CHAPTER 8

COARSE-GRAINED INTERMOLECULAR POTENTIALS DERIVED FROM THE EFFECTIVE FRAGMENT POTENTIAL: APPLICATION TO WATER, BENZENE, AND CARBON TETRACHLORIDE

GAURAV PRANAMI[1], LYUDMILA SLIPCHENKO[2], MONICA H. LAMM[1], AND MARK S. GORDON[2]

[1]*Department of Chemical and Biological Engineering, Iowa State University, Ames, IA 50011, USA*
[2]*Department of Chemistry, Iowa State University, Ames, IA 50011, USA,*
e-mail: mark@si.msg.chem.iastate.edu (M.S. Gordon)

Abstract: A force matching technique based on previous work by Voth and co-workers is developed and employed to coarse grain intermolecular potentials for three common solvents: carbon tetrachloride, benzene, and water. The accuracy of the force-matching approach is tested by comparing radial distribution functions (RDF) obtained from simulations using the atomistic and coarse-grained potentials. Atomistic molecular dynamics simulations were performed using the effective fragment potential method (EFP). The RDFs obtained from molecular dynamics simulations of EFPs for carbon tetrachloride, benzene and water are in a good agreement with the corresponding experimental data. The coarse-grained potentials reproduce the EFP molecular dynamics center-of-mass RDFs with reasonable accuracy. The biggest discrepancies are observed for benzene, while the coarse-graining of water and spherically symmetric carbon tetrachloride is of better quality

Keywords: Coarse-graining, Force-matching, Effective fragment potential method, Molecular dynamics, Radial distribution functions, Multiscale modeling

8.1. INTRODUCTION

In the molecular dynamics (MD) [1, 2] technique, a system of particles evolves in time according to the equation of motion, $F_i = m_i \ddot{x}_i$, where F_i is the net force acting on particle i, and m_i and $\ddot{x}_i$ are the mass and acceleration of particle i, respectively. In a molecular system, typical bond lengths are of the order of angstroms while bond vibrations take place at the time scale of 10^{-13} s. Therefore, the equations of motion for atoms have to be integrated with time steps on the order of 10^{-15} s. However, many important chemical and biological phenomena in macromolecules take place at much larger time scales, as shown in Table 8-1.

D.M. York and T.-S. Lee (eds.), Multi-scale Quantum Models for Biocatalysis, 197–218.
DOI 10.1007/978-1-4020-9956-4_8,

Table 8-1. Characteristic time scales of different events in macromolecular systems

System/Phenomena	Characteristic time scales
Fusion of micelles [3]	10^{-2} s
Self-assembly of diblock copolymers [4]	10^{-6} s
Entanglement of a polymer chain [5]	10^{-5} s
Protein folding [6]	10^{-6} s
DNA replication [6]	10^{-3} s
Membrane fusion [6]	10^{-1} s

Large time and length scales of characteristic events in macromolecular systems, such as polymers, lipids and nanoparticles, prohibit a molecular simulation study at the atomistic level due to enormous CPU time and memory requirements. Moreover, when studying phenomena occurring at timescales on the order of 10^{-6} to 10^{-1} s, the behavior of the fast degrees of freedom, like bond fluctuations, are not always of interest. Therefore, a systematic approach to coarse-graining is needed, in which the unimportant degrees of freedom are eliminated, but the underlying physics governing the phenomena at larger length and time scales is retained.

Thus, the aim of coarse-graining techniques is to determine the effective interaction potentials between the coarse-grained sites such that the simulation of the coarse-grained (CG) system using the CG potentials yields properties that compare favorably with those of the corresponding atomistic system. An effective coarse-graining method should be systematic, automatic, and fast. Moreover, it should be flexible enough to handle different kinds of potentials and capable of generating CG potentials that would reproduce properties matching experimental data or the properties obtained from an atomistic simulation.

In general, coarse-graining an atomistic system requires a two step process, i.e., first, grouping the atoms into CG sites, as shown in Figure 8-1, and second, determining the effective bonded and non-bonded potentials between the CG sites. Most of the coarse-graining procedures reported in the literature can be classified into three categories: (i) optimization of potential parameters by fitting them to a desired property [7–11], (ii) structure matching [5, 12–22], and (iii) force matching [23, 24]. Systematic structure matching and force matching methods are preferable to the method of ad hoc parameter guessing.

In structure matching methods, potentials between the CG sites are determined by fitting structural properties, typically radial distribution functions (RDF), obtained from MD employing the CG potential (CG-MD), to those of the original atomistic system. This is often achieved by either of two closely related methods, Inverse Monte Carlo [12–15] and Boltzmann Inversion [5, 16–22]. Both of these methods refine the CG potentials iteratively such that the RDF obtained from the CG-MD approaches the corresponding RDF from an atomistic MD simulation.

Both the inverse Monte Carlo and iterative Boltzmann inversion methods are semi-automatic since the radial distribution function needs to be re-evaluated at

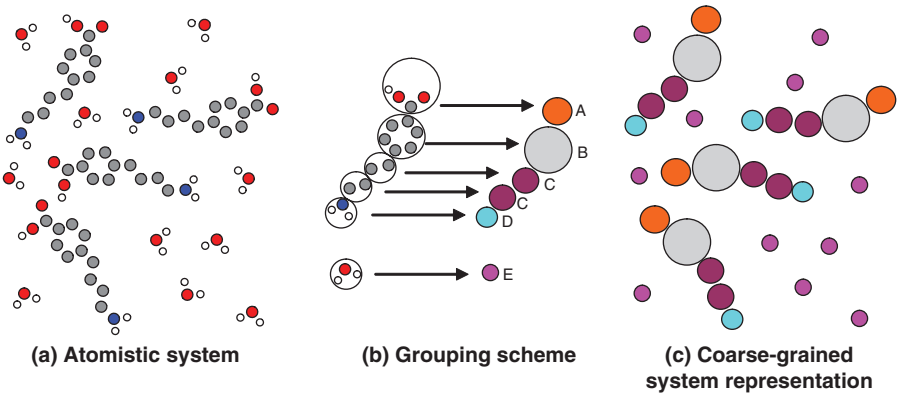

Figure 8-1. Coarse-graining procedure. (**a**) A snapshot of an atomistic system. (**b**) Groups of atoms of the atomistic system are combined into CG sites in order to reduce the number of degrees of freedom. (**c**) The atomistic system of (**a**) as represented by CG sites

each iteration. Moreover, convergence can be a problem if the potential of mean force (PMF) does not serve as a good initial guess. The inverse Monte Carlo method evaluates all the potentials at the same time and requires adequate sampling of the four-particle correlation functions during each iteration. The latter becomes time consuming if there are several different kinds of pair potentials. On the other hand, in the Boltzmann inversion method, potentials can be refined one at a time by keeping the rest of them constant. However, as the potentials depend upon each other, it is important to ensure that each potential does not change during the optimization of others. The advantage of the structure matching methods is that they result in potentials that will reproduce the correct structural properties; however, their main disadvantages are the necessity for frequent re-evaluation of radial distribution functions and the increasing complexity for a system with more than five coarse-grained sites.

In the force matching method, the effective pair-forces between coarse-grained sites are derived from the net force acting on chosen CG sites along an MD trajectory obtained from a short atomistic MD. The force-matching method has the advantage of being systematic and automatic because the CG pair forces are evaluated from the data gathered along the atomistic trajectory and there is no need to run multiple simulations. The force matching method has been successfully applied to study condensed phase liquids [23–25], ionic liquids [26, 27], C_{60} nanoparticles [28, 29] and dimyristoylphosphatidylcholine (DMPC) lipid bilayers [30, 31].

