability distribution can be thought as the observable. The minimization of the relative entropy is performed through Newton-Raphson approaches and/or stochastic optimization techniques. [106, 20]

In practice, all above numerical methods are employed to approximate a many body potential of mean force (PMF), $U_{\mathrm{PMF}}$, describing the *equilibrium* distribution of CG particles observed in simulations of atomically detailed models. Besides the above numerical parametrization schemes, more analytical approaches have also been developed for the approximation of the CG effective interaction, based on traditional liquid state theory and on pair correlation functions. [107, 108, 109, 110, 93, 111, 12]

Here we discuss an approach for estimating $U_{\mathrm{PMF}}$, and the corresponding effective CG non-bonded potential, based on cluster expansion methods. Such methods originate from the works of Mayer and collaborators [28] in the 40's. In the 60's numerous approximate expansions have been further developed [130, 131] for the study of the liquid state. Later, with the advancement of powerful computational machines, the main focus has been directed on improving the computational methods such as Monte Carlo and molecular dynamics. However, the latter are mostly bulk calculations and they get quite slow for large systems. Reducing the degrees of freedom by coarse-graining has been a key strategy to construct more efficient methods, but with many open questions with respect to error estimation, transferability and adaptivity of the suggested methods. Based on recent developments of the mathematical theory of expansion methods in the canonical ensemble [132], our purpose is to combine the two approaches and obtain powerful computational methods, whose error compared to the target atomistic calculations can be quantified via rigorous estimates. In principle, the validity of these methods is limited to the gas regime. Here we examine the accuracy of these methods in different state points. This attempt consists of the following: a priori error estimation of the approximate schemes depending on the different regimes, a posteriori error validation of the method from the coarse-grained data and design of related adaptive methods.

In previous years, we have developed CG models, based on cluster expansions, for lattice systems, obtaining higher order schemes and a posteriori error estimates [133], for both short and long range interactions [75] and designing adaptive methods [78] and investigating possible strategies for reconstruction of the atomistic information. [95] This is very much in the spirit of the polymer science literature [134, 9, 10] and in this paper we get closer by considering off-lattice models. The proposed approach is based on typical schemes that are based on isolated molecules. [38, 117, 119] Here we extend such approaches using cluster expansion tools for deriving CG effective potentials. We start from typical 2-body (pair) effective interaction, but some results can be extended to many-body interactions as well. We also present a detailed theoretical investigation about the effect of higher order terms in obtaining CG effective interaction potentials for realistic molecular systems. Then, we show some first results from the implementation of three-body terms on the effective CG potential; a more detailed work on the higher order terms will be given in a forthcoming work. [135]

The structure of the paper is as follows: In Section 5.2, we introduce the atomistic molecular system and its coarse-graining via the definition of the CG map, the n-body distribution function and the corresponding n-body potential of mean force. The cluster expansion based formulation of the CG effective interaction is presented in Section 5.3. Details about the model systems (methane and ethane) and the simulation considered here are discussed in Section 5.4. Results are presented in Section 5.5. Finally, we close with Section 5.7 summarizing the results of this work.

## 5.2 Molecular Models

### 5.2.1 Atomistic and "exact" coarse-grained (CG) description

Here we give a short description of the molecular model in the microscopic (all-atom) and mesoscopic (coarse-grained) scale. Assume a system of N (classical) atoms (or molecules) in a box $\Lambda(\ell) := (-\frac{\ell}{2}, \frac{\ell}{2}]^d \subset \mathbb{R}^d$ (for some $\ell > 0$), at temperature $T$. We will also denote the box by $\Lambda$ when we do not need to explicit the dependence on $\ell$. We consider a configuration $\mathbf{q} \equiv \{q_1, \dots, q_N\}$ of $N$ atoms, where $q_i$ is the position of the $i^{th}$ atom. The particles interact via a pair potential $V : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, which is stable and tempered. Stability means that there exists a constant $B \geq 0$ such that:

$$\sum_{1 \leq i < j \leq N} V(q_i - q_j) \geq -BN, \tag{5.1}$$

for all $N$ and all $q_1, ..., q_N$. Moreover, temperedness requires that

$$C(\beta) := \int_{\mathbb{R}^d} |e^{-\beta V(r)} - 1| dr < \infty. \tag{5.2}$$

where $\beta = \frac{1}{k_B T}$ and $k_B$ is Boltzmann's constant. The *canonical partition function* of the system is given by

$$Z_{\beta,\Lambda,N} := \frac{1}{N!} \int_{\Lambda^N} dq_1 \dots dq_N \, e^{-\beta H_\Lambda(\mathbf{q})}, \tag{5.3}$$

where $H_\Lambda$ is the Hamiltonian (total energy) of the system confined in a domain $\Lambda$:

$$H_\Lambda(\mathbf{p},\mathbf{q}) := \sum_{i=0}^{N} \frac{p_i^2}{2m} + U(\mathbf{q}). \tag{5.4}$$

By $U(\mathbf{q})$ we denote the total potential energy of the system, which for pair type potentials is:

$$U(\mathbf{q}) := \sum_{1\le i<j\le N} V(q_i - q_j), \tag{5.5}$$

where for simplicity we assume periodic boundary conditions on $\Lambda$. Integrating over the momenta in (5.3), we get:

$$Z_{\beta,\Lambda,N} = \frac{\lambda^N}{N!}\int_{\Lambda^N} dq_1\,\ldots dq_N\, e^{-\beta U(\mathbf{q})} =: \lambda^N Z^U_{\beta,\Lambda,N}, \tag{5.6}$$

where $\lambda := (\frac{2m\pi}{\beta})^{d/2}$. In the sequel, for simplicity we will consider $\lambda = 1$ and identify $Z_{\beta,\Lambda,N} \equiv Z^U_{\beta,\Lambda,N}$. Fixing the positions $q_1$ and $q_2$ of two particles, we define the two-point correlation function :

$$\rho^{(2),at}_{N,\Lambda}(q_1,q_2) := \frac{1}{(N-2)!}\int dq_3\ldots dq_N \frac{1}{Z_{\beta,\Lambda,N}} e^{-\beta U(\mathbf{q})}. \tag{5.7}$$

It is easy to see that in the thermodynamic limit the leading order is $\rho^2$, where $\rho = \frac{N}{|\Lambda|}$ and $|\Lambda|$ is the volume of the box $\Lambda$. Thus, it is common to define the following order one quantity $g(r) := \frac{1}{\rho^2}\rho^{(2),at}_{N,\Lambda}(q_1,q_2)$, for $r = |q_1 - q_2|$. More generally, for $n \le N$, we define the $n$-body version

$$\begin{aligned} g^{(n)}(q_1,\ldots,q_n) = &\frac{1}{(N-n)!\rho^n} \\ &\int_{\Lambda^{N-n}} dq_{n+1}\ldots dq_N \frac{1}{Z_{\beta,\Lambda,N}} e^{-\beta U(\mathbf{q})}, \end{aligned} \tag{5.8}$$

and from that the order $n$ potential of mean force (PMF), $U_{\mathrm{PMF}}(q_1,\ldots,q_n)$, [136, 71] given by

$$U_{\mathrm{PMF}}(q_1,\ldots,q_n) := -\frac{1}{\beta}\log g^{(n)}(q_1,\ldots,q_n). \tag{5.9}$$

We define the coarse-graining map $T : (\mathbb{R}^d)^N \to (\mathbb{R}^d)^M$ on the microscopic state space, given by $T : \mathbf{q} \mapsto T(\mathbf{q}) \equiv (T_1(\mathbf{q}),\ldots,T_M(\mathbf{q})) \in \mathbb{R}^M$, which determines the $M$ ($M < N$) CG degrees of freedom as a function of the atomic configuration $\mathbf{q}$. We call "CG particles" the elements of the coarse space with positions $\mathbf{r} \equiv \{r_1,\ldots,r_M\}$. The effective CG potential energy is defined by

$$U_{\mathrm{eff}}(r_1,\ldots,r_M) := -\frac{1}{\beta}\log\int_{\{T\mathbf{q}=\mathbf{r}\}} dq_1\,\ldots dq_N\, e^{-\beta U(\mathbf{q})}, \tag{5.10}$$

where the integral is over all atomistic configurations that correspond to a specific CG one using the coarse-graining map. Note, that $U_{\text{eff}}$ is in practice equivalent, up to a constant, to the (constraint) PMF. In the example we will deal with later, the configuration $\mathbf{r}$ will represent the centers of mass of groups of atomistic particles. This coarse graining gives rise to a series of multi-body effective potentials of one, two, up to $M$-body interactions, which are unknown functions of the CG configuration. Note also that by the construction of the CG potential in (5.10) the partition function is the same:

$$\begin{aligned} Z_{\beta,\Lambda,N} &= \int dr_1 \dots dr_M \int_{\{T\mathbf{q}=\mathbf{r}\}} d\mathbf{q} e^{-\beta U(\mathbf{q})} \\ &= \int dr_1 \dots dr_M e^{-\beta U_{\text{eff}}(r_1,\dots,r_M)} =: Z^{cg}_{\beta,\Lambda,M} \end{aligned} \tag{5.11}$$

The main purpose of this article is to give a systematic way (via the cluster expansion method) of constructing controlled approximations of $U_{\text{eff}}$ that can be efficiently computed and at the same time we have a quantification of the corresponding error for both *"structural"* and *"thermodynamic"* quantities. By structural we refer to $g(r)$, while by thermodynamic to the pressure and the free energy. Note that both depend on the partition function, but they can also be related [71] to each other as follows:

$$\beta p = \rho - \frac{\beta}{6}\rho^2 \int_0^\infty r u'(r) g(r) 4\pi r^2 dr, \tag{5.12}$$

for the general case of pair-interaction potentials u(r).

### 5.2.2 Coarse-grained approximations

As mentioned above there are several methods in the literature that give approximations to the effective (CG) interaction potential $U_{\text{eff}}$ as defined in (5.10). Below we list some of them without claim of being exhaustive:
(a) The 'correlation-based (eg. DBI, IBI and IMC) methods that use the *pair radial distribution function* $g(r)$, related to the two-body potential of mean force for the intermolecular interaction potential, as well as distribution functions of bonded degrees of freedom (e.g. bonds, angles, dihedrals) for CG systems with intramolecular interaction potential.[90, 92, 4, 5, 9, 7] These methods will be further discussed below.
(b) Force matching (FM) methods [21, 3, 88] in which the observable function is the average force acting on a CG particle. The CG potential is then determined from *atomistic force information* through a least-square minimization principle, to variationally project the force corresponding to the potential of mean force onto a force that is defined by the form of the approximate potential.
(c) Relative entropy (RE)[6, 89, 20] type methods that produce optimal CG potential parameters by minimizing the relative entropy, Kullback-Leibler

divergence between the atomistic and the CG *Gibbs measures* sampled by the atomistic model.

In addition to the above numerical methods, analytical works for the estimation of the effective CG interaction, based on integral equation theory, have also been developed [93]. A brief review and categorization of parametrization methods at equilibrium is given in references [23, 30].

The correlation-based iterative (e.g. IBI and IMC) methods use the fact that for a pair interaction $u(r)$, by plugging the virial expansion of $p$ in powers of $\rho$ into (5.12) and comparing the orders of $\rho$, one obtains that [71]

$$\begin{aligned} g(r) &= e^{-\beta u(r)}\gamma(r), \\ \gamma(r) &= 1 + c_1(r)\rho + c_2(r)\rho^2 + \ldots \end{aligned} \tag{5.13}$$

Given the atomistic *"target"* $g(r)$ from a free (i.e., without constraints) atomistic run, by inverting (5.13) and neglecting the higher order terms of $\gamma(r)$ one can obtain a first candidate for a pair coarse-grained potential $u(r)$. Then, one calculates the $g(r)$ that corresponds to the first candidate and by iterating this procedure eventually obtains the desired two-body coarse-grained potential. This iteration should in principle converge since there exists a pair interaction that can be reconstructed from a given correlation function []. However, this is only an approximation (accounting for the neglected terms of order $\rho$ and higher in the expansion of $\gamma(r)$) since we know that the "true" CG interaction potential should be multi-body, as a result of integrating atomistic degrees of freedom. Hence, having agreement on $g(r)$ does not secure proper thermodynamic behaviour and several methods have been employed towards this direction, see for example refs [5, 137, 93] and the references within.

In order to maintain the correct thermodynamic properties, our approach in this paper is based on cluster expanding (5.10) with respect to some small but finite parameter $\epsilon$ depending on the regime we are interested in. For technical reasons we will focus on low density - high temperature regime. As it will be explained in detail in the next section, the resulting cluster expansion provides us with a hierarchy of terms:

$$\begin{aligned} U_{\text{eff}} &= U^{(2)} + U^{(3)} + O(\epsilon^3), \\ U^{(2)}(r_1, \ldots, r_M) &:= \sum_{i,j} W^{(2)}(r_i, r_j), \\ U^{(3)}(r_1, \ldots, r_M) &:= \sum_{i,j,k} W^{(3)}(r_i, r_j, r_k), \quad \text{etc}, \end{aligned} \tag{5.14}$$

together with the corresponding error estimates.

The above terms can in principle be calculated independently via fast atomistic simulations of 2, 3, etc. molecules, in the spirit of the conditional reversible work CRW method. [117, 38, 99, 86] In more detail, the effective non-bonded (two-body) CG potential can be computed as follows:

(a) One method is by fixing the distance $r_{1,2} := r_1 - r_2$ between two molecules and perform molecular dynamics with such forces that maintain the fixed distance $r_{1,2}$. In this way we sample over the constrained phase space and obtain the conditioned partition function as in (5.10). Then, by integration of the constrained force the two-body effective potential can be obtained.
(b) Alternatively, by inverting $g(r)$ in (5.13) for two isolated molecules, the two-body effective potential can be directly obtained, since for such a system $\gamma(r) = 1$.

Here we examine both methods, see Figure 5.3. Note also that the validity of cluster expansion provides rigorous expansions for $g(r)$, the pressure and the other relevant quantities. Hence, with this approach we can have *a priori* estimation of the errors made in (5.13). Another benefit of the cluster expansion is that the error terms can be written in terms of the coarse-grained quantities allowing for a posteriori error estimates and the design of adaptive methods [78]; see also discussion in Section 5.7.

## 5.3 Cluster expansion

The cluster expansion method originates from the work of Mayer and collaborators, see ref. [28] for an early review, and consists of expanding the logarithm of the partition function in an absolutely convergent series of an appropriately chosen small but finite parameter. Here we will adapt this method to obtain an expansion of the conditioned partition function (5.10).

For the purpose of this article we assume that the CG map $T$ is a product $T = \otimes_{i=1}^{M} T^i$ creating $M$ groups of $l_1, \ldots, l_M$ particles each. We index the particles in the $i^{th}$ group of the coarse-grained variable $r_i$ by $k_1^i, \ldots, k_{l_i}^i$. We also denote them by $\mathbf{q}^i := (q_{k_1^i}, \ldots, q_{k_{l_i}^i})$, for $i = 1, \ldots, M$. Then (5.10) can be written as:

$$U_{\text{eff}}(r_1, \ldots, r_M) := -\frac{1}{\beta} \log \prod_{i=1}^{M} \lambda^i(\{T^i \mathbf{q}^i = r_i\}) \\ - \frac{1}{\beta} \log \int \prod_{i=1}^{M} \mu(d\mathbf{q}^i; r_i) e^{-\beta U(\mathbf{q})}, \tag{5.15}$$

where, for simplicity, we have introduced the normalized conditional measure:

$$\mu(d\mathbf{q}^i; r_i) := \frac{1}{l_i!} dq_{k_1^i} \ldots dq_{k_{l_i}^i} \frac{\mathbf{1}_{\{T^i \mathbf{q}^i = r_i\}}}{\lambda^i(\{T^i \mathbf{q}^i = r_i\})}, \tag{5.16}$$

and by $\lambda^i$ we denote the measure $\frac{1}{l_i!} dq_{k_1^i} \ldots dq_{k_{l_i}^i}$. To perform a cluster expansion in the second term of (5.15) we rewrite the interaction potential

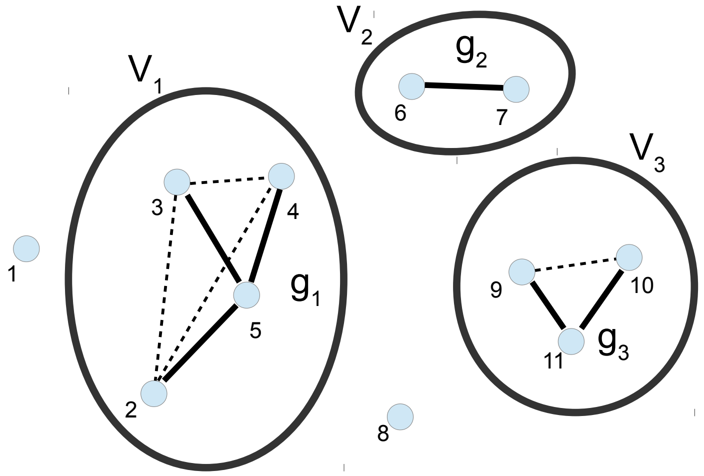

Figure 5.1: Visualization of the partition in (5.18) for non-intersecting sets $V_1 = \{2,3,4,5\}$, $V_2 = \{6,7\}$, $V_3 = \{9,10,11\}$ in each of which we display by solid lines the connected graphs $g_i \in \mathcal{C}_{V_i}$, $i = 1,2,3$.

as follows:

$$\begin{aligned} U(\mathbf{q}) &= \sum_{i<j} \bar{V}(\mathbf{q}^i,\mathbf{q}^j), \quad \text{where} \\ \bar{V}(\mathbf{q}^i,\mathbf{q}^j) &:= \sum_{m=1}^{l_i} \sum_{m'=1}^{l_j} V(|q_{k^i_m} - q_{k^j_{m'}}|). \end{aligned} \tag{5.17}$$

Then, we have

$$\begin{aligned} e^{-\beta U(\mathbf{q})} &= \prod_{i<j} \Big(1 + e^{-\beta \bar{V}(\mathbf{q}^i,\mathbf{q}^j)} - 1\Big) \\ &= \sum_{\substack{V_1,\ldots,V_m \\ |V_i|\geq 2, V_i \subset \{1,\ldots,N\}}} \prod_{l=1}^{m} \sum_{g \in \mathcal{C}_{V_l}} \prod_{\{i,j\}\in E(g)} f_{i,j}(\mathbf{q}^i,\mathbf{q}^j), \qquad (5.18) \\ &\text{where} \quad f_{i,j}(\mathbf{q}^i,\mathbf{q}^j) := e^{-\beta \bar{V}(\mathbf{q}^i,\mathbf{q}^j)} - 1, \end{aligned}$$

where for $V \subset \{1,\ldots,N\}$, we denote by $\mathcal{C}_V$ the set of *connected* graphs on the set of vertices with labels in $V$. Furthermore, for $g \in \mathcal{C}_V$, we denote by $E(g)$ the set of its edges. Since $\mu$ in (5.16) is a normalized measure, from

(5.15) we obtain:

$$\begin{aligned}
U_{\text{eff}}(r_1,\ldots,r_M) = &-\frac{1}{\beta}\log\prod_{i=1}^{M}\lambda^i(\{T^i\mathbf{q}^i = r_i\})\\
&-\frac{1}{\beta}\log\sum_{\substack{V_1,\ldots,V_m\\ |V_i|\geq 2, V_i\subset\{1,\ldots,N\}}}\prod_{l=1}^{m}\zeta(V_i)\\
=&-\frac{1}{\beta}\log\prod_{i=1}^{M}\lambda^i(\{T^i\mathbf{q}^i = r_i\})\\
&-\frac{1}{\beta}\sum_{V\subset\{1,\ldots,N\}}\zeta(V)\\
&+\frac{1}{\beta}\sum_{\substack{V,V':\\ V\cap V'=\varnothing}}\zeta(V)\zeta(V')+\ldots,
\end{aligned}\tag{5.19}$$

where $\zeta(V) := \int\sum_{g\in\mathcal{C}_V}\prod_{\{i,j\}\in E(g)} f_{i,j}(\mathbf{q}^i,\mathbf{q}^j)d\mathbf{q}_V$ with $\mathbf{q}_V := \{\mathbf{q}^i\}_{i\in V}$, is a function over the atomistic details of the system. Note that the above expression involves a sum over all possible pairs, triplets etc. which is a convergent series for values of the density $\rho = \frac{N}{|\Lambda|}$ and of the inverse temperature $\beta$ such that $\rho C(\beta) < c_0$, where $C(\beta)$ is defined in (5.2) and $c_0$ is a known small positive constant.[132] If we simplify the sum in (5.19) one can obtain [132] expansion (5.14) where

$$W^{(2)}(r_1,r_2) := -\frac{1}{\beta}\int\mu(d\mathbf{q}^1;r_1)\,\mu(d\mathbf{q}^2;r_2)\,f_{1,2}(\mathbf{q}^1,\mathbf{q}^2)\tag{5.20}$$

and

$$\begin{aligned}
W^{(3)}(r_1,r_2,r_3) := -\frac{1}{\beta}\int &\mu(d\mathbf{q}^1;r_1)\,\mu(d\mathbf{q}^2;r_2)\,\mu(d\mathbf{q}^3;r_3)\\
&f_{1,2}(\mathbf{q}^1,\mathbf{q}^2)\,f_{2,3}(\mathbf{q}^2,\mathbf{q}^3)\,f_{3,1}(\mathbf{q}^3,\mathbf{q}^1).
\end{aligned}\tag{5.21}$$

Recall also the definition of $f_{i,j}$ in (5.18).

### 5.3.1 Full calculation of the PMF

Notice that the potentials $W^{(2)}$ and $W^{(3)}$ in (5.20) and (5.21), respectively, have been expressed via the Mayer functions $f_{i,j}$. However, the full effective interaction potential between two CG particles can be directly defined as the (conditional) two-body PMF given by

$$\begin{aligned}
&W^{(2),\text{full}}(r_1,r_2) :=\\
&\qquad -\frac{1}{\beta}\log\int\mu(d\mathbf{q}^1;r_1)\,\mu(d\mathbf{q}^2;r_2)\,e^{-\beta\bar{V}(\mathbf{q}^1,\mathbf{q}^2)}.
\end{aligned}\tag{5.22}$$

By adding and subtracting 1 and expanding logarithm, we can relate it to (5.20):

$$\begin{aligned}
&-\beta W^{(2),\mathrm{full}}(r_1,r_2) = \\
&\qquad \log \int \mu(d\mathbf{q}^1;r_1)\,\mu(d\mathbf{q}^2;r_2)\,e^{-\beta \bar{V}(\mathbf{q}^1,\mathbf{q}^2)} = \\
&\qquad \log(1+\int \mu(d\mathbf{q}^1;r_1)\,\mu(d\mathbf{q}^2;r_2)\,f_{1,2}(\mathbf{q}^1,\mathbf{q}^2)) = \\
&\qquad \int \mu(d\mathbf{q}^1;r_1)\,\mu(d\mathbf{q}^2;r_2)\,f_{1,2}(\mathbf{q}^1,\mathbf{q}^2) \\
&\qquad -\frac{1}{2}\left(\int \mu(d\mathbf{q}^1;r_1)\,\mu(d\mathbf{q}^2;r_2)\,f_{1,2}(\mathbf{q}^1,\mathbf{q}^2)\right)^2 \\
&\qquad +\dots
\end{aligned} \tag{5.23}$$

Higher order terms in the above equation are expected to be less/more important in high/low temperature.

Similarly, for three CG degrees of freedom $r_1, r_2, r_3$, the full PMF is given by

$$\begin{aligned}
&W^{(3),\mathrm{full}}(r_1,r_2,r_3) := \\
&\qquad -\frac{1}{\beta}\log \int \mu(d\mathbf{q}^1;r_1)\,\mu(d\mathbf{q}^2;r_2)\mu(d\mathbf{q}^3;r_3) \\
&\qquad e^{-\beta\sum_{1\le i<j\le 3}\bar{V}(\mathbf{q}^i,\mathbf{q}^j)}.
\end{aligned} \tag{5.24}$$

By adding and subtracting 1 we can relate it to (5.20) and (5.21) (in the following we simplify notation by not explicitly showing the dependence on the atomistic configuration and neglecting the normalized conditional measure):

$$\begin{aligned}
e^{-\beta W^{(3),\mathrm{full}}} &= \int e^{-(V_{12}+V_{13}+V_{23})} \\
&= 1+\int f_{12}+\int f_{13}+\int f_{23}+\int f_{12}f_{13} \\
&\quad +\int f_{13}f_{23}+\int f_{12}f_{23}+\int f_{12}f_{23}f_{13},
\end{aligned} \tag{5.25}$$

which implies that

$$\begin{aligned}
W^{(3),\mathrm{full}} = -\frac{1}{\beta}\Bigg(&\int f_{12}+\int f_{13}+\int f_{23}+ \\
&\int f_{12}f_{23}f_{13}+\int f_{12}f_{13}+\int f_{13}f_{23}+\int f_{12}f_{23}- \\
&\left[\int f_{12}\int f_{13}+\int f_{13}\int f_{23}+\int f_{12}\int f_{23}\right]\Bigg) \\
&+\dots
\end{aligned} \tag{5.26}$$

In principle, we can rewrite (5.14) with respect to $W^{(2),\text{full}}$ and $W^{(3),\text{full}}$. Note however, that both of these terms contain the coarse-grained two-body interactions, hence in order to avoid double-counting, when we use both, we have to appropriately subtract the two-body contributions. For some related results, see also the discussion about Figure 5.11.

