THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

Critical Analysis of the Accuracy of Models Predicting or Extracting Liquid Structure Information

Marc Van Houteghem, An Rita Ghysels, Toon Verstraelen, Ward Poelmans, Michel Waroquier, and Veronique Van Speybroeck

J. Phys. Chem. B, **Just Accepted Manuscript** • DOI: 10.1021/jp411737s • Publication Date (Web): 10 Feb 2014

Critical Analysis of The Accuracy of Models Predicting or Extracting Liquid Structure Information

Marc Van Houteghem, An Ghysels, Toon Verstraelen, Ward Poelmans, Michel Waroquier, Veronique Van Speybroeck*

Center for Molecular Modeling, QCMM Alliance Ghent-Brussels, Ghent University, Technologiepark 903, B-9052 Zwijnaarde, Belgium

E-mail: veronique.vanspeybroeck@ugent.be

*To whom correspondence should be addressed

Abstract

This work aims at a critical assessment of properties predicting or extracting information on the density and structure of liquids. State-of-the-art NVT and NpT molecular dynamics (MD) simulations have been performed on five liquids: methanol, chloroform, acetonitrile, tetrahydrofuran and ethanol. These simulations allow the computation of properties based on first principles, including the equilibrium density and radial distribution functions (RDFs), characterizing the liquid structure. Refinements have been incorporated in the MD simulations by taking into account Basis Set Superposition Errors (BSSE). An extended BSSE model for an instantaneous evaluation of the BSSE corrections has been proposed, and their impact on the liquid properties has been assessed. If available, the theoretical RDFs have been compared with the experimentally derived RDFs. For some liquids significant discrepancies have been observed and a profound but critical investigation is presented to unravel the origin of these deficiencies. This discussion is focused on tetrahydrofuran where the experiment reveals some prominent peaks completely missing in any MD simulation. Experiments providing information on liquid structure consist mainly of neutron diffraction measurements offering total structure factors as the primary observables. The splitting of these factors in reciprocal space into intra- and intermolecular contributions is extensively discussed, together with their sensitivity in reproducing correct RDFs in coordinate space.

Keywords

radial distribution function, liquid density, molecular dynamics, NpT, structure factor, BSSE

Introduction

The structure of liquids is a recurring theme in the scientific literature. Liquids do not have a structure in the sense that crystals have a structure. The forces between liquid molecules are

not as strong as those holding solids together, but they show some attractive character arising from e.g. hydrogen bonding. The intermolecular forces lead to characteristic correlations, which determine the structure of the liquid and which affect many physical (thermodynamical) properties. It is important to know the microscopic structure of liquids in as much detail as possible because it determines their macroscopic properties, e.g. thermodynamic functions.

Characteristics of liquid structure are mainly governed by a repertoire of intermolecular interactions, of which hydrogen bonding constitutes the main class if present. Water is by far the most studied among the hydrogen-bonded molecules.[1–8] The liquid phase of methanol, consisting of a hydrogen-bonding hydroxyl group and a hydrophobic methyl group, is also extensively studied as it is the smallest alcohol which can be studied to characterize the hydrogen bonding in alcohols.[9–12]

Neutron diffraction measurements with isotopic substitution (NDIS) are one of the most powerful methods to extract experimental information on liquid structure. The differences in the experimental scattering patterns by replacing a particular chemical element by its isotopes (e.g. hydrogen by deuterium) in the solvent molecule enable the extraction of structure factors (SFs) determining the differential cross section of the neutron scattering process.[2,13–18] However, the interpretation of the experimental SFs is a difficult task as the measured scattering intensities are the result of a weighted summation of different pairwise spatial correlation functions formed by all atoms in the system. Disentangling the various correlations into partial SFs from the total SF remains challenging, as illustrated by the scheme in Fig. 1. In particular the separation of the experimental total SF into intramolecular and intermolecular contributions is complex even with the NDIS methodology. The total structure factor is composed of partial structure factors reflecting the pairwise spatial correlations between different atom types with multiplicative prefactors containing concentrations and scattering lengths of each atom type. In case of a solute and a solvent, e.g. aqueous solution of a polyatomic organic molecule, the atomic concentrations in the solution

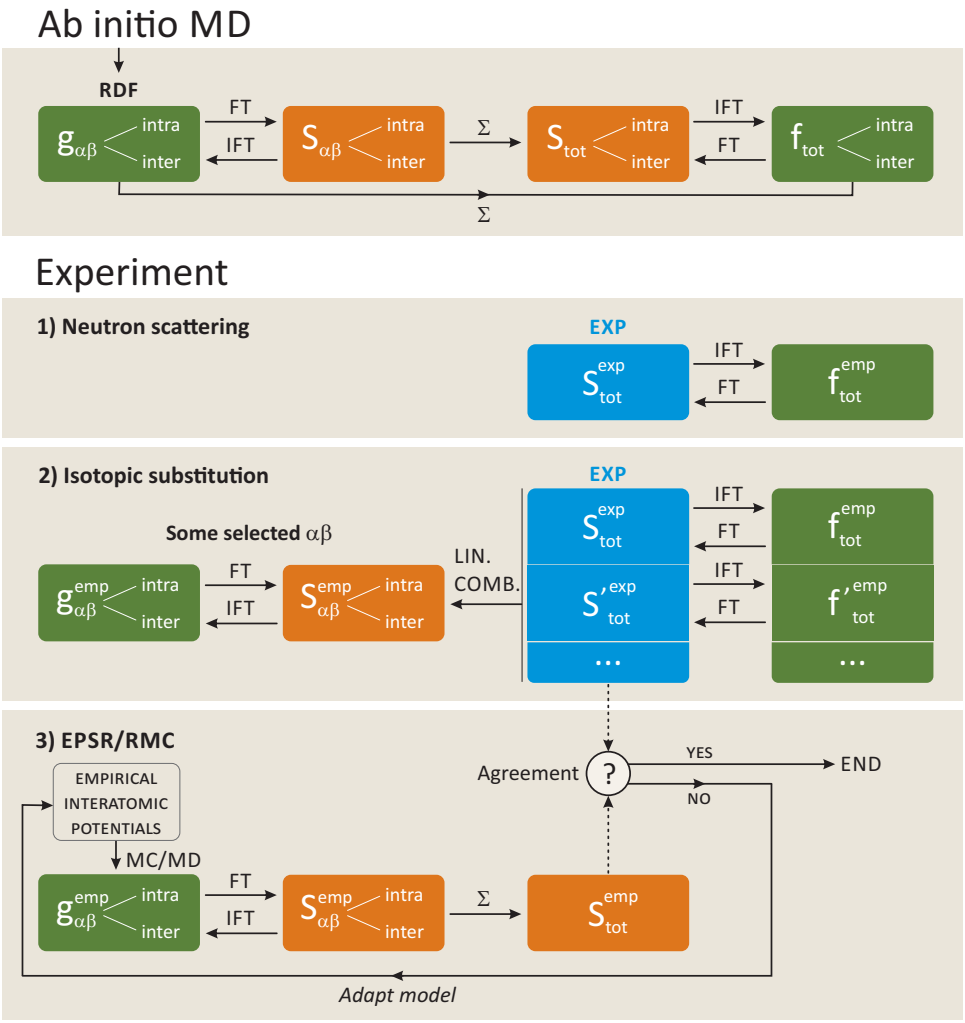


Figure 1: Procedure to derive structural quantities from ab initio molecular dynamics and from experiment. FT (IFT) refers to the (inverse) Fourier transform. Partial quantities carry the atom types $\alpha\beta$ in the subindex, while the subindex "tot" refers to total quantities. Each RDF, SF, or correlation function consists of intermolecular and intramolecular contributions. Properties in the blue boxes are measured *directly* in the experiment (and have the subscript "exp"). Properties in coordinate space (RDFs or weighted summations of RDFs) are in green boxes, and those in reciprocal space are in orange boxes. Properties can be obtained from 64simulations (no subscript) or can be *derived* from experiment (subscript "emp"). Conversions are described in the text.

can be handled as an additional degree of freedom. While the intermolecular contribution varies with concentration, the intramolecular contribution does not change. Exploiting this concentration invariance, NDIS experiments with different concentrations should in principle be able to extract intermolecular partial SFs.[16] This method has its benefits, but it can not be applied to liquids composed of one type of molecules.

There are often not enough different scattering experiments available to extract all information needed to completely characterize the liquid. Experimentalists have circumvented the incompleteness of the experimental data to extract the partial SFs by combining the experimental measurements with additional Monte Carlo simulations based on intermolecular site-site potentials.[2,10,19,20] Structure refinements are proposed as to reproduce experimental data as well as possible. The method is empirical in the sense that perturbations are induced iteratively to the various potentials until satisfactory agreement with the experimental data is reached, as shown in the scheme in Fig. 1. This methodology is called Empirical Potential Structure Refinement (EPSR) model and is closely related to the original reverse Monte Carlo simulation technique (RMC) as originally introduced by McGreevy and Pusztai.[21,22] RMC is a structure modeling method and produces a structure that is consistent with the experimental data. Both RMC and EPSR are inverse methods of structural modeling, in which building and refining of the particle configurations during the course of the simulations form an essential ingredient. Related methods, where RMC modeling is combined with MD computer simulations based on interatomic potentials, have also been applied by other authors on alcohols and haloforms.[10,11]

In the above RMC-EPSR iterative approaches the intermolecular radial distribution function (RDF) plays a central role. It is a quantity determined from intermolecular forces and forms a bridge between the microscopic structure of the liquid and macroscopic properties such as internal energy, entropy and pressure. The partial and total RDFs are however not direct observables in an experiment, as already mentioned. Hence, one needs to be cautious when using them as reference quantities when judging the quality of theoretical predictions.

Therefore, we will call these quantities "empirical", meaning that they are derived from the experimental total SF using some fitting or Monte Carlo procedure.

Even with the help of three-dimensional models of the liquid structure with empirical parameters which can be adjusted to reproduce the experimental SFs, the procedure is prone to large inaccuracies/errors. The numerical instabilities become even strengthened when inverting the SFs to real space distribution functions, as the inverse Fourier transform is error-prone.[22–25] These are serious deficiencies, and the results are not free from any bias. In a very interesting and recent review of Soper, most of the problems in data acquisition and data treatment have been reported.[2] This paper further investigates the deficiencies by comparing them with ab initio simulations.

Molecular dynamics (MD) simulations offer the appropriate theoretical tools to compute RDFs directly (Fig. 1). They can be performed with force fields[10,11,16,26,27] or based on density functional theory.[14,28,29] MD modeling has proven to be very effective in guiding assignments of peaks in these empirical RDFs. In this work we will compute theoretical (partial) RDFs and their counterparts, the (partial) SFs, which are connected by a Fourier transform. The structure of five liquids will be investigated with DFT-based MD simulations: methanol (MeOH), chloroform (TCM), acetonitrile (MeCN), tetrahydrofuran (THF) and ethanol (EtOH). The availability of first principles MD simulation studies on these liquids to derive their liquid structure is less abundant in literature than for water. The choice of the five solvents were based on (i) their importance in industry, and (ii) their classification: non-polar (TCM), polar aprotic (MeCN, THF), and polar protic (MeOH and EtOH).

A striking example is tetrahydrofuran, where the empirical intermolecular partial RDFs show sharp peaks at short distances,[19] which are missing in the MD simulations performed in this paper and in refs.[30] and.[31] This apparent discrepancy between theory and the empirical EPSR model requires a profound and thorough investigation to find the origin of the appearance of these sharp peaks in the experimental RDF of THF, and why theory based on first principles methods fails in reproducing it. This discussion should be extended to

a broader context as similar, however less pronounced, discrepancies are also observed in other liquids. In order to unravel this apparent mismatch between theory and experiment, it is better to focus on the SFs and to perform the assessment in reciprocal space than in real coordinate space. A Fourier transform to convert theoretical RDFs to theoretical SFs requires some mathematical manipulations but the accuracy of these computations is better controlled, than the extraction of empirical RDFs. The discrepancy will be discussed in Section .

Moreover, we will also investigate the impact of intramolecular contributions on the SFs, which experimentally poses a lot of complications.[16] Theoretically, separation of the intra- and intermolecular contributions of the various properties is perfectly feasible, and may assist in finding a correct interpretation of the features observed in the experimental SF.

In DFT-MD simulations, it has been shown that (i) dispersion and (ii) the basis set superposition error (BSSE) have a significant effect on the RDFs of water and methanol.

(i) Standard DFT methods do not account for dispersion interactions.[32,33] Dispersion is a (mostly attractive) long-range interaction due to the correlated electron dynamics in two well-separated parts of the wavefunction. Schmidt *et al.*[34] have found that adding the Grimme dispersion correction[35,36] improves the predicted densities of water in the isothermal-isobaric ensemble (NpT), for both BLYP[37,38] and PBE[39] functionals, compared to experiment. In addition, the first and second maxima of the oxygen-oxygen RDF of water using the BLYP-D functional instead of BLYP lead to an improved agreement with the experimental data, although the RDF for PBE-D still remains overstructured.[34]

(ii) The BSSE results from an inconsistent description of the molecular orbitals as a result of using a limited number of localized basis functions (incomplete basis set).[40] When two molecules are far apart, they can only "use" their own basis functions. When the intermolecular distance decreases, they may use each other's basis, causing an artificial strengthening of the intermolecular interactions and artificial shortening

of intermolecular distances due to the overlap of basis functions. In ref.[29] a systematic correction of the BSSE with an exponentially decaying parameterized pairwise classical force-field energy term has been proposed by the authors that can be added in subsequent MD simulations. The model was applied to methanol using MD in a canonical ensemble (NVT) with the BLYP-D functional. It was found that upon introduction of this correction, the errors on the first peak height of the relevant partial RDFs of the atoms that participate in hydrogen-bonding are reduced from 12%-16% to 0.4%-1%, compared to experiment. Another conclusion is that, for methanol, the Grimme-dispersion correction and the BSSE have the same order of magnitude but an opposite sign. Recently, Grimme proposed a semi-empirical counterpoise-type correction in the form of an atom pair-wise potential which corrects for the inter- and intra-molecular BSSE in supermolecular Hartree-Fock or periodic DFT calculations. This geometrical counterpoise correction scheme depends on the molecular geometry only.[41,42]

This paper extends the work in ref.[29] in the sense that (i) MD simulations in the NpT ensemble have been performed on the five liquids under consideration, (ii) the BSSE correction scheme as proposed in ref.[29] has been extended to all intermolecular interactions.

(i) The isothermal-isobaric ensemble (NpT) is considered to allow density fluctuations of the liquid. An RDF is directly related to the density of the molecular system, so it is vital that the density in MD simulations is properly reproduced. In ref.[29] the NVT ensemble was considered.

(ii) The BSSE-correction model proposed in ref.[29] is an exponentially decaying parametrized classical energy term for all intermolecular interactions. In ref.[29] one assumed a dominant BSSE contribution from the intermolecular oxygen-hydrogen interaction in methanol. This assumption is no longer applicable for the set of liquids under study. Therefore, in this work *all* the intermolecular interaction terms are taken into account

in the parametrisation of the correction model for BSSE.

The following section outlines the technical details of the quantum MD simulations and the computational methodology to calculate the RDF, SF and BSSE. Section reviews shortly the BSSE correction model and its extension, i.e. its mathematical form and the parameter calibration. The BSSE-correction model with calibrated parameters is then applied to a new series of NpT MD-simulations. Section presents the results of the BSSE computations, MD-runs, RDFs and SFs with the necessary discussion. Where experimental data is available, comparison with the experiment is made. In the last section the most relevant conclusions are summarized.

