entropy

MDPI

Article

Parameterization of Coarse-grained Molecular Interactions through Potential of Mean Force Calculations and Cluster Expansion Techniques: Supplementary Material

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

[1] Department of Mathematics and Applied Mathematics, University of Crete, Greece; tsourtis@uoc.gr
[2] Department of Mathematics and Applied Mathematics, University of Crete, Institute of Applied and Computational Mathematics (IACM), Foundation for Research and Technology Hellas (FORTH), GR-70013, Heraklion, Crete, Greece; harman@uoc.gr
[3] Department of Mathematics, University of Sussex, Brighton, BN1 9QH, UK; D.Tsagkarogiannis@sussex.ac.uk
* Correspondence: tsourtis@uoc.gr, harman@uoc.gr, D.Tsagkarogiannis@sussex.ac.uk

Received: 24 May 2017; Accepted: 24 July 2017; Published: date

Abstract: In this report, we present the model we used in the paper and show how to construct effective Coarse Grained (CG) potentials through atomistic simulations, using only a small number of atoms. The numerical algorithm is applied for the construction of CG potentials involving 2- and 3-body interactions; however, our proposed methodology is quite general and, in principle can be extended even for higher order terms. Then, we 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.

1. The model

In more detail, we consider N molecules of CH_4 and we denote by $\bar{\mathbf{q}} \equiv \{\bar{q}_1, \dots, \bar{q}_N\}$ the positions of the N many carbons and by $\mathbf{q}_i \equiv \{q_{i,1}, \dots, q_{i,4}\}$ 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 (LJ) 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

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

where V_{nb} is a pair potential of all possible pairs among $\bar{\mathbf{q}}, \mathbf{q}_1, \dots, \mathbf{q}_N$, all of LJ 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:

$$T(\bar{\mathbf{q}}_i) := \frac{1}{m_C + 4m_H}(m_C \bar{q}_i + m_H \sum_{j=1}^{4} q_{i,j}). \tag{2}$$

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

$$d\hat{\mu}_b(\bar{\mathbf{q}}_i; r_i) := \frac{1}{Z_b(r_i)} d\bar{\mathbf{q}}_i \mathbf{1}_{T(\bar{\mathbf{q}}_i)=r_i} e^{-\beta V_b(\bar{\mathbf{q}}_i)}, \qquad Z_b(r_i) := \frac{1}{4!}\int_{\Lambda^5} d\bar{\mathbf{q}}_i \mathbf{1}_{T(\bar{\mathbf{q}}_i)=r_i} e^{-\beta V_b(\bar{\mathbf{q}}_i)}. \tag{3}$$

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 (1) we obtain:

$$Z_{CH_4} = \frac{1}{N!}\int_{\Lambda^N} dr_1 \ldots dr_N \prod_{i=1}^{N} Z_b(r_i) \int \prod_{i=1}^{N} d\hat{\mu}_b(\bar{\mathbf{q}}_i; r_i) e^{-\beta V_{nb}(\bar{\mathbf{q}}_1, \cdots, \bar{\mathbf{q}}_N)}. \tag{4}$$

The effective free energy is defined by:

$$e^{-\beta U_{\text{eff}}(r_1, \ldots, r_M)} := \prod_{i=1}^{N} Z_b(r_i) \int \prod_{i=1}^{N} d\hat{\mu}_b(\bar{\mathbf{q}}_i; r_i) e^{-\beta V_{nb}(\bar{\mathbf{q}}_1, \cdots, \bar{\mathbf{q}}_N)}, \tag{5}$$

for which we can construct approximations following the strategy of the paper. A similar analysis holds for the case of ethane as well.

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

$$V(q) = V_{bond}(q) + V_{angle}(q) + V_{LJ}(q), \tag{6}$$

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. The parameter values of CH_4 are summarized in Table S1. Next, we also consider the ethane (the simplest non-spherically symmetric molecule) which

Table S1. Non-bonded LJ coefficients as well as bond and angle coefficients for methane. [1]

	$\epsilon_{LJ}[\frac{Kcal}{mol}]$	σ_{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 [$\frac{Kcal}{mol^2}$]	r_0 []	K_θ [$\frac{Kcal}{mol \cdot deg^2}$]	θ_0 [rad]
700	1.1	100	1.909

consists of one rigid bond connecting two united atom CH_3 beads. Table S2 summarizes this model.