### 5.3.2 Thermodynamic consistency

As already mentioned, several coarse-graining strategies lack of thermodynamic consistency, see also the discussion by Louis [138] and Guenza [93]. On the other hand, by construction, the cluster expansion approach gives quantified approximations to the correct thermodynamic behaviour. Hence, from (5.14), by considering only the two-body contribution, for the finite volume free energy we have that

$$\begin{aligned} -\frac{1}{\beta|\Lambda|}\log Z_{\beta,\Lambda,N} = -\frac{1}{\beta|\Lambda|}\log\int dr_1\dots dr_M e^{-\beta U^{(2)}} \\ + \frac{1}{\beta|\Lambda|}O(\epsilon^3)), \end{aligned} \tag{5.27}$$

where the error is uniform in $N$ and $|\Lambda|$ and negligible in the limit. Thus, the approximation $U^{(2)}$ of the CG Hamiltonian implies a good approximation of the free energy. Similarly, for the *pressure* as a function of the activity $z$, we have:

$$\begin{aligned} \frac{1}{\beta|\Lambda|}\log\sum_{N\geq 0} z^N Z_{\beta,\Lambda,N} = \frac{1}{\beta|\Lambda|}\log\sum_{N\geq 0} z^N\int dr_1\dots dr_M e^{-\beta U^{(2)}} \\ + \frac{1}{\beta|\Lambda|}O(\epsilon^3). \end{aligned} \tag{5.28}$$

Both quantities have limits given by absolutely convergent series with respect to $\rho = N/|\Lambda|$ for the first and $z$ or $\rho$ for the second. As a side remark, let us mention that in order to compute them we have two options: the first is to use (5.27) and calculate the integral $\int dr_1\dots dr_M e^{-\beta U^{(2)}}$ using molecular dynamics. Alternatively, we can use the corresponding expansions - e.g. for the free energy we would obtain [139]

$$-\frac{1}{\beta|\Lambda|}\log Z_{\beta,\Lambda,N} = \rho(\log\rho - 1) + \sum_{n\geq 1}\beta_\Lambda\rho^n + \text{finite volume errors} \tag{5.29}$$

- and compute the coefficients $\beta_\Lambda$. The latter are not bulk computations as they involve 2, 3, etc particles so they are rather efficient, at least up to some order.

### 5.3.3 Pair correlation function

Recalling the coarse-grained map $T$ from the previous section, we fix two centers of mass $r_1$ and $r_2$ and integrate over all atomistic configurations so that the first two groups $\mathbf{q}^1$ and $\mathbf{q}^2$ of atomistic configurations have the above fixed centers of mass. Partitioning the $N$ particles into $M$ groups of $l_1, \ldots, l_M$ particles and choosing two of them (indexed by 1 and 2) to be the fixed ones, we define the "projected" correlation function at the coarse-grained scale as follows:

$$\begin{aligned}
&\rho^{(2),proj}_{N,\Lambda}(r_1, r_2) := \\
&\qquad \int_{\{T_1(\mathbf{q}^1)=r_1,\, T_2(\mathbf{q}^2)=r_2\}} \prod_{i=1}^{M} \lambda^i(d\mathbf{q}^i)\, \frac{1}{Z_{\beta,\Lambda,N}} e^{-\beta U(\mathbf{q})} = \\
&\qquad \int dr_3 \ldots dr_M \int \prod_{i=1}^{M} \mu(d\mathbf{q}^i; r_i) \frac{1}{Z_{\beta,\Lambda,N}} e^{-\beta U(\mathbf{q})} = \\
&\qquad \int dr_3 \ldots dr_M \frac{1}{Z^{cg}_{\beta,\Lambda,M}} e^{-\beta U_{\mathrm{eff}}(r_1,\ldots,r_M)}.
\end{aligned}$$

Hence, using (5.14) we can construct coarse-grained approximations for the correlation functions as well. Alternatively, as a corollary of the cluster expansion, we can write (5.7) as a convergent power series with respect to the density. These are old results [131] for which the convergence has also been proved recently in the context of the canonical ensemble. [140] In the limit $N \to \infty$, $\Lambda \to \mathbb{R}^d$ such that $\frac{N}{|\Lambda|} = \rho$, we obtain:

$$\begin{aligned}
g(r) = e^{-\beta \bar{V}(\mathbf{q}^1 - \mathbf{q}^2)} \Big( 1 &+ \rho C_3(\mathbf{q}^1, \mathbf{q}^2) \\
&+ \rho^2 C_4(\mathbf{q}^1, \mathbf{q}^2) + \ldots \Big), \\
r &:= T(\mathbf{q}^1) - T(\mathbf{q}^2),
\end{aligned} \tag{5.30}$$

where

$$C_3(\mathbf{q}^1, \mathbf{q}^2) := \int_\Lambda d\mathbf{q}_3\, f_{1,3} f_{3,2}, \qquad f_{i,j} := e^{-\beta \bar{V}(\mathbf{q}_i - \mathbf{q}_j)} - 1 \tag{5.31}$$

and

$$\begin{aligned}
C_4(\mathbf{q}^1, \mathbf{q}^2) := &\int dq_3\, dq_4\, f_{1,3} f_{3,4} f_{4,2} \\
&+ 4 \int dq_3\, dq_4\, f_{1,3} f_{3,4} f_{1,4} f_{4,2} \\
&+ \int dq_3\, dq_4 f_{1,3} f_{3,2} f_{1,4} f_{4,2} \\
&+ \int dq_3\, dq_4\, f_{1,3} f_{1,4} f_{2,3} f_{2,4} f_{3,4}
\end{aligned} \tag{5.32}$$

Note that this formula could also be used at the coarse-grained level with the pair coarse-grained potential $W^{(2)}$, giving an alternative way to compute it.

## 5.4 Model and Simulations

### 5.4.1 The model

A main goal of this work, as mentioned before, is to examine the parameterization of a coarse-grained model using the cluster expansion formalism described above for simple realistic molecular systems; in this work we study liquid methane and ethane. In more detail, we consider $N$ molecules of $CH_4$ and we denote as $\bar{\mathbf{q}} \equiv \{\bar{q}_1, \dots, \bar{q}_N\}$ to be the positions of the $N$ many carbons and $\mathbf{q}_i \equiv \{q_{i,1}, \dots, q_{i,4}\}$ be the positions of the 4 hydrogens that correspond to the $i^{th}$ carbon. We have two types of interactions, namely the *bonded* with (many body) interaction potential $V_b$ and the *non-bonded* with pair interaction potential $V_{nb}$. The latter are of Lennard-Jones type between all possibilities: $C - C$, $C - H$ and $H - H$ (with different coefficients), i.e., $V_{nb} = V_{CC} + V_{CH} + V_{HH}$. In the model used here the non-bonded interactions within the same $CH_4$ molecule are excluded.

The microscopic canonical partition function is given by

$$\begin{aligned} Z_{CH_4} =& \frac{1}{N!} \int_{\Lambda^N} d\bar{\mathbf{q}} \, (\frac{1}{4!})^N \\ & \int_{\Lambda^{4N}} \prod_{i=1}^{N} d\mathbf{q}_i e^{-\beta\left(\sum_{i=1}^{N} V_b(\bar{q}_i, \mathbf{q}_i) + U_{nb}(\bar{\mathbf{q}}, \mathbf{q}_1, \dots, \mathbf{q}_N)\right)}, \end{aligned} \tag{5.33}$$

where $U_{nb}$ is a pair potential of all possible pairs among $\bar{\mathbf{q}}, \mathbf{q}_1, \dots, \mathbf{q}_N$, all of L-J type (eventually with different parameters). Note also that since only the 4 particles of $H$ are indistinguishable, we have introduced the factor $1/4!$ for each molecule.

We are interested in computing the effective Hamiltonian when only the centers of mass of the $N$ many molecules are prescribed. Hence, let us introduce a map $T : \Lambda^5 \to \Lambda$ which gives the center of mass of a molecule consisting of an atom of $C$ together with the prescribed 4 atoms of $H$ which are linked to $C$ by the bonded interactions, i.e., by denoting $\bar{\mathbf{q}}_i \equiv (\bar{q}_i, \mathbf{q}_i)$ we have:

$$T(\bar{\mathbf{q}}_i) := \frac{1}{m_C + 4m_H}(m_C\bar{q}_i + m_H \sum_{j=1}^{4} q_{i,j}). \tag{5.34}$$

We introduce the variables $r_1, \dots, r_N$ for the centers of mass. Our goal is to find the effective potential $U_{\text{eff}}(r_1, \dots, r_N)$. We define the "bonded" (normalized) prior measure by

$$
\begin{aligned}
d\hat{\mu}_b(\bar{\mathbf{q}}_i; r_i) &:= \frac{1}{Z_b(r_i)} d\bar{\mathbf{q}}_i \mathbf{1}_{T(\bar{\mathbf{q}}_i)=r_i} e^{-\beta V_b(\bar{\mathbf{q}}_i)}, \\
Z_b(r_i) &:= \frac{1}{4!} \int_{\Lambda^5} d\bar{\mathbf{q}}_i \mathbf{1}_{T(\bar{\mathbf{q}}_i)=r_i} e^{-\beta V_b(\bar{\mathbf{q}}_i)}.
\end{aligned}
\tag{5.35}
$$

Note that here we could have also included possible non-bonded interactions between atoms of the same molecule. This would be important for the case of coarse-graining a molecule with intra-molecular non-bonded interactions; for the methane molecule studied here such interactions do not exist. Then, from (5.33) we obtain:

$$
\begin{aligned}
Z_{CH_4} = \frac{1}{N!} \int_{\Lambda^N} dr_1 \dots dr_N \prod_{i=1}^{N} Z_b(r_i) \\
\int \prod_{i=1}^{N} d\hat{\mu}_b(\bar{\mathbf{q}}_i; r_i) e^{-\beta U_{nb}(\bar{\mathbf{q}}_1, \dots, \bar{\mathbf{q}}_N)}.
\end{aligned}
\tag{5.36}
$$

The effective free energy is defined by:

$$
\begin{aligned}
e^{-\beta U_{\mathrm{eff}}(r_1, \dots, r_M)} := \prod_{i=1}^{N} Z_b(r_i) \int \prod_{i=1}^{N} d\hat{\mu}_b(\bar{\mathbf{q}}_i; r_i) \\
e^{-\beta U_{nb}(\bar{\mathbf{q}}_1, \dots, \bar{\mathbf{q}}_N)},
\end{aligned}
\tag{5.37}
$$

for which we can construct approximations following formula (5.19). A similar analysis holds for ethane as well.

The total (atomistic) potential energy $V(q)$, for both methane and ethane, is defined by

$$
V(q) = V_{bond}(q) + V_{angle}(q) + V_{LJ}(q) \ . \tag{5.38}
$$

where $V_{bond}(q), V_{angle}(q)$ are quadratic intramolecular potential functions of the bonds and angles respectively. $V_{LJ}(q)$ is the non-bonded potential as defined in the previous subsection. The parameters values of $CH_4$ are summarized in Table 5.1.

The more simple, non-spherically symmetric ethane molecule consists of one rigid bond connecting two united atom $CH_3$ beads. Table 5.2 summarizes this model.

### 5.4.2 Simulations

The simplest system to simulate is the one with only two interacting methane, or ethane, molecules in vacuum. This is a reference system for which the many-body PMF is equal to the two-body one. In addition we have also simulated the corresponding liquid systems. The atomistic and CG model

|  | $\epsilon_{LJ}[\frac{Kcal}{mol}]$ | $\sigma_{LJ}$ [Å] | $r_{cut}$ [Å] |
|---|---|---|---|
| $C-C$ | 0.0951 | 3.473 | 15.0 |
| $C-H$ | 0.0380 | 3.159 | 15.0 |
| $H-H$ | 0.0152 | 2.846 | 15.0 |

| $K_b$ $[\frac{Kcal}{mol Å^2}]$ | $r_0$ [Å] | $K_\theta$ $[\frac{Kcal}{mol \cdot deg^2}]$ | $\theta_0$ [rad] |
|---|---|---|---|
| 700 | 1.1 | 100 | 1.909 |

Table 5.1: Non-bonded $LJ$ coefficients as well as bond and angle coefficients for methane. [61]

|  | $\epsilon_{LJ}[\frac{Kcal}{mol}]$ | $\sigma_{LJ}$ [Å] | $r_{cut}$ [Å] |
|---|---|---|---|
| $CH_3 - CH_3$ | 0.194726 | 3.75 | 14.0 |

Table 5.2: Non-bonded $LJ$ coefficients for ethane. [141]

methane systems were studied through molecular dynamics and Langevin dynamics (LD) simulations. All simulations were conducted in the NVT ensemble. For the MD simulations the Nose-Hoover thermostat was used. Langevin dynamics models a Hamiltonian system which is coupled with a thermostat. [35] The thermostat serves as a reservoir of energy. The densities of both liquid methane and ethane systems were chosen as the average values of NPT runs at atmospheric pressure. NVT equilibration and production runs of few $ns$ followed and the size of the systems were 512 $CH_4$ and 500 $CH_3 - CH_3$ molecules. We note here that the $BBK$ integrator used for Langevin dynamics exhibits pressure fluctuations of the order of $\pm 40$ $atm$ in the liquid phase, whereas temperature fluctuations have small variance and the system is driven to the target temperature a lot faster than with conventional MD.

In order to compute the effective non-bonded coarse-grained potential, different simulation runs have been used which are discussed below.

### Constrained runs

The first method which we use in order to estimate the effective CG potential is by constraining the intermolecular distance between two molecules, $r = r_{1,2}$, in order to compute the constrained partition function (5.11). We call it "constrained run" of two methane, or ethane, molecules and special care had to be taken in order to avoid long sampling of the low probability short distances. This method is very similar to the conditional reversible work methods in which CG degrees of freedom are constrained at fixed distances, as well as in free energy calculations. Technically, we pin the centres of mass (COM) of each CG particle in space and, on every step throughout the stochastic (Langevin) dynamics trajectory, we subtract the total force acting on each COM. Hence, we allow the atoms to move, resulting in rotations but not translations of the CG degrees of freedom (CH$_4$, COM). During these runs the constraint forces are recorded. The mean value $\langle f \rangle_{r_{12}=r}$ is calculated in the same manner and we get $W^{(2),\text{full, f}}(r)$, from $f = -\nabla W$ . Both $W^{(2),\text{full, f}}(r)$ and $W^{(2),\text{full, u}}(r)$ are based on the same trajectory. Then, the effective potential is calculated by numerical integration of the constraint force $\langle f \rangle_{r_{12}=r}$ from $r_{min}$ up to $r_{max}$.

The constrained run technique described above, accelerates the sampling for short distances but there is a caveat; the ensemble average at very short distances (left part of the potential well) is strongly affected by the geometric arrangement of specific atoms between the two molecules, and the system might be trapped in the minimum of energy. For example, the two $CH_4$ molecules are oriented according to the highly repulsive forces and rotate around the axis connecting the two COM's. Due to this specific reason, we utilized stochastic (Langevin) dynamics in order to better explore the subspace of the phase space, as a random kick breaks this alignment. We determine the minimum amount of steps needed for the ensemble average to converge, in a semi-empirical manner upon inspection of the error-bars.

### Geometric direct computation of PMF

In order to further accelerate the sampling and alleviate the noise problems at high energy regions, that might become catastrophic in the case of the non-symmetric $CH_3 - CH_3$ model, we have also calculated the two-body PMF (constraint partition function) directly, through "full sampling" of all possible configurations using a geometrical method proper for rigid bodies. In more detail, the geometric averaged constrained two-body effective potential $W^{(2),geom}(r)$, is obtained by rotating the two (methane or ethane) molecules around their COM's, through their Eulerian angles and taking account of all the possible (up to a degree of angle discretization) orientations. The main idea is to cover every possible (discretized) orientation and associate it with a corresponding weight. The Euler angles proved to be the easiest way to implement this; each possible orientation is calculated via a rotation matrix using three (Euler) angles in spherical coordinates.

The above way of sampling is more accurate (less noisy) than constrained canonical sampling and considerably faster. In addition, the nature of the computations allows massive parallelization of the procedure. We used a ZYZ rotation with $d\phi = d\psi = d\theta = \pi/20$ for $CH_4$ and simple spherical coordinate sampling with $d\phi = \pi/20, d\theta = \pi/45$ for $CH_3 - CH_3$ (as it is diagonally symmetric in the united atom description). Note however, that in this case the molecules are treated as rigid bodies; i.e., bond lengths and bond angles are kept fixed, essentially it is assumed that intra-molecular degrees of freedom do not affect the intermolecular (non-bonded potential) ones. The advantage of this method is that we avoid long (and more expensive) molecular simulations of the canonical ensemble, which might also get trapped in local minima and inadequately sample the phase space. We should also state that this method is very similar to the one used by McCoy and Curro in order to develop a $CH_4$ united-atom model from all-atom configurations. [117]

All atomistic and coarse-grained simulations have been performed using a home-made simulation package, whereas all analysis has been executed through home-made codes in Matlab and Python.

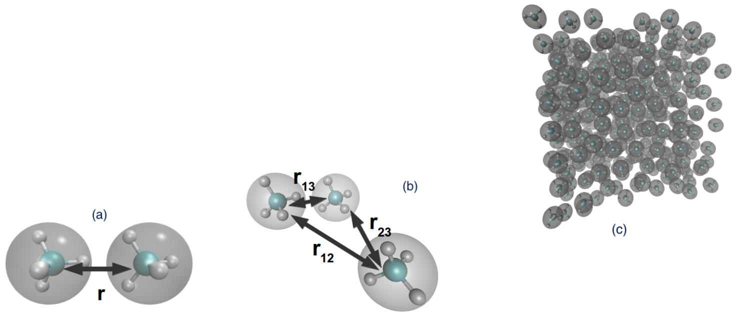

Figure 5.2: Snapshot of model systems in atomistic and coarse-grained description. (a-b) Two and three methanes used for the estimation of the CG effective potential from isolated molecules. (c) Bulk methane liquid.

## 5.5 Results

### 5.5.1 Calculation of the effective two-body CG potential

First, we present data related to the calculation of the two-body potential of mean force for the ideal system of two (isolated) molecules. For such a system the conditional M-body CG PMF is a 2-body one. In Figures 5.3a and 5.3b we provide data for the CG effective interaction between two methane and ethane molecules, through the following methods:

(a) A calculation of the PMF using the constraint force approach, $W^{(2),\text{full, f}}$, as described in section 5.4.2. In this case the constraint force required to keep two methane molecules fixed at a specific distance is computed. Then through a numerical integration the effective potential between the two molecules (CG particles), $U_{\text{CF}}^{\text{PMF}}$, is computed. This is a method that has been extensively used in the literature to estimate effective pair CG interaction between two molecules, as well as differences in the free energy between two states. Alternatively, through the same set of atomistic configurations the two-body PMF, $W^{(2),\text{full, u}}$, can be directly calculated through

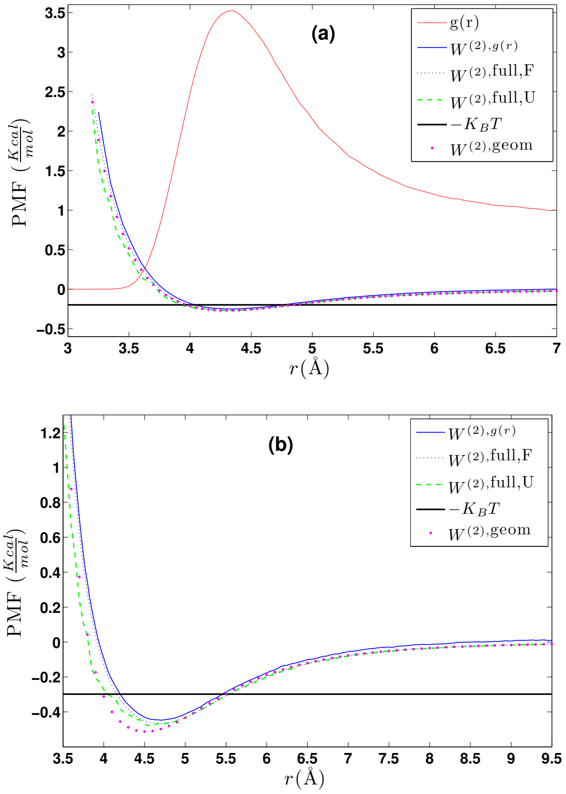

Figure 5.3: Representation of the two-body PMF, for two isolated molecules, as a function of distance $r$, through different approximations: geometric averaging, (constrained) force matching and inversion of $g(r)$. (a) $CH_4$ at $T = 100K$, (b) $CH_3$–$CH_3$ at $T = 150K$. For the methane the corresponding $g(r)$ curve is also shown.

Eq (5.22).

(b) A direct calculation of the PMF, $W^{(2),geom}$, using a geometrical approach as described in Section 5.4.2 that involves the direct calculation of the constraint partition function, treating the two molecules as rigid bodies. Note that in this case in the all-atom description bond lengths and bond angles are kept fixed.

(c) DBI method: The CG effective potential, $W^{(2),g(r)}$, is obtained by inverting the pair (radial) correlation function, $g(r)$, computed through a stochastic LD run with only two methane (or ethane) molecules freely moving in the simulation box. The pair correlation function, $g(r)$, of the two methane molecules is also shown in Figure 5.3a.

The first two of the above methods refer to the direct calculation of the constrained partition function (5.10) with constrained forces and canonical sampling, while the third uses the "Direct Boltzmann Inversion" approach. All above data correspond to temperatures in which both methane and ethane are liquid at atmospheric pressure (values of $-k_BT$ are also shown in Figure 5.3).

First, for the case of the two methane molecules (Figure 5.3a) we see very good agreement between the different methods. As expected, slightly more noisy is the $W^{(2),\text{full},U}(r_{12})$ curve as fluctuations in the $\langle e^{-\beta V(q)}\rangle$ term for a given $r_{12}$ distance in equation (5.22), are difficult to cancel out. The small probability configurations in high potential energy regimes having a large impact in the average containing the exponent, hence the corresponding plot is not as smooth as the others are. In addition, as previously mentioned, $W^{(2),\text{full, F}}$ comes from the same trajectory (run) but the integration of the $\langle f\rangle_{r_{12}}$ from $r_{\text{cutoff}}$ up to $r_{12}$ washes out any non-smoothness. Note, that for the same system recently CG effective potentials based on IBI, force matching and relative entropy methods have been derived and compared against each other. [30]

Second, for the case of the two ethane molecules (Figure 5.3b) we see a good, but not perfect, agreement between the different sets of data, especially in the regions of high potential energy (short distances). This is not surprising if we consider that high energy data from any simulation technique that samples the canonical ensemble, exhibit large error bars, due to difficulties in sampling. The latter is more important for ethane compared to methane case due to its molecular structure; indeed the atomistic structure of methane approximates much better the spherical structure of CG particles than ethane. The only method that provides a "full", within the numerical discretization, sampling at any distance is the geometric one; however as discussed before (see Section 5.4) such a method neglects the bond lengths and bond angle fluctuations.