Computational section

Ab initio MD

The same computational settings are used as in ref.[29] apart from some convergence criteria. All simulations are performed with the CP2K code.[43] For the ab initio MD, the Quickstep module of CP2K[44] is used, employing the hybrid Gaussian and plane-wave (GPW) density functional method with a BLYP gradient-corrected functional. The core electron states are represented by the norm conserving Goedecker-Teter-Hutter (GTH) pseudopotential.[5,45,46] The GPW basis set consists of a triple zeta TZVP Gaussian type orbital basis for the real space representation and an auxiliary plane wave basis to compute the long-range periodic electrostatic interactions in the reciprocal space.[23]

It is known from the work of McGrath *et al.*[24] that the convergence of the volume requires a significantly higher auxiliary basis set cut-off compared to constant-volume simulations.[24,47] Therefore, the NpT and NVT MD runs are performed with the 800 Ry plane wave cut-off. An exception is the NVT simulation of methanol, where we use the same settings of ref.[29] with a 400 Ry cut-off (with Grimme-dispersion,[36] without correction for BSSE[40]).

Liquid MeOH, TCM, MeCN, and EtOH are modeled at room temperature in a periodic

cubic simulation cell containing 129 molecules. Periodic boundary conditions are employed using the minimum image convention. Given the size of THF molecules, the unit cell for this molecular system contains 67 molecules. The simulation cell sizes for the NVT-simulations are chosen to correspond with the experimental densities ρ_{exp} of the solvents, as summarized in Table 1. The system sizes are assumed to be large enough, in view of sensitivity analyses as a function of MD simulation parameters published in literature[28,34] . The NVT/DFT-D3 MD production runs of MeOH, TCM, MeCN, THF and EtOH amount to 8.1 ps. The NpT/DFT-D3 and NpT/DFT-D3 + CP production runs (defined below) are carried out for at least 12.5 – 15.0 ps and 16.0 – 19.9 ps respectively. The exact lengths of the MD runs done for this work are given in Table S.1 of the supporting information (SI). The simulations are carried out at a temperature of 300 K with a Nosé-Hoover thermostat[53] with time constant 1 ps: the isobaric-isothermal ensemble is implemented in CP2K making use of the Martyna-Tuckerman-Tobias-Klein (MTTK) algorithm.[54,55] The stress tensor is implemented as described by Schmidt *et al.*[34] The external pressure is 1 bar and the time constant of the barostat is 1 ps. The MTTK integrator is used with an integration time step of 1 fs.

Table 1: Experimental densities (at 298 K) and cell parameters for the NVT simulations for all studied solvents.

solvent	ρ_{exp} [g/cm^3]	cell parameter [Å]
MeOH	0.787[d]	20.583
TCM	1.483[b,c]	25.865
MeCN	0.777[a]	22.445
THF	0.881[e]	20.882
EtOH	0.785[d]	23.248

[a]ref.,[48] [b]ref.[49] [c]ref.,[50] [d]ref.,[51] [e]ref.[52]

Once calibrated parameters are obtained for the classical BSSE correction energy term (see Section), a new MD simulation is performed with a combination of the Quickstep and the molecular mechanics module of CP2K (FIST). The total energy is calculated at each timestep as the sum of the DFT-energy and the classical force-field energy.

As such we perform three types of MD simulations per solvent:

- NVT/DFT-D3: dispersion-corrected DFT using the DFT-D3 scheme in the NVT-ensemble
- NpT/DFT-D3: dispersion-corrected DFT using the DFT-D3 scheme in the NpT-ensemble
- NpT/DFT-D3 + CP: dispersion- and BSSE-corrected DFT using the DFT-D3 scheme and the counterpoise (CP) correction scheme, as explained in Section

To reduce the amount of equilibration time, three measures have been taken: (i) all initial structures are generated by Packmol[56] (a software package especially developed for building initial configurations for MD simulations), (ii) all MD runs are preceded by equilibration NVT-simulations, (iii) the positions at the start of the production runs of the NpT/DFT-D3 simulations are taken as the starting geometry for the NpT/DFT-D3 + CP runs.

Radial distribution function and structure factor

The RDF is the number of particle pairs $\mathrm{d}n(r)$ in the spherical shell with radius r and $r+\mathrm{d}r$, relative to the number of particle pairs $\mathrm{d}n_0(r)$ in an uncorrelated ideal gas:

$$g(r) = \frac{\mathrm{d}n(r)}{\mathrm{d}n_0(r)}, \tag{1}$$

where r is the interparticle distance. In practice, the space around a given atom is discretized in concentric spherical shells with width δr. The number of atoms n_i in each shell is counted during the MD run, such that a histogram of the n_i values can be built. The RDF may be decomposed in an intramolecular and intermolecular contribution, $g(r) = g^{\mathrm{inter}}(r) + g^{\mathrm{intra}}(r)$, with limits $\lim_{r\to\infty} g^{\mathrm{inter}}(r) = 1$ and $\lim_{r\to\infty} g^{\mathrm{intra}}(r) = 0$.

The RDF and SF are connected through a three-dimensional Fourier transform. If $g_{\alpha\beta}(r)$ is the partial pair correlation function of atom types α and β, then the partial SF $S_{\alpha\beta}$ may

be expressed as:

$$\begin{aligned} S_{\alpha\beta}(k) - 1 &= \rho \int \exp(i\boldsymbol{k}\cdot\boldsymbol{r})(g_{\alpha\beta}(r) - 1)d\boldsymbol{r} \\ &= 4\pi\rho \int_0^\infty \frac{\sin(kr)}{kr}(g_{\alpha\beta}(r) - 1)r^2 \mathrm{d}r. \end{aligned} \tag{2}$$

Here, ρ is the average particle number density, while in the second equality the angle dependence is integrated out in the understanding that $g_{\alpha\beta}(r)$ is isotropic in space. The total structure factor $S_{\mathrm{tot}}(k)$ is a weighted sum of the partial SFs, which for neutron diffraction reads as:

$$S_{\mathrm{tot}}(k) = \sum_{\alpha\leq\beta} (2 - \delta_{\alpha\beta})\, c_\alpha b_\alpha c_\beta b_\beta \left[S_{\alpha\beta}(k) - 1\right]. \tag{3}$$

k gets a physical significance in the sense that it represents the momentum transfer in a neutron scattering experiment. The weighting factors are composed of the coherent scattering lengths b_α which characterize the interaction strength between atoms and incident radiation and which can vary from isotope to isotope, and the atomic fractions $c_\alpha = N_\alpha/N$ with N_α the number of type α atoms, and N the total number of atoms in the system. The Kronecker $\delta_{\alpha\beta}$ avoids double counting of pairs of atoms of the same type. $S_{\mathrm{tot}}(k)$ is not only a direct experimental observable, but also directly accessible by theoretical simulations, in particular molecular dynamics, from Eqs. (2) and (3). All other quantities, including the partial SFs, are not directly available from experiment. In an indirect way, however, they may be derived from the experimental total SFs obtained by changing the isotopic composition of the sample,[19] but this procedure is ill-conditioned for a liquid with one type of molecule.

By means of an inverse Fourier transform,

$$\begin{aligned} f\,(r) &= \frac{1}{\rho}\left(\frac{1}{2\pi}\right)^3 \int \exp(-i\boldsymbol{k}\cdot\boldsymbol{r}) S_{\mathrm{tot}}(k) d\boldsymbol{k} \\ &= \frac{1}{\rho}\left(\frac{1}{2\pi^2}\right) \int_0^\infty \frac{\sin(kr)}{kr} S_{\mathrm{tot}}(k) k^2 \mathrm{d}k \qquad (4) \\ &= \sum_{\alpha\leq\beta} (2 - \delta_{\alpha\beta}) c_\alpha b_\alpha c_\beta b_\beta (g_{\alpha\beta}(r) - 1), \qquad (5) \end{aligned}$$

$S_{\mathrm{tot}}(k)$ is converted to some total pair correlation function or total distribution function $f(r)$ without clear physical interpretation. Nevertheless, the equivalence of Eq. (4) and Eq. (5) can be exploited to verify the numerical accuracy of the numerical Fourier transform integrations (first equation) versus the direct computation of $f(r)$ from the partial RDFs (last equation). Note that $\lim_{r\to\infty} f(r) = 0$ since $\lim_{r\to\infty} g_{\alpha\beta}(r) = 1$.

Experimental total SFs are not free from contributions arising from intramolecular scattering (neutrons can scatter from atoms belonging to the same molecule). On the contrary, they even dominate the entire behavior of the experimental total SF. From a physical point of view, in unraveling the liquid structure, intramolecular contributions should be projected out as only intermolecular terms are of physical interest. This is no easy task from an experimental point of view. In the tetrahydrofuran liquid, for instance, intramolecular distances of up to 4.2 Å occur between hydrogen atoms, which overlaps with typical intermolecular distances in the RDF. Theory can help in estimating the impact of intramolecular parts on the various quantities, and in particular those which are inaccessible by experiment.

Counterpoise correction for BSSE

The hybrid Gaussian and plane-wave DFT method uses localized Gaussian basis sets that are centered at the nuclei of the atoms, of which relatively few basis functions are required to describe molecular orbitals. A downside of the method however is the BSSE that arises due to this limited number of basis functions, and thus incomplete basis sets.

The most widely employed method to correct for BSSE is the counterpoise (CP) method as introduced by Boys and Bernardi.[40] If a complex AB is considered, they estimated the artificial stabilization for monomer A as the energy shift ΔE_{A} when the basis for chemical system A is augmented with basis set functions of monomer B (which are located at the monomer B position), i.e. if A approaches monomer B close enough so that A can utilize the basis functions of B, and similarly for ΔE_{B}.[57] These estimated errors are then subtracted from the uncorrected interaction energy $\Delta E^{int}(\mathrm{AB})$ of a complex AB to obtain the

counterpoise-corrected interaction energy:

$$\Delta E^{\mathrm{int}}_{\mathrm{CP}}(\mathrm{AB}) = \Delta E^{\mathrm{int}}(\mathrm{AB}) - \Delta E_{\mathrm{A}} - \Delta E_{\mathrm{B}} = \Delta E^{\mathrm{int}}(\mathrm{AB}) + \Delta E_{\mathrm{CP}}, \tag{6}$$

where ΔE_{CP} is the so-called counterpoise correction. The CP-computations are carried out on pairs of molecules.

Extended force field correction model

CP model

This section develops the CP correction model for ΔE_{CP} and the calibration method for its parameters. The aim is to correct for the BSSE by adding a force field correction term $\Delta E^{\mathrm{mod}}_{\mathrm{CP}}$ to the quantum mechanical energy. In summary, the procedure comprises three stages. First, a molecular dynamics simulation is run with the BLYP functional to create a set of liquid configurations. Second, molecule pairs are selected from the configurations and the BSSE is calculation using the BLYP functional, thus creating a set of reference data $\Delta E^{\mathrm{ref}}_{\mathrm{CP}}$. The parameters of the force field correction model $\Delta E^{\mathrm{mod}}_{\mathrm{CP}}$ are determined by fitting to these reference data as closely as possible. Third, a new molecular dynamics simulation is run, including this correction model. The additional forces originating from the $\Delta E^{\mathrm{mod}}_{\mathrm{CP}}$ term are evaluated on the fly when integrating the equations of motions. The result is an MD simulation that is corrected for BSSE.

The model of ref.[29] is now briefly reviewed, and extensions, new in this work, are clarified. The proposed CP model is an additive pairwise interaction between certain atom types in two different molecules.[29] The functional form of each term is an exponential Ae^{-Br}, decaying with the separation distance r between two atoms, because BSSE is necessarily of repulsive nature ($A > 0$). The pairwise interaction is different for each pair of atom types. Fig. 2 gives the nomenclature of the atoms types of the five molecules considered in this work.

Each pair of atom types is denoted following these labels throughout the remainder of the paper. Interaction terms are only present between intermolecular atom pairs. With each

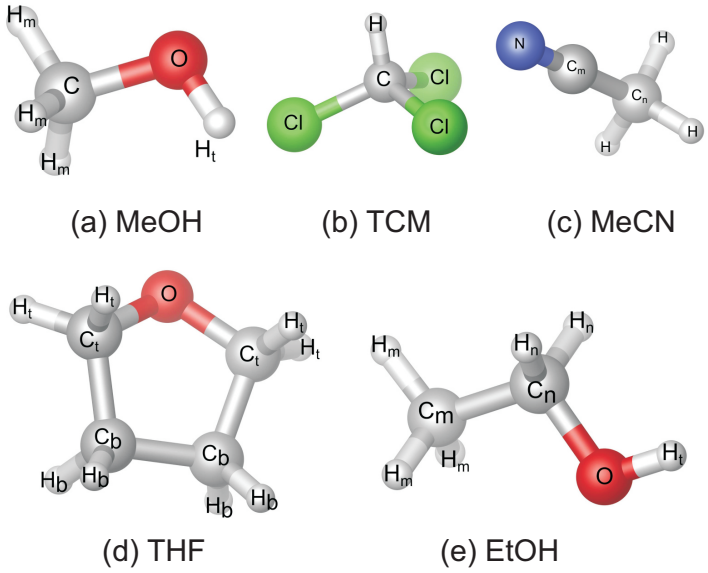


Figure 2: Atom types for the solvents MeOH, TCM, MeCN, THF, EtOH.

pair of atom types t, a unique set of parameters (A_t, B_t) is associated:

$$\Delta E_{\mathrm{CP}}^{\mathrm{mod}} = \sum_{\mu=1}^{N-1} \sum_{\nu=\mu+1}^{N} \sum_{\substack{i \in M_\nu \\ j \in M_\mu}} A_{t(ij)} \exp\left(-B_{t(ij)} r_{ij}\right), \tag{7}$$

where r_{ij} is the distance between the atoms i and j, M_ν is a set of atomic indices of molecule ν, μ and ν are molecule indices and N is the total number of molecules. To reduce the number of parameters, the $A_{t(ij)}$-parameters in this paper are computed according to the mixing-rule:

$$A_{t(ij)} = \sqrt{A_{t(ii)} \times A_{t(jj)}}, \tag{8}$$

such that only the $A_{t(ii)}$ parameters need to be determined (section). The parameter $B_{t(ij)}$ is here calculated as

$$B_{t(ij)} = \frac{1}{\tau(R_i + R_j)}, \tag{9}$$

where R_i and R_j are the van der Waals radii of the atoms that make up the atom-pair $t(ij)$. We assume that τ is a general dimensionless factor independent of the atom types. The R_i values are taken from ref.[58]

Eqs. (8) and (9) imply that the number of independent parameters is equal to the number of atom types (A parameters) plus one (τ parameter), and the other parameters are correlated. Our model does not discriminate between the different types of the same atoms, e.g. H_t and H_m in Fig. 2. For instance, methanol consists of H, C, and O, and has four independent parameters.