In order to ensure accurate CG potentials, one needs to conduct MD simulations with a reliable atomistic potential model. The most desirable theoretical approach for the atomistic-scale simulations would be to use a level of quantum mechanics (QM) that can treat both intermolecular and intramolecular interactions with acceptable accuracy. Realistically, the minimal QM levels of theory that can adequately treat all different types of chemical forces are second order perturbation theory [32] (MP2)

and, preferably, coupled cluster (CC) theory with some accounting for triples; i.e., CCSD(T) [33]. Unfortunately, a sufficiently high level of QM comes at a significant computational cost; for example, CCSD(T) scales $\sim N^7$ with a problem size, where N is the number of atomic basis functions. This places serious limitations on the sizes of accessible molecular systems. Moreover, in order to obtain the CG potentials including the three-body terms, one needs to perform an extensive sampling of the atomistic-level system, which at present is impractical even for very short time-scales. Successful examples of QM-based coarse-graining have been presented based on Car-Parrinello simulations of atomistic systems [23].

An alternative approach is to replace an accurate but expensive first-principle-based technique by a reliable model potential. Such potentials, broadly referred to as molecular mechanics (MM), generally cannot account for bond-breaking, but can, in principle, account for the range of intermolecular interactions. However, using a fitted pair-wise potential may result in losing quantitative accuracy, predictability, and the underlying physics.

This contribution pursues a different approach for preserving the accuracy of the atomistic level, by using a model potential that is exclusively derived from first principles, the effective fragment potential method (EFP). The original EFP1 method [34, 35] was developed specifically to describe aqueous solvent effects on biomolecular systems and chemical reaction mechanisms, and contains fitted parameters for the repulsive term. A general (EFP2) method [36] is applicable to any solvent; it includes all of the essential physics and has no empirically fitted parameters. A force matching technique is applied to derive a coarse-grained potential from the molecular trajectories generated with EFP MD simulations. The quality of the EFP force matching is tested on carbon tetrachloride, benzene, and water systems. This contribution is the first application of a coarse-graining procedure to the EFP method.

8.2. THEORY

8.2.1. Effective Fragment Potential Method

The effective fragment potential method is a first-principles based model potential for describing intermolecular forces. The interaction energy in EFP1, specifically designed for modeling water, consists of electrostatic, induction, and fitted exchange-repulsion terms. Presently, three different EFP1 models are available, with fitting done to represent Hartree-Fock (HF), DFT/B3LYP, and MP2 levels of theory. These models are called EFP1/HF [34], EFP1/DFT [37], and EFP1/MP2 [38], respectively. In EFP1/MP2, fitted dispersion terms are also included.

The general EFP2 model can be applied to any solvent and includes electrostatic, induction, exchange-repulsion, dispersion, and charge-transfer components, all of which are derived from first-principles using long- and short-range perturbation theory. Charge-transfer interactions are not included in this work, since they are primarily important for charged species. All of the EFP2 parameters are generated during a MAKEFP run, performed for each unique molecule; e.g., benzene and CCl_4. Once

EFP parameters for a particular fragment are generated for a given atomic basis set, they can be used in a variety of applications. The various components of the non-bonded interactions between molecules are evaluated using the EFP2 generated parameters. The procedure has been described in elsewhere [36]; only the main points are summarized below.

The electrostatic energy is calculated using the distributed multipolar expansion introduced by Stone [39, 40], with the expansion carried out through octopoles. The expansion centers are taken to be the atom centers and the bond midpoints. So, for water, there are five expansion points (three at the atom centers and two at the O-H bond midpoints), while in benzene there are 24 expansion points. The induction or polarization term is represented by the interaction of the induced dipole on one fragment with the static multipolar field on another fragment, expressed in terms of the distributed localized molecular orbital (LMO) dipole polarizabilities. That is, the number of polarizability points is equal to the number of bonds and lone pairs in the molecule. One can opt to include inner shells as well, but this is usually not useful. The induced dipoles are iterated to self-consistency, so some many body effects are included.

The Coulomb point multipole model breaks down when fragments approach too closely, since then the actual electron density on the two fragments is not well approximated by point multipoles. Thus, electrostatic interactions become too repulsive whereas the induction energy is too attractive if fragments approach each other too closely. In order to avoid this unphysical behavior, electrostatic and induction energy terms are modulated by exponential damping functions with parameters being obtained from fitting the damped multipole potential to the Hartree-Fock one [41, 42]. In EFP2, the induction energy terms are damped in a similar way [43].

The exchange repulsion energy in EFP2 is derived as an expansion in the intermolecular overlap. When this overlap expansion is expressed in terms of frozen LMOs on each fragment, the expansion can reliably be truncated at the quadratic term [44]. This term does require that each EFP carries a basis set, and the smallest recommended basis set is 6-31++G(d,p) [45] for acceptable results. Since the basis set is used only to calculate overlap integrals, the computation is very fast and quite large basis sets are realistic.

The dispersion interaction can be expressed as the familiar inverse R expansion,

$$E_{disp} = \sum_{n} C_n R^{-n} \tag{8-1}$$

The coefficients C_n may be derived from the (imaginary) frequency dependent polarizabilities summed over the entire frequency range [46]. If one employs only dipole polarizabilities the dispersion expansion is truncated at the leading term, with $n = 6$. In the current EFP2 code, an estimate is used for the $n = 8$ term, in addition to the explicitly derived $n = 6$ term. Rather than express a molecular C_6 as a sum over atomic interaction terms, the EFP2 dispersion is expressed in terms of LMO-LMO

interactions. In order to ensure that the dispersion interaction goes to zero at short distances, the damping term proposed by Tang and Toennies [47] is employed.

The effective fragment potential method is several orders of magnitude less computationally expensive than ab-initio methods because it evaluates intermolecular interactions by simplified formulas derived from perturbation series in terms of intermolecular distances and orbital overlap integrals. The most time-consuming terms in EFP2 are the charge-transfer (omitted in the current work) and the exchange-repulsion, which is evaluated using calculated on the fly orbital overlap integrals between different fragments. EFP2 can be used in MD simulations of moderately sized systems. For example, calculation of the energy and gradient for a system of 64 waters with periodic boundary conditions (PBC) requires about 2 s on one Opteron 2600 MHz processor. Despite its low computational cost, the accuracy of EFP in predicting structures and binding energies in weakly-bonded complexes and liquids is very high and comparable with that of MP2 [42].

8.2.2. Force Matching Procedure

The aim of the force matching procedure is to obtain the effective pair-force between CG sites using the force data obtained from a detailed atomistic molecular dynamics (MD) trajectory. The current implementation of the force-matching method closely follows the formulation from Refs. [23, 24].

The first step of the systematic force matching procedure is to define the CG sites, which are generally the centers of mass or geometric centers of groups of atoms, as illustrated in Figure 8-1(b), thus eliminating the group's internal degrees of freedom. Following the coarse graining scheme depicted in Figure 8-1(b), the snapshot of the MD trajectory of Figure 8-1(a) will look like the one shown in Figure 8-1(c). In the next step, the forces and positions of atoms from the detailed atomistic MD are converted to forces and positions of CG sites as depicted in Figure 8-2.