Next, we also examine an alternative method for the computation of the effective CG potential, by calculating the approximate terms from the cluster expansion approach. For the latter we use the data from the constraint runs of two methane molecules integrated over all atomistic degrees of freedom, as given in formula (5.20). In Figures 5.4a and b we demonstrate the PMF through cluster expansions and the effect of higher order terms as shown in equation (5.23), of the two isolated molecules, for $CH_4$ and $CH_3 - CH_3$ respectively. As discussed in the Section 5.3, cluster expansion is expected to be more accurate at high temperatures and/or lower densities. For this reason, we examine both systems at higher temperatures, than of the data shown in Figure 5.3; Values of $-k_BT$ are shown with full lines. Both systems show the same behavior. First, it is clear that the agreement between $W^{(2)}$ and the (more accurate) $W^{(2),\text{full}}$ is very good only at long distances, whereas there are strong discrepancies in the regions where the potential is minimum as well as in the high energy regions (short distances). Second, it is evident that adding terms up to the second order with respect to $\beta$, we obtain a

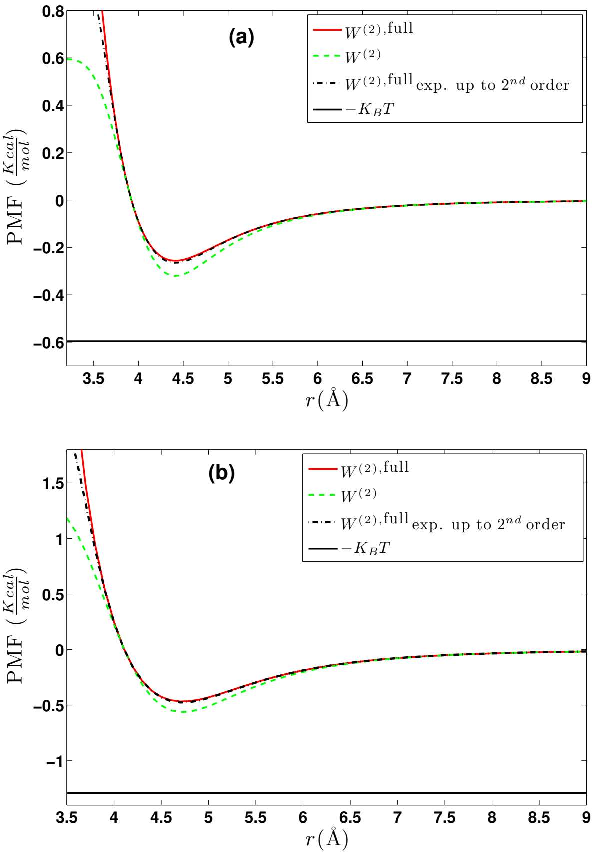

Figure 5.4: Relation of the PMF through cluster expansions and energy averaging at high temperatures, i.e., $W^{(2)}(r_1, r_2)$ and $W^{(2),\mathrm{full}}(r_1, r_2)$ through expansion over $\beta$ for (a) $CH_4$ at $T = 300K$ and (b) $CH_3{-}CH_3$ at $T = 650K$. As expected from the analytic form and the relation between the two formulas, $W^{(2)}$ and $W^{(2),\mathrm{full}}$ tend to converge to the same effective potential.

better approximation of $W^{(2),\text{full}}$.

**Effect of temperature-density**

Next, we further examine the dependence of the PMF, for the two isolated methanes, on the temperature, by studying the system at $T = 80K$,$100K$, $120K$, $300K$ and $900K$. In more detail, in Figures 5.5a and b we compare the difference between $W^{(2)}$ and $W^{(2),\text{full}}$ at different temperatures. As discussed in Section 5.3, the cluster expansion method is valid only in the high temperature regime. This is directly observed in Figure 5.5a; at high temperatures, $W^{(2)}$ is very close to $W^{(2),\text{full}}$, which is exact for the system consisting of two molecules. Note the small differences at short distances, which, as also discussed in the previous subsection, are even smaller if higher order terms are included in the calculation of $W^{(2)}$; see also Figure 5.4.

On the contrary, at low temperatures there is a strong discrepancy around the potential well as shown in Figure 5.5b. In fact, for values of $r$ close to the potential well and for rather high values of $\beta$ the contribution to the integral (5.2) is large and the latter can exceed one, rendering the expansion in (5.23) not valid. In Figure 5.5b we see that the term (5.20) is not small so the expansion (5.23) is not valid. The case for ethane is qualitatively similar.

For completeness, we also plot the potential of mean force at different temperatures for the system of two $CH_4$ molecules, see Figure 5.6. In principle, equation (5.20) is a calculation of free energy, hence it incorporates the temperature of the system and thus both approximations to the exact two-body PMF, $W^{(2)}$ and $W^{(2),\text{full}}$, are not transferable. Indeed, we observe slight differences in the CG effective interactions (free energies) for the various temperatures, which become larger for the highest temperature.

### 5.5.2 Bulk CG CH4 runs using a pair potential

In the next stage, we quantitatively examine the accuracy of the effective CG interaction potential (approximation of the two-body PMF), in the liquid state based on structural properties like $g(r)$. Here we use the different CG models (approximated pair CG interaction potentials) derived above, to predict the properties of the bulk CG methane and ethane liquids. In all cases we compare with structural data obtained from the reference all-atom bulk system, projected on the CG description.

In Figures 5.7a and b we assess the discrepancy between the CG (projected) pair distribution function, $g(r)$, taken from an atomistic run, and the one obtained from the corresponding CG run based on $W^{(2),\text{full}}$ as already seen in Figure 5.3 of methane and ethane respectively. Note that $g(r)$ is directly related to the effective CG potentials ($N = 2$ in Eq (5.8)).

It is clear that for methane (Figure 5.7a) the CG model with the $W^{(2),\text{full}}$

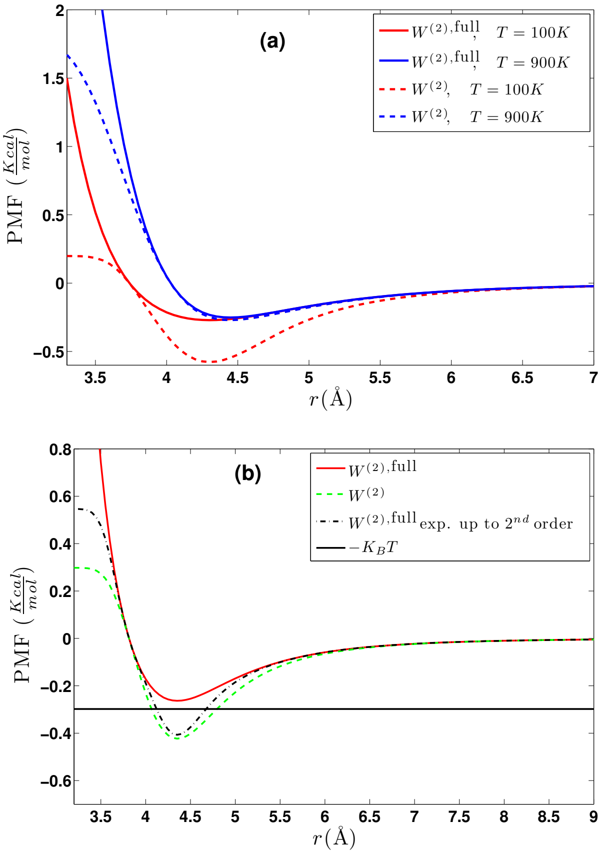

Figure 5.5: (a) PMF through cluster expansions, using (5.20) and (5.23) for different temperatures for the $CH_4$ model. (b) PMF through cluster expansions and energy averaging, i.e., $W^{(2)}(r_1, r_2)$ and $W^{(2),\mathrm{full}}(r_1, r_2)$ through expansion over $\beta$ for $CH_4$ at $T = 150K$. The expansion is not valid at this temperature.

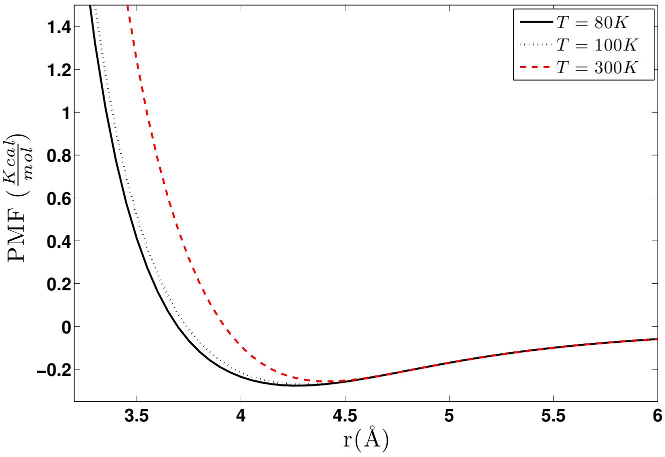

Figure 5.6: Potential of mean force at different temperatures (geometric averaging). Two CH4 molecules at T=80K, 100K, 120K, 300K

potential gives a g(r) very close to the one derived from the analysis of the all-atom data. This is not surprising if we consider that for most molecular systems, small differences in the interaction potential lead to even smaller differences in the obtained pair correlation function. Interestingly the CG model with the $W^{(2)}$ is also in good agreement with the reference one, despite the small differences in the CG interaction potential discussed above (see Figures 5.4a and 5.8b). As expected, the difference comes from the missing higher order terms of eq (5.14).

The fact that the CG effective potential, which is derived from two isolated methane molecules, gives a very good estimate for the methane structure in the liquid state is not surprising if we consider the geometrical structure of methane, which is rather close to the spherical one. On the contrary, for the case of ethane (Figure 5.7b) predictions of $g(r)$ using pair CG potential are much different compared to the atomistic one, especially for the short distances. Even larger differences would be expected for more complex systems with long-range interactions, such as water. [30]

Similar is the case also for the other temperatures ($T = 80K$) studied in this work (data not shown here).

### Effect of temperature-density

We further study the structural behavior of the CG systems at different state points; i.e., temperature/density conditions, compared to the atomistic ones. First, we examine the temperature effect by simulating the systems discussed above (see Figure 5.7) at higher temperatures; however keeping the same density. In Figures 5.8a,b we present the RDF of methane from

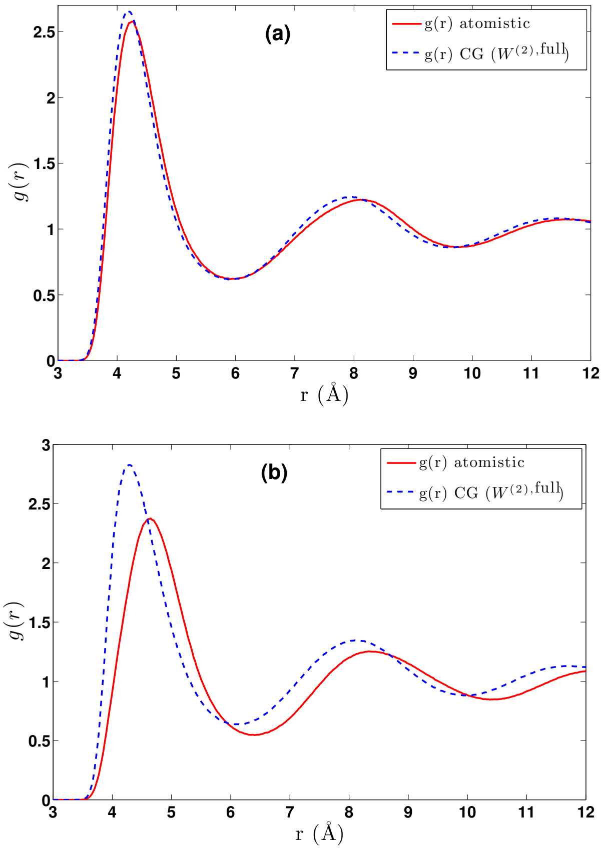

Figure 5.7: RDF from atomistic and CG using pair potential, $W^{(2)}$, for $CH_4$ system at $T = 80K$ (a) and $CH_3 - CH_3$ at $T = 150$(b). Spherical CG approximation to the non-symmetric ethane molecule induces discrepancy and implies there is more room for improvement.

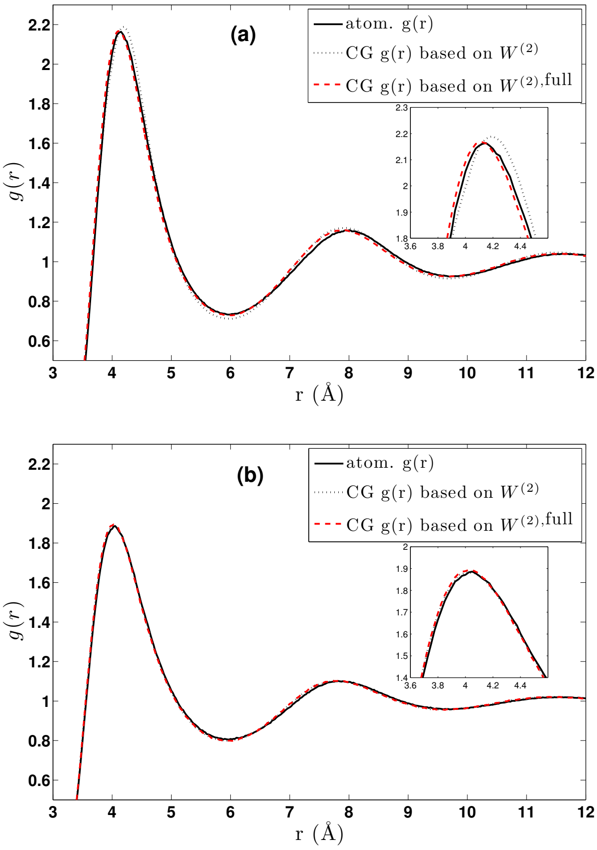

Figure 5.8: RDF of methane from atomistic data, and CG models using pair potential at different temperatures: (a) $T = 300K$, (b) $T = 900K$. In both cases the density is $\rho_1 = 0.3799\frac{gr}{cm^3}$.

atomistic and CG runs using pair potential at $T = 300K$, and $T = 900K$ respectively.

It is clear that the analysis of the CG runs using the $W^{(2),\text{full}}$ potential gives a pair distribution function $g(r)$ close to the atomistic one for both (high) temperatures, similar to the case of the $T = 80K$ shown before. In addition, the CG model with the $W^{(2)}$ potential is in very good agreement with the atomistic data at high temperature (Figure 5.8b), whereas there are small discrepancies at lower temperatures (Figure 5.8a), in particular at the maximum of $g(r)$. This is shown in the inset of Figures 5.8a,b. Note also that in this high temperature the incorporation of the higher order terms in $W^{(2)}$ leads to very similar potential as the $W^{(2),\text{full}}$ (see also Figure 5.4a), and consequently to very accurate structural $g(r)$ data as well.

Next, we examine the structural behavior of the CG systems at different

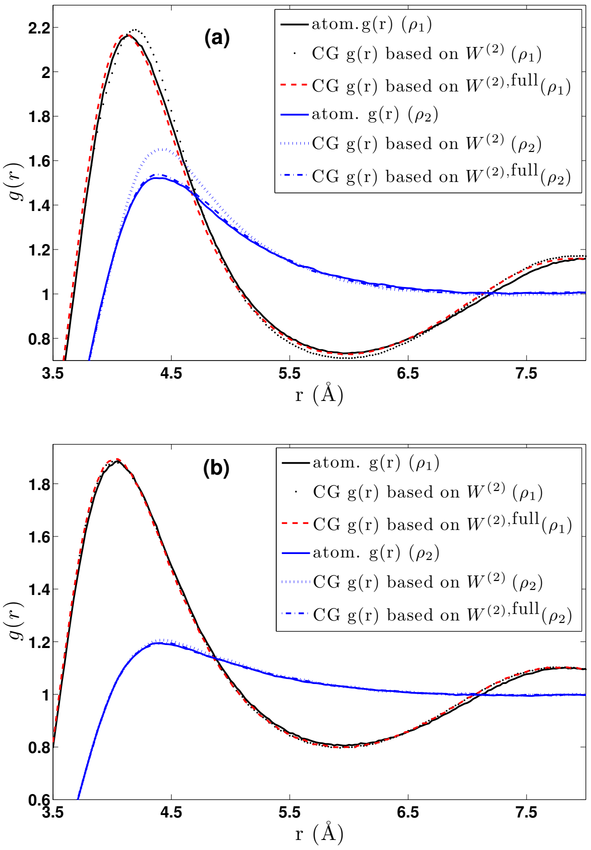

Figure 5.9: RDF of methane from atomistic and CG using pair potential at different densities $\rho_1 > \rho_2$, (a) $T = 300K$, (b) $T = 900K$.

densities. In Figure 5.9a we present the $g(r)$ from atomistic and CG runs using pair potential at different densities ($\rho_1 = 0.3799 \frac{gr}{cm^3}$ and $\rho_2 = 0.0395 \frac{gr}{cm^3}$ and $T = 300K$, and $T = 900K$). There is apparent discrepancy from the reference (atomistic) system in both densities in agreement to the data discussed above in Figure 5.8a.

For the case of higher temperature data ($T = 900K$) and the same densities, as shown in Figure 5.9b, the pair distribution function, $g(r)$, obtained from the CG model with the $W^{(2)}$ effective interaction is very close to the data derived from the $W^{(2),\text{full}}$ one, and in very good agreement to the reference, all-atom, data. This is not surprising since, as discussed before, at high temperatures the cluster expansion is expected to be more accurate, since cluster expansions hold for high $T$ and low $\rho$.

Overall, the higher the temperature the better the agreement in the $g(r)$ derived from the CG models using any of $W^{(2)}$ and $W^{(2),\text{full}}$. These data are in better agreement with the atomistic data as well.

## 5.6 Effective three-body potential

In the last part of this work we briefly discuss the direct computation of the three-body effective CG potential and its implementation in a (stochastic) dynamic simulation. More results about the three-body terms will be presented in a future work. [135]

### 5.6.1 Calculation of the effective three-body potential

In the following we present data for the 3-body potential of mean force estimated from simulation runs and geometric computations involving three isolated molecules. We have two suggestions for the 3-body PMF: (a) Formula (5.21) derived from cluster expansion formalism, which is valid for rather high temperatures and (b) another one based on the McCoy-Curro scheme given in formula (5.24). Here we present data using the latter formula, a detailed comparison of the three-body effective potentials, $W^{(3),\text{full}}$ and $W^{(3)}$, using Eqs. (5.21) and (5.24) will be given elsewhere. [135]

Similarly to the two-body potential, the corresponding calculations can be performed by constrained molecular dynamics (or any other method that performs canonical sampling). For this one needs to calculate the derivative of the three-body potential with respect to some distance. However, as previously stated, deterministic MD simulations of a constrained system might easily get trapped in local energy minima, so we utilized stochastic dynamics for the three-body case. In addition, rare events (high energy, low probability configurations) induce noise to the data, despite long equilibration (burn-in) periods or stronger heat-bath coupling in the simulations. Although smoothing could in principle have been applied, it would washout important information needed upon derivation with respect to positions

($f = -\nabla_{\mathbf{q}} W^{(3)}$). Therefore, we choose here to present results from the "direct" geometric averaging approach. The total calculations are one order of magnitude more than the two-body ones (all possible orientations of the two molecules for one of the third one), so special care was given to spatial symmetries.

The new effective three-body potential, $W^{(3),\text{full}}$, is naturally a function of three intermolecular distances: $r_{12}, r_{13}, r_{23}$. The discretization of the COM's in space is on top of the angular discretization mentioned in Section 5.4.2 and relates to the above three distances. The investigation of $W^{(3),\text{full}}$ for all possible distances is beyond the scope of this article. Here, we only study some characteristic cases, showing $W^{(3),\text{full}}$ data as a function of distance $r_{23}$ for fixed $r_{12}$ and $r_{13}$, comparing always with the sum of the corresponding two-body terms. In more detail, in Figures 5.10a-d we present simulations based on the effective three body potential $W^{(3),\text{full}}$ and the sum $\sum W^{(2),\text{full}}$ (geometric averaging) for $CH_4$ at $T = 80K$ for different COM distances [$\mathring{A}$]: (a) $r_{12} = 3.9$, $r_{13} = 3.9$, (b)$r_{12} = 4.0$, $r_{13} = 4.0$, (c) $r_{12} = 4.3$, $r_{13} = 4.0$, (d) $r_{12} = 3.8$, $r_{13} = 5.64$. At smaller distances, the potential of the triplet deviates from the sum of the three pairwise potentials and this is where improvement in accuracy can be obtained. As shown in Figure 5.10 improvement is needed for close distances around the (3 dimensional) well. We used a 3-dimensional cubic polynomial to fit the potential data (conjugate gradient method) which means that 20 constants should be determined. A lower order polynomial cannot capture the curvature of the forces upon differentiation. The benefit of this fitting methodology (over partial derivatives for instance) is the analytical solution of the forces with respect to any of $r_{12}, r_{13}, r_{23}$ in contrast to tabulated data that induce some small error.

Overall, there are clear differences between the 3-body PMF, $W^{(3),\text{full}}$, and the sum of three two-body interactions, $\sum W^{(2),\text{full}}$, at short $r_{12}$, $r_{13}$ and $r_{23}$ distances. On the contrary, for larger distances the sum of two-body interactions seems to represent the full three-body PMF very accurately. This is a clear indication of the rather short range of the three-body terms. Based on the above data, the range of the 3-body terms for this system (methane at $T = 80K$) is: $r_{12} \in [3.8 : 4.1]\mathring{A}, r_{13} \in [3.8 : 4.1]\mathring{A}$ and $r_{23} \in [3.8 : 5]\mathring{A}$; hence, the maximum distance for which three-body terms were considered, is $r_{\text{cut-off},3}$=5$\mathring{A}$. In practice we need to identify all possible triplets within $r_{\text{cut-off},3}$. Naturally, by including higher-order terms the computational cost has increased as well. More information about the numerical implementation of the three-body CG effective potential and its computational efficiency will be given elsewhere. [135] We should state here that in order to keep the temperature constant (in the BBK algorithm), due to the extra three-body terms in the CG force field, a larger coupling constant value for the heat bath was required.

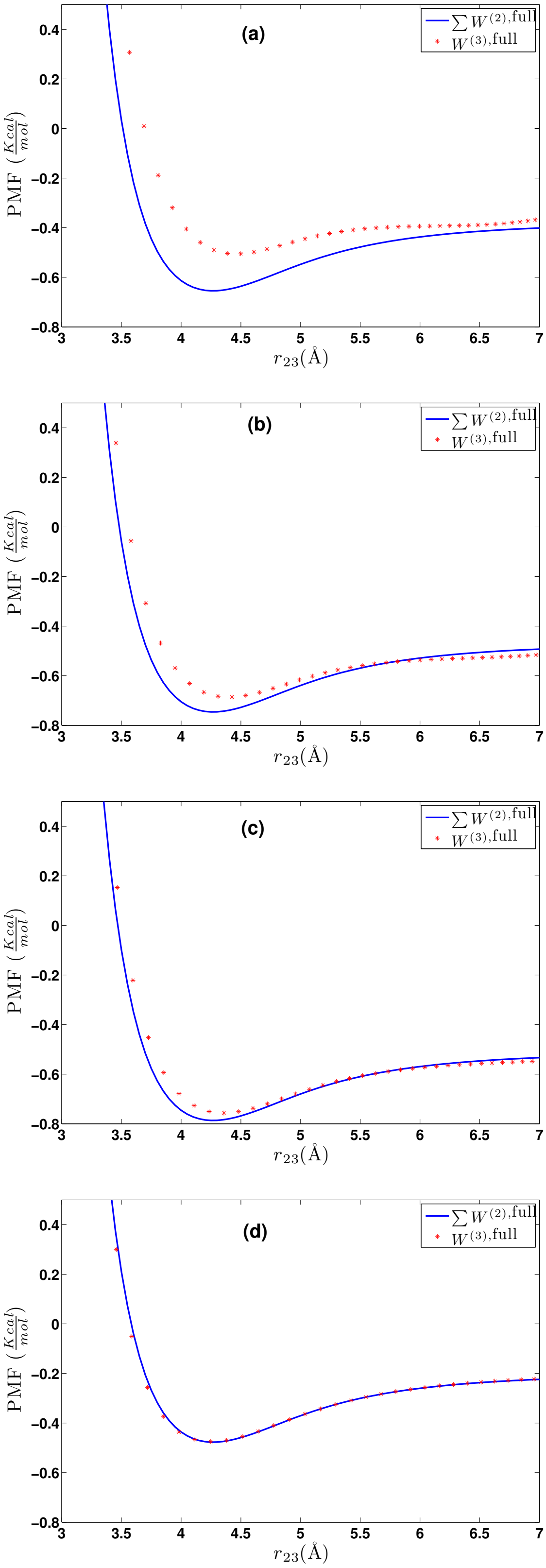

Figure 5.10: Effective potential comparison between the $W^{(3),\text{full}}$ 3-body and $\sum W^{(2),\text{full}}$ simulations (geometric averaging) for $CH_4$ at $T = 80K$ for different fixed COM distances $[\mathring{A}]$ . (a) $r_{12} = 3.9, r_{13} = 3.9$ (b)$r_{12} = 4.0, r_{13} = 4.0$ (c) $r_{12} = 4.3, r_{13} = 4.0$, (d)$r_{12} = 3.8, r_{13} = 5.64$ .