Table S2. Non-bonded LJ coefficients for ethane. [2]

	$\epsilon_{LJ}[\frac{Kcal}{mol}]$	σ_{LJ} [Å]	r_{cut} [Å]	r_0 [Å]
$CH_3 - CH_3$	0.194726	3.75	14.0	1.54

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. 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. Our goal is the construction of an effective potential from first principles, like ab initio DFT calculations, at very low density. We note that our suggested approach is a rigorous bottom up methodology, meaning that we directly sample the CG effective potential based on atomistic simulations instead of matching average forces or using any bulk information. 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 fact, 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.

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 (along the $\mathbf{r}$ vector) 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 by:

$$Z^{(2),proj}(r_1, r_2) := \int_{\{T_1(\mathbf{q_1})=r_1,\, T_2(\mathbf{q_2})=r_2\}} e^{-\beta V^{nb}(\mathbf{q})} d\mathbf{q}. \tag{7}$$

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 [3]). A schematic of the constraining is shown in Figure S1.

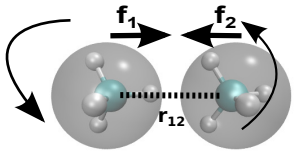

Figure S1. 2 Constrained CH_4 in space, along the X-dimension. Atomistic and CG description.

We artificially move the CG particles apart by dr, continue the simulation and repeat the procedure to get $\langle f\rangle_{r+dr}$, $\langle e^{-\beta V^{nb}}\rangle_{r+dr}$ which are canonical averages over all samples. The resulting effective potentials are denoted by $W^{(2),\text{full},F}$ and $W^{(2),\text{full},U}$ respectively (under the Cluster Expansion fomalism).

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 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 [4]. Therefore, there is an explicit temperature dependence on the observables under study.

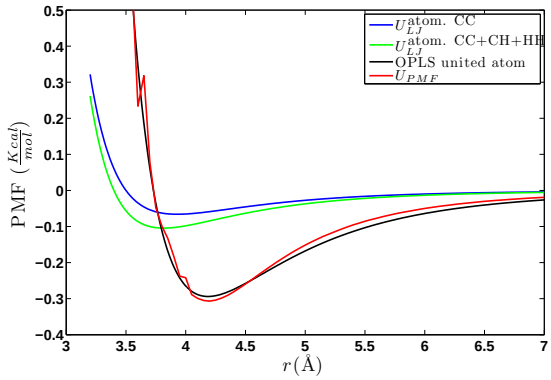

Figure S2. Our 2-body effective U_{PMF} against a United Atom model for CH_4 at $T = 300K$, the atomistic $C-C, C-H$ LJ interactions are plotted for magnitude comparison. OPLS UA forcefield [5].

In Figure S2 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*.

3. 3-Body

In the following, we are interested in constructing a higher order effective potential between three molecules. 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.

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.

3.2. COM positions

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 S3a. The next symmetry to be exploited, is along the $Y'Y$ axis in the same manner and is shown in Figure S3b, meaning that rotation of (3) around (1) is sufficient (no need to repeat around (2)).

