entropy

MDPI

Article

Parameterization of Coarse-Grained Molecular Interactions through Potential of Mean Force Calculations and Cluster Expansion Techniques

Anastasios Tsourtis [1,*], Vagelis Harmandaris [1,2,*] and Dimitrios Tsagkarogiannis [3,*]

[1] Department of Mathematics and Applied Mathematics, University of Crete, Heraklion 70013, Greece
[2] Institute of Applied and Computational Mathematics (IACM), Foundation for Research and Technology Hellas (FORTH), Heraklion 70013, Greece
[3] Department of Mathematics, University of Sussex, Brighton BN1 9QH, UK
* Correspondence: tsourtis@uoc.gr (A.T.); harman@uoc.gr (V.H.); D.Tsagkarogiannis@sussex.ac.uk (D.T.); Tel.: +30-2810-393768 (A.T.); +30-2810-393735 (V.H.); +44-1273-876824 (D.T.)

Received: 24 May 2017; Accepted: 24 July 2017; Published: 1 August 2017

Abstract: 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. In this way, the suggested model becomes computationally tractable, since no information from long n-body (bulk) simulations is required in order to develop it, while retaining the fluctuations at the coarse-grained level. 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 the 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. On the other hand, by construction, the approximate coarse-grained models retain, in principle, the thermodynamic properties of the atomistic model without the need for any further parameter fitting. We give specific examples for methane and ethane molecules in which the coarse-grained variable is the centre of mass of the molecule. We investigate different temperature (T) and density (ρ) 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 ρ 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.

Keywords: cluster expansions; PMF calculations; systematic coarse-graining; three-body effective potential

1. Introduction

The theoretical study of complex molecular systems is a very intense research area due to both basic scientific questions and technological applications [1]. A main challenge in this field is to provide a direct quantitative link between the 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 different thermodynamic conditions [1–4]. 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 [1,3,5–22].

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 [4,20]. 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, [5–9,11–13,15,17–19,23–25] and the 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 specific molecular systems without adjustable parameters and by that become truly predictive [11,14,15]. 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 [6,7,9,10,21,22]. The CG effective interactions in such methods are obtained using the direct Boltzmann inversion, or reversible work, method [10,26–28], or iterative techniques, such as the iterative Boltzmann inversion (IBI) [7,29] and the inverse Monte Carlo (IMC); or inverse Newton approach [22,30].

(b) Force matching (FM) or multi-scale CG (MSCG) methods [5,14,16,31–33] comprise a mean least squares problem that considers as the observable function the total force acting on a coarse bead.

(c) The relative entropy (RE) [8,18,34] method employs the minimization of the relative entropy, or Kullback–Leibler divergence, between the microscopic Gibbs measure μ and μ^{θ}, representing approximations to the exact coarse space Gibbs measure. In this case, the microscopic probability distribution can be thought of as the observable. The minimization of the relative entropy is performed through Newton–Raphson approaches and/or stochastic optimization techniques [19,35].

In practice, all of the above numerical methods are employed to approximate a many body potential of mean force (PMF), U_{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 [36–43]. In these methods, in order to account for the many-body contributions (from the higher order terms), a closure scheme is performed. However, although these methods do perform well in some cases, there is no rigorous quantification of the closure error in the corresponding diagrammatic expansions; see [44,45]. The closures take into account some diagrams from all orders in the expansions (and hence, can potentially give good approximations in the liquid state), but the sum of the neglected terms (i.e., the error) has not been quantified. The quantification of these terms is a challenging open problem, both from a theoretical point of view due to the complicated combinatorial structure, as well as from the computational one since it involves many-body interactions. Hence, a common idea in

all methods is to reduce this complexity. We also refer to [46], where simplifications are produced by considering the infinite dilution limit.

Here, we discuss an approach for estimating U_{PMF}, and the corresponding effective CG non-bonded potential, based on cluster expansion methods. Such methods originate from the works of Mayer and collaborators [47] in the 1940s. In the 1960s, numerous approximate expansions had been further developed [44,48,49] for the study of the liquid state. We also refer to the later textbook [50]. 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 become 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 the error estimation, transferability and adaptivity of the suggested methods. Based on recent developments of the mathematical theory of expansion methods in the canonical ensemble [51], 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 the design of related adaptive methods. We note that our suggested approach is a bottom up methodology, in the sense that we directly sample the CG effective potential based on atomistic simulations of only a few particles, instead of matching average forces or equilibrium quantities of larger n-body (bulk) systems, as in (a)–(c). By construction, the effective potential is independent of the density ρ, but naturally, it depends on the number of hierarchy terms in each computation, as well as on the temperature. Furthermore, in each regime, different expansions might be appropriate. Note that density-dependent CG effective interaction potentials, which are often used in the literature, are a non-systematic way to account for the many-body interactions [52].

In previous years, we have developed CG models, based on cluster expansions, for lattice systems, obtaining higher order schemes and a posteriori error estimates [53], for both short- and long-range interactions [54] and designing adaptive methods [55] and investigating possible strategies for reconstruction of the atomistic information [56]. This is very much in the spirit of the polymer science literature [10,11,57], 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 [26,27,58]. Here, we extend such approaches using cluster expansion tools for deriving CG effective potentials. We start from the typical two-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. In the present work, we focus on two prototypical examples of molecular systems, one with spherical symmetry (methane) and one without (ethane). In future work, we will extend these results to longer molecules, not necessarily with spherical symmetry. 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 [59].

The structure of the paper is as follows: In Section 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 3. Details about the model systems (methane and ethane) and the simulation considered here are discussed in Section 4. Results are presented in Section 5; we briefly discuss the three-body case in Section 6. Finally, we close with Section 7 summarizing the results of this work.

2. Molecular Models

2.1. Atomistic and "Exact" Coarse-Grained 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 Λ when we do not need to make explicit the dependence on ℓ. We consider a configuration $\mathbf{q} \equiv \{q_1, \ldots, q_N\}$ of N atoms, where q_i is the position of the i-th atom. The particles interact via a pair potential $V : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, for which we assume the standard conditions of stability and temperedness; namely, that there exists a constant $B \geq 0$ such that: $\sum_{1 \leq i < j \leq N} V(q_i - q_j) \geq -BN$ for all N and all $q_1, ..., q_N$ and that $C(\beta) := \int_{\mathbb{R}^d} |e^{-\beta V(r)} - 1| dr < \infty$, where $\beta = \frac{1}{k_B T}$ and k_B is the Boltzmann's constant.

The canonical partition function of the system is given by:

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

where H_Λ is the Hamiltonian (total energy) of the system confined in a domain Λ:

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

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

$$U(\mathbf{q}) := \sum_{1 \leq i < j \leq N} V(q_i - q_j), \tag{3}$$

where, for simplicity, we assume periodic boundary conditions on Λ. Integrating over the momenta in (1), we get:

$$Z_{\beta,\Lambda,N} = \frac{\lambda^N}{N!} \int_{\Lambda^N} dq_1 \ldots dq_N \, e^{-\beta U(\mathbf{q})} =: \lambda^N Z^U_{\beta,\Lambda,N}, \tag{4}$$

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:

$$\rho^{(2),at}_{N,\Lambda}(q_1, q_2) := \frac{1}{(N-2)!} \int dq_3 \ldots dq_N \frac{1}{Z_{\beta,\Lambda,N}} e^{-\beta U(\mathbf{q})}. \tag{5}$$

It is easy to see that in the thermodynamic limit, the leading order is ρ^2, where $\rho = \frac{N}{|\Lambda|}$, and $|\Lambda|$ is the volume of the box Λ. 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 $n \leq N$, we define the n-body version:

$$g^{(n)}(q_1, \ldots, q_n) = \frac{1}{(N-n)! \rho^n} \int_{\Lambda^{N-n}} dq_{n+1} \ldots dq_N \frac{1}{Z_{\beta,\Lambda,N}} e^{-\beta U(\mathbf{q})}, \tag{6}$$

and from that, the order n potential of mean force (PMF), $U_{\mathrm{PMF}}(q_1, \ldots, q_n)$ [60,61], given by:

$$U_{\mathrm{PMF}}(q_1, \ldots, q_n) := -\frac{1}{\beta} \log g^{(n)}(q_1, \ldots, q_n). \tag{7}$$