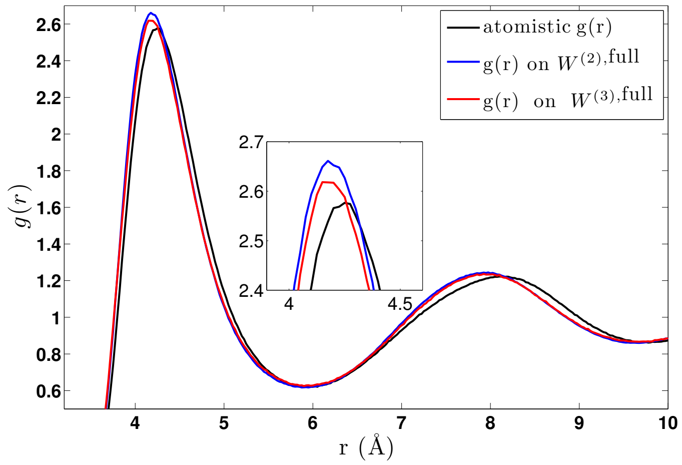

Figure 5.11: RDF from atomistic and CG using pair, $W^{(2),\text{full}}$, and three-body, $W^{(3),\text{full}}$, potential for $CH_4$ ($T = 80K$). Three dimensional cubic polynomial was used for the fitting.

### 5.6.2 CG Runs with the effective three-body potential

Next we examine the effect of the 3-body term on the CG model by performing bulk CG stochastic dynamics simulations using the new CG model with the 3-body terms described above. In this case we incorporate the 2-body CG effective potential described before for distances larger than $r_{\text{cut-off},3}$, whereas we use $W^{(3),\text{full}}$ for triplets with all distances ($r_{12}$, $r_{13}$, and $r_{23}$) below $r_{\text{cut-off},3}$. In practice, we compute all possible pair interactions and for the triplets with distances in the above defined range, we "correct" by adding the difference between $W^{(3),\text{full}}$ and the corresponding sum of $W^{(2),\text{full}}$; i.e. the difference between the data sets shown in Figures 5.10a-d.

Results on the pair distribution function, $g(r)$, for bulk (liquid) methane at $T = 80K$ are shown in Figure 5.11. In this graph data from the atomistic MD runs (projected on the CG description), the CG model involving only pair CG potentials, and the new CG model that also involves 3-body terms are shown. First, it is clear that $g(r)$ data derived from the CG model that involves only pair CG potentials show clear deviations, compared to the reference all-atom data. Second, the incorporation of the three-body terms in the effective CG potential slightly improves the prediction of the $g(r)$, mainly in the first maximum regime.

## 5.7 Discussion and conclusions

In recent years we have experienced an enormous increase of computational power due to both hardware improvements and clever CPU-architecture.

However, atomistic simulations of large complex molecular systems are still out of reach in particular when long computational times are desirable. A generic strategy in order to improve efficiency of the computational methods is to reduce the dimensionality (degrees of freedom) by considering systematic coarse-grained models. There have been many suggestions on how to compute the relevant CG effective interactions in such models; a main issue here is that even if in the microscopic (atomistic) level there are only pair interactions, after coarse-graining a multi-body effective potential (many-body PMF) is derived, which for realistic molecular complex systems cannot be calculated. Therefore, a common trend has been to approximate them by an "effective" pair potential by comparing the pair correlation function $g(r)$. This seems reasonable since given the correlation function one can solve the "inverse problem" [142] and find an interaction to which it corresponds. But, this is an uncontrolled approximation without thermodynamic consistency.

Instead, here we suggest to explicitly compute the constrained configuration integral over all atomistic configurations that correspond to a given coarse-grained state and from that suggest approximations with a quantifiable error. This is similar to the virial expansion where one needs to integrate over all positions of particles that correspond to a fixed density and it is based on the recent development of establishing the cluster expansion in the canonical ensemble. [132]; see also Ref. [140, 71] for the corresponding (in the canonical ensemble) expansions for the correlation functions and the Ornstein-Zernike equation. The main drawback that limits the applicability of these expansions is that they are rigorously valid only in the gas phase. To extend them to the liquid state is an outstanding problem and even several successful closures like the Percus-Yevick are not rigorously justified. Therefore, there is need of further developing these methods and relate them to computational strategies.

In this paper we extend the above methods by presenting an approach based on cluster expansion techniques and numerical computations of isolated molecules. As a first test we presented a detailed investigation of the proposed methodology to derive CG potentials for methane and ethane molecular systems. Each CG variable corresponds to the center-of-mass for each molecule. Below, we summarize our main findings:

(a) The hierarchy of the cluster expansion formalism allowed us to systematically define the CG effective interaction as a sum of pair, triplets, etc. interactions. Then, CG effective potentials can be computed as they arise from the cluster expansion.

(b) The two-body coarse-grained potentials can be efficiently computed via the cluster expansion giving comparable results with the existing methods, such as the conditional reversible work. In addition we present a more efficient direct geometric computation of the constrained partition function.

(c) The obtained pair CG potentials were used to model the corresponding liquid systems and the derived $g(r)$ data were compared against the all-atom ones. Clear differences between methane and ethane systems were observed; For the (almost spherical) methane, pair CG potentials seems to be a very good approximation, whereas much larger differences between CG and atomistic distribution functions were observed for ethane.

(d) We further investigated different temperature and density regimes, and in particular cases where the two-body approximations are not good enough compared to the atomistic simulations. In the latter case, we considered the next term in the cluster expansion, namely the three-body effective potentials and we found that they give a small improvement over the pair ones.

Overall, we conjecture that the cluster expansion formalism can be used in order to provide accurate effective pair and three-body CG potentials at high $T$ and low $\rho$ regimes. In order to get significantly better results in the *liquid* regime one needs to consider even higher order terms, which are in general more expensive to be computed and more difficult to be treated. A more detailed analysis of the higher-order terms will be a part of a future work. [135] Finally, another future goal is to extend this investigation in larger molecules (e.g. polymeric chains) that involve intra-molecular CG effective interactions as well, and to systems with long range (e.g. Coulombic) interactions.

# Chapter 6

# 2-Body, 3-Body calculations

## 6.1 Introduction

In this chapter, we show how to construct effective Coarse Grained (CG) potentials through atomistic simulations, using only a small number of atoms. The numerical algorithm is applied on CG potentials involving 2- and 3-body potentials; however, our proposed methodology is quite general and, in principle can be extended even for higher order terms. We later on use the obtained effective potentials in CG level simulations and compare the results (specific observables) with the corresponding atomistic (projected to the Coarse level) simulations, in order to assess the efficiency and accuracy of our findings.

The chapter is divided as follows: First, we set up the problem with the construction of 2-Body effective potential and all the issues and difficulties associated with it. Different techniques of estimating the same quantity enable us to validate our results. We then proceed to the 3-Body case which is intrinsically more complex and computationally more expensive. The challenge is the construction of accurate CG potentials that can be used in realistic production runs.

## 6.2 2-Body

Our goal is to estimate the ensemble average of an observable quantity, for instance non-bonded potential energy: $\langle V^{nb}(\mathbf{q})\rangle|_{r_{12}}$ over a subspace of the phase space. Here $\mathbf{q}$ is the vector of cartesian coordinates of all the atomic particles, $\mathbf{r_1}$, $\mathbf{r_2}$ are the projections (3 dimensional coordinates) of the first and second centres of mass (COM's) and $r_{12} = |\mathbf{r_1} - \mathbf{r_2}|$ is scalar and fixed. As mentioned in the Introduction, it is common practice that these kind of average quantities are exported from long atomistic trajectories by properly analyzing the atomistic data. Usually, a binning procedure is used by defining a discretization step, $dr$, of the variable $r$ and calculate the specific chosen quantity ($V^{nb}$) at every step $dr$ accumulating its value to the corresponding bin. After the run is over, the mean value over the grouped values that correspond to $[r, r+dr]$ converges to the desired $\langle V^{nb}(\mathbf{q})\rangle|_r$.

In other words, we end up calculating a histogram. For the case of two particles in vacuum, this is a slow process, for a number of reasons. First and foremost, the bins that correspond to longer distance values $r$ are heavily populated, due to the vacuum of the simulation box. In effect, the close $r$ value bins merely have a small number of samples, if any at all, even for long trajectories. In addition, deterministic thermostats, like the well known and frequently used Nose-Hoover fail to reach the target average temperature when not in bulk.

All of the above urged us to constrain the molecule COM's in space. This technique efficiently tackles with the problem of ”poor sampling” at the high potential energy parts of the configuration space. The ensemble average is a histogram in this case as well and the number of samples per bin is defined in the beginning of the simulation. After the number of steps (samples) is simulated, we artificially move the COM's apart by $dr$ and proceed to estimate $\langle V^{nb}(\mathbf{q})\rangle|_{r+dr}$. The two-particle (out of $N$ in total) projected constrained partition function is given:

$$Z^{(2),proj}(r_1, r_2) := \int_{\{T_1(\mathbf{q_1})=r_1,\, T_2(\mathbf{q_2})=r_2\}} e^{-\beta V^{nb}(\mathbf{q})} d\mathbf{q} \tag{6.1}$$

### 6.2.1 Constrained runs

The constraining of the molecules is performed as follows: First we select the distance $r$ between the COM's. Then we pin the COM of each CG particle in space and on every step throughout the MD trajectory, we subtract the total force acting on each COM. Hence we allow the atoms inside each CG particle to move, resulting in rotations but **not** translations of the CG degrees of freedom. Otherwise (COM's at fixed distance but mutually rotating as a rigid body apart from the individual rotation around each COM), we would have had to subtract the rotational entropy (see [38] ). Technically, this pinning procedure requires two constrains on every time step (for BBK). One is based on the momenta (3.23) and the other one on the forces. Suppose we are on the $t-th$ time step and calculate the momenta at the half step $p^{t+1/2}$ in eq (3.23). Exactly after that, we perform constraining in order to prevent velocity drift. This means, for each atom $j$ inside the CG particle $i$ we subtract the velocity of the centre of mass:

$$p_{i,j}^{t+1/2} = p_{i,j}^{t+1/2} - \frac{\sum_{j=1}^{\#\text{atoms}} p_{i,j}^{t+1/2}}{\#\text{atoms}} \tag{6.2}$$

In the fraction we don't multiply by the atomic mass because each CG particle is considered as a rigid body. On the second step of the integrator, the coordinates $q^{t+1}$ are updated and a force calculation follows. The last step in the integration scheme updates the momenta $p^{t+1}$ to a complete time step $dt$ through the total forces (conservative, drift and random), so a new constraint is needed:

$$p_{i,j}^{t+1} = p_{i,j}^{t+1} - \frac{\sum_{j=1}^{\#\text{atoms}} p_{i,j}^{t+1}}{\#\text{atoms}} \tag{6.3}$$

In the velocity Verlet case, there are two steps in the integrator so we only need one constrain for the total conservative forces, in exactly the same manner, right after $-\nabla V$ is computed. We note that the velocity centre of mass of the whole system should be set to zero once, in the beginning of the run, in the same way.

A very useful quantity (observable) derived from the constrained runs is the average force among the two particles (constraint COMs), $\langle f^{nb}\rangle|_{r_{12}}$, which is used for the calculation of the PMF. We define the vector $\mathbf{r_{12}}$ connecting the two COM's as the support of the magnitude $f$ of the force $\mathbf{f}$ between those CG particles. This force can either be positive (repulsive) or negative (attractive). In order to simplify the calculations both COM's are pinned in space as mentioned and $\mathbf{r_{12}}$ is parallel to the $X$ axis. Later on, the COM of the second particle is further moved at $\mathbf{r_{12}} + \mathbf{dr}$, along the $X$-dimension as well. Naturally, the whole procedure is implemented in a general way and works for any displacement along $XYZ$ as well. So we compute the total non-bonded force vector $\mathbf{f_i^{nb}}$ acting on COM $i$ ($\mathbf{f_1^{nb}}$=$-\mathbf{f_2^{nb}}$), then find its magnitude and project it along $\mathbf{r_{12}}$. The bonded and angular forces cancel-out, except for the very close distances where some bonds might get compressed (see section on noise 6.2.4). Note, that integration of the average force between particles 1 and 2 leads to the corresponding potential of mean force.

In figure (6.3) we compare a $CH_4$ United Atom (CG) model with given LJ parameters with our computed ensemble average of the (effective) potential between atoms, for the case of the OPLS forcefield. *Our proposed constraining method satisfactorily estimates the CG potential.*

As mentioned in the previous section, $\langle f\rangle$ at this fixed value of $r_{12}$ is a simple arithmetic mean over all samples, which follow the canonical ensemble.

Then we artificially move the CG particles apart by $dr$, continue the simulation and repeat the procedure to get $\langle f\rangle_{r+dr}$. A schematic of the constraining is shown in figure (6.1).

We stress the fact that although we have biased the dynamics of the run, sampling (ensemble average) with respect to the *proper equilibrium measure is performed.* This is due to the fact that on every step, we first sample and then constrain; correct the forces on each COM to remain in place, allowing it to rotate freely. An alternative way to avoid non-ergodicity and speed up

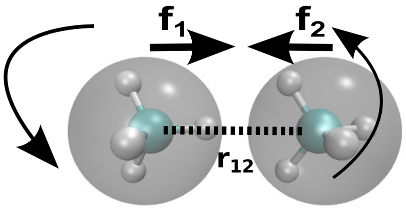

Figure 6.1: 2 Constrained $CH_4$ in space, along the $X$-dimension. Atomistic and CG description.

the sampling is through biased potentials, like umbrella sampling, conformational flooding etc. We also note that this is a *free energy calculation* type of a problem [33]. Therefore, there is an explicit temperature dependence on the observables under study.

### Stochastic Dynamics

During the constrained simulations in vacuum, we encountered severe issues when trying to place the two molecules, and sample the phase space, at close distances. Despite choosing very small timesteps, of the order of $\mathcal{O}(10^{-15})$ sec, the intermolecular bonds would either break, crashing the simulation, or rotate in parallel. The latter means that the system has found a local energy minimum and the integrator is not able to explore the phase space properly. In other words, the sampling is not performed using the correct equilibrium measure because the system cannot reach equilibrium distribution, hence the system **is not ergodic** any more.

would get trapped in every case, even for higher temperatures, and this could be seen upon simple inspection with VMD ([143]) apart from the noisy averages.

In order to overcome the above problem we used stochastic dynamics, by employing the Langevin thermostat with the BBK integrator. In this case the system is not ”trapped” in energy minima thanks to the random forces, or ”kicks” in the equations of motion, which model the heat bath. The only drawback using this integrator in the increased computational cost associated with the force evaluation. Apart from the vacuum runs, Langevin dynamics proved also to be faster in equilibrating low density bulk melts, in comparison to the deterministic MD thermostat (see figure 5.9).

|  | $\frac{maxsteps}{dr}$ | $\frac{maxsteps}{dr}$ | total # steps | $dr$ |
|---|---|---|---|---|
| $CH_4$ | $5\ 10^7 r \in [3.2 : 5.2)$ | $10^5 r \in [5.2 : r_{cut}]$ | $2\ 10^9$ | $0.05\mathring{A}$ |
| $CH_3 - CH_3$ | $15\ 10^7 r \in [3.5 : 5.7)$ | $10^5 r \in [5.7 : r_{cut}]$ | $10^{10}$ | $0.05\mathring{A}$ |

Table 6.1: run parameters

### 6.2.2 Run specifications

Practically, in all above cases our goal is to approximately estimate the constrained partition function:

$$\int\int_{T(q_i)=r_i} e^{-\beta V^{nb}(\mathbf{q})} d\mathbf{q_1} d\mathbf{q_2}, \tag{6.4}$$

which is an intrinsically extremely hard calculation (see section (6.2.3)). Every system has a different partition function and the difficulty of calculating it, monotonically increases depending on its complexity. In our case, $CH_4$ is a spherically symmetric molecule and the relative orientations needed are less than the axially symmetric $CH_3 - CH_3$ molecule. After long simulation times for the constrained run and taking into account the error bars of the observables, we can determine a lower bound of the steps required in order to obtain a smooth graph (usually the observable is the pair potential). We concluded that the discretization of $dr = 0.05\mathring{A}$ is sufficiently accurate for the length scales we are interested in. The $PMF$ calculations, described above, require numerical integration (e.g. trapezoidal or Simpson's rule) over this variable $r$, and the associated error with this discretization value is smaller than that of the order of the method scheme. We call the variable containing the number of samples (steps) needed for every fixed $r_{12}$ value in the histogram as "maxstepsdr". The variable values are summarized in table 6.1.

In figure (6.2)a we see good agreement between the method of calculating the ensemble average of the potential and the method of integration of the forces ensemble average to get the same quantity for $CH_4$. In figure (6.2)b we have the same result for the $CH_3 - CH_3$ system. In this case the molecule gets easily trapped in local energy minima and the sampling is inadequate at close distance, but the calculation of integrating the forces from $r_{cut}$ down to $r_{min}$ smooths out $\langle f \rangle$. The problem persists in higher temperatures as well, even for very large trajectories and stronger coupling (larger $\xi$ constant in BBK) with the heat bath. We address this issue in section (6.2.4).

### 6.2.3 $U^{eff}$ estimator

We need to calculate the integral

$$e^{\beta U^{(2)}_{eff}} = \int_{|q_1 - q_2| = r} e^{-\beta V^{nb}(q_1,q_2)} dq_1 dq_2 \tag{6.5}$$

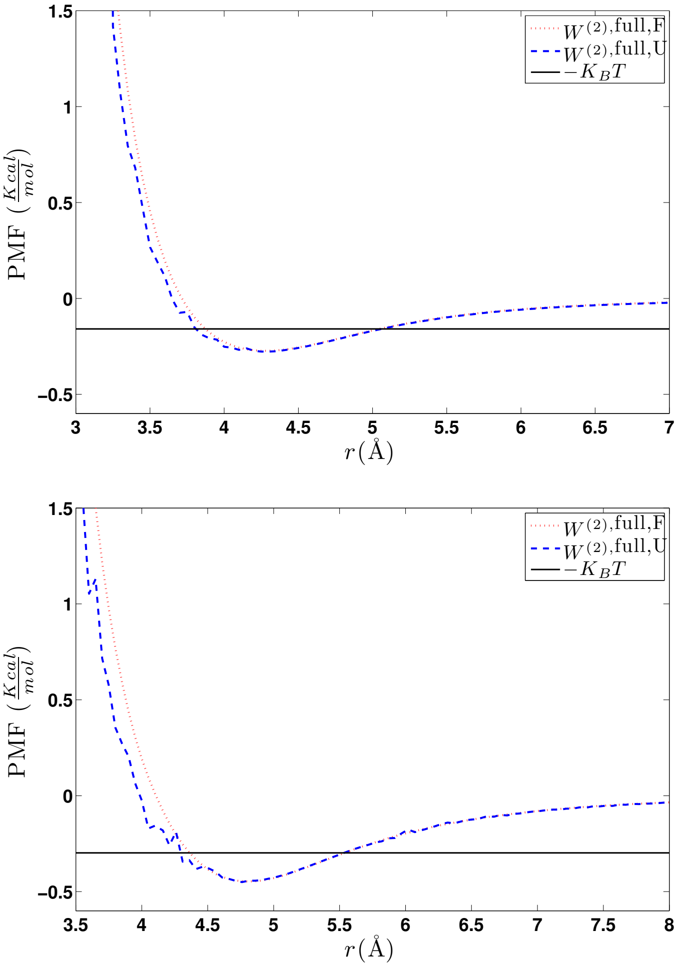

Figure 6.2: PMF $\langle V^{nb}\rangle$ and $\langle f\rangle$ for (a) $CH_4$ at $T = 80K$. Good agreement between methods. (b) same for $CH_3 - CH_3$ at $T = 150K$, noisy $\langle V^{nb}\rangle$.

numericaly, where the atomistic positions for $CH_4$ atoms are: $q_1 = \{\bar{q}_1, q_{1,1}, q_{1,2}, q_{1,3}, q_{1,4}\}$ and $q_2$ is given accordingly. On the other hand, the BBK integrator produces trajectories $q_1^{(t)}, q_2^{(t)}$ according to the stationary measure

$$\mu(\mathbf{q}) = \frac{e^{-\beta V^{bond} + V^{angle} + V^{nb}}}{\int e^{-\beta V^{bond} + V^{angle} + V^{nb}}} \tag{6.6}$$

the denominator being the atomistic partition function $Z_{atom}$, so an ergodic average $(1/nsteps \times \sum_k^{nsteps} A(q^{(k)}))$ of a quantity $A(q)$ is given by

$$\langle A(q)\rangle|_\mu = \int A(q) \frac{e^{-\beta V^{bond} + V^{angle} + V^{nb}}}{Z_{atom}} dq = \int A(q) \quad \mu(dq) \tag{6.7}$$

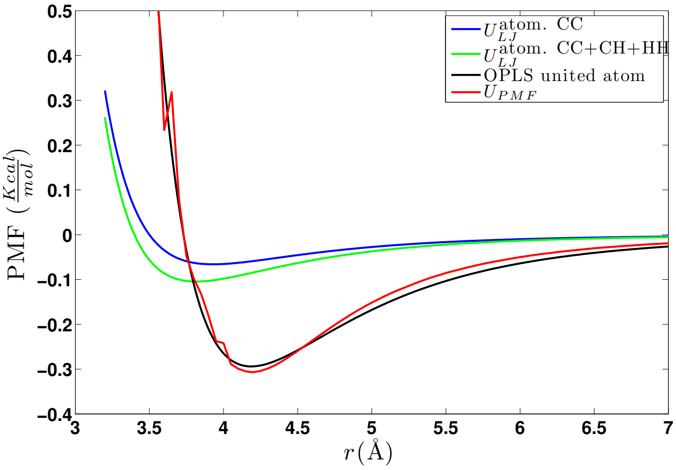

Figure 6.3: PMF $\langle V^{nb} \rangle$ and a United Atom model for $CH_4$ at $T = 300K$, the atomistic $C{-}C$, $C{-}H$ LJ interactions are plotted for magnitude comparison. OPLS UA forcefield [144].

In order to estimate the constrained integral (6.5) instead of (6.7) we need to calculate:

$$\frac{1}{\frac{1}{nsteps}\sum_{k=1}^{nsteps}\frac{1}{e^{-\beta V^{nb}}}} \tag{6.8}$$

as:

$$\begin{aligned}
\frac{1}{\int_{constr}\frac{1}{e^{-\beta V^{nb}}}\frac{e^{-\beta V^{bond}+V^{angle}+V^{nb}}}{\int e^{-\beta V^{bond}+V^{angle}+V^{nb}}}dq} &= \\
\frac{\int e^{-\beta V^{bond}+V^{angle}+V^{nb}}dq}{\int_{constr} e^{-\beta V^{bond}+V^{angle}+V^{nb}}dq} &= \\
\int_{constr} e^{-\beta V^{nb}(q)}dq + \cancel{\frac{\int_{elsewhere} e^{-\beta(V^{nb}+V^{bond}+V^{angle})dq}}{\int_{constr} e^{-\beta(V^{nb}+V^{bond}+V^{angle})}dq}} &
\end{aligned} \tag{6.9}$$

### 6.2.4 MD noise in constrained runs

Overall, for the adequate sampling of the phase space long simulations are required. This is particular problematic at very close distances of the two particles (high energy regimes); i.e. despite the fact that our 2-body constrained runs are computationally very efficient, it is not possible to accurately determine the constraint partition function for such distances. On top of that, when we are trying to compute $\langle V^{nb}(\mathbf{q})\rangle|_r$, we essentially average samples of $e^{-\beta V_k^{nb}}, k = 1$ : maxstepsdr (see section (6.2.3)). In our case, the inverse temperature $\beta$ is positive and at close distances $V^{nb}$ is large, so

$e^{-\beta V^{nb}}$ are very small numbers, in the range $\mathcal{O}(10^{-1})$ -$\mathcal{O}(10^{-4})$. Summing up very long vectors of numbers that small, produces unavoidable numerical errors, even when double precision arithmetic (and proprietary compilers) is used.

Note that the usage of typical block averaging techniques [33] do not improve the results either. This problem is further magnified when we take the logarithm of $\langle e^{-\beta V^{nb}} \rangle|_r$ at the final post processing stage of the calculation, in order to obtain $\langle V^{nb} \rangle|_r$. All of the above urged us to seek for another method to crosscheck the validity of our findings.