Assume that there are a total of N coarse-grained sites in the system for any one MD snapshot (p $=$ 1), with coordinates ($r_i = x_i, y_i, z_i$) and net forces, F_i, (where $i = 1 - N$) acting on them, and that these are known from the atomistic MD trajectory data. If $f_{ij}(r_i, r_j)$ represents the force acting on the ith CG site due to

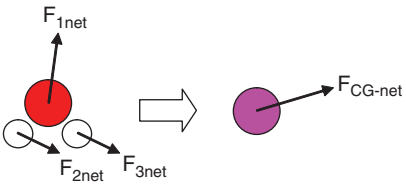

Figure 8-2. Conversion of forces from atomistic MD to forces on coarse-grained sites

the *j*th CG site, then each snapshot from the MD trajectory results in the following $n(= N$ *for one snapshot*) equations:

$$\sum_{j=1}^{N} f_{ij} = F_i \quad i = 1, 2, 3 \ldots N \tag{8-2}$$

Here the pair-force $f_{ij}(r_i, r_j)$ is unknown, so a model pair-force $f_{ij}(r_i, r_j, p_1, p_2 \ldots p_m)$ is chosen, which depends linearly upon m unknown parameters $p_1, p_2 \ldots p_m$. Consequently, the set of Eq. (8-2) is a system of linear equations with m unknowns $p_1, p_2 \ldots p_m$. The system (8-2) can be solved using the singular value decomposition (SVD) method if n $> m$ (over-determined system), and the resulting solution will be unique in a least squares sense. If $m > n$, more equations from later snapshots along the MD trajectory should be added to the current set so that the number of equations is greater than the number of unknowns. Mathematically, $n = qN > m$ where q is the number of MD snapshots used to generate the system of equations.

It is important to note that model pair-forces for the interactions *A-A, A-B, A-C*, etc. are different from each other although they may have the same functional form. If required, the interaction between two non-bonded *A* CG sites (A-$A_{non\text{-}bonded}$) can be treated differently from the interaction between two bonded *A* CG sites (A-A_{bonded}). In a system with $A, B \ldots E$ as chosen coarse-grained sites,

$$f_{ij} = \begin{cases} f_{ij}^{AA}(p_1^{AA}, p_2^{AA} \ldots p_a^{AA}) & if \quad ij = AA \\ f_{ij}^{AB}(p_1^{AB}, p_2^{AB} \ldots p_b^{AB}) & if \quad ij = AB \; or \; BA \\ \vdots & \vdots \\ f_{ij}^{EE}(p_1^{EE}, p_2^{EE} \ldots p_z^{EE}) & if \quad ij = EE \end{cases} \tag{8-3}$$

Clearly, the total number of unknowns that need to be determined is $m =$ a $+ b + \ldots + z$ and a solution set for parameters $p_1, p_2 \ldots p_m$ is determined using the singular value decomposition or any other suitable method. The mean pair-force corresponding to the "potential of mean force" can be obtained in a systematic manner by averaging a number of sets of solutions for parameters $p_1, p_2 \ldots p_m$ obtained along the atomistic MD trajectory in which the phase space is sampled extensively.

A convenient and systematic way to represent $f_{ij}(r_{ij})$ (r_{ij} is the distance between particles *i* and *j*) as a linear function of unknowns is to employ cubic splines [48], as shown in Figure 8-3. The advantage of using cubic splines is that the function is continuous not only across the mesh points, but also in the first and second derivatives. This ensures a smooth curvature across the mesh points. The distance r_{ij} is divided into 1-dimensional mesh points, thus, $f_{ij}(r_{ij})$ in the *k*th mesh ($r_k \leq r_{ij} \leq r_{k+1}$) is described by Eqs. (8-4), (8-5) and (8-6) [48].

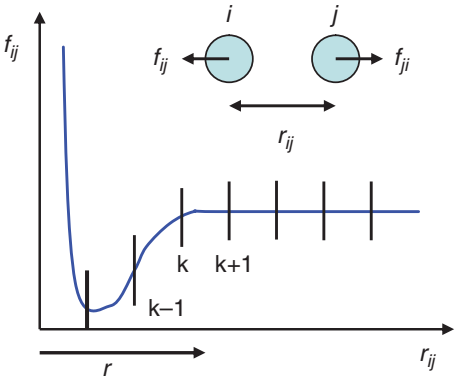

Figure 8-3. $f_{ij}(r_{ij})$ as cubic splines. The distance, r_{ij}, between atoms i and j is divided into the mesh as shown. In each mesh, the pair-force f_{ij} is modeled as a cubic polynomial

$$f_{ij}\left(r_k \le r_{ij} \le r_{k+1}\right) = A\left(r_k, r_{ij}, r_{k+1}\right) f\Big|_k + B\left(r_k, r_{ij}, r_{k+1}\right) f\Big|_{k+1} + C\left(r_k, r_{ij}, r_{k+1}\right) f''\Big|_k + D\left(r_k, r_{ij}, r_{k+1}\right) f''\Big|_{k+1} \tag{8-4}$$

$$A = \frac{r_{k+1} - r_{ij}}{r_{k+1} - r_k} \qquad B = 1 - A \qquad C = \frac{1}{6}\left(A^3 - A\right)\left(r_{k+1} - r_k\right)^2$$
$$D = \frac{1}{6}\left(B^3 - B\right)\left(r_{k+1} - r_k\right)^2 \tag{8-5}$$

Here, $f\Big|_k$, $f'\Big|_k$ *and* $f''\Big|_k$ are the values of the pair-force $f_{ij}(r_{ij})$ and its first and second derivatives, respectively, at mesh point r_k. Equation (8-4) ensures the continuity of the function and its second derivative at the mesh points. In order to make the first derivatives continuous across the mesh points r_k, an additional set of Eq. (8-6) is needed:

$$\frac{r_k - r_{k-1}}{6} f''\Big|_{k-1} + \frac{r_{k+1} - r_{k-1}}{3} f''\Big|_k + \frac{r_{k+1} - r_k}{6} f''\Big|_{k+1} = \frac{f\Big|_{k+1} - f\Big|_k}{r_{k+1} - r_k} - \frac{f\Big|_k - f\Big|_{k-1}}{r_k - r_{k-1}} \tag{8-6}$$

At the end points of the mesh, Eq. (8-6) cannot apply. Instead, one needs to introduce boundary conditions, for instance, at large r_{ij} the pair-force f_{ij} is usually zero. It is important to note that the mesh sizes should not necessarily be uniform. For example, at those separations for which the pair-force varies rapidly with distance the mesh size can be chosen to be small enough to capture all of the variations.

If some meshes do not get sampled in the set of Eq. (8-2), i.e., if the coefficients of the correspondingare $f\Big|_k$ *and* $f''\Big|_k$ are zero, then these unknowns are removed from the set of equations and set equal to zero. Equation (8-6) and boundary conditions need to be satisfied exactly, however, in this work we have solved all the equations in least squared sense. By solving the combined set of Eqs. (8-2– 8-6). Solutions obtained for a large number of such sets of equations are averaged to reduce statistical noise in the CG pair-force. Then, a suitable analytic function should be fitted to the tabulated $f_{ij}(r_{ij})$. If no distinction is made between A-A_{bonded} and A-$A_{non\text{-}bonded}$ interactions in the force matching procedure the resulting coarse-grained A-A pair-force will have the combined effect of bonded and non-bonded interactions. In a typical atomistic MD simulation, bonded and non-bonded interactions both occur at short A-A distances. Therefore, the coarse-grained A-A interaction force may not be physically correct. In general, separate treatments of bonded and non-bonded interactions are preferred even though it increases the total number of unknowns and the size of the linear least squared problem.

Once all the coarse-grained interactions are determined using the force-matching procedure, they need to be validated by running a MD simulation of the coarse-grained system. Comparing properties such as pair correlation function(s) obtained from the coarse-grained and original atomistic MD is a direct test of the quality of coarse-graining.