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 $\mathbf{r} \equiv \{r_1, \ldots, r_M\}$. We further define the effective CG potential energy by:

$$U_{\text{eff}}(r_1,\ldots,r_M) := -\frac{1}{\beta}\log\int_{\{T\mathbf{q}=\mathbf{r}\}} dq_1\ \ldots dq_N\, e^{-\beta U(\mathbf{q})}, \tag{8}$$

where the integral is over all atomistic configurations that correspond to a specific CG one using the coarse-graining map. Note that U_{eff} is in practice equivalent, up to a constant, to the (conditional) PMF. In the example we will deal with later, the configuration $\mathbf{r}$ will represent the centres 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 (8), the partition function is the same:

$$Z_{\beta,\Lambda,N} = \int dr_1\ldots dr_M \int_{\{T\mathbf{q}=\mathbf{r}\}} d\mathbf{q} e^{-\beta U(\mathbf{q})} = \int dr_1\ldots dr_M e^{-\beta U_{\text{eff}}(r_1,\ldots,r_M)} =: Z^{cg}_{\beta,\Lambda,M}. \tag{9}$$

The main purpose of this article is to give a systematic way (via the cluster expansion method) of constructing controlled approximations of U_{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 [61] to each other as follows:

$$\beta p = \rho - \frac{\beta}{6}\rho^2 \int_0^\infty r u'(r) g(r) 4\pi r^2 dr, \tag{10}$$

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

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_{eff} as defined in (8). Below, we list some of them without the claim of being exhaustive:

(a) The correlation-based (e.g., 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 [6,7,9,10,21,22]. These methods will be further discussed below.

(b) Force matching (FM) methods [5,16,31] 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)-type [8,18,19] 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 for polymers and polymer blends [40,42]. A brief review and categorization of parametrization methods at equilibrium is given in [17,62].

The correlation-based iterative methods use the fact that for a pair interaction $u(r)$, by plugging the virial expansion of p in powers of ρ into (10) and comparing the orders of ρ, one obtains that [61]:

$$g(r) = e^{-\beta u(r)}\gamma(r), \quad \text{where} \quad \gamma(r) = 1 + c_1(r)\rho + c_2(r)\rho^2 + \ldots \tag{11}$$

Given the atomistic "target" $g(r)$ from a free (i.e., without constraints) atomistic run, by inverting (11) 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 scheme is based on the theorem that if there exists a pair CG interaction that reproduces the target $g(r)$, this is unique [63]. However, this is only an approximation (accounting for the neglected terms of order ρ 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 $g(r)$, within numerical accuracy, does not secure proper thermodynamic behaviour, and several methods have been employed towards this direction; see, for example, [7,40,52] and the references within.

In order to maintain the correct thermodynamic properties, our approach in this paper is based on cluster expanding (8) with respect to some small, but finite parameter ϵ depending on the regime in which we are interested. For technical reasons we will focus here on the low density-high temperature regime. However, the cluster expansion is a perturbative method that can in principle be applied to other regimes once the appropriate target system is identified. For example, we refer to [53] where similar perturbations around mean field models for lattice systems have been implemented. In that case, the validity of the expansion can go beyond the usual high temperature regime. Extending this result to off-lattice systems, as is the case here, is more complicated, and we leave it for future investigation. As will be explained in detail in the next section, the resulting cluster expansion provides us with a hierarchy of terms. Uniformly in $\mathbf{r} \equiv r_1, \ldots, r_M$, we have:

$$\begin{aligned} &U_{\text{eff}}(\mathbf{r}) = U^{(2)}(\mathbf{r}) + O(\epsilon^2), \quad \text{or} \quad U_{\text{eff}}(\mathbf{r}) = U^{(2)}(\mathbf{r}) + U^{(3)}(\mathbf{r}) + O(\epsilon^3), \\ &U^{(2)}(\mathbf{r}) := \sum_{i,j} W^{(2)}(r_i, r_j), \quad U^{(3)}(\mathbf{r}) := \sum_{i,j,k} W^{(3)}(r_i, r_j, r_k), \quad \text{etc.}, \end{aligned} \tag{12}$$

together with the corresponding error estimates. Note that a precise choice of the error quantity ϵ, depending on the density and the temperature, will be given in Section 3.

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 [13,27,28,58]. 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 performing molecular dynamics with such forces that maintain the fixed distance $r_{1,2}$. In this way, we sample atomistic potential energy (and forces) over the constrained phase space and obtain the conditional partition function as the integral in (8). Alternatively, by integration of the constrained force ($-\int_{r_{min}}^{r_{cut}} \langle f \rangle_{r_{1,2}=R} dR$), the two-body effective potential can be obtained. These are the $W^{(2),\text{full},U}$ and $W^{(2),\text{full},F}$ terms, respectively. Note that we have not used any kind of fitting or projection over a basis as in [64,65]; the data are in tabulated form.

(b) Upon inverting $g(r)$ in (11) for two isolated molecules, the two-body effective potential can be directly obtained, since for such a system, $\gamma(r) = 1$; i.e., the low ρ regime. This method is only used for comparison with (a)as it uses the $g(r)$.

Here, we examine both of the above methods; see Section 5.1. 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 (11). 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 [55]; see also the discussion in Section 7.

2.3. Thermodynamic Consistency

As already mentioned, several coarse-graining strategies lack thermodynamic consistency; see also the discussion by Louis [46] and Guenza [40], where thermodynamically-consistent theories have been developed based on the standard closures of the liquid state theory. However, as the error in these closures has not been rigorously quantified, we first investigate here a much simpler case;

namely, what is the contribution of the three-body terms in approximating thermodynamic quantities in the regimes where the cluster expansion can be valid. Hence, defining the coarse-graining as in (12), we demonstrate below that these approximations directly imply similar approximations for thermodynamic quantities, such as the free energy and the pressure, and we also comment on the isothermal compressibility. This is a direct consequence of the validity of the cluster expansions. Thus, from (12), by considering only the two-body contribution, for the finite volume free energy $f_{\beta,\Lambda}(N)$, we have that:

$$f_{\beta,\Lambda}(N) = -\frac{1}{\beta|\Lambda|}\log Z_{\beta,\Lambda,N} = \int d\mathbf{r} e^{-\beta U_{\text{eff}}(\mathbf{r})} = -\frac{1}{\beta|\Lambda|}\log \int d\mathbf{r} e^{-\beta U^{(2)}(r)} + O(\epsilon^2), \tag{13}$$

where the error is uniform in N and $|\Lambda|$. Thus, the approximation $U^{(2)}$ of the CG Hamiltonian implies a good approximation of the free energy. By adding the next order contribution of the three-body terms $U^{(3)}$, we can further improve the error in computing the free energy and obtain that:

$$-\frac{1}{\beta|\Lambda|}\log Z_{\beta,\Lambda,N} = -\frac{1}{\beta|\Lambda|}\log \int d\mathbf{r} e^{-\beta(U^{(2)}(r)+U^{(3)}(r))} + O(\epsilon^3). \tag{14}$$

Note once again that these estimates are valid only in the low density regime where these expansions are justified.

Similarly, in order to construct approximations for the thermodynamic pressure (at finite volume) $p_{\beta,\Lambda}(z)$ as a function of the activity z, we work in the grand-canonical ensemble and obtain:

$$p_{\beta,\Lambda}(z) = \frac{1}{\beta|\Lambda|}\log \sum_{N\geq 0} z^N Z_{\beta,\Lambda,N} = \frac{1}{\beta|\Lambda|}\log \sum_{N\geq 0} z^N \int d\mathbf{r} e^{-\beta U^{(2)}(r)} + O(\epsilon^2). \tag{15}$$

Both quantities have limits given by absolutely convergent series with respect to $\rho = N/|\Lambda|$ for the first and z or ρ for the second.

In the same spirit, good approximations to the atomistic (full) partition function can give controlled approximations for other thermodynamic quantities, as well. For example, the isothermal compressibility can be related to the variance in the number of particles, which, in turn, can be expressed in terms of the derivative of the density with respect to the chemical potential (in the grand canonical ensemble). Since the density is again expressed in terms of the derivative of the logarithm of the partition function, our suggested method is also applicable here. All further manipulations needed to arrive at an expansion for the isothermal compressibility; namely, taking derivatives, as well as inverting series, can be made rigorous in the framework of the cluster expansion, but a full analysis of this and similar cases of other thermodynamic quantities is beyond the scope of the present work, so we just refer to [45] for an analysis in this direction.