### 6.2.5 Inverse $g(r)$

An alternative method to obtain accurate sampling is the Direct Boltzmann Inversion (DBI). The key idea is that we attempt to infer information from a bulk (macroscopic) quantity and construct an effective pair potential. The formula used is the *reversible work theorem*: $e^{-\beta W(q)} = g(r)$ (see section 6.2.5) which relates the radial distribution function $g(r)$ with the potential energy. In our case we have two molecules (CG particles) moving freely in vacuum, so extra care is needed in the proper normalization when calculating the $g(r)$.

$W(r)$ is the total potential of the system for every value $r$ between the COM's of the two CG particles. We simulate this system in the atomistic level, allowing the particles to move around for a very long time in vacuum which resembles the gas phase and the *limit* of the reversible work theorem. In this way, we are able to "extract" the real pair potential. The long trajectory ensures that the $g(r)$ is smooth enough. As there are only two particles involved (we map the COM's of the molecules to CG particles), the $g(r)$ form has one peak at the potential minimum and then approaches unity towards longer distances, which is typical for the gas phase.

In figure (6.4) we show a preliminary test for the non-dimensional Lennard Jones fluid case. There is not absolute agrrement between the pair potential and the inverted $g(r)$ because the latter comes from a bulk run, so there are multiple peaks up to $r_{cut}$ In figure (6.5)a we can see that for $CH_4$, the constrained run technique and the inverse $g(r)$ are almost the same, within error tolerance (less than 5% throughout the plot), so our proposed constrained run method *is accurate.* In figure (6.5)b the inverse $g(r)$ effective potential $W^{(2),g(r)}$ is similar to the $W^{(2),F}$ but the (smoothed) $W^{(2),U}$ is still not quite accurate. It is not strange as we have addressed this issue again for this less symmetric molecule. In the next section, we will use a more accurate technique. Note that in these figures, the $g(r)$ comes from two atoms moving freely in vacuum, so there's only one peak.

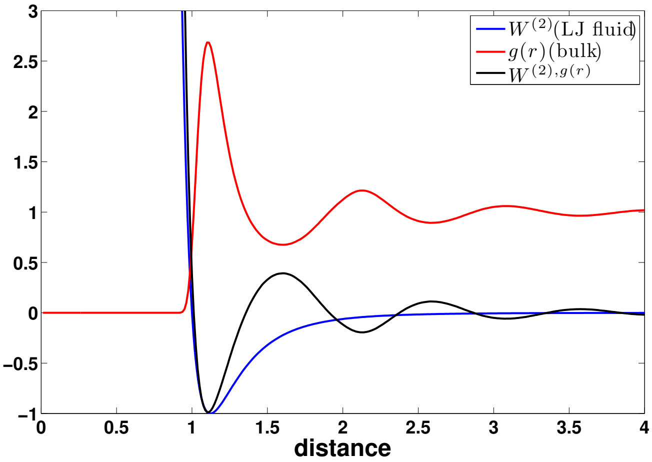

Figure 6.4: Illustration of the inverse $g(r)$ method for the non-dimensional LJ fluid case (bulk, 500 atoms).

### Reversible Work Theorem

We first define the radial distribution function for $n$ out of total $N$ identical particles and then state the theorem in the general case **??**.
Let $P^{(n)}(r_1, r_2, \ldots, r_n)$ be the joint probability distribution for finding particle 1 at $r_1, \ldots,$ particle $n$ at $r_n$.

$$P^{(n)}(r_1, r_2, \ldots, r_n) = \int \cdots \int P(r^N) dr_{n+1} \cdots dr_N \tag{6.10}$$

Assuming that the particles are identical, the probability of finding any particle at $r_1, \ldots$ etc is given by:

$$p^{(n)}(r_1, r_2, \ldots, r_n) = \frac{N!}{(N-n)!} P^{(n)}(r_1, r_2, \ldots, r_n) \tag{6.11}$$

The radial distribution function for $n$ fixed atoms (out of $N$ interacting in total through $U_N(r_1, \ldots, r_N)$) is given by (see McQuarrie):

$$\begin{aligned} g^{(n)}(r_1, \ldots, r_n) = \frac{\rho^{(n)}(r_1, \ldots, r_n)}{\rho^n} = \\ \frac{\frac{N!}{(N-n)!} P^{(n)}(r_1, \ldots, r_n)}{(\frac{N}{V})^n} = \frac{N! V^n}{(N-n)! N^n} \frac{\int \cdots \int e^{-\beta U_N} dr_{n+1} \cdots dr_N}{Z_N} \\ \approx V^n \frac{\int \cdots \int e^{-\beta U_N} dr_{n+1} \cdots dr_N}{Z_N} \end{aligned} \tag{6.12}$$

The reduced distribution functions $P^{(n)}, \rho^n$ are related to a Helmholtz free energy by [145]:

$$g^{(n)}(r_1, \ldots, r_n) \equiv e^{-\beta w^{(n)}(r_1, \ldots, r_n)} \tag{6.13}$$

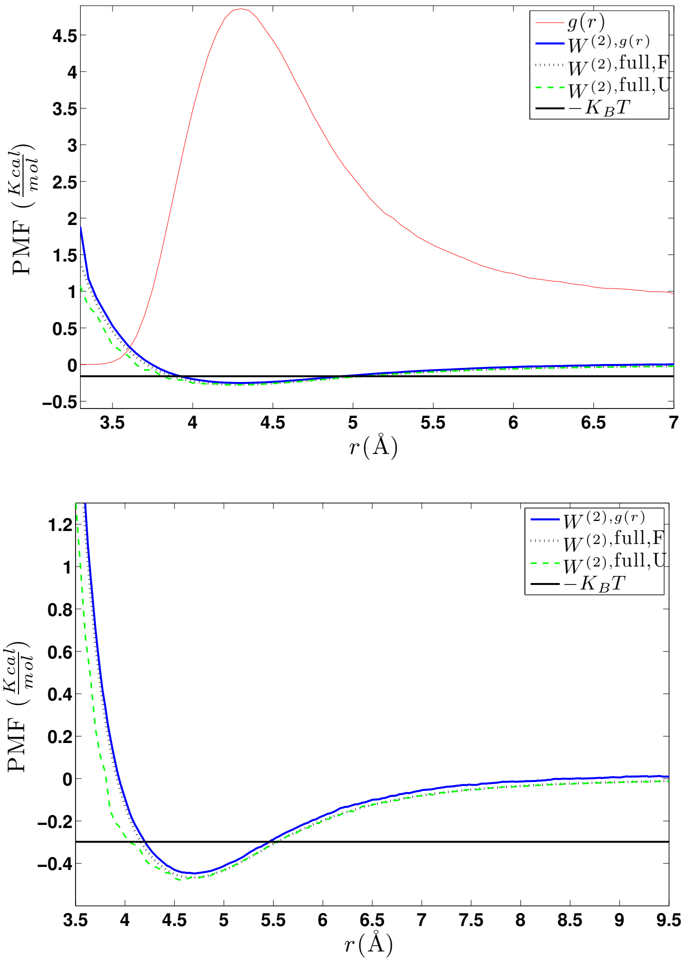

Figure 6.5: PMF $\langle V^{nb} \rangle$ and $W^{(2),g(r)}(r)$ from DBI for (a) $CH_4$ at $T = 80K$ and (b)$CH_3 - CH_3$ at $T = 150$. Good agreement between methods.

where $w(r)$ is usually termed as "*reversible work*". Upon substitution in eq.(6.12) we take the logarithm of both sides and then take the gradient w.r.t. the position of the $j$-th ($j \in \{1, \ldots, n\}$) fixed particle allowing the remaining to move.

$$-\nabla_j w^{(n)} = \frac{\int \cdots \int e^{-\beta U_N}(-\nabla_j U_N) dr_{n+1} \cdots dr_N}{\int \cdots \int e^{-\beta U_N} dr_{n+1} \cdots dr_N} \tag{6.14}$$

The term $-\nabla_j U_N$ is the force acting on particle $j$ so the rhs is the (pair) **mean force** $f_j^{(n)}$ acting on $j$ averaged over all the configuration of the other $N - n$ moving particles. So $f_n^{(n)} = -\nabla_j w^{(n)}$ which means that $w^{(n)}$ is the potential whose gradient gives the mean force acting on $j$. In the case of

$n = 2$ fixed particles at distance $r = |r_1 - r_2|$ apart, eq. (6.13) becomes $g(r) = e^{-w^{(2)}(r)}$. If we ensemble average the force acting on particle 1 we get:

$$-\left\langle \frac{d}{dr_1} U(r^N) \right\rangle_{r_1, r_2 fixed} = \frac{-\int (dU/dr_1) e^{-\beta U} dr_3 \cdots dr_N}{\int e^{-\beta U}} dr_3 \cdots dr_N \quad (6.15)$$

After straightforward calculations (see Chandler p.201) it holds:

$$-\left\langle \frac{d}{dr_1} U(r^N) \right\rangle_{r_1, r_2 fixed} = -\beta^{-1} \frac{d}{dr_1} \log g(r_1, r_2) \quad (6.16)$$

This result shows that $-k_B T \log(g|r_1 - r_2|)$ is a function whose gradient gives the force between particles 1 and 2 averaged over the equilibrium distribution **for all** the other particles. Integration of the averaged force yields the **reversible work** in eq.(6.13). When the density becomes very small, the two molecules $r$ apart, are not affected by the remaining $N - 2$ molecules and so $w^{(2)}(r) \to U(r)$ as $\rho \to 0$.

## 6.3 3-Body

In the following, we are interested in constructing a higher order effective potential between three molecules. The key idea is to assess the tradeoff between accuracy in CG calculations versus a computationally more demanding force evaluation. The extension of this framework on top of pair (2Body) interactions is neither trivial, nor computationally cheap. The reason is that we have increased the dimension of the problem; the pair potential that was a function of distance $r_{12}$ between COM's, now involves three distances $r_{12}, r_{13}, r_{23}$. Things are even more complicated when one tries to evaluate the forces between three particles. This last issue is two-fold:

i) The calculation $f_i = -\nabla_{\mathbf{q_i}} W(r_{12}, r_{13}, r_{23})$ from the data extracted from the calculations of three constrained particles in vacuum, and

ii) The identification of triplets in the CG level run and correct attribution of forces among them.

### 6.3.1 Constrained runs

We extend the notion of 2-Body constrained runs in the case of three particles in a straightforward manner. The setup for the first two CG particles remains the same and we place a third one within the cutoff range of the atomic potential. The new extra distances starting from CG particle ① and ② are $r_{13}$ and $r_{23}$ respectively. The atomistic non-bonded pair potential between atoms of CG particles ① and ② is calculated, then between ① and ③ and finally between ② and ③. The total potential energy based on this atomistic pair potential is a sample for the ensemble average $\langle W^{(3)} \rangle|_{r_{12}, r_{13}, r_{23}}$.

The same constraining methodology of subtracting COM forces (or momenta depending on the integrator) on every time step, for the CG particles to remain pinned in space, applies. When an adequate number of samples is gathered for the triplet $\{r_{12}, r_{13}, r_{23}\}$, we need to move one of the three CG particles by $dr$ and repeat this procedure. When we move, for instance $\mathbf{r_3}$ by $\mathbf{dr_3}$, two of the distances change: $r_{13}$ and $r_{23}$. As one can easily see, the dimensionality of possible discretization, the mesh, has increased by two. In the 2-body case he had a discretization of intermolecular distances $r_{12}$ by $dr$, so $\{r \in [r_{min}, r_{cut}]; dr\}$ In the 3-body case this space becomes $\{r_{12} \in [r_{min}, r_{cut}], r_{13} \in [r_{min}, r_{cut}], r_{23} \in [r_{min}, r_{cut}]; dr\} = n^3$ configurations, or "points". Each one of these $n^3$ points requires an independent long trajectory so we can run a separate simulation, provided that is has equilibrated "sufficiently" before we extract the mean. We mention again that although we insert bias in the dynamics of the run, we sample with respect to the correct equilibrium measure.

### 6.3.2 COM positions

In the previous section, we presented the problem on the huge number of points and the major computational cost associated with it. It is vital to exploit any symmetries of the vectors $\mathbf{r_{12}}, \mathbf{r_{13}}, \mathbf{r_{23}}$. Remember that in the 2-body case, we displaced the two COM's along the $x$-axis for simplicity. Here we keep those two fixed and displace ③ on a semicircle around ①, so we move along the $X - Y$ plane. Of course, the algorithm is general and the code works even if the positions employed the $Z$-dimension.

The potential energy $\langle V^{(3)} \rangle$ is a scalar quantity depending on the relative positions of the atoms or the three distances between them. So there is invariance under internal rotation of the COM indices; i.e. $W^{(3)}(r_{12}, r_{13}, r_{23}) = W^{(3)}(r_{31}, r_{32}, r_{13})$. The same holds for rotation of ③ around (by varying the $Z$ coordinate) the fixed vector $\mathbf{r_{12}}$ formed by ① and ② in the $X - Y$ plane, as seen in figure 6.6a. The next symmetry to be exploited, is along the $Y'Y$ axis in the same manner and is shown in 6.6b, meaning that rotation of ③ around ① is sufficient (no need to repeat around ②).

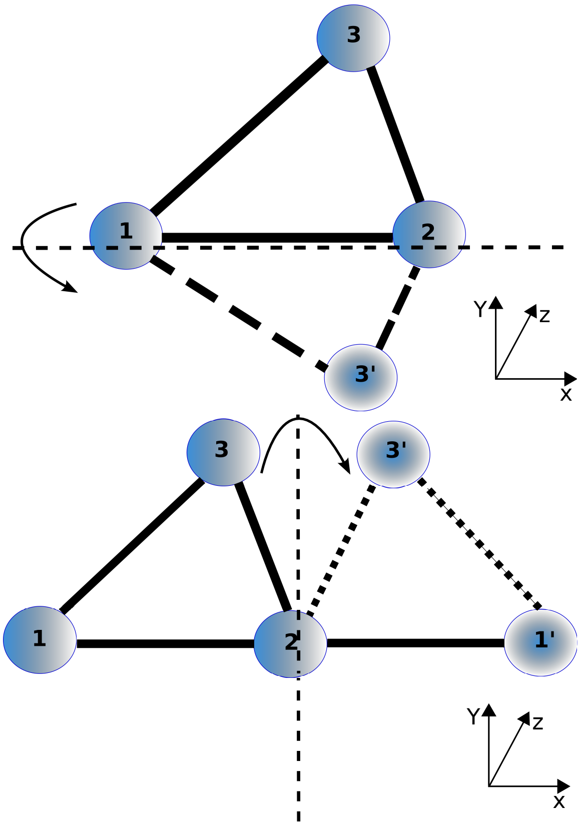

Figure 6.6: Triplet symmetries in space. Symmetry along $X'X$(upper) and $Y'Y$ (lower)

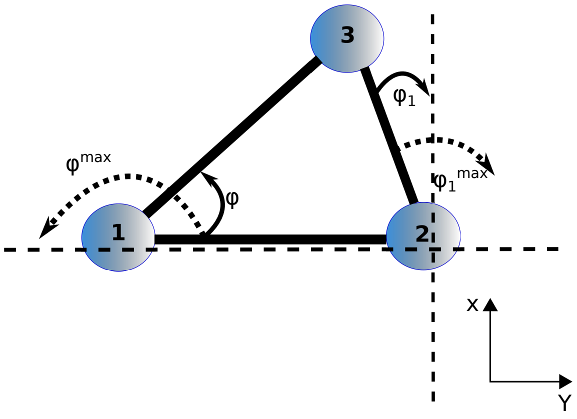

Figure 6.7: Algorithm 1 sampling strategy.

After taking account of all the above notes, we conclude that when distance $r_{12}$ is fixed (by pinning $\textcircled{1}$ and $\textcircled{2}$) and $r_{13}$ is fixed by rotating $\textcircled{3}$ counter-clockwise around $\textcircled{1}$ on a circle bow of radius $r_{13}$, the variable of distance $r_{23}$ takes all the possible values in the range $[r_{min}, r_{cut}]$. This bow starts when $r_{23}$ is at $r_{\text{min}}$ and ends when the $Y$-coefficient of $\textcircled{3}$ reaches 0, because of the symmetry across $X'X$. $\{r_{12}, r_{13}, :\}$ is sampled so we proceed to sample $\{r_{12}, r_{13} + dr, :\}$. Now $\textcircled{3}$ is moved further away from $\textcircled{1}$ and $\textcircled{2}$ so $r_{12}, r_{13}$ change.

A clever way for clarity and bookeeping of the data i.e. changing one of the variables $r_{13}, r_{13}, r_{23}$ at a time, is to rotate $\textcircled{3}$ around $\textcircled{2}$ by $d\phi_1$ clockwise. Then we recalculate the distances $r_{12}, r_{13}, r_{23}$ and proceed as before; rotate $\textcircled{3'}$ around $\textcircled{1}$. Schematically the triangle $\widehat{123}$ changes shape, but $r_{12}$ and $r_{13}$ remain the same, see figure (6.7).

On the final step we need to alter $r_{12}$ in order to sample $\{r_{12} + dr, r_{13}, :\}$, which is straightforward and then go to the first step and reiterate this procedure.

The whole above procedure is better seen in figure (6.7), while its implementation is described in algorithm 1.

Note, again that we can think of every triplet $\{r_{12}, r_{13}, r_{23}\}$ as a "separate simulation" because we bias the dynamics by pinning the COM's instead of one really long simulation. The correct equilibrium measure is maintained, provided that we throw away an initial burn-in period on every run. The space discretization for the two system COM's is found in table (6.2).

| | $dr_{12}$ | $d\phi$ | $d\phi_1$ |
|---|---|---|---|
| $CH_4$ | $0.05\mathring{A}$ | $\frac{\pi}{90}$ | $\frac{\pi}{90}$ |
| $CH_3 - CH_3$ | $0.05\mathring{A}$ | [3.8 : 5.9] | [3.8 : 6] |

Table 6.2: COM space discretization $d\theta$ in the models for algorithm 1.

**Algorithm 1** define the COM's in cartesian coordinates

```
Precondition: set COM's at $\mathbf{r_1}, \mathbf{r_2}, \mathbf{r_3}$, $d\phi$, $d\phi_1$
FIND $|r_{12}|, |r_{13}|, |r_{23}|$

for i in range=$[r_{12}^{min} : r_{12}^{max}]$ do
    FIND $\mathbf{r_2}$, $|r_{12}|$, $|r_{13}|$, $|r_{23}|$
    for $k$ in range [1 : max iterations] do
        FIND polar coordinates for $\mathbf{r_1}, \mathbf{r_2}, \mathbf{r_3}$
        ROTATE $\mathbf{r_{13}}$ by $-d\phi$ around $\mathbf{r_1}$
        if $(\mathbf{r_3})_y < (\mathbf{r_1})_y$ then          ▷ $\mathbf{r_{13}}$ parallel to $\mathbf{r_{12}}$: no new
                                                 ▷ info if rotation continues
            place $\mathbf{r_3}$ back to the original position
            find polar coordinates for $\mathbf{r_1}, \mathbf{r_2}, \mathbf{r_3}$
            ROTATE $\mathbf{r_{23}}$ by $+d\phi_1$ around $\mathbf{r_2}$
            Calculate new $|r_{13}|$
        end if
        STORE $\mathbf{r_1}, \mathbf{r_2}, \mathbf{r_3}$
    end for
end for
```

### 6.3.3 Statistical accuracy

In section 6.2.4 we mentioned the problems that arise when we try to ensemble average quantities at low probability/high energy configurations of phase space, for two CG particles in vacuum. Sampling problems can be even stronger in the 3-Body case, as the energy term $e^{-\beta(V_{12}+V_{13}+V_{23})}$, $V_{ij}$ being the atomistic level potential energy between molecules ⓘ and ⓙ, is quite small or highly improbable. In an attempt to properly visualize the 4-dimensional data $r_{12}, r_{13}, r_{23}, W^{(3)}(r_{12}, r_{13}, r_{23})$, we keep $r_{12}, r_{13}$ fixed and plot $W^{(3)}$ against $r_{23}$.

In figure (6.8) we show the effective 3-Body potential $W^{(3),MD}$ for $CH_4$ at $T = 100K$ for the set of distances $r_{12} = 3.9\mathring{A}, r_{13} = 4.0\mathring{A}$ and $r_{23} \in [3.5 : 8]\mathring{A}$ and $r_{12} = 4.1\mathring{A}, r_{13} = 4.4\mathring{A}$ and $r_{23} \in [3.5 : 8]\mathring{A}$ in conjunction with the 2-Body $W^{(2)}(r_{12}) + W^{(2)}(r_{13}) + W^{(2)}(r_{23})$ for comparison. This latter sum is essentially what all pairwise CG representations use in the last decades. In the first set, we discern a gain in information with $W^{(3),\mathrm{MD}}$, although the noise is high. As the three CG particles move away from each other (6.9), $W^{(3),\mathrm{MD}}$ becomes smooth. Even for very long trajectories of the order of $(8 \cdot 10^7)$ steps $(= 40ns)$, the fluctuations remain. We also threw out burn-in periods of length twice as much as the production run, without considerable success.

The above simulations refer to the $CH_4$ molecule. Sampling the phase space for the (spherically) asymmetric $CH_3 - CH_3$ was proved to be even

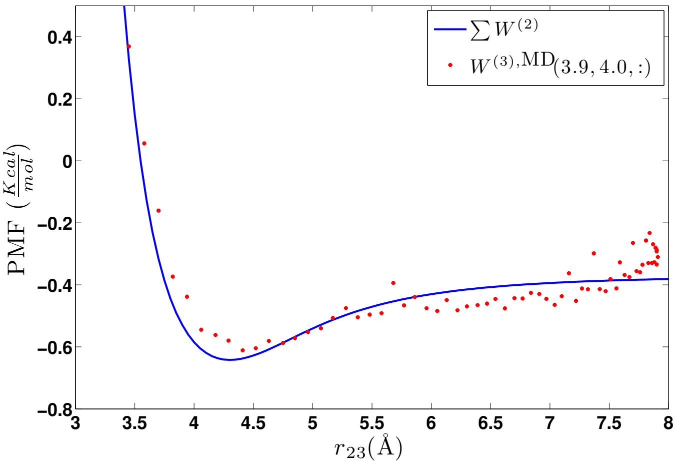

Figure 6.8: $W^{(3),MD}(3.9, 4.0, :)$ and $W^{(2)}(3.9) + W^{(2)}(4.0) + W^{(2)}(3.4 : r_{cut})$ for $CH_4$ at $T = 100K$. We see gain in information as expected, but the noise is high in the MD simulations.

harder, as we see next. We checked with the *variance of the estimator* $\langle\cdot\rangle$ and it does not decrease below a threshold dependent on the temperature and system. More specifically, we experimented with different values of the coupling coefficient of the heat-bath ($\xi$ in BBK), timestep $dt$, burn-in period (up to 60% of the trajectory) and intermolecular distances, without success.

In figure (6.10) we show the effective 3-Body potential for $CH_3 - CH_3$ at $T = 150K$ for $r_{12} = 4.2\mathring{A}, r_{13} = 4.5\mathring{A}$ and $r_{23} \in [3.5 : 8.5]\mathring{A}$, in conjunction with the 2-Body $W^{(2)}(r_{12}) + W^{(2)}(r_{13}) + W^{(2)}(r_{23})$ for comparison. There are clear fluctuations throughout the $r_{23}$ range.

In figure (6.11) we see that even for a set of relatively long distances, the estimator of the 3-Body potential $\langle W^{(3)}(4.2, 4.5, 5.54)\rangle$ has *large variance*. The problem persists for bigger/smaller timstep $dt$, larger heat-bath coupling coefficient $\xi$, and total simulation time up to $1000ns$!

Taking a closer look at the numerics of the ensemble averaging, we concluded that the reason that $\langle\cdot\rangle$ won't improve as we increase the simulation time (even if small or larger timesteps $dt$ where used) is due to **rare events**. After a couple of million steps, there would be a small number of (valid) configurations where the molecules allign with each other, giving rise to the energy of the system of circa two orders of magnitude more. In effect, this has high impact on the averages and we end up with fluctuations in the $W^{(3)}$ figure.