The reduction in the number of degrees of freedom can lead to an incorrect pressure in the simulation of the coarse-grained systems in NVT ensembles or to an incorrect density in NPT ensembles [24]. The pressure depends linearly on the pair-forces in the system, hence the effect of the reduced number of degrees of freedom can be accounted for during the force matching procedure [24]. If T is the temperature, V the volume, N the number of degrees of freedom of the system, and k_b the Boltzmann constant then the pressure P of a system is given by

$$P = \frac{Nk_bT}{3V} + \frac{1}{3V}\sum_{i<j}\vec{f}_{ij} \bullet \vec{r}_{ij} \tag{8-7}$$

In order to compensate for the reduced number of degrees of freedom in the coarse-grained system, the following constraint should be added to the set of Eq. (8-2):

$$\sum_{i<j}\vec{f}_{ij} \bullet \vec{r}_{ij} = 3P^{At-MD}V^{CG} - N^{CG}k_bT \tag{8-8}$$

Here, $P^{At\text{-}MD}$ is the pressure in the system in detailed atomistic MD and V^{CG} and N^{CG} are the volume and number of degrees of freedom of the coarse-grained system. The left side of this equation is evaluated for the coarse-grained system for each snapshot during force-matching.

8.3. COMPUTATIONAL DETAILS

8.3.1. EFP MD Simulations

Molecular dynamics simulations of liquid carbon tetrachloride, benzene, and water were performed using the effective fragment potential method, as implemented in the GAMESS (General Atomic and Molecular Electronic Structure System) electronic structure package [49, 50]. EFP2 parameters for benzene were obtained using the 6-311++G(3df,2p) basis set at the MP2/aug-cc-pVTZ [51] geometry of the benzene monomer, with C—C and C—H bond lengths of 1.3942 Å and 1.0823 Å, respectively. EFP2 parameters for CCl_4 were generated by using the 6-311++G(d,p) basis [52–54], with the monomer geometry optimized at the MP2/6-311G(d,p) level (C—Cl bond length of 1.772 Å). The EFP1/MP2 [38] potential was used for water.

The MD simulations were carried out in an NVT ensemble at ambient conditions; each simulation contained 64 molecules in a cubic box with periodic boundary conditions. The temperature is set to 300K in all the simulations. Table 8-2 summarizes details of the EFP-MD simulations used for force-matching. In particular, the type of the potential, box size, time step of integration, frequency of data sampling, total number of sampled configurations, and the total time of the equilibrated MD simulation are listed for each system. Initial equilibration of the systems was performed before recording the data for force matching. In order to ensure good energy conservation in the MD simulations, switching functions were employed for all EFP interaction terms at long distances [55]. Additionally, in simulations of water, Ewald summations were used to treat long-range electrostatic interactions (charge-charge, charge-dipole, dipole-dipole, and charge-quadrupole).

Since EFP employs frozen internal geometries of fragments, during the MD simulations, CCl_4, benzene and water molecules are treated as rigid bodies with a net force and torque acting on each center of mass (COM). Thus, the net forces acting on COMs required for force matching are directly available from the EFP MD simulations. The information about torques is not used in force-matching because each molecule is represented as a point at its COM.

Table 8-2. EFP-MD simulation parameters[a]

System	Potential	Box Length (Å)	Timestep (fs)	Frequency (fs)	Samples	Total simulation time (ps)
CCl_4	EFP2	21.77	0.3	30	1200	36
Benzene	EFP2	21.20	0.5	50	500	25
Water	EFP1/MP2	12.40	0.3	30	1000	30

[a]For each system, columns specify the type of the potential, simulation box size, time step of MD integration, frequency at which data is sampled for force-matching, total number of configurations sampled in MD simulations and the total time of equilibrated MD simulations.

8.3.2. Force Matching

Carbon tetrachloride, benzene and water molecules have been coarse-grained to their COM using the force matching technique described in Section 8.2.1. The effective COM pair-force was modeled using cubic splines over a range of distances, described by an inner cutoff and an outer cutoff, with the mesh sizes summarized in Table 8-3. The outer cutoff of the model pair-force was set such that it never exceeded half of the simulation box length and large enough to ensure that the effective pair-force obtained from force-matching naturally approaches zero at the chosen outer cutoff. The inner cut-off can be safely chosen as zero or it can be approximated as a distance which is smaller than the smallest separation between a pair of CG sites sampled in the atomistic MD. The mesh-size should be small enough to capture all the features of the effective pair-force but, as mentioned earlier, smaller mesh-sizes result in more unknowns. Consequently, a smaller mesh is used in the regions where the CG pair-force is sharply repulsive and varies rapidly with distance. A total of k meshes are used to model an interaction that is expressed in $2k + 2$ unknowns.

Table 8-3. Force matching details[a]

System	Distance (Å)	Mesh-Size (Å)	Unknowns	Configurations per Set	Number of sets averaged
CCl_4	4–6	0.05	178	3	400
	6–10.8	0.1			
Benzene	3–6	0.05	202	3	150
	6–10	0.1			
Water	2–3.5	0.025	222	4	250
	3.5–6	0.05			

[a] For each studied system, the range of distances at which pair-forces are modeled as cubic splines is given, as well as mesh sizes and the number of resulting unknowns, the number of configurations included in a set to generate an over-determined system of equations, and the number of sets for which the least squared solution is averaged.

The net forces acting on the COMs of all molecules in a given MD snapshot were equated to the corresponding net force obtained from the model pair-force; consequently, each configuration yields 64 equations since each EFP-MD simulation contains 64 molecules. Three or four (see Table 8-3) MD configurations were used to generate a set of equations such that the number of equations was greater than the number of unknowns. Solutions for a number of such sets were averaged to obtain the effective mean COM pair-force. In the results reported here, the pressure is not constrained.

At short distances, approximately equal to the excluded volume diameter, effective pair forces obtained from force matching exhibit unphysically large fluctuations. This is largely due to inadequate sampling of configurations at short distances in

the EFP-MD simulation. These short-range pair-force data were ignored in further analysis. Ignoring the force data may lead to some inconsistency in the agreement of properties of atomistic and CG systems; however, as very few CG sites exist at such small separations in the EFP-MD simulation, this should not lead to significant error if averaging is done over a large number of sets during force matching. The remaining pair-force data, *f(r)*, obtained from force-matching, are fitted to the following function:

$$F(r) = \int \sum_{n=2}^{16} \frac{A_n}{r^n} \tag{8-9}$$

The fitting coefficients A_n for CCl4, benzene, and water are listed in Table 8-4. The corresponding effective COM pair-potential, *U(r)*, was obtained by integrating *F(r)* with the condition that the potential is zero at the outer cutoff:

$$U(r) = -\int F(r)\, dr \tag{8-10}$$

The coarse-grained pair-force and pair-potential were used for carrying out the molecular dynamics simulations (coarse-grained MD, CG-MD) of 64 points, each point representing a COM of a CCl_4 or benzene or water molecule, at the same conditions as used for the corresponding atomistic MD (Table 8-2). All CG-MD simulations were run using the LAMMPS [56] (Large-scale Atomic/Molecular Massively Parallel Simulator) molecular simulation code available at http://lammps.sandia.gov. LAMMPS is capable of running MD simulations using tabulated pair-forces and

Table 8-4. Fitting coefficients A_n corresponding to Eq. (8-9). The units of *r* and *F(r)* are Å and kcal/mol-Å, respectively, for curve fitting