3. Cluster Expansion

The cluster expansion method originates from the work of Mayer and collaborators (see [47] 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 conditional partition function and consequently the desired many-body PMF.

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 (e.g., centre of mass mapping). 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, (8) can be written as:

$$U_{\text{eff}}(r_1, \ldots, r_M) := -\frac{1}{\beta}\log \prod_{i=1}^{M} \lambda^i(\{T^i\mathbf{q}^i = r_i\}) - \frac{1}{\beta}\log \int \prod_{i=1}^{M} \mu(d\mathbf{q}^i; r_i) e^{-\beta U(\mathbf{q})}, \tag{16}$$

where, for simplicity, we have introduced the normalized conditional measure, related to the specific CG map:

$$\mu(d\mathbf{q}^i; r_i) := \frac{1}{l_i!} dq_{k_1^i} \dots dq_{k_{l_i}^i} \frac{\mathbf{1}_{\{T^i\mathbf{q}^i = r_i\}}}{\lambda^i(\{T^i\mathbf{q}^i = r_i\})}, \tag{17}$$

and by λ^i, we denote the measure $\frac{1}{l_i!} dq_{k_1^i} \dots dq_{k_{l_i}^i}$.

To perform a cluster expansion in the second term of (16), we rewrite the interaction potential as follows:

$$\begin{aligned} U(\mathbf{q}) &= \sum_{i<j} \bar{V}(\mathbf{q}^i, \mathbf{q}^j), \quad \text{where} \\ \bar{V}(\mathbf{q}^i, \mathbf{q}^j) &:= \sum_{m=1}^{l_i} \sum_{m'=1}^{l_j} V(|q_{k_m^i} - q_{k_{m'}^j}|). \end{aligned} \tag{18}$$

Then, letting $f_{i,j}(\mathbf{q}^i, \mathbf{q}^j) := e^{-\beta \bar{V}(\mathbf{q}^i, \mathbf{q}^j)} - 1$, we have:

$$e^{-\beta U(\mathbf{q})} = \prod_{i<j} \left(1 + e^{-\beta \bar{V}(\mathbf{q}^i, \mathbf{q}^j)} - 1\right) = \sum_{\substack{V_1, \dots, V_m \\ |V_i| \geq 2, V_i \subset \{1, \dots, N\}}} \prod_{l=1}^{m} \sum_{g \in \mathcal{C}_{V_l}} \prod_{\{i,j\} \in E(g)} f_{i,j}(\mathbf{q}^i, \mathbf{q}^j), \tag{19}$$

where for $V \subset \{1, \dots, N\}$, 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 μ in (17) is a normalized measure, from (16), we obtain:

$$\begin{aligned} U_{\text{eff}}(r_1, \dots, r_M) &= -\frac{1}{\beta} \log \prod_{i=1}^{M} \lambda^i(\{T^i\mathbf{q}^i = r_i\}) - \frac{1}{\beta} \log \sum_{\substack{V_1, \dots, V_m \\ |V_i| \geq 2, V_i \subset \{1, \dots, N\}}} \prod_{l=1}^{m} \zeta(V_i) \\ &= -\frac{1}{\beta} \log \prod_{i=1}^{M} \lambda^i(\{T^i\mathbf{q}^i = r_i\}) - \frac{1}{\beta} \sum_{V \subset \{1, \dots, N\}} \zeta(V) + \frac{1}{\beta} \sum_{\substack{V, V': \\ V \cap V' = \varnothing}} \zeta(V)\zeta(V') + \dots, \end{aligned} \tag{20}$$

where:

$$\zeta(V) := \int \sum_{g \in \mathcal{C}_V} \prod_{\{i,j\} \in E(g)} f_{i,j}(\mathbf{q}^i, \mathbf{q}^j) \prod_{i \in V} \mu(d\mathbf{q}^i; r_i) \tag{21}$$

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 the values of the density $\rho = \frac{N}{|\Lambda|}$ and of the inverse temperature β such that $\rho \bar{C}_\beta(r) < c_0$, where $\bar{C}_\beta(r) := \int |f_{i,j}(\mathbf{q}^i, \mathbf{q}^j)| \mu(d\mathbf{q}^i; 0) \mu(d\mathbf{q}^i; r)$ and c_0 is a known small positive constant [51]. Here, we do not give the full proof, which can be easily obtained by a slight modification of the one in [51], but in order to motivate the error estimates in (12), we note that the sum over triplets will give a contribution of the order $\binom{N}{3}$#{trees on 3 labels} $\frac{1}{|\Lambda|^3} \bar{C}_\beta(r)^2 \sim \rho(\rho \bar{C}_\beta(r))^2$, i.e., here, $\epsilon \sim \rho \bar{C}_\beta(r)$. If we simplify the sum in (20), one can obtain [51] Expansion (12), where:

$$W^{(2)}(r_1, r_2) := -\frac{1}{\beta} \int \mu(d\mathbf{q}^1; r_1)\, \mu(d\mathbf{q}^2; r_2)\, f_{1,2}(\mathbf{q}^1, \mathbf{q}^2) \tag{22}$$

and:

$$W^{(3)}(r_1, r_2, r_3) := -\frac{1}{\beta} \int \mu(d\mathbf{q}^1; r_1)\, \mu(d\mathbf{q}^2; r_2)\, \mu(d\mathbf{q}^3; r_3)\, f_{1,2}(\mathbf{q}^1, \mathbf{q}^2)\, f_{2,3}(\mathbf{q}^2, \mathbf{q}^3)\, f_{3,1}(\mathbf{q}^3, \mathbf{q}^1). \tag{23}$$

Recall also the definition of $f_{i,j}$ in (19).

3.1. Full Calculation of the PMF

Notice that the potentials $W^{(2)}$ and $W^{(3)}$ in (22) and (23), 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:

$$W^{(2),\text{full}}(r_1, r_2) := -\frac{1}{\beta} \log \int \mu(d\mathbf{q}^1; r_1)\, \mu(d\mathbf{q}^2; r_2)\, e^{-\beta \bar{V}(\mathbf{q}^1, \mathbf{q}^2)}. \tag{24}$$

By adding and subtracting one and expanding the logarithm, we can relate it to (22):

$$\begin{aligned} -\beta W^{(2),\text{full}}(r_1, r_2) &= \log \textstyle\int \mu(d\mathbf{q}^1; r_1)\mu(d\mathbf{q}^2; r_2) e^{-\beta \bar{V}(\mathbf{q}^1, \mathbf{q}^2)} \\ &= \log(1 + \textstyle\int \mu(d\mathbf{q}^1; r_1)\mu(d\mathbf{q}^2; r_2) f_{1,2}(\mathbf{q}^1, \mathbf{q}^2)) \\ &= \textstyle\int \mu(d\mathbf{q}^1; r_1)\, \mu(d\mathbf{q}^2; r_2) f_{1,2}(\mathbf{q}^1, \mathbf{q}^2) - \frac{1}{2} \left(\int \mu(d\mathbf{q}^1; r_1)\, \mu(d\mathbf{q}^2; r_2) f_{1,2}(\mathbf{q}^1, \mathbf{q}^2) \right)^2 + \dots \end{aligned} \tag{25}$$

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:

$$W^{(3),\text{full}}(r_1, r_2, r_3) := -\frac{1}{\beta} \log \int \mu(d\mathbf{q}^1; r_1)\, \mu(d\mathbf{q}^2; r_2) \mu(d\mathbf{q}^3; r_3)\, e^{-\beta \sum_{1 \le i < j \le 3} \bar{V}(\mathbf{q}^i, \mathbf{q}^j)}. \tag{26}$$

By adding and subtracting one, we can relate it to (22) and (23) (in the following, we simplify notation by not explicitly showing the dependence on the atomistic configuration and neglecting the normalized conditional measure):

$$\begin{aligned} e^{-\beta W^{(3),\text{full}}} &= \int e^{-(V_{12} + V_{13} + V_{23})} \\ &= 1 + \int f_{12} + \int f_{13} + \int f_{23} + \int f_{12} f_{13} + \int f_{13} f_{23} + \int f_{12} f_{23} + \int f_{12} f_{23} f_{13}, \end{aligned} \tag{27}$$

which implies that:

$$\begin{aligned} W^{(3),\text{full}} = -\frac{1}{\beta} \Bigg(& \int f_{12} + \int f_{13} + \int f_{23} + \int f_{12} f_{23} f_{13} + \int f_{12} f_{13} + \int f_{13} f_{23} + \int f_{12} f_{23} - \\ & \left[\int f_{12} \int f_{13} + \int f_{13} \int f_{23} + \int f_{12} \int f_{23} \right] \Bigg) + \dots \end{aligned} \tag{28}$$

In principle, we can rewrite (12) 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 11.