We have further examined different smoothing techniques, such as the moving average with various values for the moving window. However, all of them proved to be useful in the range of $W^{(3)}() - \sum W^{(2)}()$, so any

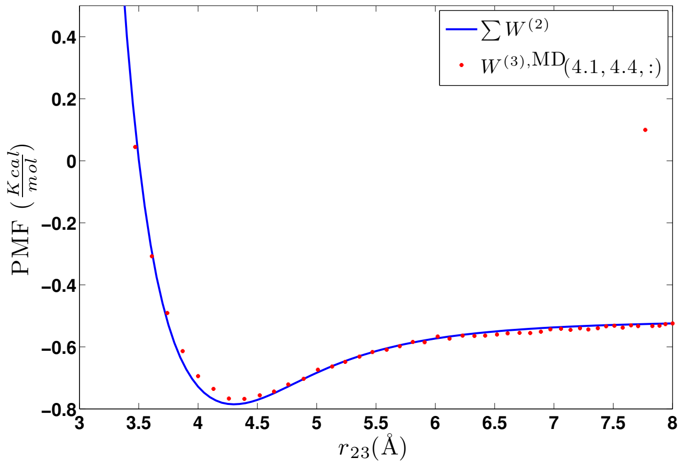

Figure 6.9: $W^{(3),MD}(4.1, 4.4, :)$ and $W^{(2)}(4.1) + W^{(2)}(4.4) + W^{(2)}(3.4 : r_{cut})$ for $CH_4$ at $T = 100K$. There is less noise as well as information gain as CG particles move away.

information from our detailed atomistic constrained runs could be washed out in this manner.

One would suggest that an alternative way of calculating the ensemble average, which in practice is:

$$\langle U(\mathbf{q})\rangle|_{\mu} = \int U(\mathbf{q})\mu(d\mathbf{q}) \tag{6.17}$$

and so far we approximate by:

$$\langle U(\mathbf{q})\rangle|_{\mu} \approx \lim_{N\to\infty} \frac{1}{N}\sum_{t=1}^{N} U(\mathbf{q^n}) \tag{6.18}$$

by Monte Carlo (MC) simulation. Note, that for ergodic trajectories, MD and MC results should are equivalent with respect to the calculation of statistical averages (apart from dynamic related results like the MSD). That calculation would work provided that we model the bond potentials correctly and define appropriate MC moves. In principle, the result would be slower averaging convergence because the problem of sampling the low probability areas remains. The proposal distribution based on the configurational energy from the intermolecular potential, by which the MC moves are accepted, is still the same: $e^{-\beta U^{nb}(r_{12})}$. It is very small and low probability configurations are rarely accepted (remember that $u \sim U[0, 1]$ and $u < e^{-\beta U^{nb}}$).

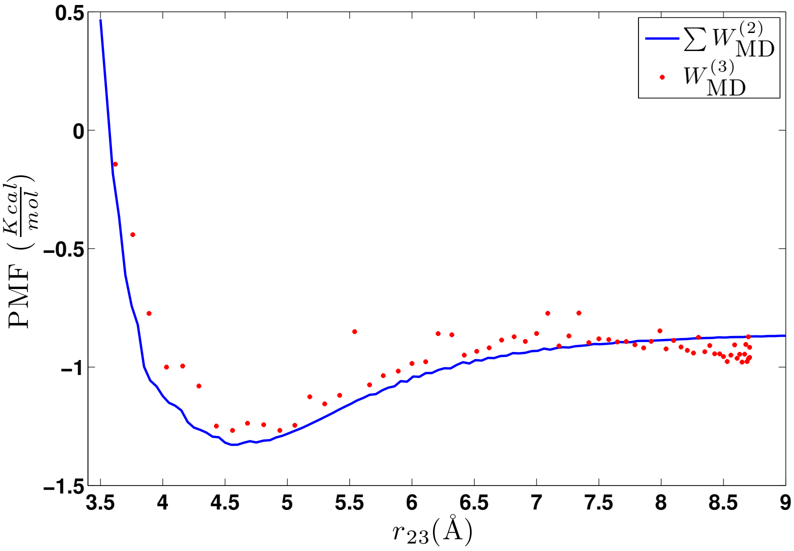

Figure 6.10: $W^{(3)}(4.2, 4.5, :)$ and $W^{(2)}(4.2) + W^{(2)}(4.5) + W^{(2)}(3.4 : r_{cut})$ for $CH_3 - CH_3$ at $T = 150K$. We see some gain in information around the well as expected, but the noise is still high in the MD simulations.

### 6.3.4 Geometric averaging

All of the sampling issues discussed above were resolved with the geometric averaging technique **??**. At this point we do not perform molecular dynamics for the 2-body and 3-body systems any more. On the contrary, we displace (rotate more precisely) the molecules around their COM, taking account all possible orientations based on their appropriate probability weight.

In more detail, for the 2-body system, we pin the COM of each CG particle in space and place the atoms of that particle by defining their Cartesian coordinates. Then, instead of integrating the equations of motion, we rotate $\textcircled{2}$ while keeping $\textcircled{1}$ still. The rotation is done by using the Euler angle formulation; the axes of the original cartesian frame are rotated by three angles: $\alpha, \beta, \gamma$ [34]. Each one is formed by rotation of $X'X$ towards $Y'Y$, $Y'Y$ towards $Z'Z$ and $Z'Z$ towards $X'X$ respectively. There are six possible rotation sequences for full coverage of the sphere surface and we used the $ZYZ$ one.

The angle discretization was $d\theta = \pi/20$ for $CH_4$ ($d\theta = \pi/90, d\phi = \pi/45$ for $CH_3 - CH_3$) and was determined by separate sequential runs until convergence on the ensemble average was reached. We note that in this method, we do not take into account bond vibrations, i.e. the molecule is rigid. We stress that we found no differentiation in the results. Furthermore, tests on constrained MD and SD runs agree with this finding as well, because the intermolecular forces cancel out on the average and the bond and angular

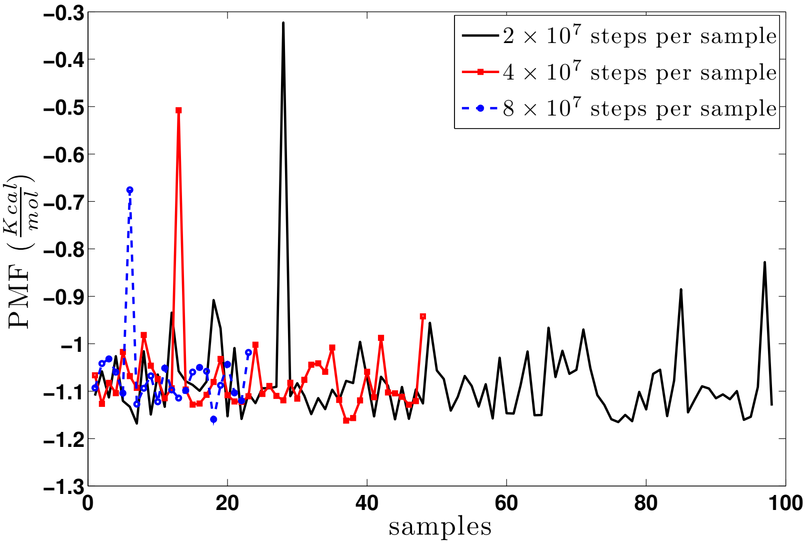

Figure 6.11: MD noise in $\langle W^{(3)}(4.2, 4.5, 5.54)\rangle$ for varying the number of timesteps per sample of the same simulation of length $1000ns$. $CH_3 - CH_3$ at $T = 150K$

| | $\alpha$ | $\beta$ | $\gamma$ | $d\theta$ |
|---|---|---|---|---|
| $CH_4$ | $[0 : \pi]$ | $[0 : \frac{\pi}{2}]$ | $[0 : \frac{\pi}{2}]$ | $\frac{\pi}{20}$ |
| | $\theta$ | $\phi$ | $d\theta$ | $d\phi$ |
| $CH_3 - CH_3$ | $[0 : \pi]$ | $[0 : 2\pi]$ | $\frac{\pi}{45}$ | $\frac{\pi}{20}$ |

Table 6.3: Euller angle range and discretization $d\theta$ in the models.

potential energies are not taken into account in the ensemble averages.

The same procedure was extended for the 3-body case. The computational cost increases by an order of magnitude as there are in total $n^3$ ($= n \times n \times n$, $n$ is the number of orientations per molecule) orientations. In the $CH_3 - CH_3$ case, we need to take into account more orientations (despite the fact that the atom-atom computations are fewer) because of the asymmetry of the molecule. By exploiting the symmetries of all possible orientations for this system, we were able to perform the computations relatively fast; i.e. full sampling can be performed in a couple of days for $CH_4$ and five days for $CH_3 - CH_3$, on a system of 20 cores (intel Xeon @ 2.6GHz).

In algorithm 2 we sketch the geometric averaging method for the case of 3 molecules. This computation includes a triple loop over orientations, so $d\theta$ is the discretization variable that defines the order of the computational cost. The $CH_4$ system takes about 1 hour (serial runs on 8 cores) for $d\theta = \frac{\pi}{20}$

**Algorithm 2** Geometric Averaging for 3 CG particles

```
Precondition: Use algorithm 1 to define the COM positions: $COM_1(), COM_2(), COM_3()$

# define orientations once i.e. rotations about each COM
for $\alpha$ in $[0, 2\pi]$, $\alpha = \alpha + d\theta$ do    ▷ rotation of coordinate frame along $\alpha$ angle
    for $\beta$ in $[0, \pi]$, $\beta = \beta + d\theta$ do
        for $\gamma$ in $[0, 2\pi]$, $\gamma = \gamma + d\theta$ do
            # $ZYZ$ orientation of a CG particle at the origin $(0, 0, 0)$ according to $\alpha, \beta, \gamma$
            $orient(1 : n\_atoms, 1 : 3, idx\_orientations) = rot\_matrix(\alpha, \beta, \gamma)$
            $idx\_orientations = idx\_orientations + 1$
        end for
    end for
end for

# calculate atomistic positions at $COM_1(), COM_2(), COM_3()$
for $i$ in the set of COM's do
    # calculate atomistic positions for $COM_1$
    $q_1(1 : n\_atoms, 1 : 3, i) = COM_1(i) + orient(1 : n\_atoms)$
    $q_2(1 : n\_atoms, 1 : 3, i) = COM_2(i) + orient(1 : n\_atoms)$
    $q_3(1 : n\_atoms, 1 : 3, i) = COM_3(i) + orient(1 : n\_atoms)$
end for
# main loops for sampling the potential on every $\{r_{12}, r_{13}, r_{23}\}$
$\beta_{kb} = \frac{1}{K_b T}$
for $i$ in $[1 : idx\_orientations]$ do
    for $j$ in $[1 : idx\_orientations]$ do
        for $k$ in $[1 : idx\_orientations]$ do
            Calculate $U_{ij}^{\text{atom}}, U_{ik}^{\text{atom}}, U_{jk}^{\text{atom}}$    ▷ atom-atom
            $U_{cur} = U_{ij}^{\text{atom}} + U_{ik}^{\text{atom}} + U_{jk}^{\text{atom}}$
            $U_{total} = U_{total} + e^{-\beta_{kb} U_{cur}}$    ▷ Weight included!
        end for
    end for
end for
$U_{total} = \frac{U_{total}}{idx\_orientations^3}$    ▷ Normalize
$U_{CG} = \frac{\log(U_{total})}{-\beta_{kb}}$
```

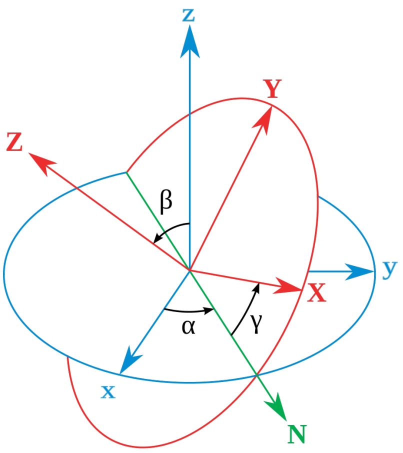

Figure 6.12: Euller angles and corresponding rotation of the coordinate frame (Wikipedia)

and 11 hours (intel Xeon @ 2.6GHz) for a discretization of $d\theta = \frac{\pi}{30}$, as the number of orientations (per CG particle) has increased from 2542 to 7935. On the other hand, the accuracy was negligible meaning that $d\theta = \frac{\pi}{20}$ is sufficient for this model.

For $CH_3 - CH_3$ we used plain spherical coordinate sampling due to the geometry of the model; rod-like. Of course the Euler angle discretization can be applied with an extra computational cost. In order to speed up the computations, we parallelized all above computations. In more detail, we employed OMP paradigm, since all different configurations are independent. The dimensions of the problem scaled up to 20 CPU's (dependent on angle discretization). As a reference, we used 8 OMP threads per run and the number of orientations where 1886, so it took about 3 days (on 7*8=56 cores). The angle ranges and discretizations are summarized in table (6.3).

At this point we mention that the $CH_4$ system is faster, as we exploited (experimentally) one of its properties. We converted the triple "for loop" into a double while we oriented the CG particles in an inverse order. More specifically, in algorithm 2, line 21, we replaced the loop with $j = idx\_orientations - i + 1$ and the ensemble average remained the same, while the computational cost dropped by days!

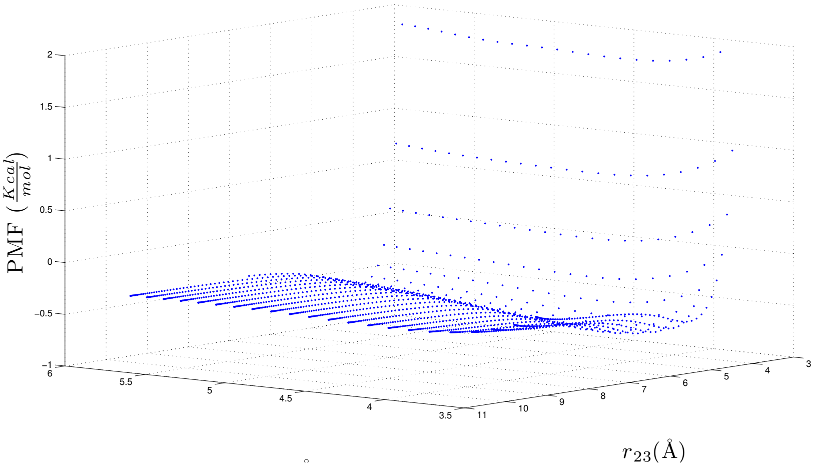

Figure 6.13: 3-dimensional representation of $W^{(3),\text{geom}}$ data for $\mathbf{r_{12} = 4.3}$ fixed, and $r_{13} \in [3.8 : 5.8]$, $r_{13} \in [3.2 : r_{cut}]$ for $CH_4$ at $T = 80K$. The double-well is clear.

### 6.3.5 4-Body

For completeness purposes, we include a rough description of how complex the 4-Body effective potential might be. We need 4 additional variables to uniquely determine the positions of the COM's as shown in figure (6.14). Note that we employ torsional angles as only ① and ② reside on the $Z = 0$ plane. All of these mean that the level of complexity in constructing, as well as using, the potential has risen dramatically. In the case of $CH_4$ where the triple for loop over neighbours is computationally less demanding, the $W^{(3),\text{geom}}$ might be feasible in days.

Our assumption is that in the CG level simulation, a 4-Body potential is computationally unfeasible because of the 7 parameters. First and foremost the bottleneck of any MD code is the force evaluation and a quad "for loop" over neighbours is more expensive by one order. Inside the loops, there will be external trigonometric function calls for angle estimations which have a detrimental impact in performance. Last and most importantly, the fitting of the data is hard and prone to numerical errors, so the overall benefit might not even be quantifiable; the examination of such terms could be a part of a future work.

## 6.4 $W^{(3)}$ representation

After the collection of $W^{(3)}$ data has finished we are able to use them in the CG level simulation. In order for $W^{(3)}(r_{12}, r_{13}, r_{23})$ to be in a usable form, we need either a functional form of the potential and of the forces, as we

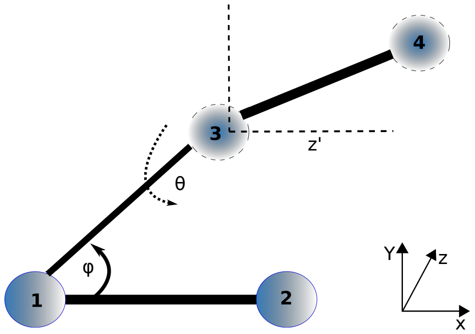

Figure 6.14: 4-Body COM constraining. $W^{(4)}(r_{12}, r_{13}, \phi, \theta, r_{34}, r_{14}, r_{24})$. reside on the $Z = 0$ plane.

normally have in the atomistic simulations, or a tabulated (up to a degree of discretization) form for the potential and forces. Both methodologies have advantages and disadvantages, so we focus on each one separately.

### 6.4.1 Cubic polynomial

The simplest, handy and usual methodology is fitment of the data to a usable formula. By usable we mean a tradeoff between low complexity and accuracy for less and correct calculations. In principle, as we employ higher dimensional functions containing more terms, the mean squared error of the fitting is reduced. Then, after $W^{(3)}_{cubic}$ is determined, we have to take the spatial gradient with respect to each cartesian position: $-\nabla_{\mathbf{q_i}} W^{(3)}_{cubic}$ for the calculation of the forces. This is done analytically, once, for the specific functional form.

We note at this point, as it was our first failed attempt, that the functional should be at least a three dimensional cubic polynomial. Three dimensional as of the parameters $(r_{12}, r_{13}, r_{23})$ and cubic, because we require its gradient, which is quadratic, to be able to capture the curvature of the force well.

The form of the cubic polynomial, containing constants $P_{\_\_\_}$ to be determined, is:

$$
\begin{aligned}
f(\mathbf{x},\mathbf{y},\mathbf{z}) = P_{000} &+ P_{100}\mathbf{x} + P_{010}\mathbf{y} + P_{001}\mathbf{z} \\
&+ P_{200}\mathbf{x^2} + P_{020}\mathbf{y^2} + P_{002}\mathbf{z^2} \\
&+ P_{110}\mathbf{xy} + P_{101}\mathbf{xz} + P_{011}\mathbf{yz} \\
&+ P_{300}\mathbf{x^3} + P_{030}\mathbf{y^3} + P_{003}\mathbf{z^3} \\
&+ P_{111}\mathbf{xyz} + P_{210}\mathbf{x^2y} + P_{201}\mathbf{x^2z} \\
&+ P_{021}\mathbf{y^2z} + P_{012}\mathbf{yz^2} + P_{120}\mathbf{xy^2} + P_{102}\mathbf{xz^2}
\end{aligned}
\tag{6.19}
$$

where

$$
\begin{aligned}
\mathbf{x} &= |\mathbf{r_{ij}}| = |\mathbf{q_i} - \mathbf{q_j}| \\
\mathbf{y} &= |\mathbf{r_{ik}}| = |\mathbf{q_i} - \mathbf{q_k}| \\
\mathbf{z} &= |\mathbf{r_{jk}}| = |\mathbf{q_j} - \mathbf{q_k}| \\
\mathbf{x}^2 &= (q_i^{(1)} - q_j^{(1)})^2 + (q_i^{(2)} - q_j^{(2)})^2 + (q_i^{(3)} - q_j^{(3)})^2
\end{aligned}
\tag{6.20}
$$

The gradient of the $X$-dimension with respect to $\mathbf{q_i}$ is:

$$
\begin{aligned}
\frac{\partial f}{\partial q_i^{(1)}} = 0 &+ P_{100}\frac{q_i^{(1)} - q_j^{(1)}}{|\mathbf{x}|} + P_{010}\frac{q_j^{(1)} - q_k^{(1)}}{|\mathbf{y}|} + P_{001}0 \\
&+ P_{200}2(q_i^{(1)} - q_j^{(1)}) + P_{020}2(q_i^{(1)} - q_k^{(1)}) + 0 \\
&+ P_{110}\Big((q_i^{(1)} - q_j^{(1)})\frac{|\mathbf{y}|}{|\mathbf{x}|} + (q_i^{(1)} - q_k^{(1)})\frac{|\mathbf{x}|}{|\mathbf{y}|}\Big) \\
&+ P_{101}\Big((q_i^{(1)} - q_j^{(1)})\frac{|\mathbf{z}|}{|\mathbf{x}|}\Big) + P_{011}\Big((q_i^{(1)} - q_k^{(1)})\frac{|\mathbf{z}|}{|\mathbf{y}|}\Big) \\
&+ P_{300}\big(3|\mathbf{x}|(q_i^{(1)} - q_j^{(1)})\big) + P_{030}3|\mathbf{y}|(q_i^{(1)} - q_k^{(1)}) \\
&+ P_{111}\Big((q_i^{(1)} - q_j^{(1)})\frac{|\mathbf{y}|}{|\mathbf{x}|} + (q_i^{(1)} - q_k^{(1)})\frac{|\mathbf{x}|}{|\mathbf{y}|}\Big)|\mathbf{z}| \\
&+ P_{210}\Big(2(q_i^{(1)} - q_j^{(1)})|\mathbf{y}| + (q_i^{(1)} - q_k^{(1)})\frac{|\mathbf{x}|^2}{|\mathbf{y}|}\Big) \\
&+ P_{021}\Big(2(q_i^{(1)} - q_k^{(1)})|\mathbf{z}|\Big) + P_{012}\Big((q_i^{(1)} - q_k^{(1)})\frac{|\mathbf{z}|^2}{|\mathbf{y}|}\Big) \\
&+ P_{120}\Big(q_i^{(1)} - q_j^{(1)})\frac{|\mathbf{y}|^2}{|\mathbf{x}|} + 2(q_i^{(1)} - q_k^{(1)})|\mathbf{x}|\Big) \\
&+ P_{201}\big(2(q_i^{(1)} - q_j^{(1)})|\mathbf{z}|\big) + P_{102}\Big((q_i^{(1)} - q_j^{(1)})\frac{|\mathbf{z}|^2}{|\mathbf{x}|}\Big)
\end{aligned}
\tag{6.21}
$$

and after tedious, error-prone chain rule differentiation

$$
\frac{\partial f}{\partial q_j^{(1)}}, \quad \frac{\partial f}{\partial q_k^{(1)}}
\tag{6.22}
$$

are calculated in the same manner. Fortunately, there is symmetry with respect to the $^{(2)}$ and $^{(3)}$ coordinates. As one can see, the cost and complexity increases dramatically if we move to a polynomial of order four.

At this point we need to fit the data from the constraint (or geometric) runs. The main idea of the fitting is to solve the minimization problem:

$$\min_{\mathbf{x},\mathbf{y},\mathbf{z}} \Big| f(\mathbf{x},\mathbf{y},\mathbf{z}) - data \Big| = \min_{\mathbf{x},\mathbf{y},\mathbf{z}} G(\mathbf{x},\mathbf{y},\mathbf{z}) \tag{6.23}$$

with the *Conjugate Gradient* method, where:

$$G(\mathbf{x},\mathbf{y},\mathbf{z}) = \frac{1}{2} X^T A X - X^T b \tag{6.24}$$

where matrix $A \in \mathbb{R}^{n\times 20}$:

$$A = \begin{pmatrix} & \cdots & & & & & & & & \\ & & & & & & & & & \\ 1 & \mathbf{x^{(m)}} & \mathbf{y}^{(m)} & \mathbf{z^{(m)}} & (\mathbf{x^{(m)}})^2 & (\mathbf{y^{(m)}})^2 & (\mathbf{z^{(m)}})^2 & (\mathbf{x^{(m)}})^2\mathbf{y^{(m)}} \cdots & \mathbf{x^{(m)}}(\mathbf{z^{(m)}})^2 \\ & \vdots & & & & & & & \\ & & & & & & & & \end{pmatrix}$$

$\mathbf{x} = |\mathbf{r_{ij}}|$ and the $m$-th row contains the terms for the coefficients in eq. (6.19) corresponding to $data^{(m)}$.

$$X = \begin{pmatrix} P_{000} \\ P_{100} \\ P_{010} \\ P_{001} \\ \vdots \\ P_{102} \end{pmatrix} \in \mathbb{R}^{20\times 1}$$

$$b = \begin{pmatrix} data^{(1)} \\ data^{(2)} \\ \vdots \\ data^{(n)} \end{pmatrix} \in \mathbb{R}^{n\times 1}$$

and $data^{(m)}$ is the $W^{(3)}(\cdot,\cdot,\cdot)$ of the $m$-th triplet $\{r_{12}^{(m)}, r_{13}^{(m)}, r_{23}^{(m)}\}$. We convert $A$ to a symmetric matrix, i.e. its *normal form*, by:

$$\begin{aligned} \tilde{A} &= A^T A \\ \tilde{b} &= A^T b \end{aligned} \tag{6.25}$$

The conjugate gradient algorithm takes up to a hundred steps to converge, depending on the data set.