	Carbon Tetrachloride	**Benzene**	**Water**
A_2	−1.969703152101180E + 21	1.244993048322720E+21	2.175775867955310E+17
A_3	4.306989690407250E+21	−2.620983313167290E + 21	−8.551913708210820E + 17
A_4	−4.333455273390900E + 21	2.545558326687080E+21	1.550778775883470E+18
A_5	2.659644152606080E+21	−1.511552672176440E + 21	−1.719444254190780E + 18
A_6	−1.112575488728160E + 21	6.129876548482810E+20	1.302147676537040E+18
A_7	3.356022412317150E+20	−1.795710744899950E + 20	−7.124434785792410E + 17
A_8	−7.528548090161490E + 19	3.918212052328320E+19	2.903928257920110E+17
A_9	1.276021907548210E+19	−6.468394131217870E + 18	−8.957222760163820E + 16
A_{10}	−1.642124116314870E + 18	8.117691446975380E+17	2.100873205120480E+16
A_{11}	1.596839968332750E+17	−7.706208900534920E + 16	−3.728428139785960E + 15
A_{12}	−1.155063457121670E + 16	5.446871839354590E+15	4.928243298286980E+14
A_{13}	6.026936982240990E+14	−2.779449283288870E + 14	−4.704569177419660E + 13
A_{14}	−2.144530055247100E + 13	9.678942266368310E+12	3.066010958528650E+12
A_{15}	4.658131620368480E+11	−2.058787501037060E + 11	−1.221016058567130E + 11
A_{16}	−4.660182908774720E + 09	2.018075773802740E+09	2.241814292894480E+09

pair-potentials. Therefore, the CG pair-forces obtained from force matching can be directly used to run the CG MD simulations. The integration time step was 1 fs. Each equilibrated CG-MD simulation was 3 ns long and the position data was collected every 1 ps. In order to test the ability of coarse-grained potentials to reproduce properties of atomistic systems, RDFs obtained from CG-MD are compared below to the corresponding COM-COM RDFs from atomistic EFP-MD.

8.4. RESULTS AND DISCUSSION

8.4.1. EFP-MD Simulations

EFP radial distribution functions for liquid CCl_4, benzene, and water are presented in Figures 8-4, 8-5 and 8-6. EFP2 and experimental [57] C—Cl and Cl—Cl RDFs for liquid carbon tetrachloride are shown in Figure 8-4. The Cl—Cl EFP2 RDFs are in good agreement with the experimental data. The discrepancies in the C—Cl RDF curves are more significant, although the qualitative features of the experimental RDF are reproduced. It is possible that the strong structural enhancement observed in the experimental CCl_4 RDFs is an artifact that arises due to numerical instabilities when specific atom-atom RDFs are obtained from X-ray and neutron analysis data [58]. To confirm this the so-called $G_d(r)$ functions were calculated. The $G_d{}^X(r)$ and $G_d{}^n(r)$ functions are Fourier transforms of the X-ray and neutron diffraction distinct structure functions, respectively; the latter are unambiguously determined experimentally [57]. For CCl_4, $G_d{}^X(r)$ and $G_d{}^n(r)$ functions are connected to specific atom-atom RDFs in the following way:

$$G_d^X(r) \approx 0.00 g_{CC}(r) - 0.12 g_{CCl}(r) - 0.88 g_{ClCl}(r)$$
$$G_d^n(r) \approx 0.02 g_{CC}(r) + 0.25 g_{CCl}(r) + 0.75 g_{ClCl}(r),$$

where $g_{CC}(r)$, $g_{CCl}(r)$, and $g_{ClCl}(r)$ are C—C, C—Cl, and Cl—Cl RDFs, respectively.

Experimental and EFP-MD $G_d{}^n(r)$ and $G_d{}^X(r)$ functions are shown in Figure 8-4c and 8-4d, respectively. The agreement between the EFP2 and experimental G_d-functions is better than that between specific RDFs, although some discrepancies remain. EFP2 overestimates the heights of the peaks at 4.0 Å in both G_d graphs, and the peaks at 6.2 Å are slightly shifted to longer distances.

Figure 8-5 shows the EFP2 and experimental [59] C—C RDFs for liquid benzene. The EFP2 RDF is in reasonable agreement with the experimental curve, with three distinct peaks in the 4–7 Å region. These peaks might correspond to different orientations of neighboring benzene molecules in solution, e.g., T-shaped-like and parallel-displaced configurations are possible. Compared to experiment, the EFP2 RDF features are slightly more pronounced, suggesting that EFP2 over-structures

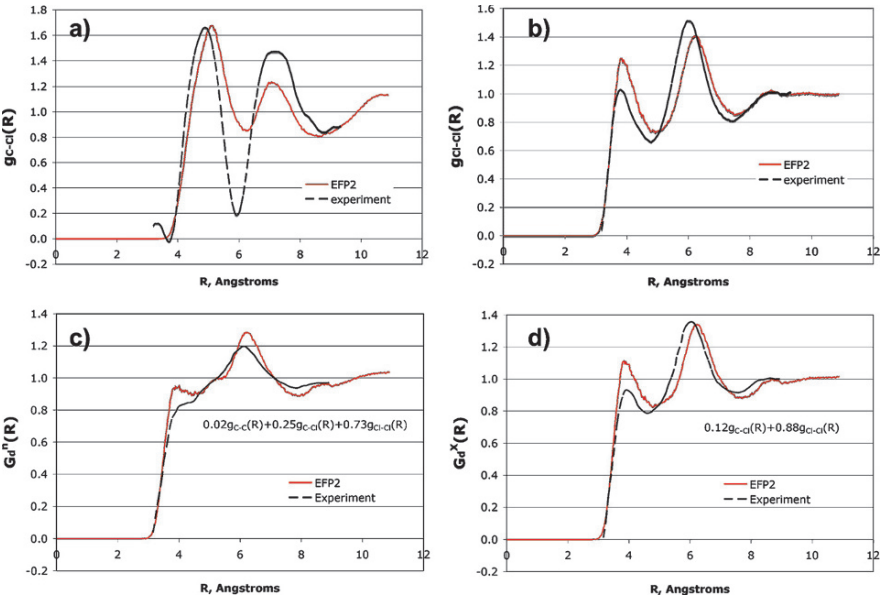

Figure 8-4. Comparison of EFP2 and experimental RDFs and G_d-functions for liquid carbon tetrachloride: (**a**) C−Cl, and (**b**) Cl−Cl RDFs, (**c**) $G_d{}^n$, (**d**) $G_d{}^X$

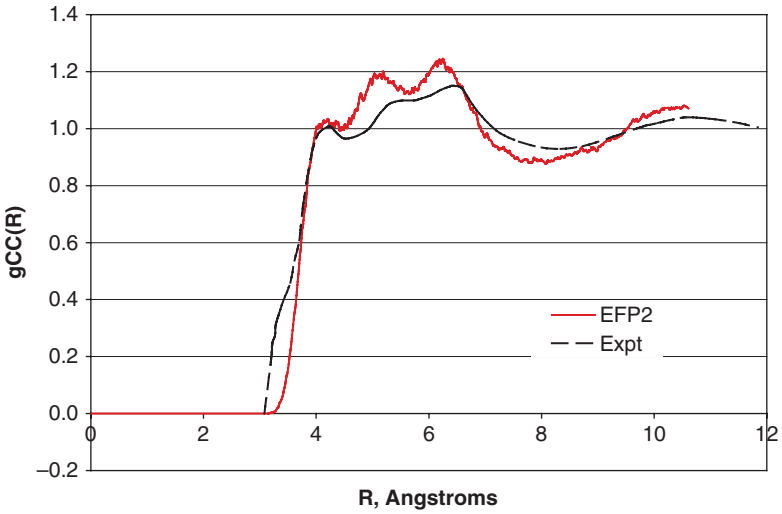

Figure 8-5. Comparison of EFP2 and experimental C−C RDFs for liquid benzene

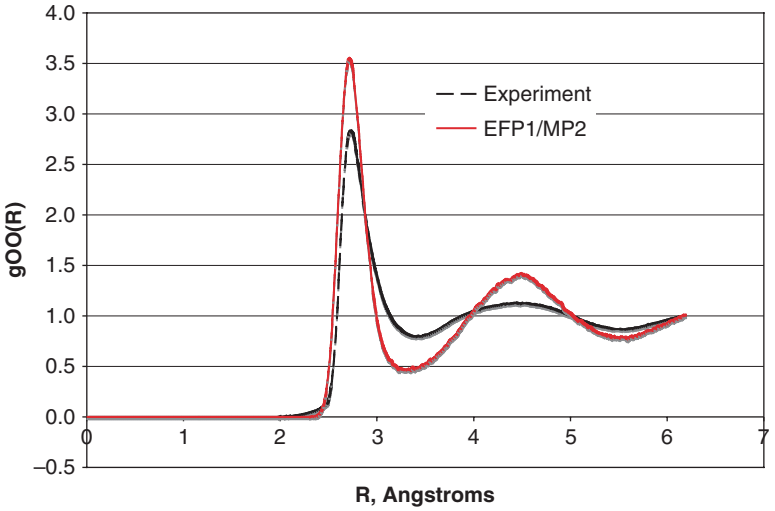

Figure 8-6. Comparison of EFP1/MP2 and experimental O—O RDFs for liquid water

liquid benzene. This may be due to the fact that EFP2 slightly overestimates the interactions between benzene molecules [42].