Note finally that in the proposed approach, the hierarchy of CG effective potential (two-body, three-body, etc.) terms is not fitted or included via some uncontrolled closure; they can be computed, through a conditional sampling of a few (two, three, etc.) CG particles. Furthermore, our work distinguishes from those using the standard liquid state theory also by the fact that we do not use any simplified models (PRISM, thread model, etc.), but integrate over the atomistic degrees of freedom.

4. Model and Simulations

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 methane and ethane in various density and temperature

regimes. We refer the reader to the Supplementary Material for the atomistic models, as well as their coarse-graining through the cluster expansion formalism. In order to clarify the notation, we refer to Figure 1; for the three-body clustering case, each set is composed by exactly three interacting CG molecules (dots connected with solid lines). The centres of mass (COM) of the set i is located at r_i and associated with $\mu(d\mathbf{q}^{\mathbf{i}}, r_i)$.

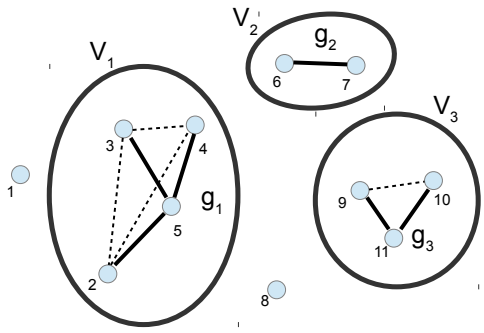

Figure 1. Visualization of the partition in (19) for non-intersecting sets $V_1 = \{2,3,4,5\}$, $V_2 = \{6,7\}$, $V_3 = \{9,10,11\}$ in each of which we display by solid lines the connected graphs $g_i \in \mathcal{C}_{V_i}$, $i = 1,2,3$.

4.2. Simulations

The simplest system to simulate is the one with only two interacting methane, or ethane, molecules in a 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 systems in various densities, as well as in the liquid regime. The atomistic and CG model 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 that is coupled with a thermostat [66]. The thermostat serves as a reservoir of energy. The densities of both the liquid methane and ethane systems were chosen as the average values of NPT runs at atmospheric pressure. NVT equilibration and production runs of a few *ns* followed, and the sizes 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 ± 40 atm in the liquid phase, whereas temperature fluctuations have small variance, and the system is driven to the target temperature much faster than with conventional MD. Long production atomistic and CG simulations have been performed, from 10 ns up to 100 ns, depending on the system. The error bars (computed via the standard deviation) in all reported data are about 1–2% of the actual values. Details about the simulation parameters are shown in Table 1.

Table 1. Simulation parameters for CH_4.

System	N (Molecules)	T/K	Simulation Time/ns
CH_4, CH_3–CH_3	2, 3	100–900	10–20
CH_4	512	80, 100, 120, 300, 900	100
CH_3–CH_3	500	150, 300, 650	100

In order to compute the hierarchy of CG effective non-bonded potential terms, discussed above, different simulation runs have been performed, which are discussed below.

4.2.1. Constrained Runs

The first method that 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 (9). We call it the "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 notion is similar to the free energy calculation methods in which CG degrees of freedom are constrained at fixed distances. 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 OM. 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 $f = -\nabla W$, as explained in Section 2.2. Both $W^{(2),\text{full, F}}(r)$ and $W^{(2),\text{full, U}}(r)$ are based on the same trajectory.

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 COMs. 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.

4.2.2. Geometric Direct Computation of PMF

In order to further accelerate the sampling and alleviate the noise problems at high energy regions, which 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 $W^{(2),geom}(r)$ given in (24) is obtained by rotating the two (methane or ethane) molecules around their COMs, through their Eulerian angles and taking account of all of 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. However, for this system, there is no considerable difference in the resulting non-bonded effective potential. 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 [58].

All atomistic and coarse-grained simulations have been performed using a home-made simulation package that has been extensively used in the past [62,67], whereas all analysis has been executed through home-made codes in MATLAB and Python. In addition, long simulation runs have been performed resulting in error bars that are of the order of the width of the lines in all figures presented

below. Characteristic snapshots from two-body, three-body and bulk simulations are shown in Figure 2.

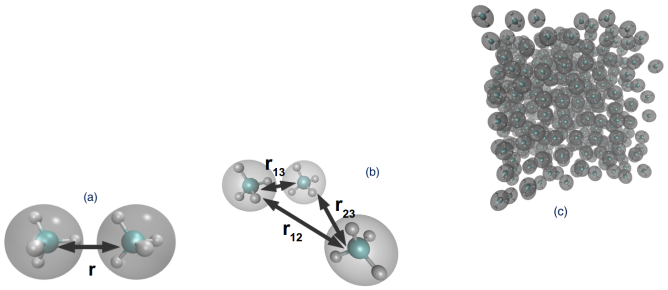

Figure 2. Snapshot of model systems in atomistic and coarse-grained description. (**a**,**b**) Two and three methanes used for the estimation of the coarse-graining (CG) effective potential from isolated molecules; (**c**) bulk methane liquid.

5. Results

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 two-body one. In Figure 3a,b, 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 4.2.1. Alternatively, through the same set of atomistic configurations, the two-body PMF, $W^{(2),\text{full, u}}$, can be directly calculated through Equation (24).
(b) A direct calculation of the PMF, $W^{(2),geom}$, using a geometrical approach as described in Section 4.2.2.
(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 3a.

The first two of the above methods refer to the direct calculation of the constrained partition function (8) with constrained forces and canonical sampling, while the third uses the "direct Boltzmann inversion" approach. All of the 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 3).

First, for the case of the two methane molecules (Figure 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 (24) are difficult to cancel out. The small probability configurations in high potential energy regimes have a large impact on 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_{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 [62].

Second, for the case of the two ethane molecules (Figure 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 the ethane compared to the 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 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 (22). In Figure 4a,b, we demonstrate the PMF through cluster expansions and the effect of higher order terms as shown in Equation (25), of the two isolated molecules, for CH_4 and CH_3–CH_3, respectively. As discussed in Section 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 3; values of $-k_BT$ are shown with full lines.

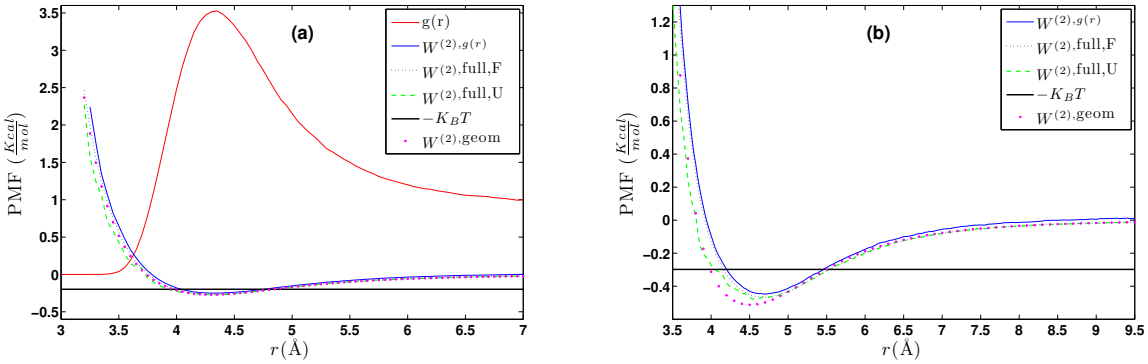

Figure 3. Representation of the two-body PMF, for two isolated molecules, as a function of distance r, through different approximations: geometric averaging, (constrained) force matching and inversion of $g(r)$. (**a**) CH_4 at T = 100 K; (**b**) CH_3–CH_3 at T = 150 K. For the methane, the corresponding $g(r)$ curve is also shown.