Parameter calibration

The training data is obtained from CP computations on snapshots from the NpT/DFT-D3 MD simulations (Section), and is used to fit the parameters of the model (Eq. (7)). The procedure, schematically presented in Fig. 3, consists of the following steps:

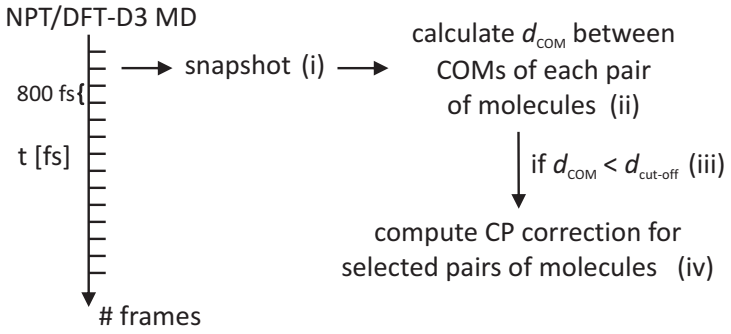


Figure 3: Flowscheme for obtaining training data from MD simulations.

(i) A snapshot of the liquid structure is selected every 800 steps (= 800 fs) of the NpT/DFT-D3 MD simulation, corresponding to a time greater than the velocity auto correlation time of the center of mass (COM) of the molecules to ensure statistically independent samples. The correlation times are computed with the MD-TRACKS program,[31,59] and are all below 270 fs (see Table S.2 of the SI). In this way a respectable number of

frames $(15 - 20)$ of the MD simulation are selected. The exact number depends on the type of the solvent.

(ii) At each snapshot, the distance d_{COM} between the COMs of every possible pair of molecules is computed with the minimum image convention.

(iii) The pairs of molecules for which this distance is smaller than a cut-off distance $d_{\text{cut-off}}$, are selected for the computation of training data.

(iv) We then calculate the CP-correction $\Delta E_{\mathrm{CP}}^{\mathrm{ref}}$ with CP2K for each of the selected pairs, which serves as training data.

The same specifications (basis set, electronic structure method, 800 Ry cut-off, ...) are used for the CP-computations as those for the MD simulations of the solvents (Section). The box size has been chosen such that there is a margin of minimally 5 Å between the dimer and the edge of the simulation box. The sizes of the unit cells for the five solvents are given in Table S.3 of the SI.

We use the COM RDF as a criterium to determine a proper $d_{\text{cut-off}}$ for selecting molecule pairs for the training CP-computations, i.e. $d_{\text{cut-off}}$ is chosen as the distance that contains the whole first peak of the COM RDF determined from the NpT/DFT-D3 simulations. Using this approach, it is ensured that at least the first layer of intermolecular interactions is represented in the training set. The COM RDF of methanol is shown in Fig. 4 as an example. Based on this RDF the $d_{\text{cut-off}}$ for methanol is set at 6.2 Å. Following this criterium, the $d_{\text{cut-off}}$ of 4.0 Å chosen in ref.[29] is improved. This latter value for $d_{\text{cut-off}}$ resulted in 3850 molecule pairs, while the higher cut-off distance in this paper results in a total number of 15714 pairs. The $d_{\text{cut-off}}$ values and the COM RDFs for the other solvents are given in the SI (Table S.3 and Fig. S.1, respectively).

The parameters (A_i, τ) in Eq. (7) are estimated from the training data set of CP-computations $(\Delta E_{\mathrm{CP}}^{\mathrm{ref}})$ by a non-linear least-squares fit. In this paper, the CMA-evolution strategy[60] (CMA-ES) is applied to minimize the cost function. CMA is a population-based

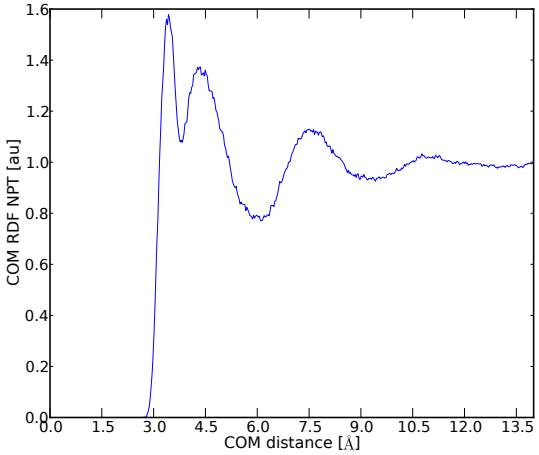


Figure 4: The COM RDF of methanol derived from the NpT/DFT-D3 MD simulation.

stochastic search algorithm for parameter optimization. It neither approximates nor uses gradients, making it an efficient optimization method for problems where derivative-based methods may fail due to a rugged search landscape presenting multiple discontinuities, sharp bends, noise and local optima.

The considered cost function is the root-mean-square deviation (RMSD) between the natural logarithms of the model $\Delta E_{\mathrm{CP}}^{\mathrm{mod}}$ and the training $\Delta E_{\mathrm{CP}}^{\mathrm{ref}}$ corrections, and is further discussed in the SI.

Once average calibrated parameters (A_i, τ) are obtained, a new MD simulation is performed with the Quickstep and FIST modules of CP2K. The energy in the MD run is computed as the DFT-energy (Quickstep module) plus the CP model term (Eq. (7)) (FIST module).

Results and Discussion

CP-computations on NpT/DFT-D3 MD

From the procedure as outlined in Section , 16 – 19 snapshots from the NpT/DFT-D3 MD simulations are used to select 7982 – 15111 molecule pairs, with values for $d_{\text{cut-off}}$ ranging between 6.2 and 7.9 Å. CP-computations with CP2K on these selected pairs serve as training data for the five considered liquids. The logarithms $\ln A_i$ and $\ln \tau$ are regarded as parameters in the algorithm to avoid that A_i become negative. To assess the impact of the stochastic factor inherent to the CMA-algorithm, the procedure is conducted 200 times for each solvent. This results in 200 estimates for each of the parameters. For instance, MeOH has 4 parameters ($\ln \tau$, $\ln A_{\text{H-H}}$, $\ln A_{\text{C-C}}$, $\ln A_{\text{O-O}}$) that are varied in the CMA algorithm, and other parameters (B parameters, other A parameters) are calculated with Eq. (8) and Eq. (9). The final parameters are then averaged over the 200 parameter estimates, and are shown in Table 2.

Table 2: Final A_i, τ and B_i parameters for MeOH, TCM, MeCN, THF and EtOH. The last columns contain the average BSSE per molecule in kJ/mol, calculated with Eqs. (12) and (13).

solvent	interaction	A [kJ/mol]	τ [/]	B [Å^{-1}]	$\langle \Delta E^{\text{mod}}_{\text{CP},\alpha\beta} \rangle$	%	$\langle \Delta E^{\text{mod}}_{\text{CP}} \rangle$
MeOH	H-H	3.1689	0.1796	1.3514	1.21	17	7.31
	C-C	0.0779		0.8416	0.01	0	
	O-O	7.9499		0.6626	3.25	44	
	H-C	0.4968		1.0372	0.26	4	
	H-O	5.0192		1.1116	2.38	33	
	C-O	0.7869		0.8899	0.19	3	
TCM	H-H	0.0491	0.1949	1.2455	0.00	0	1.64
	C-C	0.3867		0.7757	0.03	2	
	Cl-Cl	1.6284		0.7802	1.18	72	

Continued on next page

Table 2 – *Continued from previous page*

solvent	interaction	A [kJ/mol]	τ [/]	B [Å^{-1}]	$\langle \Delta E^{\mathrm{mod}}_{\mathrm{CP},\alpha\beta} \rangle$	%	$\langle \Delta E^{\mathrm{mod}}_{\mathrm{CP}} \rangle$
	H-C	0.1377		0.9560	0.01	1	
	H-Cl	0.2827		0.9594	0.06	4	
	C-Cl	0.7935		0.7780	0.36	22	
MeCN	H-H	0.3212	0.2387	1.0169	0.15	16	0.93
	C-C	0.0084		0.6334	0.01	1	
	N-N	0.6938		0.6884	0.16	17	
	H-C	0.0519		0.7805	0.09	10	
	H-N	0.4720		0.8210	0.42	46	
	C-N	0.0763		0.6597	0.09	10	
THF	H-H	0.8028	0.1927	1.2596	0.69	17	4.14
	C-C	0.4101		0.7846	0.50	12	
	O-O	3.9314		0.8801	0.18	4	
	H-C	0.5737		0.9669	1.29	31	
	H-O	1.7764		1.0362	0.82	20	
	C-O	1.2696		0.8296	0.64	16	
EtOH	H-H	1.7398	0.1757	1.3816	0.87	20	4.43
	C-C	0.7422		0.8605	0.25	6	
	O-O	9.0476		0.9653	0.75	17	
	H-C	1.1363		1.0606	1.02	23	
	H-O	3.9674		1.366	0.79	18	
	C-O	2.5913		0.9099	0.75	17	

Fig. 5 shows the correlation between training $\Delta E^{\mathrm{ref}}_{\mathrm{CP}}$ and model $\Delta E^{\mathrm{mod}}_{\mathrm{CP}}$ data for methanol. The CP model-energies are computed with Eq. (7), using the parameters as listed in Table 2.

The plots for the other four solvents look similar (Fig. S.2 of the SI). The correlation between

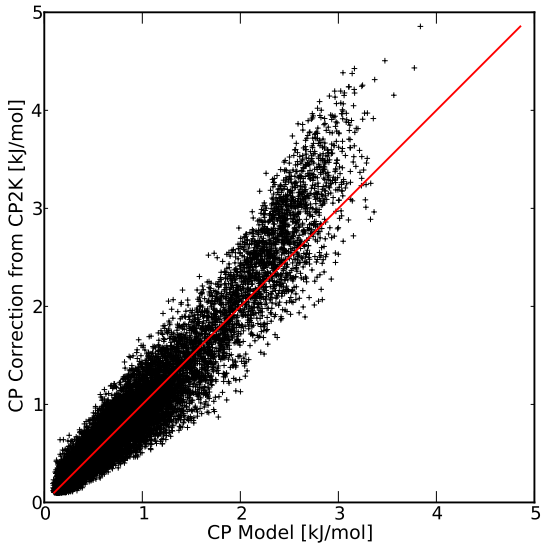


Figure 5: CP-energies computed with CP2K ($\Delta E_{\mathrm{CP}}^{\mathrm{ref}}$) versus model CP-energies ($\Delta E_{\mathrm{CP}}^{\mathrm{mod}}$) in methanol.

the two data sets (Fig. 5) is not perfect but sufficient. As for the model parameters, no clear trend is observed from their values. Given the importance of hydrogen bonds in methanol and ethanol, one would expect that the O-H parameters are dominant. The $A_{\text{O-O}}$ parameter however is stronger than the $A_{\text{O-H}}$ parameter. This suggests that the O-O distances are the determining factor for the BSSE corrections, but this is not the correct interpretation. The reason for the strong $A_{\text{O-O}}$ parameter originates from the correlation between the parameters as imposed by Eq. (8): the $A_{\text{O-H}}$ parameter necessarily lies between the $A_{\text{H-H}}$ parameter and the $A_{\text{O-O}}$ parameter. A significant $A_{\text{O-H}}$ parameter is thus only possible in combination with an even stronger $A_{\text{O-O}}$ parameter. A plain comparison of the parameters thus does not give much insight in the origin of the BSSE. Statistical correlations between the independent parameters are illustrated by the large off-diagonal elements in the sampling covariance matrix. This is also confirmed by the high condition number (see Table S.4 of the SI).

As a significant trend in the final parameters (A_i,τ,B_i) is lacking in Table 2, we evaluate the importance of BSSE by estimating each BSSE contribution. The average BSSE

contribution per atom is given by

$$\langle \Delta E_{\mathrm{CP}} \rangle = \frac{1}{N} \frac{\int \mathrm{d}\boldsymbol{r}^N \Delta E_{\mathrm{CP}}(\boldsymbol{r}^N) e^{-\beta E(\boldsymbol{r}^N)}}{\int \mathrm{d}\boldsymbol{r}^N e^{-\beta E(\boldsymbol{r}^N)}} \tag{10}$$

where $E(\boldsymbol{r}^N)$ is the energy of the N particle system including the BSSE correction. With the pairwise force field model for the BSSE, this quantity may be derived conveniently from the partial RDFs, which are at our disposal,

$$\begin{aligned} \left\langle \Delta E^{\mathrm{mod}}_{\mathrm{CP},\alpha\beta} \right\rangle &= \frac{1}{2} \rho \, c_\alpha c_\beta (2 - \delta_{\alpha\beta}) \, 4\pi \int \Delta E^{\mathrm{mod}}_{\mathrm{CP},\alpha\beta}(r) g_{\alpha\beta}(r) r^2 \, \mathrm{d}r & (11) \\ &= \frac{1}{2} \rho \, c_\alpha c_\beta (2 - \delta_{\alpha\beta}) \, 4\pi \int A_{\alpha\beta} e^{-B_{\alpha\beta} r} g_{\alpha\beta}(r) r^2 \, \mathrm{d}r & (12) \end{aligned}$$

and

$$\left\langle \Delta E^{\mathrm{mod}}_{\mathrm{CP}} \right\rangle = \sum_{\alpha \leq \beta} \left\langle \Delta E^{\mathrm{mod}}_{\mathrm{CP},\alpha\beta} \right\rangle . \tag{13}$$

The average BSSE per molecule is shown in Table 2 using the intermolecular RDFs from the NPT/DFT-D3+CP simulations. The BSSE per molecule is most important for methanol (7.31 kJ/mol/atom), followed by ethanol and THF. The BSSE is small for TCM and MeCN. The largest corrections are correlated with the capability of forming hydrogen bonds.

NVT versus NpT MD

This work encompasses in total 15 MD simulations (NVT, NpT and NpT + CP for five solvents). The duration of the production runs depends on the type of the solvent but especially on the type of the dynamics. NpT dynamics require run times of at least 20 ps or even more. The instantaneous densities of the five considered solvents found in our NpT MD simulations, including the equilibration run, are presented in Fig. 6.

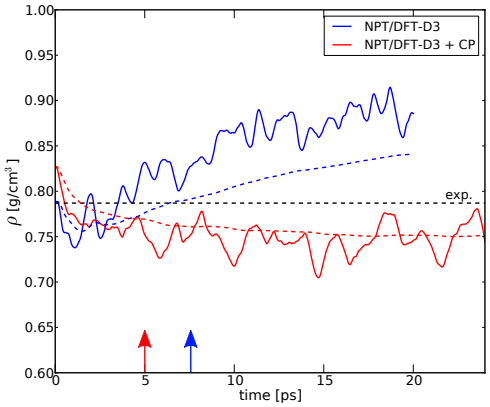
It should be stressed that NpT dynamics of molecular liquids within the DFT methodology has a much larger complexity than standard NVT simulations. Their outcome largely depends on the accuracy of the DFT functionals. It is well-known that the convergence of

the pressure requires a significantly higher basis set cutoff than normally used for standard NVT MD simulations.[34,47] Another problem is that the equilibration time for NpT dynamics can be significantly larger, as the volume fluctuations are slow compared to molecular vibrations. A comprehensive DFT study on the performance of two popular gradient-corrected exchange correlation functionals on the structure and density of liquid water is presented in[34] in the NpT ensemble. Both PBE and BLYP functionals underestimate the density by about 25 % and 12 % respectively. Adding Grimme corrections of the type D1,[35] the authors of ref.[34] found a significant improvement of the densities. Our results are in line with the findings of this previous work on water.

