ELSEVIER

Chemical Physics 202 (1996) 295–306

Chemical Physics

Empirical potential Monte Carlo simulation of fluid structure

A.K. Soper *,[1]

ISIS Facility, Rutherford Appleton Laboratory, Chilton, Didcot, Oxon, OX11 0QX, UK

Received 18 May 1995; in final form 2 October 1995

Abstract

It is shown that data on the site–site pair correlation functions for a fluid of molecules can be used to derive a set of empirical site–site potential energy functions. These potential functions reproduce the fluid structure accurately but at the present time do not reproduce thermodynamic information on the fluid, such as the internal energy or pressure. The method works in an iterative manner, starting from a reference fluid in which only Lennard-Jones interactions are included, and generates, by Monte Carlo simulation, successive corrections to those potentials which eventually lead to the correct site–site pair correlation functions. Using this approach the structure of water as determined from neutron scattering experiments is compared to the structure of water obtained from the simple point charge extended (SPCE) model of water interactions. The empirical potentials derived from both experiment and SPCE water show qualitative similarities with the true SPCE potential, although there are quantitative differences. The simulation is driven by a set of potential energy functions, with equilibration of the *energy* of the distribution, and not, as in the reverse Monte Carlo method, by equilibrating the value of χ^2, which measures how closely the simulated site–site pair correlation functions fit a set of diffraction data. As a result the simulation proceeds on a true random walk and samples a wide range of possible molecular configurations.

1. Introduction

Since its inception in 1988, reverse Monte Carlo simulation (RMC) [1] has become a powerful and widely used tool for the analysis of structure in disordered materials [2]. The primary objective of RMC is to set up a distribution of atoms and molecules which is consistent with known or measured structural information for a particular material, normally in the form of structure factors obtained by X-ray and neutron diffraction. The power of the method arises from the fact that the available data are modelled in terms of a physical distribution of particles which has the same density as the material under question, and the particles themselves are not permitted to overlap in a manner representative of the real system. Various other constraints can be built into the simulation, such as the expected coordination and arrangement of atoms in molecules. In that sense therefore the RMC technique for disordered materials is analogous to the Rietveld refinement technique [3] used in powder crystallography.

However questions have often been raised about the utility of RMC simulation in particular cases. The main stumbling block to it becoming completely accepted as a valid tool for analysing disordered materials structure appears to be the question of uniqueness [2]. It is quite possible that for a given set of diffraction data there is more than one particle distribution which is consistent with those data, even

* Corresponding author.

[1] Also affiliated with Department of Physics and Astronomy, University College London, Gower Street, London, WC1E 6BT, UK.

SSDI 0301-0104(95)00357-6

before allowance is made for measuring uncertainties. For example a certain peak at a distance r in the site–site pair correlation function, $g(r)$, may have an integrated coordination number of say four, but $g(r)$ by itself cannot distinguish between the case where every atom is coordinated by *exactly* four atoms at this distance, and the case where four represents the *average* coordination, with some atoms coordinated perhaps with three atoms, other atoms have four atoms, while yet others may have five atoms, all at the same distance. In order to circumvent this kind of difficulty *additional* information about the local coordination of atoms is required.

In fact the lack of uniqueness in the RMC solution should really be regarded as a strength, since the RMC method allows the user to define exactly what constraints the solution is expected to satisfy. The imposed constraints can be tested against the data: if no solution can be found, within the limits imposed by experimental error, it is likely the constraints are not supported by the data. At the same time the fact that the RMC solution does not have an expected three-dimensional arrangement of atoms for a particular material probably indicates that the $g(r)$ data either do not support that structure at all or else are simply not sensitive to that particular structure. The results of the analysis are therefore much less likely to contain the subjective bias that sometimes is associated with direct interpretation of the peaks in measured site–site pair correlation functions.

It has been shown (see for example Refs. [4,5]) that if the particles in a fluid interact by purely pairwise additive forces, then there is a unique relationship between the site–site pair potential and the site–site pair correlation function. Even in systems where many-body forces are important, considerable progress can be made with pairwise forces by inventing an *effective* site–site pair potential which takes account of many-body effects in an average way. If, therefore, a set of effective site–site pair potentials can be found that reproduces a given set of site–site pair correlation functions, then on the basis of the uniqueness theorem, that potential should also produce a realistic 3-dimensional distribution of atoms or molecules. This would go at least some way towards assuaging fears about the uniqueness of the result. Given that computer simulation provides a numerical, but exact, route to solving the many-body problem in disordered systems, it is perhaps surprising that remarkably little work appears to have been done on the problem of inverting a set site–site pair correlation functions by computer simulation to produce an effective site–site pair potential for the system of interest.