Both systems show the same behaviour. First, it is clear that the agreement between $W^{(2)}$ and the (more accurate) $W^{(2),\mathrm{full}}$ is very good only at long distances (not surprisingly, since the logarithm expansion holds for every β, as $\bar{V} \to 0$ in $C(\beta)$), 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 β, we obtain a better approximation of $W^{(2),\mathrm{full}}$. All of the above suggest that geometric averaging is the most accurate and computationally less expensive.

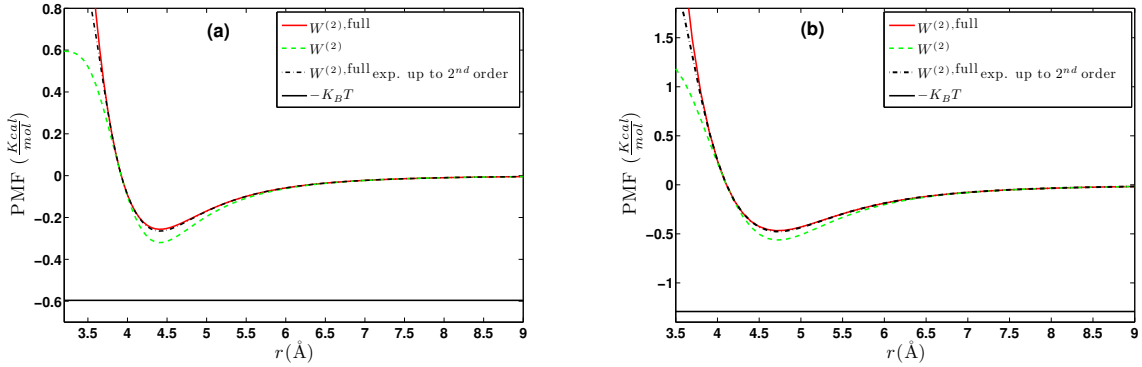

Figure 4. Relation of the PMF through cluster expansions and energy averaging at high temperatures, i.e., $W^{(2)}(r_1, r_2)$ and $W^{(2),\text{full}}(r_1, r_2)$, through expansion over β for (**a**) CH_4 at T = 300 K; (**b**) CH_3–CH_3 at T = 650 K. As expected from the analytic form and the relation between the two formulas, $W^{(2)}$ and $W^{(2),\text{full}}$ tend to converge to the same effective potential.

5.1.1. 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 = 80 K , 100 K, 120 K, 300 K and 900 K. In more detail, in Figure 5a,b, we compare the difference between $W^{(2)}$ and $W^{(2),\text{full}}$ at different temperatures. As discussed in Section 3, the cluster expansion method is valid only in the high temperature regime. This is directly observed in Figure 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 4.

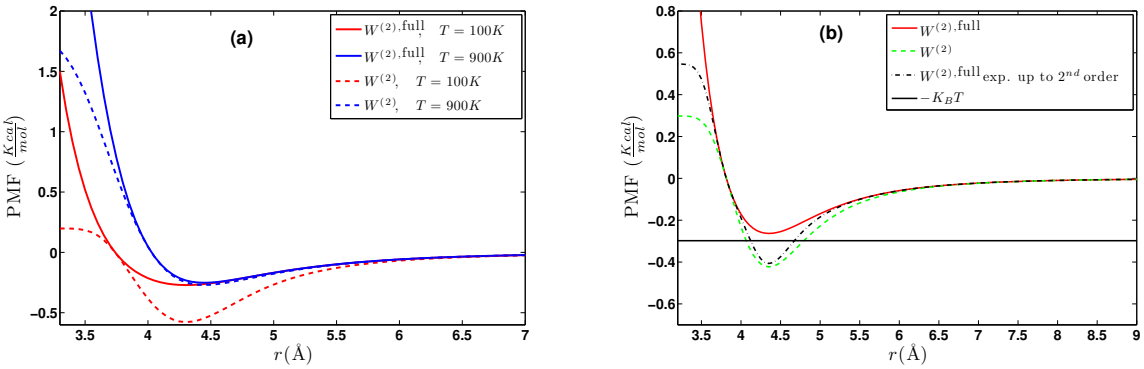

Figure 5. (**a**) PMF through cluster expansions, using (22) and (25) for different temperatures for the CH_4 model; (**b**) PMF through cluster expansions and energy averaging, i.e., $W^{(2)}(r_1, r_2)$ and $W^{(2),\text{full}}(r_1, r_2)$ through expansion over β for CH_4 at T = 150 K. The expansion is not valid at this temperature.

On the contrary, at low temperatures, there is a strong discrepancy around the potential well, as shown in Figure 5b. In fact, for values of r close to the potential well and for rather high values of β, the contribution to the integral $C(\beta)$ is large, and the latter can exceed one, rendering the expansion in (25) not valid. In Figure 5b, we see that the term (22) is not small, so the expansion (25) 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 6. In principle, Equation (22) 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.

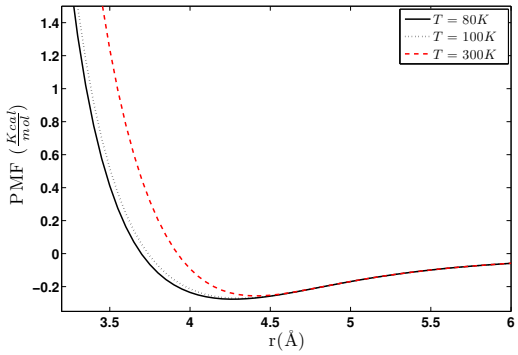

Figure 6. Potential of mean force at different temperatures (geometric averaging). Two CH_4 molecules at T = 80 K, 100 K, 300 K.

5.2. Bulk CG CH_4 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 Figure 7a,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 3 of methane and ethane, respectively. Note that $g(r)$ is directly related to the effective CG potentials ($N = 2$ in Equation (6)).

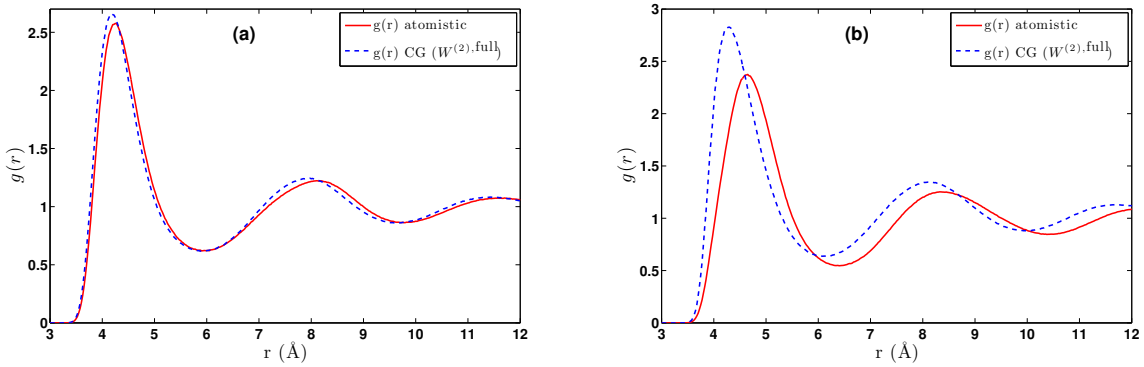

Figure 7. RDF from atomistic and CG using pair potential, $W^{(2)}$, for (**a**) CH_4 at T = 80 K and (**b**) CH_3–CH_3 at T = 150 K. Spherical CG approximation to the non-symmetric ethane molecule induces discrepancy and implies that there is more room for improvement.

It is clear that for methane (Figure 7a), the CG model with the $W^{(2),\text{full}}$ 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 4a and 8b). As expected, the difference comes from the missing higher order terms of Equation (12).