The EFP1/MP2 oxygen-oxygen RDF for water is shown in Figure 8-6. The positions of the peaks in the EFP1/MP2 and experimental RDFs are in excellent agreement [60], but the intensities of the EFP peaks are overestimated, i.e., EFP1/MP2 over-structures the water RDF. Some degree of over-structuring has been attributed to omitting quantum affects [61], although such affects are likely to be very small when no H atoms are involved. Over-structuring could arise due to intrinsic inaccuracies in the EFP1/MP2 potential, for example, water-water interactions that are too strong. A detailed analysis of the performance of different EFP models for liquid water can be found elsewhere [62].

8.4.2. Coarse-Graining

Because carbon tetrachloride is a spherically symmetric molecule, it is logical to coarse-grain it to its COM and represent it as a single point. The effective pair-force and pair-potential for the CCl_4 COM obtained from force matching are shown in Figure 8-7. The CG pair-potential was obtained by integrating the pair-force according to Eq. (8-10). The potential, *U(r)*, becomes sharply repulsive below r $\sim$ 4Å indicating that the CCl_4 excluded volume corresponds to the diameter $\sim$ 4Å. This is reasonable given that the C—Cl bond length in CCl_4 is 1.767Å and the excluded volume diameter should be slightly larger than twice the C—Cl bond length ($\approx$ 3.534 Å). Therefore, the force matching method has taken excluded volume into account. The potential *U(r)* is smooth and slowly varying with a wide minimum at r $\sim$ 7Å.

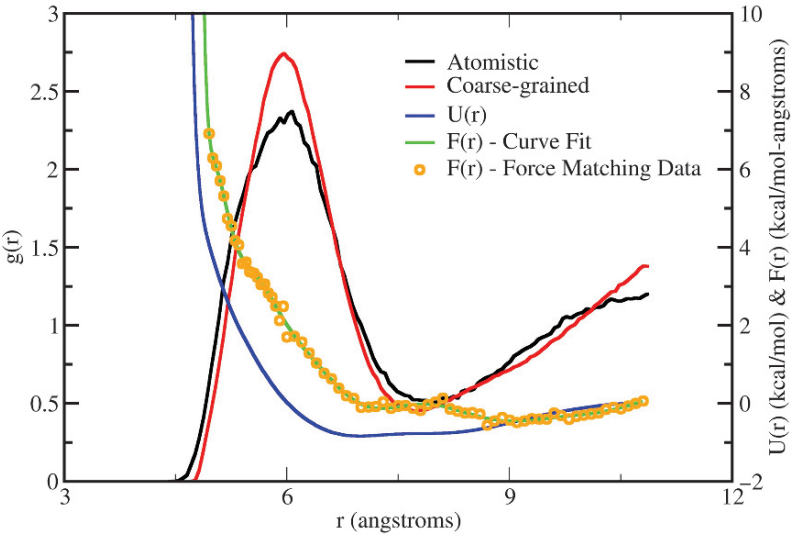

Figure 8-7. Coarse-graining of CCl_4. (**a**) *Black*: COM-COM RDF from EFP-MD, (**b**) *red*: RDF from CG-MD, (**c**) *orange circles*: the effective COM pair-force, (**d**) *green*: polynomial fit of the force matching data, (**e**) *blue*: the effective COM pair potential

The comparison of the CG and EFP RDFs clearly indicates that the coarse-grained potential is able to reproduce the liquid structure of CCl_4 reasonably well. The locations of the CG RDF peaks are in good agreement with those in the EFP RDF, although the CG peaks are a bit higher. This may be attributed to an overly steep repulsive CG pair-force (at $r \sim 4.5$ Å) used in the CG MD. There is an uncertainty about the nature of the repulsive CG pair-force at short distances, where the pair-force obtained from force-matching exhibits large unphysical fluctuations due to insufficient sampling of pairs at short separations ($r \sim 4.5$ Å) in the atomistic MD. Ignoring the data with large unphysical fluctuations in CG pair-force and replacing it with a fit through the remaining pair-force data may make the pair-force strongly repulsive at $r \sim 4.5$ Å as shown in Figure 8-7. This repulsion may be stronger than the repulsion in the corresponding EFP-MD. Strong repulsion at close separations ($r \sim 4.5$ Å) in the CG MD at the same density as the EFP-MD probably results in a more structured liquid, so sharper peaks are observed.

Favorable coarse-graining results for CCl_4 are not surprising because this molecule is spherically symmetric. Planar benzene presents a more stringent test for the force-matching approach. Figure 8-8 shows the effective pair-force and pair-potential for the benzene COM.

The locations of the peaks in the benzene CG RDF are in good agreement with the EFP RDF, but the heights of the peaks consistently exceed that of the EFP RDF. Moreover, the first peak in the CG RDF is not as broad as the first peak in the EFP RDF. This indicates that the coarse-grained pair-potential produces a more structured

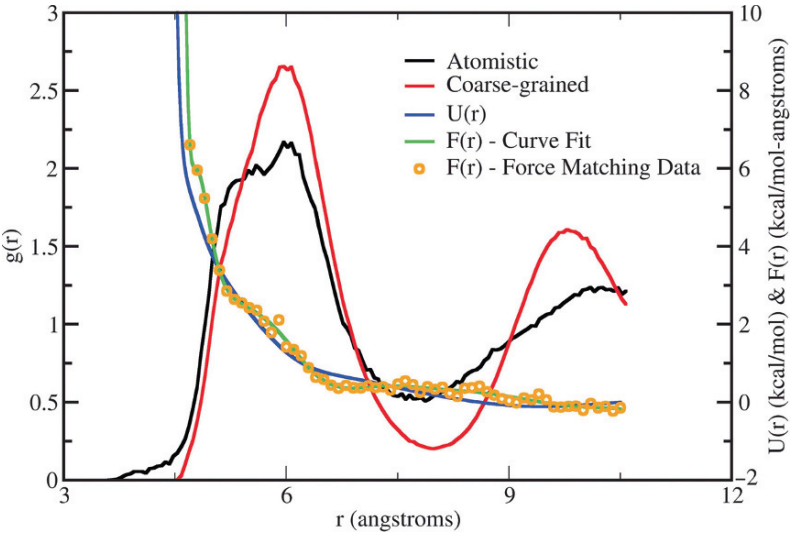

Figure 8-8. Coarse-graining of benzene. (**a**) *Black*: COM-COM RDF from EFP-MD, (**b**) *red*: RDF from CG-MD, (**c**) *orange circles*: the effective COM pair-force, (**d**) *green*: polynomial fit of the force matching data, (**e**) *blue*: the effective COM pair potential

liquid compared to that of the EFP. It is possible that these discrepancies arise due to the use of a system (64 molecules) that is too small and a 25 ps EFP MD run that is too short. This might result in inadequate configuration sampling. For example, as noted above for CCl_4, inadequate configuration sampling at short distances, r ~ 4–4.5 Å, the CG leads to unphysically large fluctuations in the CG pair force. Therefore the data in this range has to be neglected. This leads to the loss of information relating to the minimum energy parallel displaced benzene dimer configuration, see Figure 8-9(b). Fitting a curve using the remaining *f(r)* data makes the interaction at short distances more repulsive, resulting in a larger excluded volume and a narrower first peak. The 5.0 Å shoulder in the first peak of the EFP RDF can be associated with the T-shaped benzene dimer structure, Figure 8-9(a). Due to inadequate sampling and repulsion at short distances, the peak in the CG RDF lacks this shoulder and is narrower and higher than the corresponding EFP RDF peak. Additionally, due to the short EFP run, the coarse-grained potential has been averaged over only 150 sets, compared to 250 and 400 sets for water and CCl_4, respectively (Table 8-3). This is because the EFP MD simulations of liquid benzene are more computationally demanding than the simulations for water or CCl_4 (see Table 8-5).