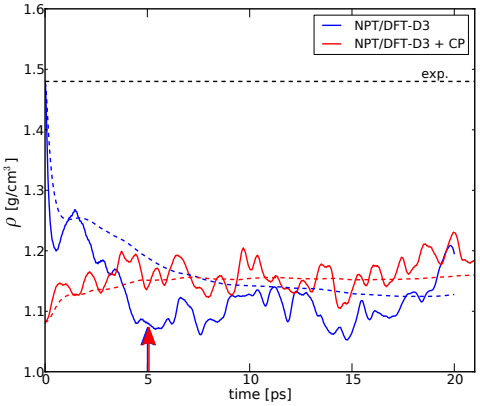
The densities computed from the NpT runs for the five liquids under study show large fluctuations (Fig. 6). To visualize the convergence of the average density, the running average (average over all previous time steps[34]) is plotted. Some of the running averages in Fig. 6 still slightly change with time, suggesting that even longer simulation times (> 20 ps) would be needed to reach absolute convergence, yet the current results are sufficient to support our conclusions.

The NpT ensemble and the BSSE correction model affect the density of the five solvents differently. The plots give the impression that the agreement with experiment becomes better when including the BSSE-CP corrections, although we need to be cautious in generalizing these conclusions as they are not statistically founded . This is remarkably the case in methanol and ethanol. In only one case – chloroform – the predicted density is largely underestimated. In the absence of hydrogen bonds the remaining intermolecular forces – which are weaker – become now decisive in determining the liquid density. However, they are prone to the interaction potentials, involved in the DFT functionals by construction, and therefore one could expect a large dependence of the predicted density on the choice of functional. Within this context we also examine the geometry of each separate liquid molecule (the monomer). When averaging the C-H and C-Cl bond lengths of all chloroform molecules in the snapshots taken during the NpT simulations, an average bond length of 1.82 Å is found

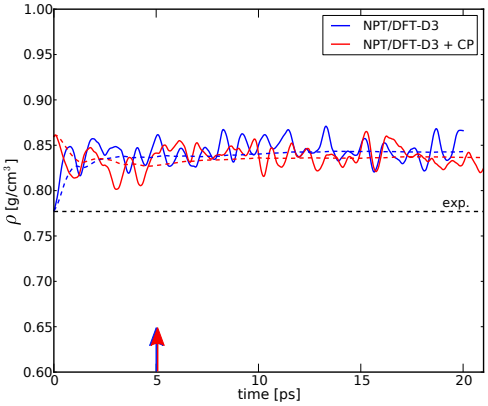
Figure 6: Evolution of the density as a function of time for the NpT/DFT-D3 and NpT/DFT-D3 + CP MD simulations: instanteneous density (solid lines) and running average density (dashed lines). The production run starts at a time indicated by an arrow. The horizontal line gives the experimental density, which has been imposed in the NVT/DFT-D3 runs.



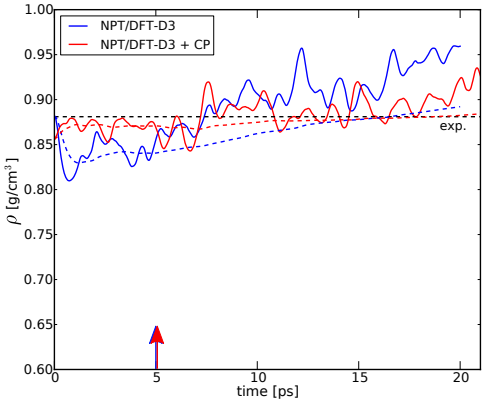
(a) MeOH



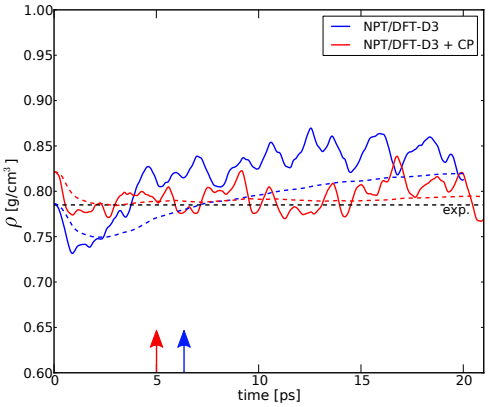
(b) TCM



(c) MeCN



(d) THF



(e) EtOH

for C-Cl and 1.09 Å for C-H. A geometry optimization of the monomer in gas phase with the B3LYP/311+g(d,p) level of theory, yields 1.79 Å and 1.08 Å respectively. The average "volume" taken by a single chloroform molecule (pyramidal structure) in the simulation is thus 7.5 % larger than that in gas phase. It partially explains the underestimation by 20 % for the density measured in the NpT simulation with BLYP compared to experiment. It also implies that in this specific case of chloroform, the BLYP functional is not adequate enough to predict the correct intramolecular geometry. Moreover, since hydrogen bond interactions are absent in chloroform liquid, the weak intermolecular forces are mainly determining the liquid structure, and DFT functionals are not optimized to reproduce these accurately. The interplay of all these small effects has direct consequences in the reproduction of the correct density.

For the convenience of the reader, the average densities ρ_{av} for the five liquids under study are tabulated in Table 3. These averages are taken over the production runs only, starting from 5 ps unless otherwise stated (arrow in Fig. 6).

Table 3: Average densities ρ_{av} for the NpT/DFT-D3 and NpT/DFT-D3 + CP MD simulations, compared to experiment (NVT/DFT-D3, see Table 1), for the five considered solvents. The values between brackets show the deviation compared to experiment.

solvent	ρ_{av} [g/cm^3]		
	exp NVT/DFT-D3	NpT/DFT-D3	NpT/DFT-D3 + CP
MeOH	0.787	0.872 (0.085)	0.746 (-0.041)
TCM	1.483	1.107 (-0.372)	1.162 (-0.317)
MeCN	0.777	0.845 (0.068)	0.839 (0.062)
THF	0.881	0.909 (0.028)	0.889 (0.008)
EtOH	0.785	0.837 (0.052)	0.795 (0.01)

Radial distribution functions

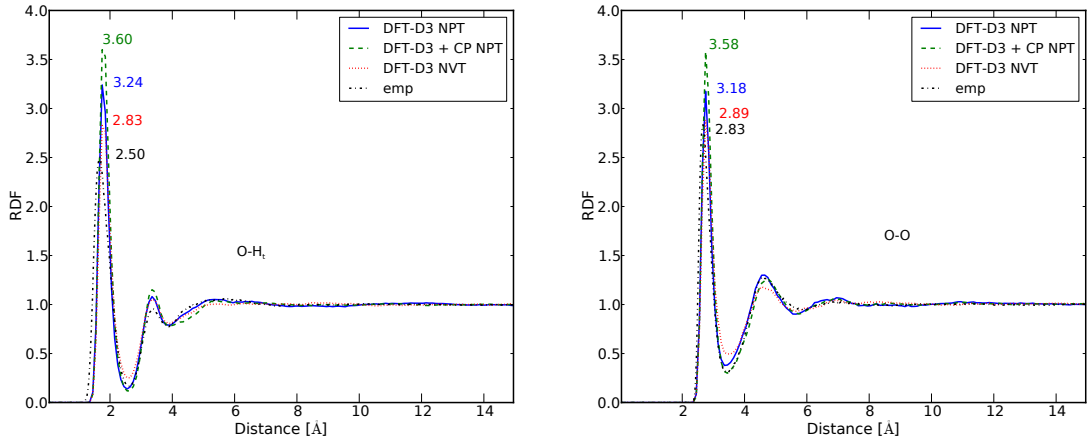
The previous subsection showed that overall the liquid density in the NpT simulation improves, when taking into account the BSSE (CP correction model). However, a good repro-

duction of the liquid density does not necessarily lead to a correct description of the structure of the liquid. This is illustrated by the case of water where PBE has the tendency toward overstructuring, whereas the obtained density in NpT dynamics is closer to the experimental value than with BLYP.[34]

We calculated the intermolecular partial RDFs for all five liquids from our NpT simulations. We only display a selection of them in Fig. 7, using the atom types as defined in Fig. 2; a complete list of the RDFs is given in the SI (Figure S.3-S.7). The most characteristic features of the RDFs, determining the structure of the liquid, are given by the height and position of the first peak. Table IV shows the amplitude and position of the first peaks of the RDFs from Fig. 7. For comparison, structure properties found in constant volume NVT simulations are also included in Table IV, where the density is set to the experimental value.

FIG. 7. Intermolecular partial radial distribution functions, between atom types defined in Fig. 2. Experimental data from refs.[61,62] (MeOH), ref.[20] (TCM), and ref.[19] (THF).

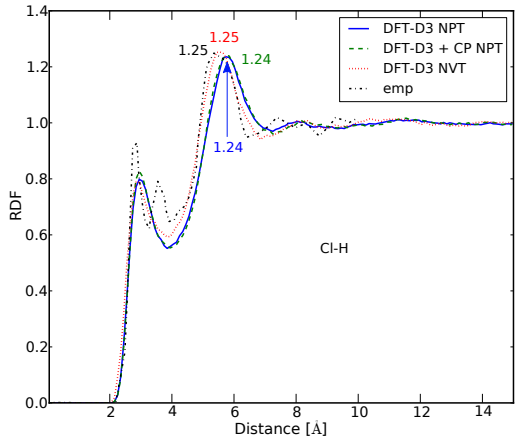
MeOH

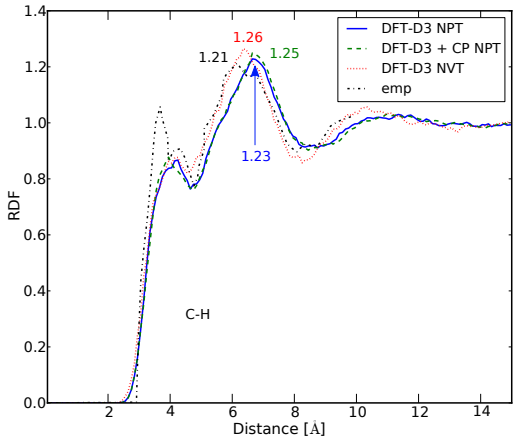


Continued on next page

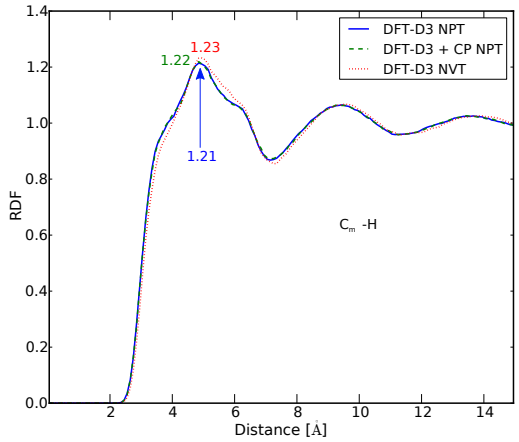
FIG. 7 – *Continued from previous page*

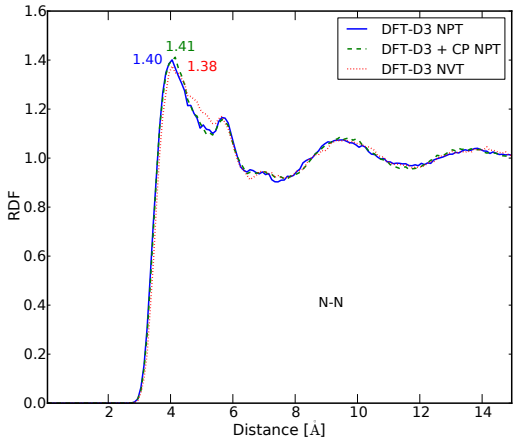
TCM



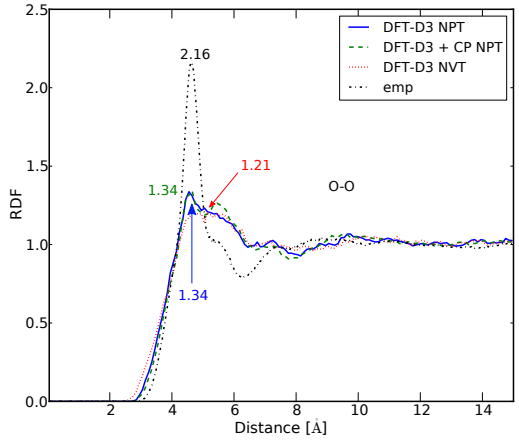


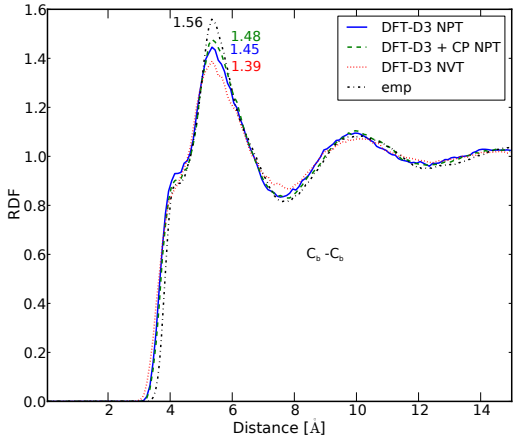
MeCN





THF

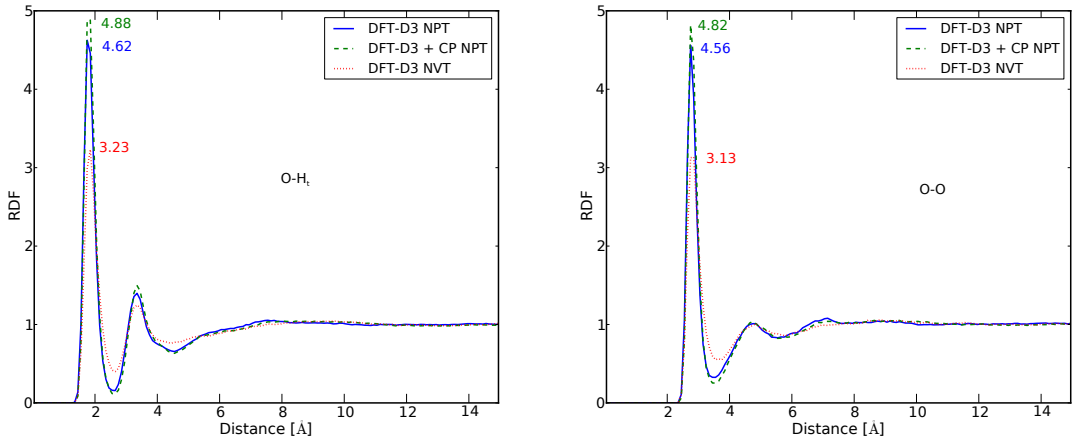




Continued on next page

FIG. 7 – *Continued from previous page*

EtOH



(i) NpT simulations clearly give rise to an overstructured **methanol** liquid. The heights for the first oxygen-hydrogen (O-H_t) and oxygen-oxygen (O-O) peak are largely overestimated. In contrast to NVT simulations,[29] the BSSE-CP corrections even strengthen the discrepancy. This is surprising, as the BSSE switches the density to a slightly underestimated value (0.746 g/cm^3), but closer to the experimental value of 0.787 g/cm^3 (see Fig. 6). On the other hand, in similar work on water,[34] it is observed that an underestimation of the density is generally accompanied by a locally overstructured liquid. Many other theoretical studies on water[7,8,29,63–67] overestimate the first intermolecular peak. Experiments indeed do not measure pair interactions directly, as explained in the Section . The various manipulations in extracting empirical partial RDFs from the experimental total SF are not free from inaccuracies, which we will discuss in detail in the remainder of this paper, and one should be cautious in drawing final conclusions.