We validated the solution with Cholesky decomposition of $AX = b$ as well. In order to have an estimate of the wellness of fit, apart from inspection, we use the root mean squared error (RMSE):

$$RMSE = \frac{1}{n}\sqrt{\sum_{i}^{n}(f^{(i)} - data^{(i)})^2} \tag{6.26}$$

here $f^{(i)} = W^{(3)}_{cubic}(r^{(i)}_{12}, r^{(i)}_{13}, r^{(i)}_{23})$ is the determined cubic polynomial value and $data^{(i)} = W^{(3)}(r^{(i)}_{12}, r^{(i)}_{13}, r^{(i)}_{23})$ at the same COM distances. For better inspection, we define the normalized "local error" $E_{loc}$, as the curvature is steeper towards $r_{min}$:

$$E^{(i)}_{loc} = \frac{|f^{(i)} - b^{(i)}|}{|b^{(i)}|} \tag{6.27}$$

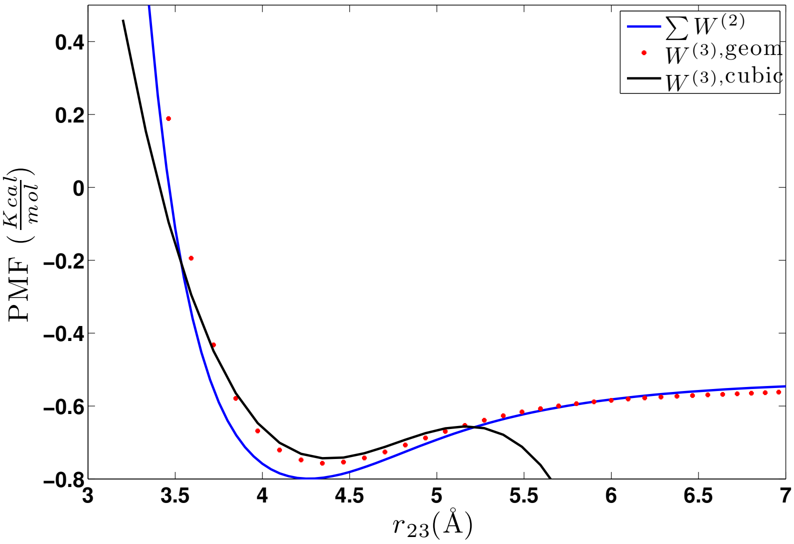

Figure 6.15: Comparison between the fitted cubic polyomial PMF $W^{(3),\text{cubic}}(4.0, 3.9, :)$, geometric averaging data $W^{(3),\text{geom.}}(4.0, 3.9, :)$ and $W^{(2)}(4.0) + W^{(2)}(3.9) + W^{(2)}(3.4 : r_{cut})$ for $CH_4$ at $T = 80K$, $\mathbf{r_{23}} \in [\mathbf{3.8} : \mathbf{5.0}]$. The fit captures the potential well, so we are in agreement with the data and can proceed to the 3-Body CG run.

In figure (6.15) we see the resulting fitted cubic polynomial to the geometric averaging $W^{(3),\text{geom.}}(4.0, 3.9, :)$ data for $CH_4$. The polynomial successfully captures most of the area of interest around the potential well. The range of $r_{23}$ for this system was determined experimentally by eq.'s (6.26) and (6.27).

|  | $r_{12}$ | $r_{13}$ | $r_{23}$ | $r_{cut}$ |
|---|---|---|---|---|
| $CH_4$ | [3.8 : 4.1] | [3.8 : 4.1] | [3.8 : 5] | $12\mathring{A}$ |
| $CH_3 - CH_3$ | [3.8 : 4.4] | [3.8 : 5.9] | [3.8 : 6] | $14\mathring{A}$ |

Table 6.4: Fitting range for distances $r_{12}, r_{13}$ and $r_{23}$ in the models.

### 6.4.2 Numerical calculation of partial derivatives

Next, we examine the usage of the $W^{(3)}$ data in the CG simulations, using numerical calculation of partial derivatives. We term this partial derivatives because we used central differences in order to evaluate the forces:

$$-\frac{\partial W^{(3)}(r_{12}, r_{13}, r_{23})}{\partial \mathbf{q_1}}, -\frac{\partial W^{(3)}(r_{12}, r_{13}, r_{23})}{\partial \mathbf{q_2}}, -\frac{\partial W^{(3)}(r_{12}, r_{13}, r_{23})}{\partial \mathbf{q_3}} \tag{6.28}$$

$$r_{12} = |\mathbf{q_1} - \mathbf{q_2}| \tag{6.29}$$

on the triplets, meaning:

$$\begin{aligned} \frac{\partial W^{(3)}(r_{12}, r_{13}, r_{23})}{\partial \mathbf{q_1}} = & \frac{\partial W^{(3)}(r_{12}, r_{13}, r_{23})}{\partial r_{12}} \frac{\partial r_{12}}{\partial \mathbf{q_1}} \\ & + \frac{\partial W^{(3)}(r_{12}, r_{13}, r_{23})}{\partial r_{13}} \frac{\partial r_{13}}{\partial \mathbf{q_1}} \\ & + \frac{\partial W^{(3)}(r_{12}, r_{13}, r_{23})}{\partial r_{12}} \frac{\partial r_{23}}{\partial \mathbf{q_1}} \end{aligned} \tag{6.30}$$

$$\frac{\partial r_{12}}{\partial \mathbf{q_1}} = \frac{1}{r_{12}} \mathbf{r_{12}} \tag{6.31}$$

Note that $\frac{\partial W^{(3)}}{\partial \mathbf{q_1}}$ contains information from $\textcircled{2}$ and $\textcircled{3}$. We use the notation $1, 2, 3$ instead of $i, j, k$ because we require $r_{12} < r_{13} < r_{23}$.

The central differences scheme reads:

$$\frac{\partial W^{(3)}(r_{12}, r_{13}, r_{23})}{\partial r_{12}} = \frac{W^{(3)}(r_{12} + dr_{12}, r_{13}, r_{23}) - W^{(3)}(r_{12} - dr_{12}, r_{13}, r_{23})}{2dr_{12}} \tag{6.32}$$

So we end up with three tables containing the partial derivatives of $W^{(3)}$, on the discretization of the triplet positions.

In the CG level run, we look up the 3-Body potential $W^{(3),\text{p.d.}}$, as well as the force magnitude $f^{(3)}(r_{12}, r_{13}, r_{23})$. We try to be more precise in the simulation by using the average value between staggered points, for instance $W^{(3),\text{p.d.}}(4.02, 5, 5) = (W^{(3),\text{p.d.}}(4.0, 5, 5) + W^{(3),\text{p.d.}}(4.05, 5, 5))/2$ and the same holds for the forces.

The differentiation was not a straightforward task because $r_{12}, r_{13}, r_{23}$ have different ranges. On top of that, $r_{23}$ values slightly vary when we place COM $\textcircled{3}$ on a circle bow ($d\phi$=fixed) for every $r_{13}$ value. We imposed different tolerance values for central differences (6.32) across the range of $r_{12}, r_{13}, r_{23}$. We mention again that the reason for not allowing $r_{ij}$ in the range $[r_min, r_{cut}]$ is (with the aid of symmetries): a) to limit the computational cost of sampling $W^{(3)}$ b) keep the tables small for quick access in the CG simulations.

The different range of the three distances induces a little complexity to the CG level code. We need to check if the triplet of particles is within $r_{cut}$ range and then sort the distances (in ascending order) before we look them up in the partial derivatives tables, which slows down the performance in comparison with the polynomial fit.

### 6.4.3 Comparison

At this point we are required to assess both methods. We cannot directly compare $W^{(3)}$, as it is in tabulated form in the partial derivatives method (geometric averaging data). So we directly compare the forces between the two methods.

We did opt for the partial derivatives when we suspected accuracy issues at close distances for the cubic polynomial in the $CH_3-CH_3$ model. All the potential curvatures are very steep at the repulsive (left part of the well) in comparison to the attractive, for instance see figure (6.15). This means that it is harder for the polynomial to be accurate in that region. This would not pose a problem if it had not been for the force calculation. As mentioned in another chapter, there is a small fraction of configurations at close distances, so that would affect the total potential energy by a small constant. But it is more complicated than that. The fitting procedure **evenly** captures the whole dataset of the geometric averaging sampling $W^{(3),\text{geom}}$ with bigger local errors $E_{loc}$ at closer $\{r_{12}, r_{13}, r_{23}\}$ distances. In effect, although the differentiation is done **analytically**, giving an accurate value for each triplet, it is skewed in the closer regions. At this point, we suspected that the partial derivatives is more accurate, despite the discretization errors (tabulated form and central differences $\mathcal{O}(dr^2)$ error). In figure (6.16) we see the comparison of the force evaluation between $W^{(3),\text{cubic}}$ and partial derivatives at close distances $r_{12} = 3.9, r_{13} = 3.8, r_{23} \in [3.2 : 5]$. We can see differences towards $r_{min}$ and we see believe that the partial derivatives are more accurate. In the next section, we are able to assess the effect of this claim in the actual CG simulation.

## 6.5 CG runs

In the previous sections, we have been constructing effective potentials that describe the interactions between CG particles. In this section, we assess the accuracy by inspection of the CG simulation results. More specifically, we will assess the accuracy of the effective potentials with respect to the

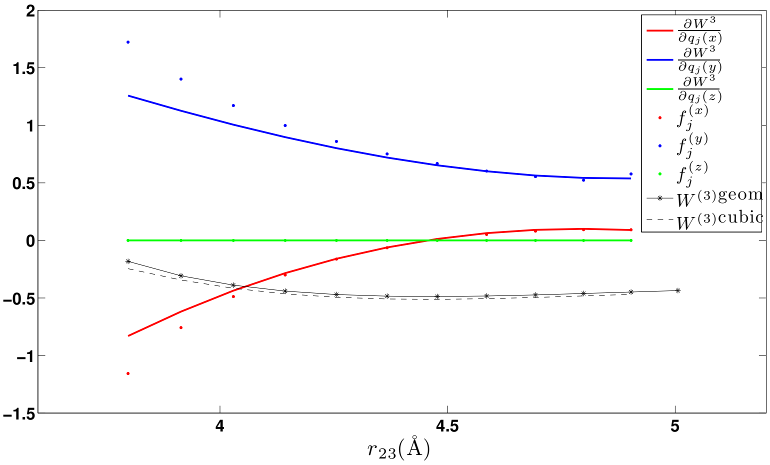

Figure 6.16: Comparison between the forces taken from fitted cubic polynomial $W^{(3),\text{cubic}}(3.9, 3.8, :)$ and partial derivatives (doted lines) in the same range, in all three dimensions. $CH_4$ at $T = 80$. We see good agreement as we move to longer $r_{23}$ values, though there is minor difference at close distances.

thermodynamic observable $g(r)$. A representation of the procedure we have followed so far is depicted in (6.19).

We note here that the evaluation of pressure is not correct. This is because every MD algorithm has a tail correction for the pressure $p_{trunc}$, which is analogous to the tail correction of the potential energy ($U_{\text{tail}} = (1/2)\int_{rcut}^{\infty} 4\pi r^2 \rho(r)U(r)dr$) due to the cutoff radius $r_{cut}$ [33]. In our construction of the 2-Body potential, we end up with a tabulated form of the effective potential $W^{(2)}$ and $p_{trunc}$ cannot be evaluated. On top of that, the stress tensor associated with pressure evaluation, is written with respect to pair interactions, so it is not clear how to include three body interactions. Nevertheless, we still have an estimate of the relative pressure (without $p_{trunc}$) for the 2-Body runs, but not a direct comparison with the reference (atomistic) systems.

As seen in (6.19), we insert the effective potential (as a table or parameterized formula) in the CG level simulation, along with the CG coordinates through the mapping operator $T$ (see Chapter 4). Our aim is to let the system equilibrate (by monitoring average quantities like the energy, temperature, pressure $g(r)$ and others) and export statistics and observables.

In the case of the $W^{(2),\text{geom}}$ 2-Body CG potential for both systems, the $g(r)$ between the reference and CG system is shown in figure (6.17)a for both systems. As expected, the effective pair potential does not predict the structure correctly. In the $CH_4$ case, the difference between the CG and reference $g(r)$ is smaller and it gets even smaller as temperature rises. This

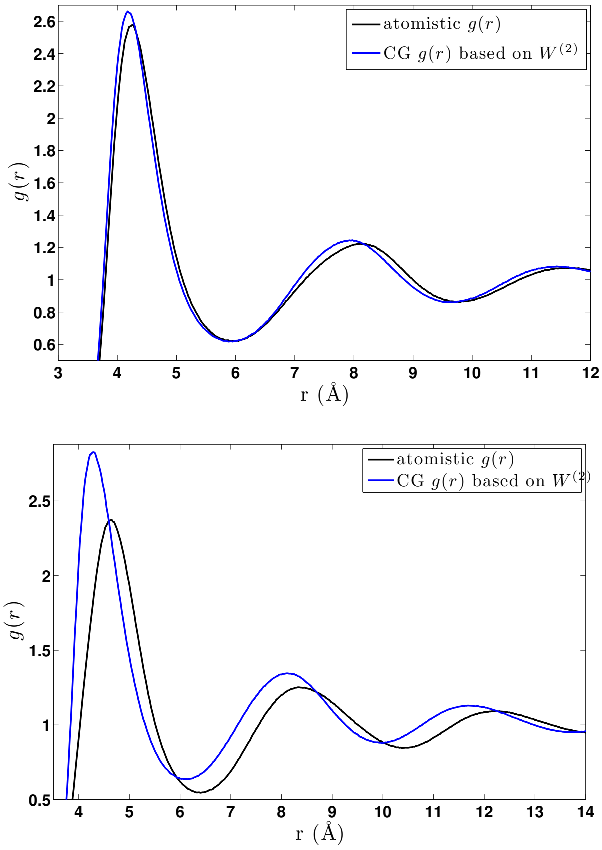

Figure 6.17: Comparison between the reference (atomistic) and CG $W^{(2)}$ $g(r)$ for a) $CH_4$ at $T = 80K$ and b) $CH_3 - CH_3$ at $T = 150K$.

small difference, in conjunction with the $CH_3 - CH_3$ system, can be justified from the close packing of this symmetric molecule; the mapping of one $CH_4$ to a spherical superatom of the same mass is a good approximation. So the CG particles tend to come even closer than the molecules in the atomistic description and this can be extracted from the height of the first peak.

In figure (6.17)b, the $CH_3 - CH_3$ sytem is not properly described by a spherical superatom, as it is rod-like and symmetric only in two dimensions. In effect, we see a significant difference between the two $g(r)$ plots. This difference persists even in higher temperatures and in lower densities. So it is due to the mapping and less because of the higher order terms in the pairwise potential approximation.

In figure (6.18) we show the improvement due to the more accurate 3-Body potentials $W^{(3),\text{cubic}}$ and $W^{(3),\text{p. d.}}$. Both methods, slightly vary because of the different approximation of the forces over the triplets of atom, as we argued in section (6.4.3). Overall, both methods improve on the

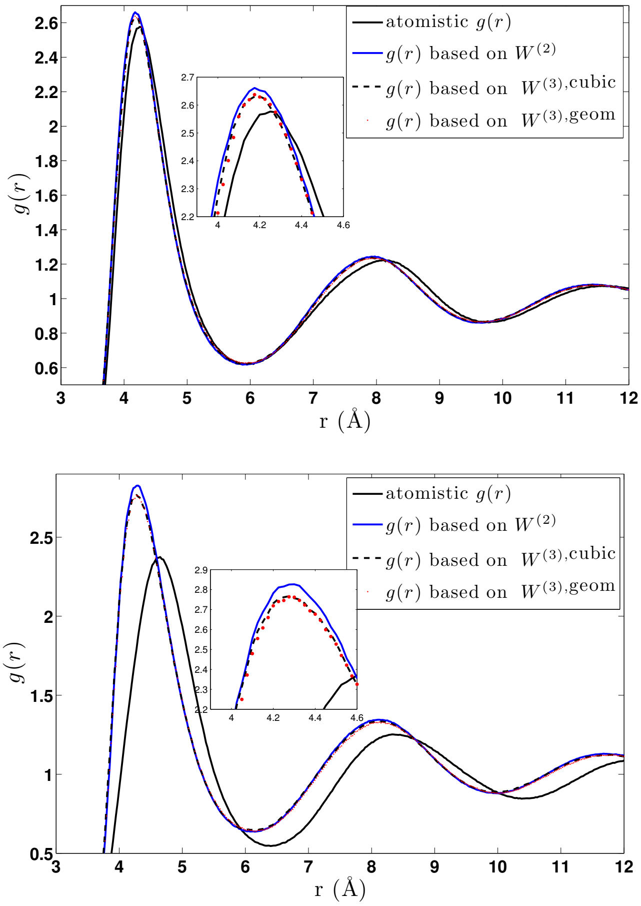

Figure 6.18: Comparison between the reference (atomistic) and CG $W^{(2)}$ $g(r)$ for a) $CH_4$ at $T = 80K$ and b) $CH_3 - CH_3$ at $T = 150K$.

estimation of the CG $g(r)$ at the cost of extra computations.

We note that the extra forcing in the system, required stronger coupling with the heat bath (dissipation), because the temperature was higher as a result of the extra kinetic energy.

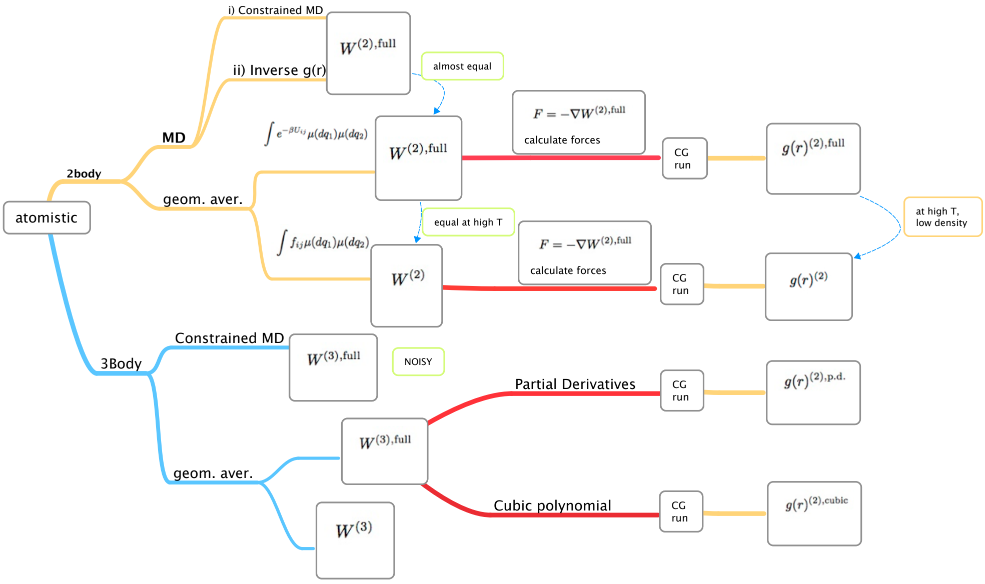

Figure 6.19: Schematic representation of the strategies used in the entire 3-body computations.

# Chapter 7

# Conclusions

In this work, we focused on: (a) the sensitivity analysis of molecular models, and (b) systematic coarse graining of different molecular systems. In chapter 3 we used the relative entropy, RE, metric in order to assess, qualitatively and quantitatively, which model (force field) parameters are more significant to small variations. We validated our findings against thermodynamic and structural observable quantities, which indicated that RE indeed predicts the impact that certain parameters have. The simulations were conducted on the path space, in equilibrium and non-equilibrium steady-state regimes. This method is quite general and can be used in models of stochastic nature whose random component is under mild assumptions.

The Relative Entropy Rate and Fisher Information matrix are the components per unit time of RE which we used in order to infer sensitivity. The RER and observables chosen, are related through the Pinsker inequality which bounds differences before and after perturbations in the parameter nominal values. RER is based on force differences so calculations are performed on the fly. FIM is independent of the parameter values and can be calculated *once* in order to screen out the most important parameters in high-dimensional systems. Our proposed methodology is independent of the numerical scheme, and provides a gradient-free approach for parametric SA.

In the LJ fluid system, the most sensitive potential parameter was $\sigma^{LJ}$, whereas in the more complex $CH_4$ system, which totals 10 bonded and non-bonded parameters, the bond length $r_0$ is the most sensitive one. The MSD and rdf observables agree with this finding. We also concluded that from a computational efficiency perspective, less steps are required for an accurate estimate than the typical number of steps need for the calculation of an observable.

In chapter 4 RE is used as a measure of loss of information in Coarse-Graining. We adapted the RE method to quantify and indicate the most efficient CG interaction potential in systems of increasing complexity. Also, we compared this way of determining Coarse-Graining with force matching and Iterative Boltzmann Inversion. The probabilistic formalism provides a generalized FM formula as a CG minimization problem, both for linear and non-linear CG maps. It also proves that RE and FM are in principle asymptotically equivalent. In practice, we found that the numerical implementations in the three methods are not bound to converge to the same solution, as there are differences in the derived numerical schemes. We used the derived pair potentials in CG simulations and checked that the structural properties are in good agreement with the reference data, for the case of the $CH_4$ system. The dynamic properties, like the MSD, are more sensitive to slight differences in the three derived CG potentials. For the case of the water model, there are larger differences in the derived potentials.

The applicability of FM and RE depends heavily on the basis functional used and it is of vital importance that the samples in high-energy, low-probability areas at close distances are adequately populated. We tackled with this issue by a preliminary detailed analysis in the all-atom configurations, and proper extrapolation and extra basis nodes in those regions.

In chapters 5 and 6 we derived 2-body and 3-body effective potentials using cluster expansions and potential of mean force techniques. We proposed a rigorous, systematic Coarse Graining strategy, both theoretical and algorithmic, in which we construct a hierarchy of CG Hamiltonians. We quantified the accuracy of the proposed CG potentials in terms of the rdf in CG simulations. For the 2-body potentials, we used four different methods which converged (within minor fluctuations) to similar form, for the $CH_4$ model and with a small difference in the non-spherically symmetric ethane model. The 3-body potential is computationally more expensive but slightly more accurate than the 2-body. We highlighted the issues and complexity associated with such computations. Finally, we conjecture that the cluster expansion formalism can be used in order to provide accurate effective pair and 3-body CG potentials at high temperature and low density regimes.

The above methods, that were developed during this PhD were applied in a few characteristic molecular fluids, like methane and ethane as well as in water and in small alkanes. However, the methods are quite general and in principle can be applied to more complex molecular systems, such as macomolecular (polymeric) fluids, biomolecular systems (e.g. peptides and proteins) and hybrid polymer/nanoparticles nanocomposites or graphene based hybrid systems. The application of these techniques on such complex systems will be a main direction in the future.

## .1 Definitions

- **simulation box** the volume in our simulations where particles move, subjected to periodic boundary conditions

- **periodic boundary conditions** when a particle moves towards the walls of the simulation volume it does not bounce back with opposite velocity sign but disappears and enters through the wall across it instead.

- **thermostat** method of controlling the temperature in a simulation to remain fixed. It can either be extra degrees of freedom in the equations of motion, rescaling of the velocities with respect to kinetic energy, random and dissipative forcing according to distributions etc.

- **integrator** the numerical scheme that solves the Newtonian equations of motion of the system in time according to time discretization *dt*

- **ensemble average** collection of particle configurations that correspond to macroscopic quantities $N, V, T, P, E$ and are associated with a probability measure. In principle this set is not countable.

- **trajectory** set of cartesian coordinates for each particle in the system, for every (discretized by *dt*) time point in the interval [0:T].

- **Coarse Graining** the mapping of atoms, through an operator, to a larger particle "superatom" in order to decrease the system total degrees of freedom and reduce the computational cost.

- **Burn-in period** the part of the trajectory (or number of timesteps) that we throw out when calculating averages, or properties in a simulation. It is an empirical value, dependent on the complexity of the system or stochastic process at study.

- $\mathbf{r_{cut}}$ cutoff distance (or radius) beyond which the intermolecular potential is set to zero, instead of asymptotically approaching zero. This approximation reduces the bulk simulation computational cost significantly and appropriate (potential) tail corrections are applied.

# Bibliography

[1] https://www.top500.org/.

[2] nvidia. https://en.wikipedia.org/wiki/cuda.

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

[4] 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.

[5] F. Müller Plathe. Coarse-graining in polymer simulation: From the atomistic to the mesoscopic scale and back. *Chem Phys. Chem.*, 3(9):754–769, 2002.

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

[7] W. J. Briels and R. L. C. Akkermans. Coarse-grained interactions in polymer melts: a variational approach. *J. Chem. Phys.*, 115:6210, 2001.