Water is a very important and widely used solvent. Many of the unique properties of water are the result of the complex interactions that occur among water molecules. A water molecule is planar and highly polar, so it is an important system to test with the force matching approach. The coarse-graining results for water are summarized

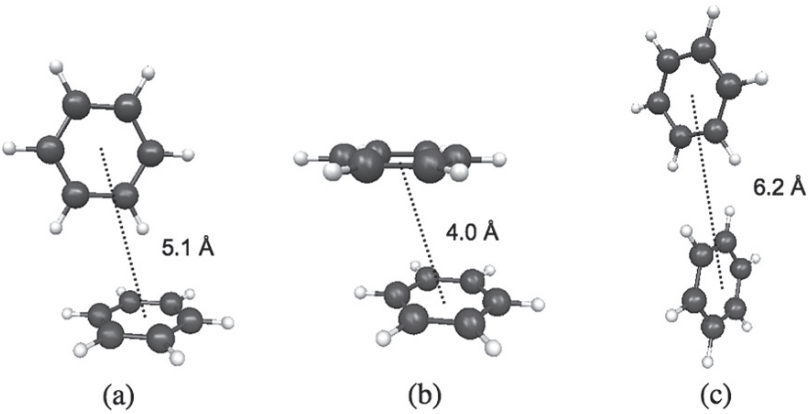

Figure 8-9. The minimum energy configurations of benzene dimer: (**a**) T-shaped, (**b**) parallel-diplaced, and (**c**) edge-to-edge structures

in Figure 8-10. The CG pair-force and pair potentials match qualitatively with the results reported in the literature [24].

The CG RDF is in reasonable agreement with the EFP MD COM-COM RDF. In particular, the first peak is in excellent agreement, while the second peak is slightly off (4.2 vs. 4.5 Å and slightly lower compared to the EFP RDF). After about 5 Å there is almost no structure in the CG RDF. Water is a complicated molecule to coarse-grain to a single site due to the presence of Van der Waals and coulombic interactions. Moreover, as it is a highly polar molecule, coarse-graining it to a single point at its COM may not be the best choice. Despite these shortcomings, the one-site coarse-grained potential is able to reproduce the first and second peaks, indicating that the force matching technique works reasonably well even for polar non-symmetric molecules.

The real advantage of coarse-graining is the speed up due to the reduction in the number of degrees of freedom and due to substituting a complex EFP potential by

Table 8-5. CPU time per timestep for EFP-MD and CG-MD simulations and CPU speed-up due to coarse-graining[a]

	EFP-MD		CG-MD		
	CPU time (s)	Timestep (fs)	CPU time (s)	Timestep (fs)	Speed-up[b]
CCl_4	24.80	0.3	0.000238	1.0	$3.47 \cdot 10^5$
Benzene	117.30	0.5	0.000249	1.0	$9.42 \cdot 10^5$
Water	1.03	0.3	0.000239	1.0	$1.29 \cdot 10^4$

[a] All CPU times reported are for MD simulations carried out on one AMD 280 Opteron 2.4 GHz processor.
[b] Speed-up = EFP-MD (CPU time/time step)/CG-MD (CPU time/time step)

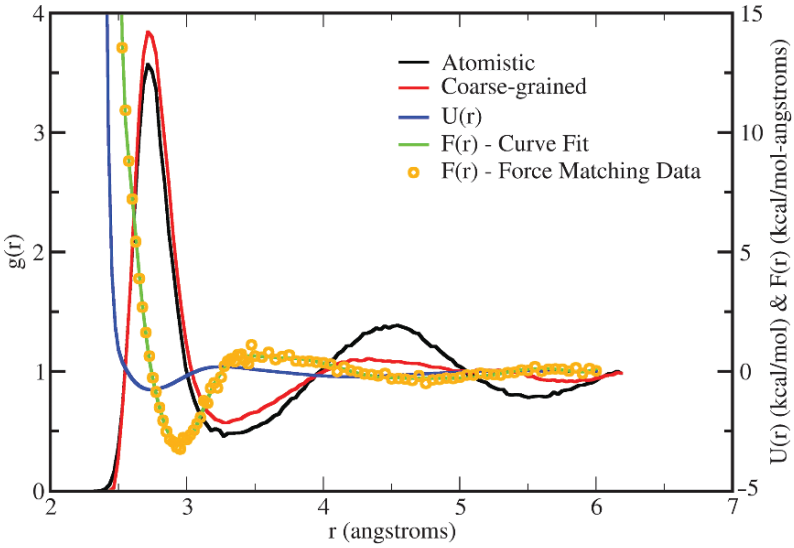

Figure 8-10. Coarse-graining of water. (**a**) *Black*: COM-COM RDF from EFP-MD, (**b**) *red*: RDF from CG-MD, (**c**) *orange circles*: the effective COM pair-force, (**d**) *green*: polynomial fit of the force matching data, (**e**) *blue*: the effective COM pair potential

a simpler polynomial one. The speed-ups achieved for the three studied systems are listed in Table 8-5. Additional speed-up can be achieved by using larger timestep for running CG-MD, which is reasonable to do because the CG potentials vary less rapidly with distance compared to underlying interaction potentials used in atomistic MD.

8.5. CONCLUSIONS

In the work presented here, the force matching technique is used to coarse-grain three typical solvents, carbon tetrachloride, benzene, and water, to their centers of mass. The accuracy of the force-matching is tested by comparing structural properties, namely, the radial distribution functions, of the underlying atomistic and coarse-grained systems. The atomistic MD simulations were performed using the effective fragment potential method. The EFP is a first-principles-based method designed for describing intermolecular interactions.

The RDFs for all three systems obtained from coarse-grained MD compare favorably with RDFs from EFP MD. Owing to its spherical symmetry, CCl_4 was found to be the most amenable to coarse-graining using the force matching method. For benzene, the coarse-grained MD produced a more structured liquid than that obtained with the atomistic MD. This might be attributed to a limited sampling of configuration space in the atomistic MD that is required for force

matching. This issue needs to be explored further. On the other hand, the coarse-grained RDF of water is in reasonably good agreement with the corresponding atomistic RDF.

The quality of coarse-grained potentials critically depends on the accuracy of the underlying atomistic MD from which the data required for force matching are generated. EFP, used for "atomistic" MD simulations in this work, is a promising technique for capturing the chemistry of liquids and solvents. Most of the previous applications of the EFP method focused on analysis of reactions and properties in complexes and clusters. This work presents the results of EFP MD simulations on liquid CCl_4, benzene, and water. In all cases, the EFP RDFs are in reasonable agreement with the available experimental data. EFP does tend to produce sharper peaks in RDFs, suggesting that some overstructuring of liquids may occur. For coarse-graining, the quality of the sampling of conformational space in an EFP MD simulation can be an issue. For example, more extensive sampling of conformational space in benzene could potentially improve the quality of its coarse-graining. Because it is a first principles-based technique, EFP is significantly more expensive than other force fields. This makes long EFP-MD simulations computationally demanding. These issues will be addressed in future work. Future contributions will also extend the methodology presented here to coarse-graining polymers, in order to study the mechanisms of their aggregation.