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

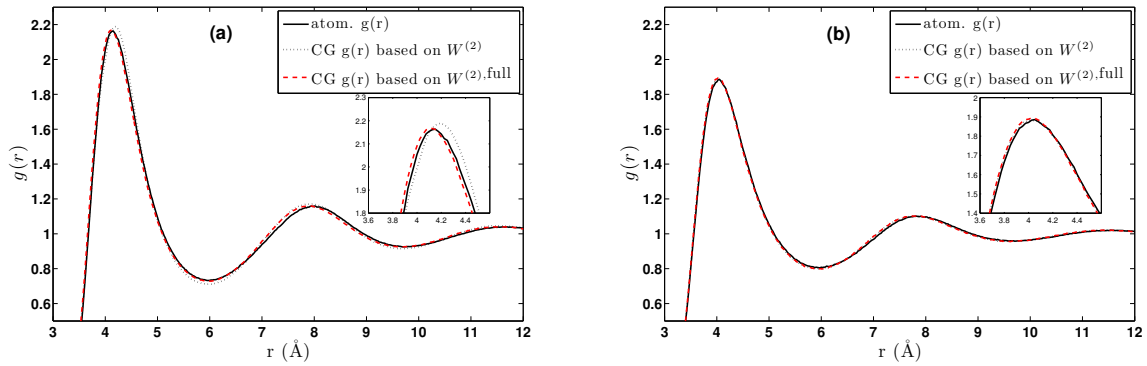

Figure 8. RDF of methane from atomistic data and CG models using pair potential at different temperatures: (**a**) T = 300 K; (**b**) T = 900 K. In both cases, the density is $\rho_1 = 0.3799 \frac{\text{g}}{\text{cm}^3}$.

5.2.1. Effect of Temperature-Density

We further study the structural behaviour 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 7) at higher temperatures; however, keeping the same density. In Figure 8a,b, we present the RDF of methane from atomistic and CG runs using pair potential at T = 300 K and T = 900 K 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 $T = 80$ K 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 8b), whereas there are small discrepancies at lower temperatures (Figure 8a), in particular at the maximum of $g(r)$. This is shown in the inset of Figure 8a,b. Note also that at 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 4a) and consequently to very accurate structural $g(r)$ data, as well.

Next, we examine the structural behaviour of the CG systems at different densities. In Figure 9a, we present the $g(r)$ from atomistic and CG runs using pair potential at different densities (ρ_1 = 0.3799 $\frac{\text{g}}{\text{cm}^3}$ and ρ_2 = 0.0395 $\frac{\text{g}}{\text{cm}^3}$ and T = 300 K and T = 900 K). There is apparent discrepancy from the reference (atomistic) system in both densities in agreement with the data discussed above in Figure 8a.

For the case of higher temperature data (T = 900 K) and the same densities, as shown in Figure 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 with 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 ρ.

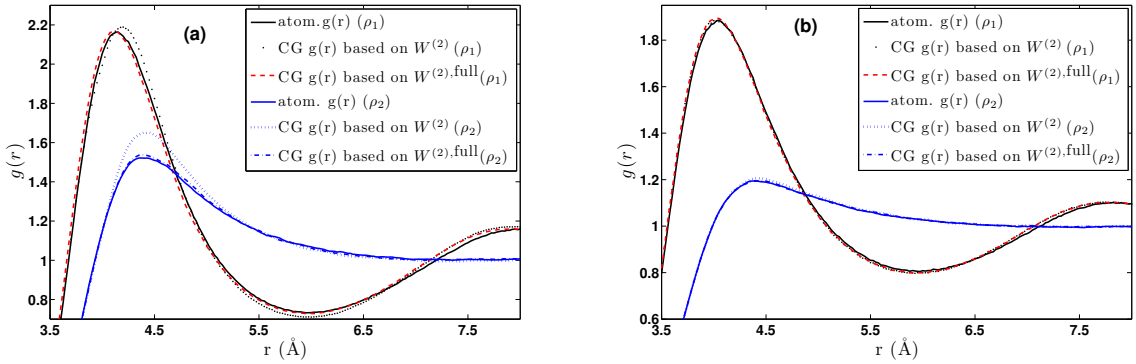

Figure 9. RDF of methane from atomistic and CG using pair potential at different densities $\rho_1 > \rho_2$. (**a**) $T = 300K$; (**b**) T = 900 K. For this model, the pair approximation is sufficient, and in low ρ, high T conditions, $W^{(2)}$ converges to the reference $g(r)$.

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.

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 [59].

6.1. Calculation of the Effective Three-Body Potential

In the following, we present data for the three-body potential of mean force estimated from simulation runs and geometric computations involving three isolated molecules. We have two suggestions for the three-body PMF: (a) Formula (23) derived from the cluster expansion formalism, which is valid for rather high temperatures; (b) another one based on the McCoy–Curro scheme given in Formula (26). 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 Formulas (23), and (26) will be given elsewhere [59].

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 wash-out 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 COMs in space is on top of the angular discretization mentioned in Section 4.2.2 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 Figure 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 = 80 K for different COM distances (): (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 10, improvement is needed for close distances around the (three-dimensional) well. We used a three-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.

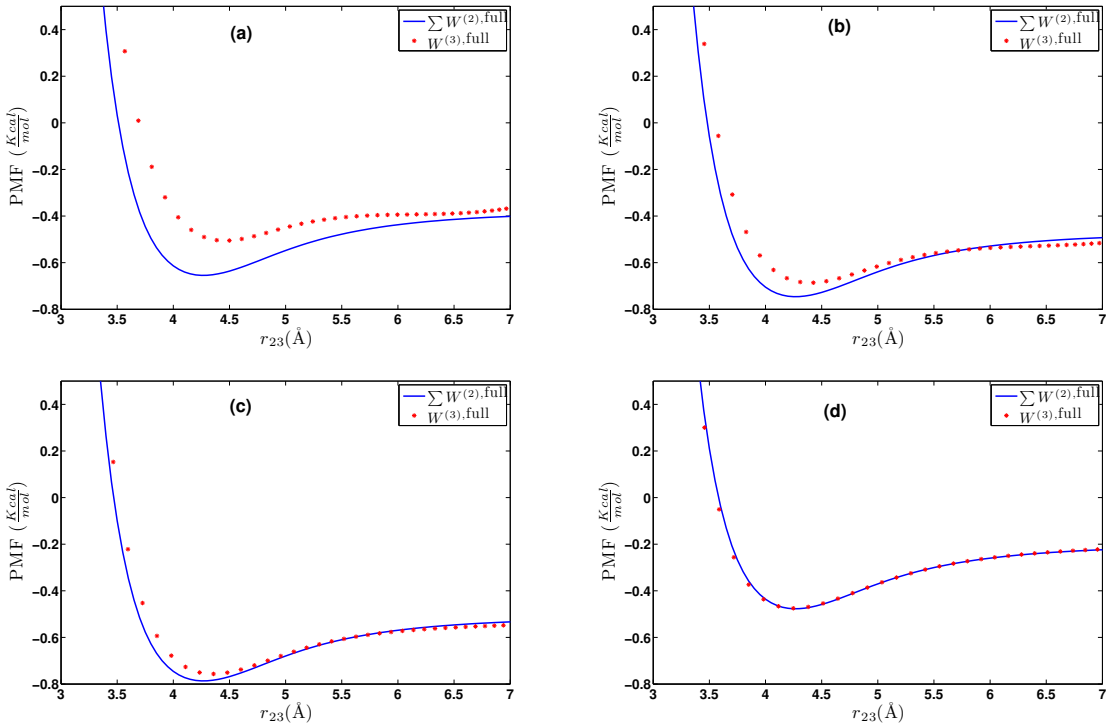

Figure 10. Effective potential comparison between the $W^{(3),\text{full}}$ three-body and $\sum W^{(2),\text{full}}$ simulations (geometric averaging) for CH_4 at T = 80 K for different fixed centre of mass (COM) distances. (**a**) $r_{12} = 3.9, r_{13} = 3.9$; (**b**) $r_{12} = 4.0, r_{13} = 4.0$; (**c**) $r_{12} = 4.3, r_{13} = 4.0$; (**d**) $r_{12} = 3.8, r_{13} = 5.64$.

Overall, there are clear differences between the three-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 three-body terms for this system (methane at T = 80 K) is: $r_{12} \in [3.8 : 4.1], r_{13} \in [3.8 : 4.1]$ and $r_{23} \in [3.8 : 5]$; hence, the maximum distance for which three-body terms were considered is $r_{\text{cut-off},3}$ = 5. 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 [59]. 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.

6.2. CG Runs with the Effective Three-Body Potential

Next, we examine the effect of the three-body term on the CG model by performing bulk CG stochastic dynamics simulations using the new CG model with the three-body terms described above.

In this case, we incorporate the two-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 datasets shown in Figure 10a–d.

Results on the pair distribution function, $g(r)$, for bulk (liquid) methane at T = 80 K are shown in Figure 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 three-body terms are shown. First, it is clear that $g(r)$ data derived from the CG model that involves only pair CG interactions deviate, 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.