The RDFs of **ethanol** show large similarities with those of methanol. The first peak rises significantly when switching from NVT to NpT; correcting for BSSE even enhances this trend.

(ii) Despite the serious underestimation of the density in **chloroform**, we now observe a lower first peak than experiment in most of the partial RDFs (see also Fig. S.4 of the SI). BSSE corrections on the RDFs are small. The experiment reveals a second peak at around 3.1 Å in the partial RDF of Cl-H, but this peak is not reproduced by the simulations. Similar observations may be made in the other partial RDFs, and in particular in C-H and H-H. Prominent peaks at short distances appear, but are completely missing in our simulations and other MD studies[68–70] are in line with our results. RMC calculations performed by Pothoczki *et al.*[20] without additional constraints imposed by diffraction data also give no evidence for an additional peak in the region 3.5 Å – 4.0 Å.

(iii) No experimental data are available for **acetonitrile**. All simulations predict the same RDFs. BSSE corrections are negligible.

(iv) The situation is problematic for **THF** where the experiment predicts sharp prominent peaks at short distances that are not reproduced by any theoretical MD simulation, neither in this work nor elsewhere in literature. This strongly deviant behavior, also noticed in the case of TCM, requires special attention and a serious in depth discussion to understand the discrepancies between theory and experiment. As the primary observable is the total structure factor, the discussion should also focus on this property which gives information in the reciprocal space. What characteristic in the (partial) structure factor in momentum space can cause such a prominent peak at short distances in coordinate space? This will be unraveled in the next subsection (Section).

TABLE IV. Amplitudes and positions of the first intermolecular peak of the RDFs shown in Fig. 7. The positions are in units Å.

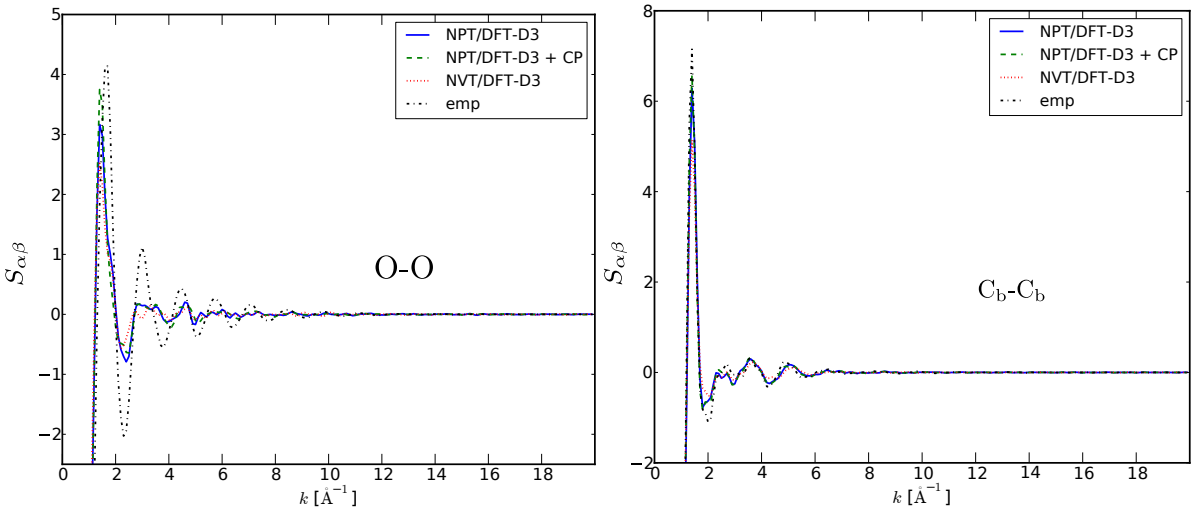
solvent	interaction	property	exp	NpT/DFT-D3	NpT/DFT-D3 + CP	NVT/DFT-D3
MeOH	O-H_t	amplitude	2.50	3.24	3.60	2.83
		position	1.68	1.75	1.75	1.75
	O-O	amplitude	2.83	3.18	3.58	2.89
		position	2.67	2.75	2.75	2.75
TCM	Cl-H	amplitude	0.93/0.79/1.25	0.79/1.24	0.83/1.24	0.79/1.25
		position	2.87/3.80/5.36	2.95/5.75	2.95/5.75	2.85/5.55
	C-H	amplitude	1.06/0.90/1.21	0.87/1.23	0.87/1.25	0.87/1.26
		position	3.68/4.02/6.15	4.25/6.65	3.95/6.65	4.25/6.35
MeCN	C_m-H	amplitude		1.21	1.22	1.23
		position		4.85	4.85	4.95
	N-N	amplitude		1.40	1.41	1.38
		position		4.05	4.15	4.05
THF	O-O	amplitude	2.16	1.34	1.34	1.21
		position	4.62	4.55	4.65	5.15
	C_b-C_b	amplitude	1.56	1.45	1.48	1.39
		position	5.34	5.35	5.35	5.35
EtOH	O-H_t	amplitude		4.62	4.88	3.23
		position		1.75	1.75	1.85
	O-O	amplitude		4.56	4.82	3.13
		position		2.75	2.75	2.85

Structure factors of THF

Observations in experimental and theoretical SF.

The example of tetrahydrofuran is selected for the discussion of the structure factors. To unravel the origin of the large peaks at short distances in the experimentally derived RDFs it is convenient to analyze the liquid structure in the reciprocal domain, because the *total SF* (S_{tot}) is the sole quantity that is directly accessible by experiment and appropriate for a direct comparison between theory and experiment; all other quantities are empirical. We start by clarifying the origin of all RDF and SF quantities that are displayed in Figs. 8, 9, 10, and their relationship as shown in the scheme in Fig. 1.

Figure 8: Intermolecular partial SFs $S^{\mathrm{inter}}_{\mathrm{O\text{-}O}}(k)$ and $S^{\mathrm{inter}}_{\mathrm{C_b\text{-}C_b}}(k)$ for THF. Theoretical partial SFs are constructed from partial RDFs with Eq. 2, empirical partial SFs (label "emp") are computed as the Fourier transform of the empirical partial RDFs reported by Bowron,[19] see text.



(a) $S_{\mathrm{O\text{-}O}}$ (b) $S_{\mathrm{C_b\text{-}C_b}}$

First, consider the *theoretical predictions* of SFs from the MD simulations. Our theoretical partial RDFs $g_{\alpha\beta}(r)$ may be transformed to reciprocal space by a Fourier transform. Using Eq. (2), we obtain not only theoretical partial SFs $S_{\alpha\beta}(k)$, but also their intermolecular and intramolecular contributions. The transformation of the theoretical RDFs from r-space to k-space can be accurately performed as each RDF is a well-determined function. We focus

on the partial SF between the intermolecular oxygen-oxygen and carbon-carbon atom pairs to explore the possible ingredients lying at the origin of the strong empirical RDF peak at 4.6 Å in real space (Fig. 7). The theoretical partial $S^{\text{inter}}_{\text{O-O}}(k)$ and $S^{\text{inter}}_{\text{C}_\text{b}\text{-C}_\text{b}}(k)$ are displayed in Fig. 8 for the three MD runs; other partial SFs are given in the SI (Fig. S.8). These partial SFs show the typical behavior of a spectral function in reciprocal space with damped oscillations at increasing k (Fig. S.8). A weighted summation of the partial SFs, as given in Eq. (3), yields the theoretical total SF S_{tot}, which is displayed in Fig. 9. The coherent neutron scattering lengths for Eq. (3) are taken from ref.[71] (Table 5).

Table 5: Coherent neutron scattering lengths.[71]

isotope	b_α [fm]
C	6.646
O	5.803
H	-3.739
D	6.671

Next, consider the *experimental* SFs as measured in neutron scattering experiments.[19] The paper by Bowron *et al.* reports the experimental total SF $S^{\text{exp}}_{\text{tot}}$ and the empirical intermolecular partial RDFs $g^{\text{emp,inter}}_{\alpha\beta}$. We remind the reader that the label *empirical* refers to the fact that the reported partial RDFs are constructed from the experimental total SF with a rather complex procedure.[19] Fig. 9 compares $S^{\text{exp}}_{\text{tot}}$ with our theoretical total SF. In addition, we added in Fig. 8 the empirical intermolecular partial SF $S^{\text{emp,inter}}_{\text{O-O}}$ obtained by calculating the inverse Fourier transform of Bowron's empirical RDF $g^{\text{emp,inter}}_{\text{O-O}}$. We also display the weighted sum $S^{\text{emp,inter}}_{\text{tot}}$ of all partial SFs[72] in Fig. 9(a) (labeled: Bowron-adapted).

We include the SFs for the isotope variant, i.e. the deuterium sample C_4D_8O, in Fig. 10 to visualize the effect of an isotope substitution. There are indeed some fundamental differences between the two liquids, which are atttributed to the negative scattering length of the proton versus the positive scattering length of the deuterium (Table 5).

Comparison of empirical/experimental and theoretical curves gives several observations:

- *Comparison of partial SFs* in Figs. 8 and S.8. The empirical intermolecular partial

SF $S_{\text{O-O}}^{\text{emp,inter}}$ of THF shows a quasi-perfect damped harmonic oscillator behavior, completely due to the sharp peak at 4.6 Å observed in the empirical $g_{\text{O-O}}^{\text{emp,inter}}$ (Fig. 4. This is a typical feature in FT. The theoretical SFs are dominated by anharmonicities. Also the amplitudes of the oscillations are much smaller. It is the result of a complete absence of prominent peaks in the theoretical RDF $g_{\text{O-O}}^{\text{inter}}$.

- *Comparison of total SFs.* In C_4H_8O (Fig. 9), the theoretical simulations predict the first peak of S_{tot} at the correct position, but its amplitude is barely half the experimental estimate. The remaining part of the spectrum (high k-region) is fairly well reproduced by theory. This is a very important conclusion as it is the only plot where theoretical results can be directly compared with a real measurable quantity. The same conclusions hold for the total structure factor for C_4D_8O (Fig. 10). The most striking discrepancy in the S_{tot} spectrum is the appearance of a shoulder structure in $S_{\text{tot}}^{\text{exp}}$ for C_4H_8O at around 2.0 Å^{-1} which is not present in our MD simulations (Fig. 9). The impact of this shoulder structure on other derived quantities will be further discussed in the remainder of this paper. Note that this shoulder is not present in C_4D_8O (Fig. 10).

- *Comparison of total intermolecular SFs* in Fig. 9 and 10. By summing all empirical partial SFs according to Eq. (3) we are able to compare the experimentally derived property with the theoretical prediction. The agreement is surprisingly good in view of the significant differences observed in the partial SFs (Figs. 8 and S.8). Inspection of the various contributions leading to $S_{\text{tot}}^{\text{inter}}$ reveals that this is mainly due to cancellation effects. It is a strong indication that multiple sets of partial SFs can yield the same total structure factor. Note that the intermoleculer spectrum is short ranged. All oscillations in the high k-region ($k > 2.0\,\text{Å}^{-1}$) fade out.

- *Comparison of pair correlation function or total radial distribution function* $f(r)$ in Figs. 9 and 10. To estimate the impact of peaks in $S_{\text{tot}}^{\text{exp}}$ in real space, the inverse Fourier transform is computed leading to the total pair correlation function $f_{\text{tot}}^{\text{emp}}(r)$

using Eq. (4) and plotted in Fig. 9(d). Typical FT features appear: the long tail of S_{tot} gives rise to well pronounced peaks concentrated in a small window in real space. A structure factor with signals restricted in the low k-region like $S_{\mathrm{tot}}^{\mathrm{inter}}$ causes a more structured pattern in r-space but the amplitudes of the fluctuations are one order of magnitude smaller than in the case of inter/intra f_{tot}. In C_4H_8O, a prominent peak appears at a distance of 0.75 Å. However, there are no atom pairs available in the THF liquid at such a short distance, so this peak must be spurious.

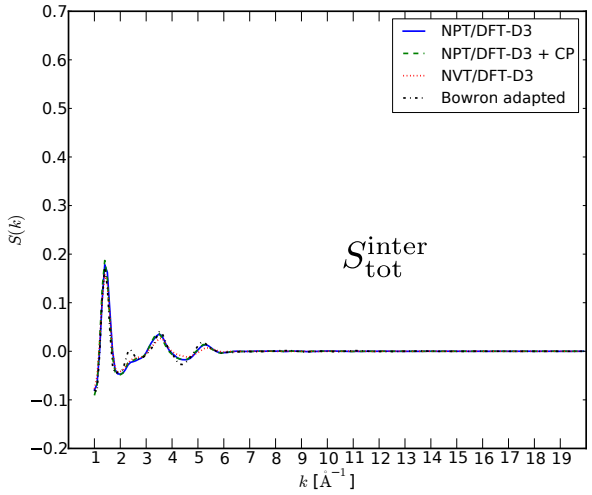
Summarizing, we stress the nice reproduction of the total structure factor by the NpT MD simulations from first principles, but emphasize the striking differences noticed in the partial radial distribution functions (RDFs), and in the corresponding partial SFs, between theory and experiment. The disentanglement between spurious and genuine structural features in the RDFs is a serious issue that has received since long a lot of attention in literature[19,27] and further discussed in a recent review of Soper.[2]

Inaccuracies in empirical procedure.

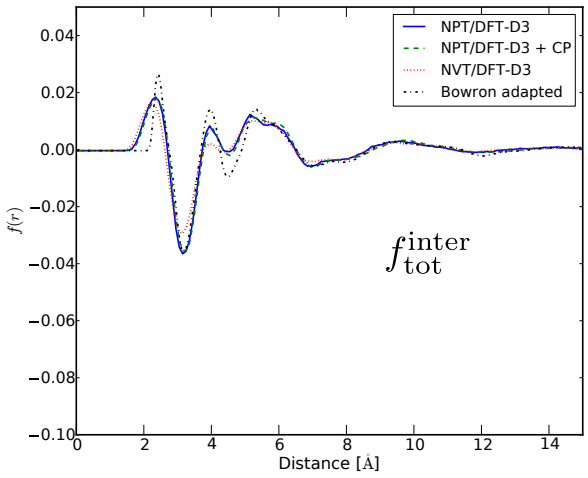
Apart from the total SF, all other experimental quantities are empirical and are obtained through a series of manipulations (Fig. 1). Inaccuracies can infiltrate the experimental analysis at different stages; we discriminate between three types: (1) windowing effects in Fourier transform, (2) extraction of partials, and (3) decomposition in intermolecular and intramolecular contributions.