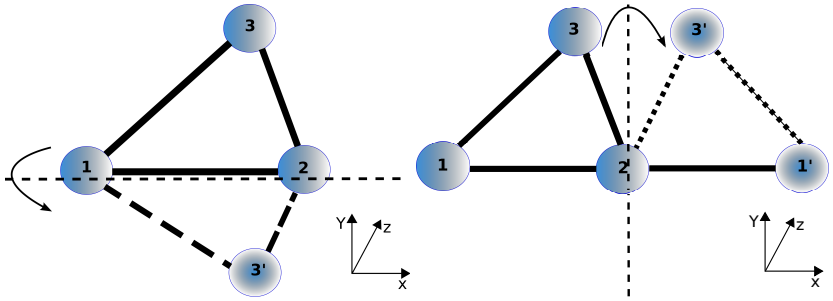

Figure S3. Triplet symmetries in space. Symmetry along $X'X$(upper) and $Y'Y$ (lower)

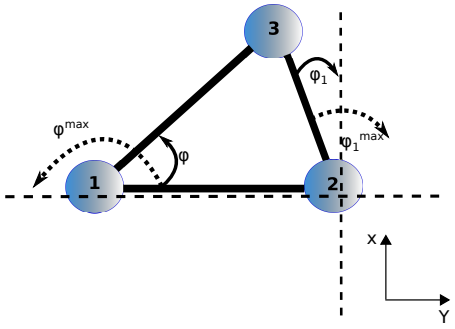

Figure S4. Algorithm 1 sampling strategy.

The implementation is described in Algorithm S1.

3.3. Statistical accuracy

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 (i) and (j), 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} (see main paper Figure 10 as well).

In Figure S5 we show the effective 3-Body potential $W^{(3),MD}$ for CH_4 at $T = 100K$ for the set of distances $r_{12} = 3.9, r_{13} = 4.0$ and $r_{23} \in [3.5 : 8]$ and $r_{12} = 4.1, r_{13} = 4.4$ and $r_{23} \in [3.5 : 8]$ 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, $W^{(3),\mathrm{MD}}$ 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.

Algorithm S1 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

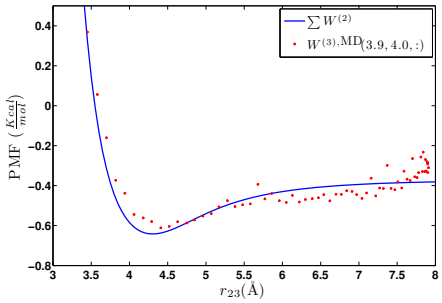

Figure S5. $W^{(3),MD}(3.9, 4.0, :)$ and $W^{(2)}(3.9) + W^{(2)}(4.0) + W^{(2)}(3.4 : r_{cut})$ for CH_4 at $T = 100K$. We see gain in information as expected, but the noise is high in the MD simulations.

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 (2) while keeping (1) still. The rotation is done by using the Euler angle formulation; the axes of the original cartesian frame are rotated by three angles: α, β, γ [6]. 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 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 Algorithm S2 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

Algorithm S2 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 α in $[0, 2\pi]$, $\alpha = \alpha + d\theta$ **do** ▷ rotation of coordinate frame along α angle
  for β in $[0, \pi]$, $\beta = \beta + d\theta$ **do**
    for γ in $[0, 2\pi]$, $\gamma = \gamma + d\theta$ **do**
      # ZYZ orientation of a CG particle at the origin $(0, 0, 0)$ according to α, β, γ
      $orient(1 : n\_atoms, 1 : 3, idx\_orientations) = rot\_matrix(\alpha, \beta, \gamma)$
      $idx\_orientations = idx\_orientations + 1$
    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
$U_{CG} = \frac{\log(U_{total})}{-\beta_{kb}}$

order of the computational cost. The CH_4 system takes about 1 hour (serial runs on 8 cores) for $d\theta = \frac{\pi}{20}$ 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.

We should also state that this method is very similar to the one used by McCoy and Curro in order to develop a CH4 united-atom model from all-atom configurations. [7]