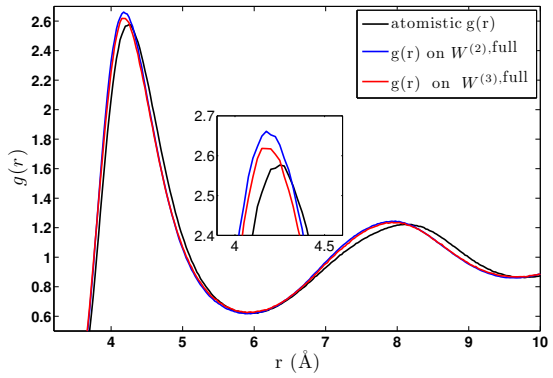

Figure 11. RDF from atomistic and CG using pair, $W^{(2),\text{full}}$, and three-body, $W^{(3),\text{full}}$, potential for CH_4 (T = 80 K). The three-dimensional cubic polynomial was used for the fitting.

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 the 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 at 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" [68] and find an interaction to which it corresponds. However, 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 [51]; see also [45,61] for the corresponding (in the canonical ensemble) expansions for the correlation functions and the Ornstein–Zernike equation. See also the references therein for a more comprehensive list of the literature on the diagrammatic expansions. 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 the need for further developing these methods and relating them to computational strategies. Hence, in this paper, working in the rigorous framework of the cluster expansion, we seek quantifiable coarse-grained approximations, which we compute using molecular simulations. 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 centre 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. Note, that for this estimation, no information from long simulations of n-body (bulk) systems is required.
(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, which also alleviates sampling noise issues. No basis function is needed in any of these methods.
(c) The obtained pair CG potentials were used to model methane and ethane systems in various regimes. 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 seem 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 in the liquid state.

Finally, note that the proposed approach becomes computationally efficient, since the hierarchy of the effective non-boned CG potentials terms is developed by the conditional sampling of a few CG particles, and no information from long n-body (bulk) simulations is required. However, for dense systems and in the liquid regime, numerical estimation of higher, than the three-body, order terms is required; this is a major computational challenge.

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 ρ 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 compute and more difficult to treat. Furthermore, expansions in the correlation functions can also be considered, as was already suggested in one of the original works of Hiroike and Morita [48]. Finally, another future goal is to extend this investigation to 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.

Supplementary Materials: The following are available online at www.mdpi.com/1099-4300/19/8/395/s1: More details about (a) the model; (b) the two-body and (c) the three-body simulation algorithms.

Acknowledgments: We acknowledge support by the European Union (ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the NSRF-Research Funding Programs: THALES and ARISTEIA II.

Author Contributions: Dimitrios Tsagkarogiannis and Vagelis Harmandaris conceived and designed the research work; Anastasios Tsourtis performed the simulations; All authors contributed equally in writing the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Frenkel, D.; Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications*, 2nd ed.; Academic Press: New York, NY, USA, 2002.
2. Allen, M.P.; Tildesley, D.J. *Computer Simulation of Liquids*; Oxford University Press: Oxford, UK, 1987.
3. Harmandaris, V.A.; Mavrantzas, V.G.; Theodorou, D.; Kröger, M.; Ramírez, J.; Öttinger, H.C.; Vlassopoulos, D. Dynamic crossover from Rouse to entangled polymer melt regime: Signals from long, detailed atomistic molecular dynamics simulations, supported by rheological experiments. *Macromolecules* **2003**, *36*, 1376–1387.
4. Kotelyanskii, M.; Theodorou, D.N. *Simulation Methods for Polymers*; Taylor & Francis: Abingdon, UK, 2004.
5. Izvekov, S.; Voth, G.A. Multiscale coarse-graining of liquid-state systems. *J. Chem. Phys.* **2005**, *123*, 134105.
6. Tschöp, W.; Kremer, K.; Hahn, O.; Batoulis, J.; Bürger, T. Simulation of polymer melts. I. Coarse-graining procedure for polycarbonates. *Acta Polym.* **1998**, *49*, 61–74.
7. Müller-Plathe, F. Coarse-Graining in Polymer Simulation: From the Atomistic to the Mesoscopic Scale and Back. *ChemPhysChem* **2002**, *3*, 754–769.
8. Shell, M.S. The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *J. Chem. Phys.* **2008**, *129*, 144108.
9. Briels, W.J.; Akkermans, R.L.C. Coarse-grained interactions in polymer melts: A variational approach. *J. Chem. Phys.* **2001**, *115*, 6210.
10. Harmandaris, V.A.; Adhikari, N.P.; van der Vegt, N.F.A.; Kremer, K. Hierarchical Modeling of Polystyrene: From Atomistic to Coarse-Grained Simulations. *Macromolecules* **2006**, *39*, 6708.
11. Harmandaris, V.A.; Kremer, K. Dynamics of Polystyrene Melts through Hierarchical Multiscale Simulations. *Macromolecules* **2009**, *42*, 791.
12. Harmandaris, V.A.; Kremer, K. Predicting polymer dynamics at multiple length and time scales. *Soft Matter* **2009**, *5*, 3920.
13. Johnston, K.; Harmandaris, V. Hierarchical simulations of hybrid polymer–solid materials. *Soft Matter* **2013**, *9*, 6696–6710.
14. Noid, W.G.; Chu, J.W.; Ayton, G.S.; Krishna, V.; Izvekov, S.; Voth, G.A.; Das, A.; Andersen, H.C. The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.* **2008**, *128*, 4114.
15. Lu, L.; Izvekov, S.; Das, A.; Andersen, H.C.; Voth, G.A. Efficient, Regularized, and Scalable Algorithms for Multiscale Coarse-Graining. *J. Chem. Theor. Comput.* **2010**, *6*, 954–965.
16. Rudzinski, J.F.; Noid, W.G. Coarse-graining entropy, forces, and structures. *J. Chem. Phys.* **2011**, *135*, 214101.
17. Noid, W.G. Perspective: Coarse-grained models for biomolecular systems. *J. Chem. Phys.* **2013**, *139*, 090901.
18. Chaimovich, A.; Shell, M.S. Anomalous waterlike behaviour in spherically-symmetric water models optimized with the relative entropy. *Phys. Chem. Chem. Phys.* **2009**, *11*, 1901–1915.
19. Bilionis, I.; Zabaras, N. A stochastic optimization approach to coarse-graining using a relative-entropy framework. *J. Chem. Phys.* **2013**, *138*, 044313.
20. Coifman, R.R.; Kevrekidis, I.G.; Lafon, S.; Maggioni, M.; Nadler, B. Diffusion Maps, Reduction Coordinates, and Low Dimensional Representation of Stochastic Systems. *Multiscale Model. Simul.* **2008**, *7*, 842–864.
21. Soper, A.K. Empirical potential Monte Carlo simulation of fluid structure. *Chem. Phys.* **1996**, *202*, 295–306.
22. Lyubartsev, A.P.; Laaksonen, A. On the Reduction of Molecular Degrees of Freedom in Computer Simulations. In *Novel Methods in Soft Matter Simulations*; Karttunen, M., Lukkarinen, A., Vattulainen, I., Eds.; Springer: Berlin, Germany, 2004; Volume 640, pp. 219–244.
23. Harmandaris, V.A. Quantitative study of equilibrium and non-equilibrium polymer dynamics through systematic hierarchical coarse-graining simulations. *Korea Aust. Rheol. J.* **2014**, *26*, 15–28.
24. Espanol, P.; Zuniga, I. Obtaining fully dynamic coarse-grained models from MD. *Phys. Chem. Chem. Phys.* **2011**, *13*, 10538–10545.
25. Padding, J.T.; Briels, W.J. Uncrossability constraints in mesoscopic polymer melt simulations: Non-Rouse behaviour of $C_{120}H_{242}$. *J. Chem. Phys.* **2001**, *115*, 2846–2859.
26. Deichmann, G.; Marcon, V.; van der Vegt, N.F.A. Bottom-up derivation of conservative and dissipative interactions for coarse-grained molecular liquids with the conditional reversible work method. *J. Chem. Phys.* **2014**, *141*, 224109.