ACKNOWLEDGEMENT

This work was supported in part by a National Science Foundation NIRT grant to Iowa State University, and in part by a Department of Energy SciDAC (Scientific Discovery through Advanced Computing) grant administered by the Ames Laboratory. The authors are very grateful to Professor Greg Voth for enlightening discussions and advice.

REFERENCES

1. Allen MP, Tildesley DJ (1996) Computer simulations of liquids. Oxford University Press.,Oxford
2. Frenkel D, Smit B (2001) Understanding molecular simulations. Academic Press, San Diego.
3. Smit B, Esselink K, Hilbers PAJ, Vanos NM, Rupert LAM, Szleifer I (1993) Langmuir, 9(1):9–11
4. Chushak Y, Travesset A (2005) J Chem Phys 123(23)
5. Baschnagel J, Binder K, Paul W, Laso M, Suter UW, Batoulis I, Jilge W, Burger T (1991) J Chem Phys 95(8):6014–6025
6. Nielsen SO, Lopez CF, Srinivas G, Klein ML (2004) J Phys Condens Matter 16(15):R481–R512
7. Shelley JC, Shelley MY, Reeder RC, Bandyopadhyay S, Klein ML (2001) J Phys Chem B 105(19):4464–4470
8. Shelley JC, Shelley MY, Reeder RC, Bandyopadhyay S, Moore PB, Klein ML (2001) J Phys Chem B 105(40):9785–9792
9. Marrink SJ, de Vries AH, Mark AE (2004) J Phys Chem B 108(2):750–760

10. Marrink SJ, Mark AE (2003) J Am Chem Soc 125(49):15233–15242
11. Harmandaris VA, Adhikari NP, van der Vegt NFA, Kremer K (2006) Macromolecules 39(19): 6708–6719
12. Lyubartsev AP, Laaksonen A (1995) Phys Rev E 52(4):3730–3737
13. Lyubartsev AP, Laaksonen A (1999) J Chem Phys 111(24):11207–11215
14. Murtola T, Falck E, Patra M, Karttunen M, Vattulainen I (2004) J Chem Phys 121(18):9156–9165
15. Murtola T, Falck E, Karttunen M, Vattulainen I (2007) J Chem Phys 126(7):075101-1–075101-14
16. Soper AK (1996) Chem Phys 202(2–3):295–306
17. Reith D, Putz M, Muller-Plathe F (2003) J Comput Chem 24(13):1624–1636
18. Tschop W, Kremer K, Batoulis J, Burger T, Hahn O (1998) Acta Polymerica 49(2–3):61–74
19. Li XJ, Kou DZ, Rao SL, Liang HJ (2006) J Chem Phys 124(20):204909-1–204909-7
20. Li XJ, Ma XJ, Huang L, Liang HJ (2005) Polymer 46(17):6507–6512
21. Depa PK, Maranas JK (2005) J Chem Phys 123(9):094901-1–094901-7
22. Depa PK, Maranas JK (2007) J Chem Phys 126(5):054903-1–054903-8
23. Izvekov S, Parrinello M, Burnham CJ, Voth GA (2004) J Chem Phys 120(23):10896–10913
24. Izvekov S, Voth GA (2005) J Chem Phys 123(13):134105-1–134105-13
25. Izvekov S, Voth GA (2005) J Phys Chem B 109(14):6573–6586
26. Wang YT, Izvekov S, Yan TY, Voth GA (2006) J Phys Chem B 110(8):3564–3575
27. Wang YT, Voth GA (2006) J Phys Chem B 110(37):18601–18608
28. Violi A, Voth GA (2005) In High Performance Computing and Communications, Proceedings, pp 938–947
29. Izvekov S, Violi A, Voth GA (2005) J Phys Chem B 109(36):17019–17024
30. Izvekov S, Voth GA (2005) J Phys Chem B 109(7):2469–2473
31. Izvekov S, Voth GA (2006) J Chem Theory Comput 2(3):637–648
32. Moller C, Plesset S (1934) Phys Rev 46:618
33. Raghavachari K, Trucks GW, Pople JA, Head-Gordon M (1989) Chem Phys Lett 157(6): 479–483
34. Gordon MS, Freitag MA, Bandyopadhyay P, Jensen JH, Kairys V, Stevens WJ (2001) J Phys Chem A 105(2):293–307
35. Jensen JH, Day PN, Gordon MS, Basch H, Cohen D, Garmer DR, Kraus M, Stevens WJ (1994) Modeling the Hydrogen Bond 569:139–151
36. Gordon MS, Slipchenko LV, Li H, Jensen JH (2007) Ann Rep Comp Chem 3:177–193
37. Adamovic I, Freitag MA, Gordon MS (2003) J Chem Phys 118(15):6725–6732
38. Song J, Gordon MS unpublished
39. Stone AJ (1981) Chem Phys Lett 83(2):233–239
40. Stone AJ (1996) The theory of intermolecular forces, Oxford University Press, Oxford
41. Freitag MA, Gordon MS, Jensen JH, Stevens WJ (2000) J Chem Phys 112(17):7300–7306
42. Slipchenko LV, Gordon MS (2007) J Comput Chem 28(1):276–291
43. Slipchenko LV, Gordon MS unpublished results
44. Jensen JH, Gordon MS (1996) Mol Phys 89(5):1313–1325
45. Jensen JH, Gordon MS (1998) J Chem Phys 108(12):4772–4782
46. Adamovic I, Gordon MS (2005) Mol Phys 103(2–3):379–387
47. Tang KT, Toennies JP (1984) J Chem Phys 80(8):3726–3741
48. William H Press SAT, William T Vetterling, Brian P Flannery (2002) Numerical recipes: the art of scientific computing, Cambridge University Press, Cambridge
49. Gordon MS, Schmidt MW (2005) In: Dykstra CE, Frenking G, Kim KS, Scuseria GE (eds) Theory and applications of computational chemistry, Ch 41, Elsevier, Amsterdam

50. Schmidt MW, Baldridge KK, Boatz JA, Elbert ST, Gordon MS, Jensen JH, Koseki S, Matsunaga N, Nguyen KA, Su SJ, Windus TL, Dupuis M, Montgomery JA (1993) J Comput Chem 14(11): 1347–1363
51. Woon DE, Dunning TH (1993) J Chem Phys 98(2):1358–1371
52. Hariharan PC, Pople JA (1973) Theoretica Chimica Acta 28(3):213–222
53. Krishnan R, Binkley JS, Seeger R, Pople JA (1980) J Chem Phys 72(1):650–654
54. Clark T, Chandrasekhar J, Spitznagel GW, Schleyer PV (1983) J Comput Chem 4(3):294–301
55. Li H, Netzloff HM, Gordon MS (2006) J Chem Phys 125(19):194103-1–194103-9
56. Plimpton S (1995) J Comput Phys 117(1):1–19
57. Narten AH (1976) J Chem Phys 65(2):573–579
58. Steinhauser O, Neumann M (1980) Mol Phys 40(1):115–128
59. Narten AH (1977) J Chem Phys 67(5):2102–2108
60. Sorenson JM, Hura G, Glaeser RM, Head-Gordon T (2000) J Chem Phys 113(20):9149–9161
61. Allesch M, Schwegler E, Gygi F, Galli G (2004) J Chem Phys 120(11):5192–5198
62. Netzloff HM, Gordon MS (2004) J Chem Phys 121(6):2711–2714