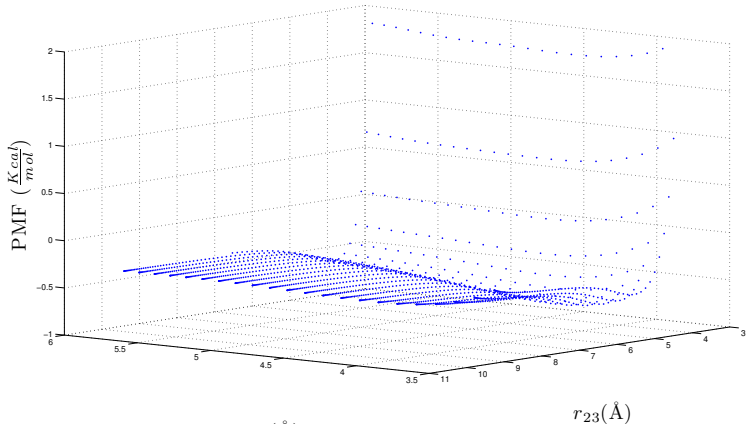

Figure S6. 3-dimensional representation of $W^{(3),\text{geom}}$ data for $\mathbf{r_{12}} = \mathbf{4.3}$ fixed, and $r_{13} \in [3.8 : 5.8]$, $r_{13} \in [3.2 : r_{cut}]$ for CH_4 at $T = 80K$. The double-well is clear.

3.5. $W^{(3)}$ 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 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.

3.6. Cubic polynomial

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.

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:

$$\begin{aligned} f(\mathbf{x},\mathbf{y},\mathbf{z}) = P_{000} &+ P_{100}\mathbf{x} + P_{010}\mathbf{y} + P_{001}\mathbf{z} + P_{200}\mathbf{x^2} + P_{020}\mathbf{y^2} + P_{002}\mathbf{z^2} \\ &+ P_{110}\mathbf{xy} + P_{101}\mathbf{xz} + P_{011}\mathbf{yz} + P_{300}\mathbf{x^3} + P_{030}\mathbf{y^3} + P_{003}\mathbf{z^3} \\ &+ P_{111}\mathbf{xyz} + P_{210}\mathbf{x^2y} + P_{201}\mathbf{x^2z} + P_{021}\mathbf{y^2z} + P_{012}\mathbf{yz^2} + P_{120}\mathbf{xy^2} + P_{102}\mathbf{xz^2} \end{aligned} \quad (8)$$

where

$$\mathbf{x} = |\mathbf{r_{ij}}| = |\mathbf{q_i} - \mathbf{q_j}|, \quad \mathbf{y} = |\mathbf{r_{ik}}| = |\mathbf{q_i} - \mathbf{q_k}|, \quad \mathbf{z} = |\mathbf{r_{jk}}| = |\mathbf{q_j} - \mathbf{q_k}|$$
$$\mathbf{x}^2 = (q_i^{(1)} - q_j^{(1)})^2 + (q_i^{(2)} - q_j^{(2)})^2 + (q_i^{(3)} - q_j^{(3)})^2$$

We omit the partial differentiation needed for the forces. 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:

$$\min_{\mathbf{x},\mathbf{y},\mathbf{z}} \Big| f(\mathbf{x},\mathbf{y},\mathbf{z}) - data \Big| = \min_{\mathbf{x},\mathbf{y},\mathbf{z}} G(\mathbf{x},\mathbf{y},\mathbf{z}) \tag{9}$$

with the *Conjugate Gradient* method, where:

$$G(\mathbf{x},\mathbf{y},\mathbf{z}) = \frac{1}{2} X^T A X - X^T b \tag{10}$$

where matrix $A \in \mathbb{R}^{n \times 20}$ contains the values of the data at the points $\mathbf{x}, \mathbf{y}, \mathbf{z}$, X is the vector with the polynomial constants to be determined and b is the vector with the data.

3.7. 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:

$$-\frac{\partial W^{(3)}(r_{12}, r_{13}, r_{23})}{\partial \mathbf{q_1}}, -\frac{\partial W^{(3)}(r_{12}, r_{13}, r_{23})}{\partial \mathbf{q_2}}, -\frac{\partial W^{(3)}(r_{12}, r_{13}, r_{23})}{\partial \mathbf{q_3}} \tag{11}$$

$$r_{12} = |\mathbf{q_1} - \mathbf{q_2}| \tag{12}$$

on the triplets, meaning:

$$\frac{\partial W^{(3)}(r_{12}, r_{13}, r_{23})}{\partial \mathbf{q_1}} = \frac{\partial W^{(3)}(r_{12}, r_{13}, r_{23})}{\partial r_{12}} \frac{\partial r_{12}}{\partial \mathbf{q_1}} + \frac{\partial W^{(3)}(r_{12}, r_{13}, r_{23})}{\partial r_{13}} \frac{\partial r_{13}}{\partial \mathbf{q_1}} + \frac{\partial W^{(3)}(r_{12}, r_{13}, r_{23})}{\partial r_{12}} \frac{\partial r_{23}}{\partial \mathbf{q_1}}$$

$$\frac{\partial r_{12}}{\partial \mathbf{q_1}} = \frac{1}{r_{12}} \mathbf{r_{12}} \tag{13}$$

Note that $\frac{\partial W^{(3)}}{\partial \mathbf{q_1}}$ contains information from (2) and (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:

$$\frac{\partial W^{(3)}(r_{12}, r_{13}, r_{23})}{\partial r_{12}} = \frac{W^{(3)}(r_{12} + dr_{12}, r_{13}, r_{23}) - W^{(3)}(r_{12} - dr_{12}, r_{13}, r_{23})}{2dr_{12}} \tag{14}$$

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 $f^{(3)}(r_{12}, r_{13}, r_{23})$.

3.8. 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 thermodynamic observable $g(r)$.

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 S7a for both systems, together with the ones with the added 3-Body interactions

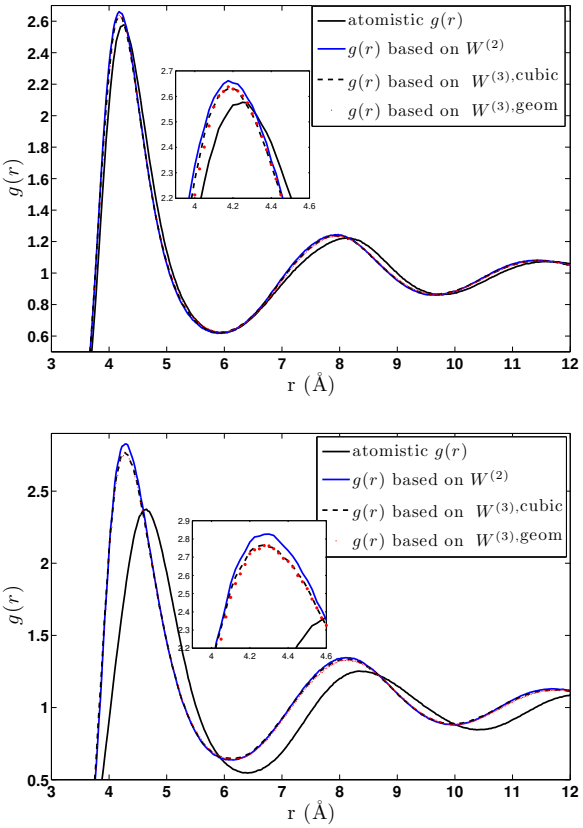

Figure S7. Comparison between the reference (atomistic) and CG $W^{(2)}$ $g(r)$ for a) CH_4 at $T = 80K$ and b) $CH_3 - CH_3$ at $T = 150K$.

Overall, both methods improve on the estimation of the CG $g(r)$ at the cost of extra computations of the 3-Body case as can be seen in Figure S7. 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.

References

1. S. Mayo, B. Olafson, and W. Goddard. Dreiding: a generic force field for molecular simulations. *J. Phys. Chem.*, 94(26):8897–8909, 1990.
2. 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.
3. 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.
4. Daan Frenkel and B. Smit. *Understanding Molecular Simulation, Second Edition: From Algorithms to Applications (Computational Science)*. Academic Press, 2001.
5. 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.
6. Herbert Goldstein. *Classical Mechanics*. Addison-wesley, 1950.

7. J.D. McCoy and J.G. Curro. Mapping of explicit atom onto united atom potentials. *Macromolecules*, 31:9352–9368, 1998.