27. Fritz, D.; Harmandaris, V.; Kremer, K.; van der Vegt, N. Coarse-grained polymer melts based on isolated atomistic chains: Simulation of polystyrene of different tacticities. *Macromolecules* **2009**, *42*, 7579–7588.
28. Brini, E.; Algaer, E.A.; Ganguly, P.; Li, C.; Rodriguez-Ropero, F.; van der Vegt, N.F.A. Systematic Coarse-Graining Methods for Soft Matter Simulations—A Review. *Soft Matter* **2013**, *9*, 2108–2119.
29. Reith, D.; Pütz, M.; Müller-Plathe, F. Deriving effective mesoscale potentials from atomistic simulations. *J. Comput. Chem.* **2003**, *24*, 1624–1636.
30. Lyubartsev, A.P.; Laaksonen, A. Calculation of effective interaction potentials from radial distribution functions: A reverse Monte Carlo approach. *Phys. Rev. E* **1995**, *52*, 3730–3737.
31. Izvekov, S.; Voth, G.A. Effective force field for liquid hydrogen fluoride from ab initio molecular dynamics simulation using the force-matching method. *J. Phys. Chem. B* **2005**, *109*, 6573–6586.
32. Noid, W.G.; Liu, P.; Wang, Y.; Chu, J.; Ayton, G.S.; Izvekov, S.; Andersen, H.C.; Voth, G.A. The multiscale coarse-graining method. II. Numerical implementation for coarse-grained molecular models. *J. Chem. Phys.* **2008**, *128*, 244115.
33. Lu, L.; Dama, J.F.; Voth, G.A. Fitting coarse-grained distribution functions through an iterative force-matching method. *J. Chem. Phys.* **2013**, *139*, 121906.
34. Katsoulakis, M.A.; Plechac, P. Information-theoretic tools for parametrized coarse-graining of non-equilibrium extended systems. *J. Chem. Phys.* **2013**, *139*, 4852–4863.
35. Chaimovich, A.; Shell, M.S. Coarse-graining errors and numerical optimization using a relative entropy framework. *J. Chem. Phys.* **2011**, *134*, 094112.
36. Cho, H.M.; Chu, J.W. Inversion of radial distribution functions to pair forces by solving the Yvon–Born–Green equation iteratively. *J. Chem. Phys.* **2009**, *131*, 134107.
37. Noid, W.G.; Chu, J.; Ayton, G.S.; Voth, G.A. Multiscale coarse-graining and structural correlations: Connections to liquid-state theory. *J. Phys. Chem. B* **2007**, *111*, 4116–4127.
38. Mullinax, J.W.; Noid, W.G. Generalized Yvon–Born–Green theory for molecular systems. *Phys. Rev. Lett.* **2009**, *103*, 198104.
39. Mullinax, J.W.; Noid, W.G. Generalized Yvon–Born–Green theory for determining coarse-grained interaction potentials. *J. Phys. Chem. C* **2010**, *114*, 5661–5674.
40. McCarty, J.; Clark, A.J.; Copperman, J.; Guenza, M.G. An analytical coarse-graining method which preserves the free energy, structural correlations, and thermodynamic state of polymer melts from the atomistic to the mesoscale. *J. Chem. Phys.* **2014**, *140*, 204913.
41. Li, Z.; Bian, X.; Li, X.; Karniadakis, G.E. Incorporation of memory effects in coarse-grained modeling via the Mori-Zwanzig formalism. *J. Chem. Phys.* **2015**, *143*, 243128.
42. McCarty, J.; Clark, A.J.; Lyubimov, I.Y.; Guenza, M.G. Thermodynamic Consistency between Analytic Integral Equation Theory and Coarse-Grained Molecular Dynamics Simulations of Homopolymer Melts. *Macromolecules* **2012**, *45*, 8482–8493.
43. Clark, A.J.; McCarty, J.; Guenza, M.G. Effective potentials for representing polymers in melts as chains of interacting soft particles. *J. Chem. Phys.* **2013**, *139*, 124906.
44. Stell, G. The Percus-Yevick equation for the radial distribution function of a fluid. *Physica* **1963**, *29*, 517–534.
45. Kuna, T.; Tsagkarogiannis, D. Convergence of density expansions of correlation functions and the Ornstein-Zernike equation. *arXiv* **2016**, arXiv:1611.01716.
46. Bolhuis, P.G.; Louis, A.A.; Hansen, J.P. Many-body interactions and correlations in coarse-grained descriptions of polymer solutions. *Phys. Rev. E* **2001**, *64*, 021801.
47. Mayer, J.E.; Mayer, M.G. *Statistical Mechanics*; John Wiley & Sons: Hoboken, NJ, USA, 1940.
48. Morita, T.; Hiroike, K. The statistical mechanics of condensing systems. III. *Prog. Theor. Phys.* **1961**, *25*, 537.
49. Frisch, H.; Lebowitz, J. *Equilibrium Theory of Classical Fluids*; W.A. Benjamin: New York, NY, USA, 1964.
50. Hansen, J.P.; McDonald, I.R. *Theory of Simple Lipquids*; Academic Press: London, UK, 1986.
51. Pulvirenti, E.; Tsagkarogiannis, D. Cluster expansion in the canonical ensemble. *Commun. Math. Phys.* **2012**, *316*, 289–306.
52. Louis, A.A. Beware of density dependent pair potentials. *J. Phys. Condens. Matter* **2002**, *14*, 9187–9206.
53. Katsoulakis, M.; Plecháč, P.; Rey-Bellet, L.; Tsagkarogiannis, D. Coarse-graining schemes and a posteriori error estimates for stochastic lattice systems. *ESAIM Math. Model. Numer. Anal.* **2007**, *41*, 627–660.
54. Katsoulakis, M.; Plecháč, P.; Rey-Bellet, L.; Tsagkarogiannis, D. Coarse-graining schemes for stochastic lattice systems with short and long range interactions. *Math. Comput.* **2014**, *83*, 1757–1793.

55. Katsoulakis, M.; Plecháč, P.; Rey-Bellet, L.; Tsagkarogiannis, D. Mathematical strategies and error quantification in coarse-graining of extended systems. *J. Non Newton. Fluid Mech.* **2008**, *152*, 101–112.
56. Trashorras, J.; Tsagkarogiannis, D. Reconstruction schemes for coarse-grained stochastic lattice systems. *SIAM J. Numer. Anal.* **2010**, *48*, 1647–1677.
57. Kremer, K.; Müller-Plathe, F. Multiscale problems in polymer science: Simulation approaches. *MRS Bull.* **2001**, *26*, 205–210.
58. McCoy, J.D.; Curro, J.G. Mapping of Explicit Atom onto United Atom Potentials. *Macromolecules* **1998**, *31*, 9352–9368.
59. Tsourtis, A.; Harmandaris, V.; Tsagkarogiannis, D. Effective coarse-grained interactions: The role of three-body terms through cluster expansions, under preparation.
60. Kirkwood, J.G. Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.* **1935**, *3*, 300–313.
61. McQuarrie, D.A. *Statistical Mechanics*; University Science Books: Sausalito, CA, USA, 2000.
62. Kalligiannaki, E.; Chazirakis, A.; Tsourtis, A.; Katsoulakis, M.; Plecháč, P.; Harmandaris, V. Parametrizing coarse grained models for molecular systems at equilibrium. *Eur. Phys. J.* **2016**, *225*, 1347–1372.
63. Henderson, R. A uniqueness theorem for fluid pair correlation functions. *Phys. Lett. A* **1974**, *49*, 197–198.
64. Larini, L.; Lu, L.; Voth, G.A. The multiscale coarse-graining method. VI. Implementation of three-body coarse-grained potentials. *J. Chem. Phys.* **2010**, *132*, 164107.
65. Das, A.; Andersen, H.C. The multiscale coarse-graining method. IX. A general method for construction of three body coarse-grained force fields. *J. Chem. Phys.* **2012**, *136*, 194114.
66. Lelièvre, T.; Rousset, M.; Stoltz, G. *Free Energy Computations: A Mathematical Perspective*; Imperial College Press: London, UK, 2010.
67. Tsourtis, A.; Pantazis,Y.; Katsoulakis, M.A.; Harmandaris, V. Parametric sensitivity analysis for stochastic molecular systems using information theoretic metrics. *J. Chem. Phys.* **2015**, *143*, 014116.
68. Kuna, T.; Lebowitz, J.; Speer, E. Realizability of point processes. *J. Stat. Phys.* **2007**, *129*, 417–439.