(1) A first source of errors lies in the finite windowing effect when taking the Fourier transform. This is a familiar issue in Fourier analysis well-known as spectral leakage. Applied to the Fourier transform of the RDF , it implies that a numerical integration over a finite r-range (from 0 up to R_{max}) convolutes the real spectrum with a $\mathrm{sinc}(kR_{\mathrm{max}})$ function. The computed spectrum has not only the original peak, but also a series of spurious damped peaks at distances $\Delta k = 2\pi/R_{\mathrm{max}}$. For instance, a window of size $R_{\mathrm{max}} = 15\,\text{Å}$ creates spurious peaks at distances $\Delta k = 0.42\,\text{Å}^{-1}$ of the original peak. Similarly, a window size $k_{\mathrm{max}} = 20\,\text{Å}^{-1}$ in reciprocal space creates spurious peaks in real space at distances $\Delta r =$

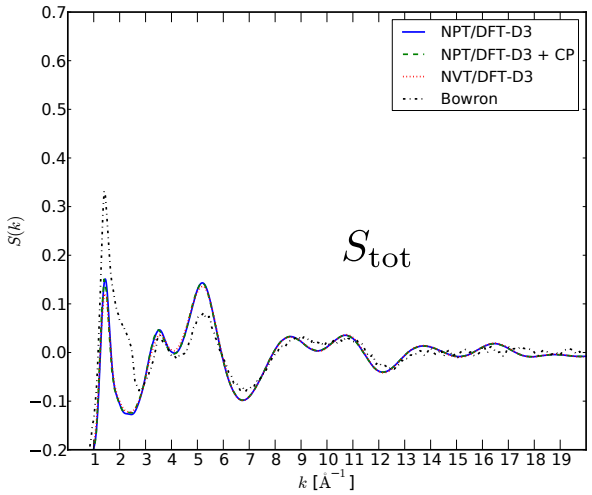
Figure 9: Total structure factor for THF C_4H_8O obtained as a weighted sum of the partial SFs limited to intermolecular contributions (upper left) and to inter- and intramolecular contributions (bottom left). Theoretical total SF is constructed from partial SFs with Eq. (3), experimental total SF from Bowron et al.[19] The corresponding total pair correlation functions $f(r)$ are given in the right column, as defined in Eq. (4) for the empirical curve and Eq. (5) for the theoretical curves. Units of $S(k)$ and $f(r)$ are barn/sr/atom.



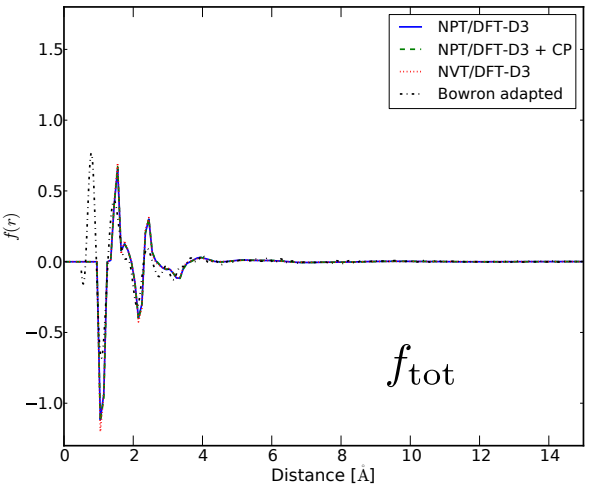
(a) inter



(b) inter

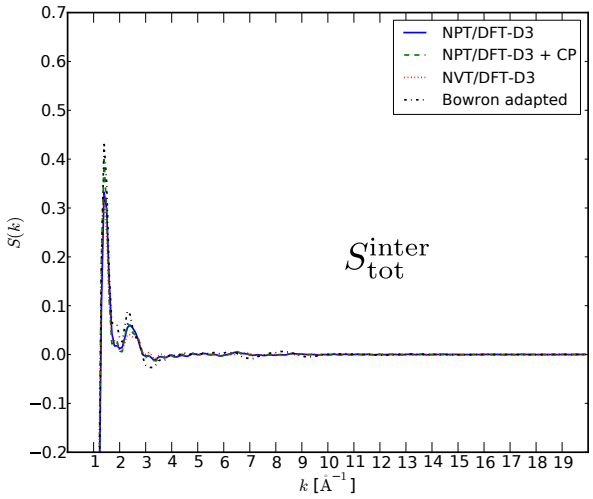


(c) inter + intra

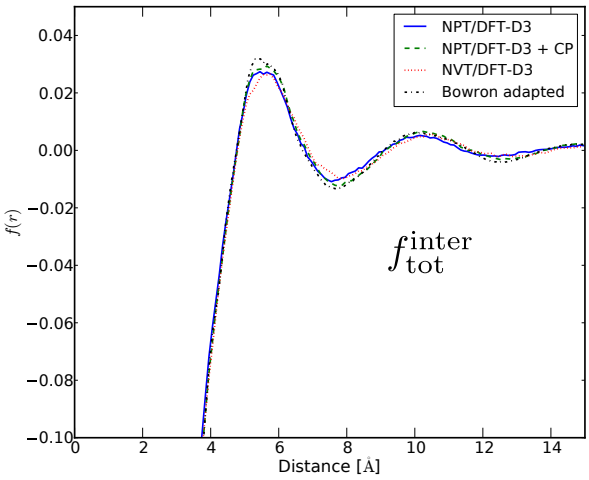


(d) inter + intra

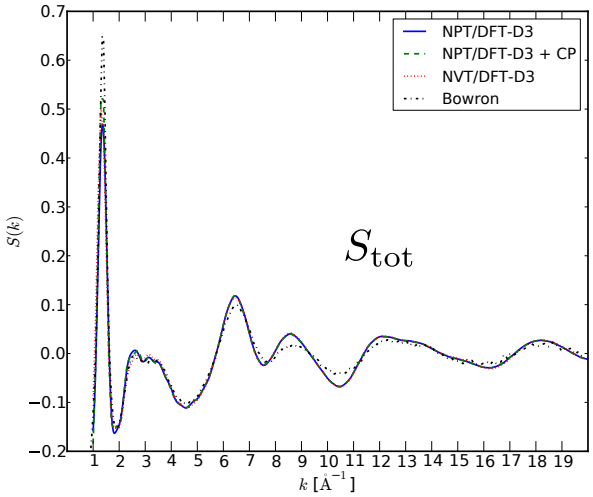
Figure 10: Total structure factor for C_4D_8O obtained as a weighted sum of the partial SFs limited to intermolecular contributions (upper left) and to inter- and intramolecular contributions (bottom left). Theoretical total SF is constructed from partial SFs with Eq. (3), experimental total SF from Bowron et al.[19] The corresponding total pair correlation functions $f(r)$ are given in the right column, as defined in Eq. (4) for the empirical curve and Eq. (5) for the theoretical curves. Units of $S(k)$ and $f(r)$ are barn/sr/atom.



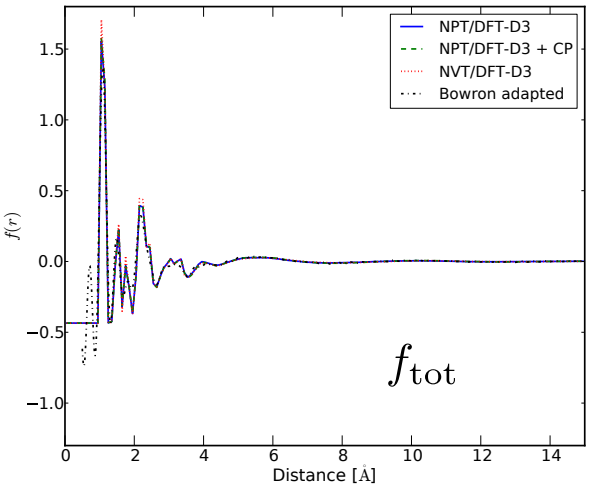
(a) inter



(b) inter



(c) inter + intra



(d) inter + intra

0.31 Å. In addition, the limit of the SF to $k = 0$ is not calculated from the theoretical RDFs using the Fourier transform, because the numerical integration diverges for $k \to 0$. SFs are only evaluated starting from $k = 1\,\text{Å}^{-1}$, and therefore a subsequent inverse Fourier transform from the SFs to the pair correlation function may experience additional numerical errors.

The ghost peak appearing at a distance of 0.75 Å in $f_{\text{tot}}^{\text{emp}}(r)$ is a neat example of a windowing error, as no atom pairs exist in the THF liquid at that distance (Fig. 9). The ghost peak is an artifact caused by taking the inverse Fourier transform of the experimental total SF ($S_{\text{tot}}^{\text{exp}}$). We have visualized the effect of subsequent Fourier transforms (Eqs. (2) and (4)) in Fig. S.11 of the SI (indirect note). In the direct route the total distribution function $f(r)$ is obtained from direct summation of theoretical partial RDFs (Eq. (5)). Comparison of the results obtained in the two routes, shows that the peaks in Fig. S.11 below 1 Å are indeed spurious. However, these windowing errors are not responsible for the large discrepancies observed in the RDFs and may easily be resolved by removing all peaks appearing at unphysically short distances.

(2) A second and probably the major source of uncertainties entering the analysis lies in the extraction of partial SFs from the total SF. Theoretically, the separate contributions to the RDFs and SFs can be easily computed from the MD simulations, whereas experimentally, the extractions of partials is a non-trivial task and requires extra experimental procedures such as neutron diffraction measurements with isotopic substitution (NDIS) as applied in ref.[19] (see Fig. 1). However, for complex systems, with more than two atom types, one has often not performed enough isotope variation experiments to extract all of the partial SFs. Experimentalists therefore introduce a three-dimensional model of the liquid structure. By means of e.g. Monte Carlo and intermolecular potentials (RMC, EPSR), liquid structure is modeled iteratively by adapting the intermolecular potentials until the modeled SFs are in agreement with the available experimental SFs (loop in Fig. 1).[25,73] However, RMC/EPSR explores a broad range of structural models that are consistent with the available scattering data. The method does not ensure unique partial SFs. In addition, the functional form of

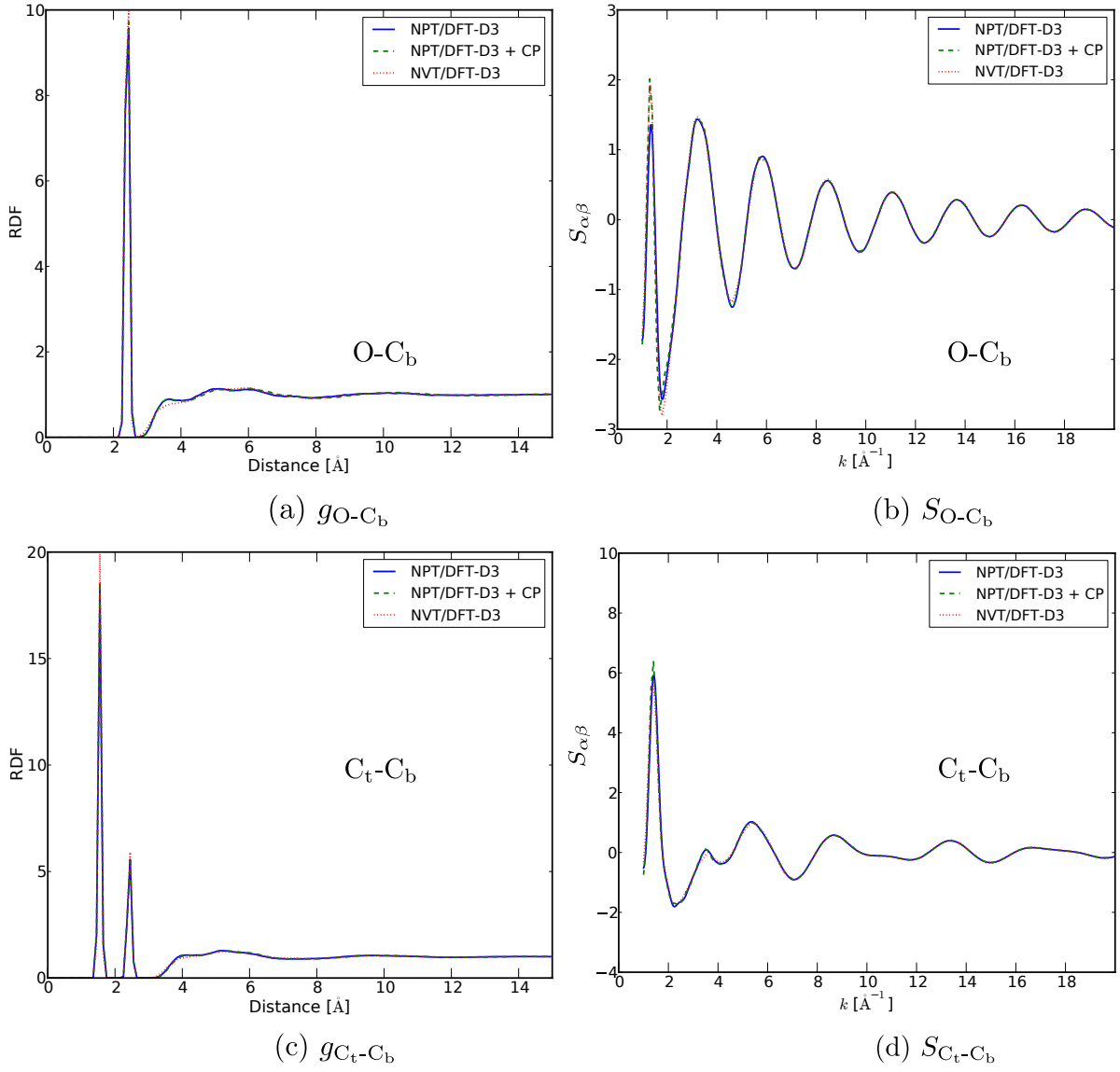
the potential can bias the model, e.g. by preventing a priori certain atomic configurations. These artifacts are well-known by the experimentalists doing neutron scattering,[2] but have not yet been emphasized so strongly as in the case of THF. This procedure is ill-conditioned, because multiple structure models may give the same total SF. The empirical partial RDFs and those predicted from the MD simulations show manifestly different structural features. Nevertheless, Fig. 9 is a clear example of how two different sets of RDFs succeed in reproducing the same total intermolecular structure factor.

(3) A third possible source to induce inaccuracies lies in the decomposition into intermolecular and intramolecular contributions. In the particular case of THF, three measured total SFs (for C_4H_8O, C_4D_8O and the 1:1 mixture) are at the disposal of the experimentalists to decompose into intermolecular and intramolecular contributions.[19] So for THF the interatomic partial hydrogen-hydrogen SF may be derived by taking a linear combination of these three experimental total SFs (arrow labeled "lin. comb." in Fig. 1). Again, the missing intermolecular partial SFs are guessed by means of an iterative RMC/ESPR method. The most ideal procedure, however, is to succeed into a separation of the two contributions, completely free of any bias, in other words not established with the help of interatomic potentials, but extracted solely from the measured total structure factors of the different samples. In the EPSR-procedure, the distribution of interatomic distances is mostly fixed a priori.

Intermolecular versus intramolecular structure factors.

The influence of subtracting intramolecular contributions from the total SF is immense (Figs. 9 and 10). A first observation is that the intermolecular SF ($S_{\mathrm{tot}}^{\mathrm{inter}}$) and the combined inter/intra SF (S_{tot}) look significantly different. The typical diffraction pattern in the intermolecular $S_{\mathrm{tot}}^{\mathrm{inter}}$ fades out at relatively low values for the scattering vector ($k = 6.0\,\text{Å}^{-1}$ for C_4H_8O and $k = 3.0\,\text{Å}^{-1}$ for C_4D_8O). The intermolecular total SF is thus short-ranged in reciprocal space. In contrast, the combined intermolecular/intramolecular $S_{\mathrm{tot}}^{\mathrm{inter}}$ has fluctuations extending over the whole k range. Fluctuations at larger k-values in S_{tot}

Figure 11: Total (intra + inter) partial RDFs $g_{\mathrm{O\text{-}C_b}}(r)$ and $g_{\mathrm{C_t\text{-}C_b}}(r)$ (left column) in THF. Also given are the corresponding partial structure factors (right column).