[8] Harmandaris V. A., Mavratzas V., Theodorou D., Kröger M., Ramìrez J., őttinger H.C., and Vlassopoulos D. Dynamic crossover from rouse to entangled polymer melt regime: Signals from long, detailed atomistic molecular dynamics simulations, supported by rheological experiments. *Macromolecules*, 36:1376–1387, 2003.

[9] 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.

[10] Vagelis A. Harmandaris and Kurt Kremer. Dynamics of polystyrene melts through hierarchical multiscale simulations. *Macromolecules*, 42:791, 2009.

[11] Carmen Hijon, Pep Español, Eric Vanden-Eijnden, and Rafael Delgado-Buscalioni. Mori-Zwanzig formalism as a practical computational tool. *Faraday Discuss.*, 144:301–322, 2010.

[12] A. J. Clark, J. McCarty, and M. G. Guenza. Effective potentials for representing polymers in melts as chains of interacting soft particles. *J. Chem. Phys.*, 139:124906, 2013.

[13] Grigorios A Pavliotis and Andrew M Stuart. *Multiscale methods, volume 53 of Texts in Applied Mathematics*. Springer, New York, 2008.

[14] M.J. Kotelyanskii and D.N. Theodorou. *Simulation Methods for Polymers*, volume Chapter "Molecular Dynamics Simulations of Polymers". Marcel Dekker, New York, 2004.

[15] R. R. Coifman, I. G. Kevrekidis, S. Lafon, M. Maggioni, and B. Nadler. Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems. *Multiscale Modeling & Simulation*, 7(2):842–864, 2008.

[16] A. Papavasiliou, G.A. Pavliotis, and A.M. Stuart. Maximum likelihood drift estimation for multiscale diffusions. *Stochastic Processes and their Applications*, 119(10):3173 – 3210, 2009.

[17] Dominik Fritz, Claudia R. Herbers, Kurt Kremer, and Nico F. A. van der Vegt. Hierarchical modeling of polymer permeation. *Soft Matter*, 5:4556–4563, 2009.

[18] V. Johnston and V. Harmandaris. Hierarchical simulations of hybrid polymer/solid materials. *Soft Matter*, 9:6696–6710, 2013.

[19] W. G. Noid, Jhih-Wei Chu, Gary S. Ayton, Vinod Krishna, Sergei Izvekov, Gregory A. Voth, Avisek Das, and Hans C. Andersen. The multiscale coarse-graining method. i. a rigorous bridge between atomistic and coarse-grained models. *The Journal of Chemical Physics*, 128(24):244114, 2008.

[20] I. Bilionis and N. Zabaras. A stochastic optimization approach to coarse-graining using a relative-entropy framework. *J. Chem. Phys.*, 138(4), 2013.

[21] 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.

[22] W. G. Noid, Pu Liu, Yanting Wang, Jhih-Wei Chu, Gary S. Ayton, Sergei Izvekov, Hans C. Andersen, and Gregory A. Voth. The

multiscale coarse-graining method. ii. numerical implementation for coarse-grained molecular models. *The Journal of Chemical Physics*, 128(24):244115, 2008.

[23] W. G. Noid. Perspective: Coarse-grained models for biomolecular systems. *The Journal of Chemical Physics*, 139(9):090901, 2013.

[24] S. K. Rao, R. Imam, K. Ramanathan, and S. Pushpavanam. Sensitivity analysis and kinetic parameter estimation in a three way catalytic converter. *Industrial & Engineering Chemistry Research*, 48:3779–3790, 2009.

[25] A. F. Emery and A. V. Nenarokomov. Optimal experiment design. *Measurement Science and Technology*, 9(6):864, 1998.

[26] B. Cooke and S. C. Schmidler. Statistical prediction and molecular dynamics simulation. *Biophysical Journal*, 95:4497–4511, November 2008.

[27] Y. Pantazis, M. Katsoulakis, and D. Vlachos. Parametric sensitivity analysis for biochemical reaction networks based on pathwise information theory. *BMC*, 2013.

[28] J. E. Mayer and M. G. Mayer. *Statistical Mechanics*. John Wiley and Sons, 1940.

[29] A. Tsourtis, Y. Pantazis, M. Katsoulakis, and V. Harmandaris. Parametric sensitivity analysis for stochastic molecular systems using information theoretic metrics. *The Journal of Chemical Physics*, 143:014116, 2015.

[30] 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.

[31] A. Tsourtis, D. K. Tsagkarogiannis, and Harmandaris V. Parameterization of coarse-grained molecular interactions through potential of mean force calculations and cluster expansions techniques. *J Chem. Phys. (submitted)*, 2016.

[32] M. P. Allen and D. J. Tildesley. *Computer simulation of liquids*. Clarendon Press, New York, NY, USA, 1987.

[33] Daan Frenkel and B. Smit. *Understanding Molecular Simulation, Second Edition: From Algorithms to Applications (Computational Science)*. Academic Press, 2001.

[34] Herbert Goldstein. *Classical Mechanics*. Addison-wesley, 1950.

[35] T. Lelievre, M. Rousset, and G. Stoltz. *Free Energy Computations: A Mathematical Perspective*. Imperial College Press, 2010.

[36] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman J.C Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3):327 – 341, 1977.

[37] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J Salmon, Y. Shan, and W. Wriggers. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330(6002):341–346, 2010.

[38] D. Fritz, V. A. Harmandaris, K. Kremer, and N. Van der Vegt. Coarse-grained polymer melts based on isolated atomistic chains: simulation of polystyrene of different tacticities. *Macromolecules*, 42(19):7579–7588, 2009.

[39] A. Chernatynskiy, S. Phillpot, and R. LeSar. Uncertainty quantification in multiscale simulation of materials: A prospective. *Annual review of Materials Research*, 43:157–182, July 2013.

[40] P. Angelikopoulos, C. Papadimitriou, and P. Koumoutsakos. Bayesian uncertainty quantification and propagation in molecular dynamics simulations: A high performance computing framework. *J. Chem. Phys*, 137(14), 2012.

[41] H. Liu, A. Sudjianto, and W. Chen. Relative entropy based method for probabilistic sensitivity analysis in engineering design. *Journal of Mechanical Design*, 128(2):326–336, 2005.

[42] F. Cailliez and P. Pernot. Statistical approaches to forcefield calibration and prediction uncertainty in molecular simulation. *J Chem. Phys.*, 134(5), 2011.

[43] F. Rizzi, H. N. Najm, B. J. Debusschere, K. Sargsyan, M. Salloum, H. Adalsteinsson, and O. M. Knio. Uncertainty quantification in md simulations. part i: forward propagation. *SIAM Multiscale Model. Simul.*, 10(4):1428–1459, 12 2012.

[44] Y. Pantazis and M. Katsoulakis. A relative entropy rate method for path space sensitivity analysis of stationary complex stochastic dynamics. *J. Chem. Phys*, 138, 2013.

[45] R. D. Braatz, R. C. Alkire, E. Seebauer, E. Rusli, R. Gunawan, T.O. Drews, X. Li, and Y. He. Perspectives on the design and control of multiscale systems. *J. Proc. Control*, 16:193–204, 2006.

[46] R. Gunawan, Y. Cao, L. Petzold, and F. J. III Doyle. Sensitivity analysis of discrete stochastic systems. *Biophysical Journal*, 88:2530–2540, 2005.

[47] M. Rathinam, P. W. Sheppard, and M. Khammash. Efficient computation of parameter sensitivities of discrete stochastic chemical reaction networks. *J. Chem. Phys.*, 132:034103–(1–13), 2010.

[48] D. F. Anderson. An efficient finite difference method for parameter sensitivities of continuous-time Markov chains. *SIAM J. Numerical Analysis*, 50(5):2237–2258, 2012.

[49] G. Arampatzis and M. A. Katsoulakis. Goal-oriented sensitivity analysis for lattice kinetic monte carlo simulations. *J. Chem. Phys.*, 12(140):124108, 2014.

[50] P. W. Glynn. Likelihood ratio gradient estimation for stochastic systems. *Communications of the ACM*, 33(10):75–84, 1990.

[51] S. Plyasunov and A. P. Arkin. Efficient stochastic sensitivity analysis of discrete event systems. *J. Comput. Phys.*, 221(2):724–738, February 2007.

[52] P. B. Warren and R. J. Allen. Steady-state parameter sensitivity in stochastic modeling via trajectory reweighting. *J. Chem. Phys.*, 136(10), 2012.

[53] T. Iordanov, G. Schenter, and B. Garret. Sensitivity analysis of thermodynamic properties of liquid water: a general approach to improve empirical potentials. *J. Phys. Chem*, A(110):762–771, 2006.

[54] A. Majda and B. Gershgorin. Quantifying uncertainty in climate change science through empirical information theory. *PNAS*, 107(34):14958–14963, 2010.

[55] 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.

[56] K. R. Haas, H. Yang, and J-W Chu. Fisher information metric for the Langevin equation and least informative models of continuous stochastic dynamics. *J. Chem. Phys.*, 139(12), SEP 28 2013.

[57] V. Johnston and V. Harmandaris. Hierarchical multiscale modeling of polymer-solid interfaces: Atomistic to coarse-grained description and structural and conformational properties of polystyrene-gold systems. *Macromolecules*, 46:5741–5750, 2013.

[58] A. Rissanou and V. Harmandaris. Dynamics of various polymer/graphene interfacial systems through atomistic molecular dynamics simulations. *Soft Matter*, 42(10):2876–2888, 2014.

[59] V. Harmandaris. Quantitative study of equilibrium and non-equilibrium polymer dynamics through systematic hierarchical coarse-graining simulations. *Korea-Aust. Rheol. J.*, 26:15–28, 2014.

[60] B. Smit. Phase diagrams of lennard-jones fluids. *J. Chem. Phys*, 96(11), 6 1992.

[61] S. Mayo, B. Olafson, and W. Goddard. Dreiding: a generic force field for molecular simulations. *J. Phys. Chem.*, 94(26):8897–8909, 1990.

[62] T Cover and J. Thomas. *Elements of Information Theory.* John Wiley and Sons, 1991.

[63] B. Oksendal. *Stochastic Differential Equations: An introduction with applications.* Springer-Verlag, 5th edition, 2000.

[64] V. Harmandaris, V. Mavrantzas, and D. Theodorou. Atomistic molecular dynamics simulation of stress relaxation upon cessation of steady-state uniaxial elongational flow. *Macromolecules*, 33(21):8062–8076, 2000.

[65] B. Leimkuhler and C. Matthews. Robust and efficient configurational molecular sampling via langevin dynamics. *J. Chem. Phys*, 138(17), 2013.

[66] V. Bally and D. Talay. The law of the Euler scheme for stochastic differential equations. I. Convergence rate of the distribution function. *Probab. Theory Related Fields*, 104:43–60, 1996.

[67] V. Bally and D. Talay. The law of the Euler scheme for stochastic differential equations. II. Convergence rate of the density. *Monte Carlo Methods Appl.*, 2:93–128, 1996.

[68] J. C. Mattingly, A. M. Stuart, and M. V. Tretyakov. Convergence of numerical time-averaging and stationary measures via Poisson equations. *SIAM J. Numer. Anal.*, 48:552–577, 2010.

[69] 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.

[70] D. A. Hensher and K. J. Button. *Handbook of Transport and the Environment.* Elsevier, 2003.

[71] D. McQuarrie. *Statistical Mechanics.* Harper Collins Publishers, 1976.

[72] S. B. Zhu and C. F. Wong. Sensitivity analysis of a polarizable water model. *J. Chem. Phys*, 98:4695–4701, 1994.

[73] H. Heinz, W. Paul, and K. Binder. Calculation of local pressure tensors in systems with many-body interactions. *Phys. Rev. E*, 72(6):066704–066714, 2005.

[74] M. S. Shell. Systematic coarse-graining of potential energy landscapes and dynamics in liquids. *J. Chem. Phys*, 137(8), 2012.

[75] M. A. Katsoulakis, P. Plecháč, L. Rey-Bellet, and D. K. Tsagkarogiannis. Coarse-graining schemes and a posteriori error estimates for stochastic lattice systems. *ESAIM: Mathematical Modelling and Numerical Analysis*, 41:627–660, 5 2007.

[76] M. A. Katsoulakis, A. J. Majda, and D. G. Vlachos. Coarse-grained stochastic processes for microscopic lattice systems. *Proc. Natl. Acad. Sci. USA*, 100(3):782–787, 2003.

[77] M. A. Katsoulakis, A. J. Majda, and D. G. Vlachos. Coarse-grained stochastic processes and monte carlo simulations in lattice systems. *J. Comp. Phys.*, 112:250–278, 2003.

[78] M. A. Katsoulakis, L. Rey-Bellet, P. Plecháč, and D. K. Tsagkarogiannis. Mathematical strategies in the coarse-graining of extensive systems: error quantification and adaptivity. *J. Non Newt. Fluid Mech.*, 2008.

[79] M.A. Katsoulakis and P. Plecháč. Information-theoretic tools for parametrized coarse-graining of non-equilibrium extended systems. *J. Chem. Phys.*, 139(arXiv:1304.7700), 2013.

[80] A. Rissanou, E. Georgilis, M. Kasotaskis, A. Mitraki, and V. Harmandaris. Effect of solvent on the self-assembly of dialanine and diphenylalanine peptides. *J. Phys. Chem. B*, 117:3962–3975, 2013.

[81] Evangelia Kalligiannaki, Vagelis Harmandaris, Markos A. Katsoulakis, and Petr Plechac. The geometry of generalized force matching and related information metrics in coarse-graining of molecular systems. *The Journal of Chemical Physics*, 143(8), 2015.

[82] E. Kalligiannaki, V. Harmandaris, M. Katsoulakis, and P. Plecháč. Parametrization of coarse grained models for non equilibrium complex systems. *to be submitted*.

[83] B. Leimkuhler, E. Noorizadeh, and F. Theil. A gentle stochastic thermostat for molecular dynamics. *J. Stat. Phys.*, (135):261–277, 2009.

[84] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 2009.

[85] Vagelis A. Harmandaris and Kurt Kremer. Predicting polymer dynamics at multiple length and time scales. *Soft Matter*, 5:3920, 2009.

[86] Karen Johnston and Vagelis Harmandaris. Hierarchical simulations of hybrid polymer– solid materials. *Soft Matter*, 9:6696–6710, 2013.

[87] Lanyuan Lu, Sergei Izvekov, Avisek Das, Hans C. Andersen, and Gregory A. Voth. Efficient, regularized, and scalable algorithms for multiscale coarse-graining. *Journal of Chemical Theory and Computation*, 6(3):954–965, 2010.

[88] Joseph F. Rudzinski and W. G. Noid. Coarse-graining entropy, forces, and structures. *The Journal of Chemical Physics*, 135(21):214101, 2011.

[89] 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.

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

[91] A.P. Lyubartsev and A. Laaksonen. Calculation of effective interaction potentials from radial distribution functions: A reverse Monte Carlo approach. *Phys. ReV. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.*, 52:3730–3737, 1995.

[92] 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.

[93] J. McCarty, A. J. Clark, J. Copperman, and M. G. Guenza. An analytical coarse-graining method which preserves the free energy, structural correlations, and thermodynamic state of polymer melts from the atomistic to the mesoscale. *The Journal of Chemical Physics*, 140(20):–, 2014.

[94] Markos A. Katsoulakis, Petr Plecháč, and Luc Rey-Bellet. Numerical and statistical methods for the coarse-graining of many-particle stochastic systems. *J. Sci. Comput.*, 37(1):43–71, 2008.

[95] Jose Trashorras and Dimitrios K. Tsagkarogiannis. From Mesoscale Back to Microscale: Reconstruction Schemes for Coarse-Grained Stochastic Lattice Systems. *SIAM Journal on Numerical Analysis*, 48(5):1647–1677, 2010.

[96] Vagelis Harmandaris, Evangelia Kalligiannaki, Markos A. Katsoulakis, and Petr Plecháč. Path-space variational inference for non-equilibrium coarse-grained systems. *Journal of Computational Physics*, 314:355–383, 2016.

[97] Evangelia Kalligiannaki, Markos A. Katsoulakis, Petr Plechac, and Dionisios G. Vlachos. Multilevel coarse graining and nano-pattern discovery in many particle stochastic systems. *J. Comp. Phys.*, 231(6):2599–2620, Mar 2012.

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

[99] Emiliano Brini, Elena A. Algaer, Pritam Ganguly, Chunli Li, Francisco RodrìguezRopero, and Nico F. A. van der Vegt. Systematic coarse-graining methods for soft matter simulations–a review. *Soft Matter*, 9:2108–2119, 2013.

[100] Christine Peter and Kurt Kremer. Multiscale simulation of soft matter systems - from the atomistic to the coarse-grained level and back. *Soft Matter*, 5(22):4357–4366, 2009.

[101] Paola Carbone, Hossein Ali Karimi Varzaneh, Xiaoyu Chen, and Florian Mŭller-Plathe. Transferability of coarse-grained force fields: The polymer case. *J. Chem. Phys.*, 128:064904, 2008.

[102] Y. N. Pandey, A. Brayton, C. Burkhart, G. J. Papakonstantopoulos, and Doxastakis M.J. Multiscale modeling of polyisoprene on graphite. *. Chem. Phys.*, 140:054908, 2014.

[103] J. T. Padding and W. J. Briels. Uncrossability constraints in mesoscopic polymer melt simulations: Non-rouse behavior of c120h242. *The Journal of Chemical Physics*, 115(6):2846–2859, 2001.

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

[105] 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.

[106] A. Chaimovich and M. S. Shell. Coarse-graining errors and numerical optimization using a relative entropy framework. *J. Chem. Phys.*, 134(9), 2011.

[107] H.M. Cho and J.W. Chu. Inversion of radial distribution functions to pair forces by solving the yvon–born–green equation iteratively. *J. Chem. Phys.*, 131:134107, 2009.

[108] W. G. Noid, G. S. Ayton J. Chu, and G. A. Voth. Multiscale coarse graining and structural correlations: Connections to liquid-state theory. *J. Phys. Chem. B*, 111:4116–4127, 2007.

[109] J.W. Mullinax and W. G. Noid. Generalized yvon–born–green theory for molecular systems. *Phys. Rev. Lett.*, 103:198104, 2009.

[110] J.W. Mullinax and W. G. Noid. Generalized yvon–born–green theory for determining coarse-grained interaction potentials. *J. Phys. Chem. C*, 114:5661–5674, 2010.

[111] Zhen Li, Xin Bian, Xiantao Li, and George Em Karniadakis. Incorporation of memory effects in coarse-grained modeling via the mori-zwanzig formalism. *J. Chem. Phys.*, 143(-):243128, 2015.

[112] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes*. Cambridge University Press, 2007.

[113] and Lanyuan Lu Avisek Das, Hans C. Andersen, and Gregory A. Voth. The multiscale coarse-graining method. x. improved algorithms for constructing coarse-grained potentials for molecular systems. *The Journal of Chemical Physics*, 136:194115, 2012.

[114] M. Maiolo, A. Vancheri, R. Krause, and A. Danani. Wavelets as basis functions to represent the coarse-graining potential in multiscale coarse graining approach. *Journal of Computational Physics*, 300(-):592–604, 2015.

[115] A. P. Bartók, M. C. Payen, R. Kondor, and G. Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical Review Letters*, 104(-):136403, 2010.

[116] Luca Larini, Lanyuan Lu, and Gregory A. Voth. The multiscale coarse-graining method. vi. implementation of three-body coarse-grained potentials. *The Journal of Chemical Physics*, 132(16):164107, 2010.

[117] J.D. McCoy and J.G. Curro. Mapping of explicit atom onto united atom potentials. *Macromolecules*, 31:9352–9368, 1998.

[118] Vikram Reddy Ardham, Gregor Deichmann, Nico F. A. van der Vegt, and Frédéric Leroy. Solid-liquid work of adhesion of coarse-grained models of n-hexane on graphene layers derived from the conditional reversible work method. *J. Chem. Phys.*, 143:243135, 2015.

[119] Gregor Deichmann, Valentina Marcon, and Nico F. A. van der Vegt. Bottom-up derivation of conservative and dissipative interactions for coarse-grained molecular liquids with the conditional reversible work method. *J. Chem. Phys.*, 141:224109, 2014.

[120] T. Murtola, A. Bunker, I. Vattulainen, M. Deserno, and M. Karttunen. Multiscale modeling of emergent materials: biological and soft matter. *Phys. Chem. Chem. Phys.*, 11:1869–1892, 2009.

[121] Victor Ruhle, Christoph Junghans, Alexander Lukyanov, Kurt Kremer, and Denis Andrienko. Versatile object-oriented toolkit for coarse-graining applications. *J. Chem. Theory Comput.*, 5(-):3211–3223, 2009.

[122] S. P. Carmichael and M. S. Shell. A new multiscale algorithm and its application to coarse-grained peptide models for self-assembly. *The Journal of Physical Chemistry B*, 116(29):8383–8393, 2012.

[123] James C. Spall. *Introduction to Stochastic Search and Optimization.* John Wiley & Sons, Inc., New York, NY, USA, 1 edition, 2003.

[124] Léon Bottou. Stochastic learning. In Olivier Bousquet and Ulrike von Luxburg, editors, *Advanced Lectures on Machine Learning*, Lecture Notes in Artificial Intelligence, LNAI 3176, pages 146–168. Springer Verlag, Berlin, 2004.

[125] Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *J. Mach. Learn. Res.*, 14(1):1303–1347, May 2013.

[126] Han Wang, Christoph Junghans, and Kurt Kremer. Comparative atomistic and coarse-grained study of water: What do we lose by coarse-graining? *Eur. Phys. J. E*, 28:221–229, 2009.

[127] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma. The missing term in effective pair potentials. *J. Phys. Chem.*, 91(-):6269–6271, 1987.

[128] W.L. Jorgensen, Maxwell D.S., and J. Tirado-Rives. Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.*, 118(-):11225–11236, 1996.

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

[130] T. Morita and K. Hiroike. The statistical mechanics of condensing systems. III. *Prog. Theor. Phys.*, 25:537, 1961.

[131] G. Stell. *Cluster expansion for classical systems in equilibrium in H. Frisch and J. Lebowitz, ed., Classical Fluids.* New York: Benjamin, 1964.

[132] E. Pulvirenti and D. Tsagkarogiannis. Cluster expansion in the canonical ensemble. *Comm. Math. Phys.*, 316(2):289–306, 2012.

[133] M. Katsoulakis, P. Plecháč, L. Rey-Bellet, and D. Tsagkarogiannis. Coarse-graining schemes and a posteriori error estimates for stochastic lattice systems. *ESAIM: Math. Model. and Num. Analysis*, 41(3):627–660, 2007.

[134] K. Kremer and F. Müller-Plathe. Multiscale problems in polymer science: simulation approaches. *MRS Bull.*, -:205, 2001.

[135] A. Tsourtis, V. Harmandaris, and D. Tsagkarogiannis. Effective coarse-grained interactions: The role of three-body terms through cluster expansions. *under preparation.*

[136] J. G. Kirkwood. Statistical mechanics of fluid mixtures. *J. Chem. Phys.*, 3(5):300–313, 1935.

[137] A. A. Louis. Beware of density dependent pair potentials. *J.Phys.: Condens. Matter*, 14:9187–9206, 2002.

[138] P. G. Bolhuis, A. A. Louis, and J. P. Hansen. Many-body interactions and correlations in coarse-grained descriptions of polymer solutions. *Phys. Rev. E*, 64:021801, 2001.

[139] E. Pulvirenti and D. Tsagkarogiannis. Finite volume corrections and decay of correlations in the canonical ensemble. *J. Stat. Phys.*, 159(5):1017–1039, 2014.

[140] T. Kuna and D. Tsagkarogiannis. Convergence of density expansions of correlation functions and the Ornstein-Zernike equation. *preprint arXiv:1611.01716*, 2016.

[141] C. D. Wick, M. G. Martin, and J. I. Siepmann. Transferable potentials for phase equilibria. 4. united-atom description of linear and branched alkenes and alkylbenzenes. *J. Phys. Chem. B*, 104:8008–8016, 2000.

[142] T. Kuna, J. Lebowitz, and E. Speer. Realizability of point processes. *J. Stat. Phys.*, 129:417–439, 2007.

[143] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[144] Marcus .G Martin and J.Ilja Siepmann. Transferable potentials for phase equilibria. 1. united-atom description of n-alkanes. *J. Phys. Chem.*, B(102 (14)):2569–2577, 1998.

[145] David Chandler. *Introduction to Modern Statistical Mechanics.* Oxford University Press, 1987.