(a) $g_{\mathrm{O\text{-}C_b}}$ (b) $S_{\mathrm{O\text{-}C_b}}$

(c) $g_{\mathrm{C_t\text{-}C_b}}$ (d) $S_{\mathrm{C_t\text{-}C_b}}$

have completely disappeared in $S_{\mathrm{tot}}^{\mathrm{inter}}$, so these fluctuations may be fully attributed to intramolecular contributions. Within this context, we present a schematic model wherein the intramolecular contribution to the partial RDF $g_{\alpha\beta}(r)$ can be modeled as a sum of one or more Dirac delta functions:

$$g_{\alpha\beta}(r) = g_{\alpha\beta}^{\mathrm{inter}}(r) + g_{\alpha\beta}^{\mathrm{intra}}(r) = g_{\alpha\beta}^{\mathrm{inter}}(r) + \sum_i A_{\alpha\beta,i}\delta(r - R_{\alpha\beta,i}) \tag{14}$$

Multiple intramolecular distances may be present between atom types α and β; they are discriminated by means of the index i. Eq. (14) is valid for rigid liquid molecules, but may be regarded as a realistic approximation in molecules like THF where one may expect nearly constant intra-atomic distances $R_{\alpha\beta,i}$ in the same molecule during the simulations. A better representation would be to replace the Dirac delta functions by some function with a finite width, implying that the molecule is not rigid and the interactomic distances may fluctuate during the simulation. But this picture will not significantly change the global feature. Neglecting this finite width and working with the Dirac delta functions, the intramolecular SF is obtained by Fourier transforming the last term of Eq. (14):

$$S_{\alpha\beta}^{\mathrm{intra}}(k) = 4\pi\rho \sum_i A_{\alpha\beta,i} R_{\alpha\beta,i}^2 \frac{\sin(kR_{\alpha\beta,i})}{kR_{\alpha\beta,i}} \tag{15}$$

Each RDF peak due to an intramolecular contribution gives a sinc function in the SF, with a large central peak at $k = 0$ and damped oscillations with a period of $2\pi/R_{\alpha\beta,i}$. The shorter the atomic distance $R_{\alpha\beta,i}$, the more the oscillations are spread out in reciprocal space. Strong peaks at short distances ($\approx 1\,\text{Å}$) in the partial RDFs $g_{\alpha\beta}(r)$ thus give rise to large damping oscillations covering the whole k-region. A clear example is the intermolecular/intramolecular partial SF $S_{\text{O-C}_\text{b}}$ in Fig. 11(b), which has oscillations ranging from $1\,\text{Å}^{-1}$ to more than $20\,\text{Å}^{-1}$.

A second observation when comparing the intermolecular ($S_{\mathrm{tot}}^{\mathrm{inter}}$) and inter/intra ($S_{\mathrm{tot}}$) spectra is that the position of the first peak is maintained at about $k = 1.5\,\text{Å}^{-1}$, while

theoretically its strength is mainly preserved, but substantially reduced in experiment (Fig. 9 and 10). This peak at low k is thus fed by both inter- and intramolecular contributions. This entanglement of contributions through the spectrum makes the decomposition a delicate task. Depending on the number of available total structure factors from NDIS measurements, the experimental data are incomplete to extract an accurate decomposition. The EPSR method, as developed by Soper,[2,25,73] assists in acquiring a plausible separation, as for the decomposition into the various partial SFs, by means of iterative refinement of empirical interatomic potentials (see Fig. 1). This method is not bias free.

Intermolecular versus intramolecular radial distribution functions.

A third observation is that the corresponding pair correlation functions - also called total radial distribution functions - differ significantly. The intermolecular $f_{\mathrm{tot}}^{\mathrm{inter}}(r)$ is one order of magnitude smaller than the combined inter/intra $f_{\mathrm{tot}}(r)$. This means that the intramolecular contribution is dominating the intermolecular contribution in $f(r)$: intra $\gg$ inter. Consequently, changes in the intermolecular partial RDFs are not visible in the total radial distribution function $f_{\mathrm{tot}}(r)$, as they are completely masked by the intramolecular contributions. This conclusion is not new and already reported in literature[2,16] and supports the general conclusion that the empirical procedure is prone to large inaccuracies. Total SFs are very insensitive to changes in partial intermolecular RDFs and this is illustrated too in Fig. 8 where the partial SFs $S_{\mathrm{O\text{-}O}}^{\mathrm{inter}}$ and $S_{\mathrm{C_b\text{-}C_b}}^{\mathrm{inter}}$ are hardly affected by the type of the MD simulation (and thus by the interactions, e.g. BSSE), while the corresponding RDFs exhibit clear structural differences in Fig. 7. This explicitly shows that the derivation of empirical partial intermolecular RDFs from a small number of total SFs from isotope variation experiments, is ill-conditioned: there may be multiple possibilities for the RDFs that match these total SFs, as spectacularly demonstrated by our simulations.

Returning to the observed shoulder structure in the C_4H_8O spectrum $S_{\mathrm{tot}}^{\mathrm{exp}}$ in Fig. 9(c), we notice that the shoulder is due to intramolecular interactions, as it is not present in $S_{\mathrm{tot}}^{\mathrm{emp,inter}}$. In addition, as the shoulder structure appears in the short-range region of reciprocal space,

its effect in real space is most likely small, and the discrepancy between experiment and theory is here thus probably of little importance.

Finally an interesting exercise is the comparison of the intermolecular partial SFs ($S_{\alpha\beta}^{\text{inter}}$) and the inter/intra partial SFs ($S_{\alpha\beta}$) (see Fig. 8, Fig. S.8 of SI versus Fig. 11, Fig. S.9 and Fig. S.10 of the SI). The figures confirm manifestly an intramolecular contribution at the first peak at small k values. This is not really surprising. The typical high peaks at very short distances in the intramolecular RDFs can be regarded as Dirac delta peaks, generating a sinc like function in reciprocal space as proposed in the model of Eq. (15). In the presence of one single RDF peak, the partial SF clearly shows a one-component damped harmonic oscillation with a central peak at about $1\,\text{Å}^{-1}$. The peaks in intermolecular RDFs are less sharp, and create anharmonicities in reciprocal space. Now, let us look back at $S_{\alpha\beta}^{\text{emp,inter}}$ (see Fig. 9, Fig. S.8 of SI). We expect anharmonicities in these intermolecular SFs. This is indeed the case for the theoretical curves derived from our ab initio molecular dynamics, but the empirical curves surprisingly describe quasi-perfect damped oscillations. This could support the hypothesis that the empirical intermolecular SFs are contaminated with some residual intramolecular contribution, and consequently the empirical RDFs are contaminated with intramolecular peaks. But this hypothesis does not match the observation that a weighted sum of the empirical intermolecular SFs coincides with that predicted by the MD simulations, as already mentioned. Concluding, the theoretical simulations give evidence that the extraction of partial SFs (error type 2) and the decomposition in intermolecular/intramolecular partial RDFs (error type 3) from experimental data are error-prone, as multiple solutions fit with the experimental data. Our study thus clearly shows how ab initio molecular dynamics may assist in distinguishing genuine from spurious structural features in the intermolecular partial RDFs.

A similar profound analysis can be performed on chloroform where visible spurious spikes are also observed in the region of short hydrogen-hydrogen distances (below 3.5 Å).[20] This maximum is reproduced neither by any of our MD simulations nor the RMC study of

Pothoczki.[20] Despite H-H partial SFs only contain intermolecular terms, the appearance of the spurious peak might be the result of some Fourier transform induced noise (error type 1) on the experimental total SF for $CHCl_3$ and $CDCl_3$, and a small perturbation may give rise to serious changes of the RDFs in real space, as clearly demonstrated in the case of THF.

Conclusions

This work comprises the study of the structure of five liquids: methanol, chloroform, acetonitrile, tetrahydrofuran and ethanol by means of DFT-based MD simulations in both a NVT and NpT ensemble. First, the influence of BSSE on the various liquid properties has been investigated. For this purpose the BSSE correction scheme as proposed in ref.[29] has been extended to all intermolecular interactions. MD simulations in a NpT ensemble show that they have a beneficial effect on the reproduction of the density in all five liquids under study. In only one case – chloroform – a serious underestimation of the density is observed. We ascribe this deficiency to the inappropriate description of the employed DFT functional BLYP-D3 in describing the intermolecular interactions in the absence of hydrogen bonds, as was confirmed by Schmidt *et al.* in NpT dynamics on water.[34] Second, properties have been computed from the simulations for a structural determination of the five liquids. All partial RDFs for the three methods under consideration (NpT/DFT-D3, NpT/DFT-D3+CP and NVT/DFT-D3) have been calculated, and compared with empirical data extracted from experiment if available. BSSE corrections turn out to overstructure the liquid in NpT simulations, but for the majority of the liquids under study such corrections remain minor and of no real importance in the further discussion on the overall quality of the theoretical RDFs.

In the liquids TCM and THF – where hydrogen bonds are not occurring – the theoretical RDFs do not reproduce experimentally derived properties. Peaks appear at short distances, which are completely missing in any MD simulation. The case of THF has been selected for further investigation as all necessary experimental data are available to allow for an in-

depth comparative study with the state-of-the-art MD simulations. As partial RDFs are not directly measurable but derived after an inverse Fourier transform of the partial structure factors, which in turn are determined from an iterative protocol combining neutron scattering experiments on isotope variations and MD runs with empirical interatomic potentials (RMC and EPSR methods), these properties are not free from cumulating inaccuracies generated in the various steps in the procedure. A fair assessment can only be performed in reciprocal space, since the total SF is the only direct experimental observable. Our simulations succeed in reproducing the total structure factor fairly well apart from a shoulder structure at small k, which turns out to originate from intramolecular interactions and thus of no importance for the partial intermolecular RDFs. An important conclusion we may draw from this study is that despite substantial structural differences in the partial RDFs or SFs between theory and experiment, they succeed in reproducing the same total intermolecular structure factor. This can only be achieved when large cancellation effects may occur between the various partial contributions in the weighted sum. There are thus multiple sets of partials SFs that yield the same total (intermolecular) structure factor, and in principle all of them can be regarded as plausible. There is no reason to favor or disfavor one set of partial RDFs over the other when they reproduce the correct intermolecular structure factor, unless additional partial SFs can be extracted from the neutron diffraction experiment, e.g. by changing the isotopic composition of the sample. The property that the intermoleculer spectrum is only noticeable in the low k-region, and hence characterized by a small number of genuine structural features, makes the decomposition of the total structure factor into its partial components even more ill-conditioned.

As well-known in literature we confirm that intramolecular interactions largely dominate the total structure factor in the whole range of reciprocal space, including the low k-region where we expect that it is mainly controlled by intermolecular contributions. This dominance is even more pronounced in real space, where intramolecular contributions to the total radial distribution function are systematically one order of magnitude larger than their

intermolecular counterparts.

The most delicate operation in the experimental procedure is the separation of the inter from the total structure factor; this procedure is highly error prone. From theoretical viewpoint, it is important that this decomposition can be accomplished with great care and with a minimum of inaccuracies. It constitutes the only intermolecular property which stands very closely to the measurements, as it may be derived from NDIS data, without relying on refined potentials, provided sufficient isotopic substitution can be performed. The intermolecular total structure factor can then serve as a reliable reference quantity for the RDFs resulting from various MD simulations.

The isotopic substitution technique is very powerful, but if the isotopic composition of the sample can only be changed by replacing hydrogens by deuteriums, the number of diffraction measurements generating independent information is small. To overcome this incompleteness, EPSR-refined models are constructed determining all the partial structure factors, of which a weighted sum leads to the $S_{\mathrm{tot}}^{\mathrm{inter}}$. The review by Soper reported most of the problems in data acquisition and data treatment.[2]

We tried to highlight the origin of errors by a detailed comparison of the experimental data with accurate ab initio MD. Our conclusion is that the experimentally derived properties are prone to large artifacts. To improve the reliability of the experimentally derived properties, we need an unbiased model-independent intermolecular total SF without the help of interatomic potentials. The lack of such a bias free quantity hinders a fully reliable comparison with theory. Summarizing, our ab-initio MD simulations succeed in reproducing fairly well all structure factors in THF available by experiment. The proton sample displays a shoulder at about $2\,\mathrm{\AA}^{-1}$, which is not present in the MD simulations, but originates from intramolecular interactions and consequently has no influence on the RDFs. What we finally may conclude is that MD simulations from first principles - and in particular NpT - are ideal tools to elucidate the liquid structure and to help in identifying spurious peaks, which may appear in RDFs resulting from RMC/EPSR procedures.

Acknowledgement

This work is supported by the Fund for Scientific Research – Flanders (FWO), the Research Board of Ghent University (BOF) and BELSPO in the frame of IAP/6/27, the Belgian Prodex office and ESA. The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by Ghent University, the Hercules Foundation and the Flemish Government – department EWI. VVS acknowledges the European Research Council under the European Community's Seventh Framework Programme (FP7(2007-2013) ERC grant agreement number 240483). AG and TV are post-doctoral researchers funded by the Foundation of Scientific Research - Flanders. Funding was also received from the Research Board of Ghent University (BOF), and BELSPO in the frame of IAP/7/05.

The authors would like to thank D. T. Bowron and A. K. Soper to kindly provide their experimental scattering data and experimentally derived partial radial distribution functions.

Supporting Information Available

More details on Molecular Dynamics runs; more details on BSSE and CP corrections; intermolecular radial distribution functions (RDFs) and structure factors for the five liquids; total inter- and intramolecular structure factors. The material is available free of charge via the Internet at http://pubs.acs.org. This material is available free of charge via the Internet at `http://pubs.acs.org/`.

References

(1) Clark, G. N. I.; Cappa, C. D.; Smith, J. D.; Saykally, R. J.; Head-Gordon, T. The Structure of Ambient Water. *Mol. Phys.* **2010**, *108*, 1415–1433.

(2) Soper, A. K. The Radial Distribution Functions of Water as Derived from Radiation Total Scattering Experiments: Is There Anything We Can Say for Sure? *ISRN Physical Chemistry* **2013**, *2013*, 279463–279530.

(3) Soper, A. K. The Quest for The Structure of Water and Aqueous Solutions. *J. Phys.: Condens. Matter* **1997**, *9*, 2717–2730.

(4) Soper, A. K. The Radial Distribution Functions of Water and Ice from 220 to 673 K and at Pressures up to 400 MPa. *Chem. Phys.* **2000**, *258*, 121–137.

(5) VandeVondele, J.; Mohamed, F.; Krack, M.; Hutter, J.; Sprik, M.; Parrinello, M. The Influence of Temperature and Density Functional Models in Ab Initio Molecular Dynamics Simulation of Liquid Water. *J. Chem. Phys.* **2005**, *122*, 014545.

(6) Todorova, T.; Seitsonen, A. P.; Hutter, J.; Kuo, I.-F. W.; Mundy, C. J. Molecular Dynamics Simulation of Liquid Water: Hybrid Density Functionals. *J. Phys. Chem. B* **2006**, *110*, 3685–3691.

(7) Lin, I.-C.; Seitsonen, A. P.; Coutinho-Neto, M. D.; Tavernelli, I.; Rothlisberger, U. Importance of van der Waals Interactions in Liquid Water. *J. Phys. Chem. B* **2009**, *113*, 1127–1131.

(8) McGrath, M. J.; Kuo, I.-F. W.; Siepmann, J. I. Liquid Structures of Water, Methanol, and Hydrogen Fluoride at Ambient Conditions from First Principles Molecular Dynamics Simulations with a Dispersion Corrected Density Functional. *Phys. Chem. Chem. Phys.* **2011**, *13*, 19943–19950.

(9) Yadav, V. K.; Chandra, A. Dynamics of Hydrogen Bonds and Vibrational Spectral Diffusion in Liquid Methanol From First Principles Simulations with Dispersion Corrected Density Functional. *Chem. Phys.* **2013**, *415*, 1–7.

(10) Vrhovšek, A.; Gereben, O.; Pothoczki, S.; Tomšič, M.; Jamnik, A.; Kohara, S.; Pusztai, L. An Approach Towards Understanding the Structure of Complex Molecular Systems: the Case of Lower Aliphatic Alcohols. *J. Phys.: Condens. Matter* **2010**, *22*, 404214.

(11) Vrhovšek, A.; Gereben, O.; Jamnik, A.; Pusztai, L. Hydrogen Bonding and Molecular Aggregates in Liquid Methanol, Ethanol, and 1-Propanol. *J. Phys. Chem. B* **2011**, *115*, 13473–13488.

(12) Zoranić, L.; Sokolić, F.; Perera, A. Microstructure of Neat Alcohols: A Molecular Dynamics Study. *J. Chem. Phys.* **2007**, *127*, 024502.

(13) Trouw, F. R.; Price, D. L. Chemical Applications of Neutron Scattering. *Annu. Rev. Phys. Chem.* **1999**, *50*, 571–601.

(14) Hull, S.; Norberg, S. T.; Eriksson, S. G.; Mohn, C. E. Neutron Powder Diffraction and Molecular Dynamics Study of Superionic $SrBr_2$. *J. Phys.: Condens. Matter* **2013**, *25*, 454205.

(15) Proffen, T.; Egami, T.; Billinge, S. J. L.; Cheetham, A. K.; Louca, D.; Parise, J. B. Building a High Resolution Total Scattering Powder Diffractometer – Upgrade of NPD At MLNSC. *Appl. Phys. A-Mater.* **2002**, *74*, S163–S165.

(16) Mason, P. E.; Neilson, G. W.; Dempsey, C. E.; Price, D. L.; Saboungi, M.-L.; Brady, J. W. Observation of Pyridine Aggregation in Aqueous Solution Using Neutron Scattering Experiments and MD Simulations. *J. Phys. Chem. B* **2010**, *114*, 5412–5419.

(17) Bowron, D. T.; Finney, J. L.; Soper, A. Structural Investigation of SoluteSolute Interactions in Aqueous Solutions of Tertiary Butanol. *J. Phys. Chem. B* **1998**, *102*, 3551–3563.

(18) Bowron, D. T.; Soper, A. K.; Finney, J. L. Temperature Dependence of The Structure of A 0.06 Mole Fraction Tertiary Butanol-Water Solution. *J. Chem. Phys.* **2001**, *114*, 6203.

(19) Bowron, D. T.; Finney, J. L.; Soper, A. K. The Structure of Liquid Tetrahydrofuran. *J. Am. Chem. Soc.* **2006**, *128*, 5119–5126.

(20) Pothoczki, S.; Temleitner, L.; Kohara, S.; Jóvári, P.; Pusztai, L. The Liquid Structure of Haloforms $CHCl_3$ and $CHBr_3$. *J. Phys.: Condens. Matter* **2010**, *22*, 404211.

(21) McGreevy, R. L.; Pusztai, L. Reverse Monte Carlo Simulation: A New Technique for the Determination of Disordered Structures. *Mol. Simulat.* **1988**, *1*, 359–367.

(22) McGreevy, R. L. Reverse Monte Carlo modelling. *J. Phys.: Condens. Matter* **2001**, *13*, R877.

(23) Lippert, G.; Hutter, J.; Ballone, P.; Parrinello, M. Response Function Basis Sets: Application to Density Functional Calculations. *J. Phys. Chem.* **1996**, *100*, 6231–6235.

(24) McGrath, M. J.; Siepmann, J. I.; Kuo, I.-F. W.; Mundy, C. J.; VandeVondele, J.; Hutter, J.; Mohamed, F.; Krack, M. Simulating Fluid-Phase Equilibria of Water from First Principle. *J. Phys. Chem. A* **2006**, *110*, 640–646.

(25) Soper, A. K. Partial Structure Factors From Disordered Materials Diffraction Data: An Approach Using Empirical Potential Structure Refinement. *Phys. Rev. B* **2005**, *72*, 104204.

(26) Thomas, J. L.; Tobias, D. J.; Jr., A. D. M. Direct Comparisons of Experimental and Calculated Neutron Structure Factors of Pure Solvents as a Method for Force Field Validation. *J. Phys. Chem. B* **2007**, *111*, 12941–12944.

(27) Gereben, O.; Pusztai, L. Conformational Analysis of Bis(methylthio)methane and Diethyl Sulfide Molecules in The Liquid Phase: Reverse Monte Carlo Studies Using Classical Interatomic Potential Functions. *J. Phys.: Condens. Matter* **2013**, *25*, 454201.

(28) Kühne, T. D.; Krack, M.; Parrinello, M. Static and Dynamical Properties of Liquid Water from First Principles by a Novel CarParrinello-like Approach. *J. Chem. Theory Comput.* **2009**, *5*, 235–241.

(29) Houteghem, M. V.; Verstraelen, T.; Ghysels, A.; Vanduyfhuys, L.; Waroquier, M.; Speybroeck, V. V. Analysis of The Basis Set Superposition Error in Molecular Dynamics of Hydrogen-Bonded Liquids: Application to Methanol. *J. Chem. Phys.* **2012**, *137*, 104506.

(30) Bedard-Hearn, M. J.; Larsen, R. E.; Schwartz, B. J. Understanding Nonequilibrium Solute and Solvent Motions through Molecular Projections: Computer Simulations of Solvation Dynamics in Liquid Tetrahydrofuran (THF). *J. Phys. Chem. B* **2003**, *107*, 14464–14475.

(31) Verstraelen, T.; Houteghem, M. V.; Speybroeck, V. V.; Waroquier, M. MD-TRACKS: A Productive Solution for the Advanced Analysis of Molecular Dynamics and Monte Carlo simulations. *J. Chem. Inf. Model.* **2008**, *48*, 2414–2424.

(32) Kristyán, S.; Pulay, P. Can (Semi)Local Density Functional Theory Account for the London Dispersion Forces? *Chem. Phys. Lett.* **1994**, *229*, 175–180.

(33) Hobza, P.; Šponer, J.; Reschel, T. Density Functional Theory and Molecular Clusters. *J. Comput. Chem.* **1995**, *16*, 1315–1325.

(34) Schmidt, J.; VandeVondele, J.; Kuo, I.-F. W.; Sebastiani, D.; Siepmann, J. I.; Hutter, J.; Mundy, C. J. IsobaricIsothermal Molecular Dynamics Simulations Utilizing Density Functional Theory: An Assessment of the Structure and Density of Water at Near-Ambient Conditions. *J. Phys. Chem. B* **2009**, *113*, 11959–11964.

(35) Grimme, S. Accurate Description of Van der Waals Complexes by Density Functional Theory Including Empirical Corrections. *J. Comput. Chem.* **2004**, *25*, 1463–1473.

(36) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for The 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.

(37) Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior. *Phys. Rev. A* **1988**, *38*, 3098–3100.

(38) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula Into a Functional of The Electron Density. *Phys. Rev. B* **1988**, *37*, 785–789.

(39) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–386.

(40) Boys, S. F.; Bernardi, F. The Calculation of Small Molecular Interactions by The Differences of Separate Total Energies. Some Procedures with Reduced Errors. *Mol. Phys.* **1970**, *19*, 553–566.

(41) Kruse, H.; Grimme, S. A Geometrical Correction for The Inter- and Intra-molecular Basis Set Superposition Error in Hartree-Fock and Density Functional Theory Calculations for Large systems. *J. Chem. Phys.* **2012**, *136*, 154101.

(42) Brandenburg, J. G.; Alessio, M.; Civalleri, B.; Peintiger, M. F.; Bredow, T.; Grimme, S. Geometrical Correction for the Inter- and Intramolecular Basis Set Superposition Error in Periodic Density Functional Theory Calculations. *J. Phys. Chem. A* **2013**, *117*, 9282–9292.

(43) CP2K, 2011; `http://cp2k.berlios.de`.

(44) VandeVondele, J.; Krack, M.; Mohamed, F.; Parrinello, M.; Chassaing, T.; Hutter, J. Quickstep: Fast and Accurate Density Functional Calculations Using a Mixed Gaussian and Plane Waves Approach. *Comput. Phys. Commun.* **2005**, *167*, 103–128.

(45) Goedecker, S.; Teter, M.; Hutter, J. Separable Dual-Space Gaussian Pseudopotentials. *Phys. Rev. B* **1996**, *54*, 1703–1710.

(46) Harwigsen, C.; Goedecker, S.; Hutter, J. Relativistic Separable Dual-Space Gaussian Pseudopotentials From H to Rn. *Phys. Rev. B* **1998**, *58*, 3641–3662.

(47) McGrath, M. J.; Siepmann, J. I.; Kuo, I.-F. W.; Mundy, C. J.; VandeVondele, J.; Hutter, J.; Mohamed, F.; Krack, M. IsobaricIsothermal Monte Carlo Simulations from First Principles: Application to Liquid Water at Ambient Conditions. *Chem. Phys. Chem.* **2005**, *6*, 1894–1901.

(48) Cunningham, G. P.; Vidulich, G. A.; Kay, R. L. Several Properties of Acetonitrile-Water, Acetonitrile-Methanol, and Ethylene Carbonate-Water Systems. *J. Chem. Eng. Data* **1967**, *12*, 336–337.

(49) Riddick, J. A.; Bunger, W. B.; Sakano, T. K. *Organic Solvents: Physical Properties and Methods of Purification*, 3rd ed.; Wiley-Interscience, 1970.

(50) Yaws, C. L. *Thermodynamic and Physical Property Data*; Golf Publishing, 1992.

(51) Gao, J.; Habibollazadeh, D.; Shao, L. A Polarizable Intermolecular Potential Function for Simulation of Liquid Alcohols. *J. Phys. Chem.* **1995**, *99*, 16460–16467.

(52) Franks, F.; Quickenden, M. A. J.; Reid, D. S.; Watson, B. Calorimetric and Volumetric Studies of Dilute Aqueous Solutions of Cyclic Ether Derivatives. *Trans. Faraday Soc.* **1970**, *66*, 582–589.

(53) Nosé, S. A Unified Formulation of The Constant Temperature Molecular Dynamics Methods. *J. Chem. Phys.* **1984**, *81*, 511.

(54) Martyna, G. J.; Tuckerman, M. E.; Tobias, D. J.; Klein, M. L. Explicit Reversible Integrators for Extended Systems Dynamics. *Mol. Phys.* **1996**, *87*, 1117–1157.

(55) Tuckerman, M. E.; Alejandre, J.; López-Rendón, R.; Jochim, A. L.; Martyna, G. J. A Liouville-Operator Derived Measure-Preserving Integrator for Molecular Dynamics Simulations in The Isothermal-Isobaric Ensemble. *J. Phys. A: Math. Gen.* **2006**, *39*, 5629–5651.

(56) Martínez, L.; Andrade, R.; Birgin, E. G.; Martínez, J. M. PACKMOL: A Package For Building Initial Configurations For Molecular Dynamics Simulations. *J. Comput. Chem.* **2009**, *30*, 2157–2164.

(57) van Duijneveldt, F. B.; van Duijneveldt - van de Rijdt, J. G. C. M.; van Lenthe, J. H. State of the Art in Counterpoise Theory. *Chem. Rev.* **1994**, *94*, 1873–1885.

(58) Rowland, R. S.; Taylor, R. Intermolecular Nonbonded Contact Distances in Organic Crystal Structures: Comparison with Distances Expected from van der Waals Radii. *J. Phys. Chem.* **1996**, *100*, 7384–7391.

(59) Center for Molecular Modeling, `http://molmod.ugent.be/software`.

(60) Hansen, N.; Ostermeier, A. Completely Derandomized Self-Adaptation in Evolution Strategies. *Evol. Comput.* **2001**, *9*, 159–195.

(61) Yamaguchi, T.; Hidaka, K.; Soper, A. The Structure of Liquid Methanol Revisited: A Neutron Diffraction Experiment At −80 °C and +25 °C. *Mol. Phys.* **1999**, *96*, 1159–1168.

(62) Yamaguchi, T.; Hidaka, K.; Soper, A. Erratum: The Structure of Liquid Methanol Revisited: A Neutron Diffraction Experiment At −80 °C and +25 °C. *Mol. Phys.* **1999**, *97*, 603–605.

(63) Handgraaf, J.-W.; Meijer, E. J.; Gaigeot, M.-P. Density-Functional Theory-Based Molecular Simulation Study of Liquid Methanol. *J. Chem. Phys.* **2004**, *121*, 10111.

(64) Martín, M. E.; Sánchez, M. L.; del Valle, F. J.; Aguilar, M. A. *J. Chem. Phys.* **2002**, *116*, 1613.

(65) Ishiyama, T.; Sokolov, V. V.; Morita, A. A Theoretical Study of Liquid Alcohols Using Averaged Solvent Electrostatic Potentials Obtained From Molecular Dynamics Simulations: Methanol, Ethanol and Propanol. *J. Chem. Phys.* **2011**, *134*, 024509.

(66) Yu, H.; Geerke, D. P.; Liu, H.; Gunsteren, W. F. V. Molecular Dynamics Simulations of Liquid Methanol and Methanol-Water Mixtures with Polarizable Models. *J. Comput. Chem.* **2006**, *27*, 1494–1504.

(67) Patel, S.; Brooks, C. L. A Nonadditive Methanol Force Field: Bulk Liquid and Liquid-Vapor Interfacial Properties Via Molecular Dynamics Simulations Using a Fluctuating Charge Model. *J. Chem. Phys.* **2005**, *122*, 024508.

(68) Dietz, W.; Heinzinger, K. A Molecular Dynamics Study of Liquid Chloroform. *Ber. Bunsenges. Phys. Chem.* **1985**, *89*, 968–977.

(69) Böhm, H. J.; Ahlrichs, R. Molecular Dynamics Simulation of Liquid CH_2Cl_2 and $CHCl_3$ With New Pair Potentials. *Mol. Phys.* **1985**, *54*, 1261–1274.

(70) Torii, H. Atomic Quadrupolar Effect in Intermolecular Electrostatic Interactions of Chloroalkanes: The Cases of Chloroform and Dichloromethane. *J. Mol. Liq.* **2005**, *119*, 31–39.

(71) Sears, V. F. Neutron Scattering Lengths and Cross Sections. *Neutron News* **1992**, *3*, 26–37.

(72) Bowron, D. T. Private Communication.

(73) Soper, A. K. Empirical Potential Monte Carlo Simulation of Fluid Structure. *Chem. Phys.* **1996**, *202*, 295–306.

Figure 12: TOC figure