Water is an excellent example of such a case. The simple point charge (SPC) model of water interactions, and its extension, SPCE [6] is widely used in the computer simulation of water and aqueous systems. Although simple in form, this potential predicts the liquid–vapour coexistence curve of water with good accuracy [7]. It is also able to predict quite accurately the site–site pair correlation functions, $g_{HH}(r)$, $g_{OH}(r)$ and $g_{OO}(r)$, for water under ambient conditions, although there is now some evidence that it may not do so well as the critical point of water is approached [8]. Frustratingly, however, from the point of view of the experimentalist, none of the existing water potentials predict the site–site pair correlation functions entirely quantitatively: there are always small differences in peak positions and heights, which are outside the sometimes significant measuring errors. The same comment applies to almost every material which has been studied by diffraction and computer simulation. In that situation it not possible to establish quantitatively how accurately the simulation is actually reproducing a material's structure: it is possible for example that differences between simulated and experimental site–site pair correlation functions arise from particular assumptions in the model's potential energy function.

The principle behind RMC simulation is to set up distributions of particles which are consistent with measured quantities, such as the site–site pair correlation function. To do this it is normally assumed that short range hard-core interactions exist between the particles to prevent atomic overlap. The simulation follows that of a standard Monte Carlo simulation with random moves. Providing a move does not violate atomic overlap, the site–site pair correlation function is calculated and compared with the data via the χ^2 statistics, which measures how well the simulated site–site pair correlation fits the data, given some assumed error bar on the data, $\epsilon(r)$:

$$\chi^2 = \sum_r \left(\frac{g(r) - g^D(r)}{\epsilon(r)} \right)^2, \qquad (1)$$

where $g(r)$ is the simulated site–site pair correlation function and $g^{D}(r)$ is the correlation function obtained from diffraction data. If the change in χ^2 is negative the move is always accepted, but if positive is accepted only with a Boltzmann-type of probability distribution. The "temperature" in this distribution controls how small the value of χ^2 becomes, as in a conventional simulation.

However the procedure of constraining χ^2 contains two inherent weaknesses. Firstly χ^2 is unable to distinguish between two configurations of particles, one of which has a large statistical uncertainty but lies through the supplied data, while the other has lower statistical uncertainty but significantly misfits the peaks. Secondly for any computer simulation the site–site pair correlation function will have its own statistical uncertainty as a function of r which arises from the necessarily finite number of particles involved in the simulation. If the number of particles in the simulation is N, and the density is ρ, then the statistical uncertainty in $g(r)$ at distance r is of order $1/\sqrt{4\pi\rho N r^2 \Delta r}$, where Δr is the bin width used to define $g(r)$. For $N = 1000$, $\rho = 0.05$ per Å^3, and $\Delta r = 0.02$ Å for example, this uncertainty is of order 10% at $r = 3$ Å. In particular it has nothing to do with the quality of the supplied data, but is defined only by the parameters of the simulation. On the other hand modern diffraction data is frequently available with statistical uncertainties much better than 1%. Thus any attempt to fit $g(r)$ to better than this intrinsic uncertainty by reducing the size of $\epsilon(r)$ may result in the simulation becoming locked up in a local minimum and not proceeding on a true random walk.

The ideal situation therefore would be to develop a simulation where the individual distributions produce $g(r)$s with large statistical deviations but lying through the data. However when these individual distributions are averaged over many configurations the statistical deviations should become negligible. In this way it would be seen that the simulation was indeed sampling a wide range of phase space and therefore correctly mimicking the experiment.

With these thoughts in mind the question arises as to whether it is possible to extend the RMC algorithm so that instead simply deriving a distribution of particles which is consistent with the structure data, it also establishes empirically an effective site–site pair potential (or set of such potentials for multicomponent systems) which, when used in a computer simulation of the fluid, reproduce quantitatively the site–site pair correlation function obtained from diffraction data. A significant advantage of such a scheme over conventional RMC is that since the energy would be constrained instead of χ^2 the simulation would be less likely to become locked in a local configuration, and the individual simulated site–site pair correlation functions would have the correct statistical uncertainties. Furthermore the potentials would not be limited to Monte Carlo simulation and could be used in molecular dynamics calculations as well.

The rest of this paper is devoted to an exploration of how this might be achieved, with some preliminary results for water being shown at the end.

2. Empirical potential Monte Carlo method

It is assumed at the outset that a set of "data", the site–site pair correlation functions for the material of interest, $g^{D}_{\alpha\beta}(r)$, exist. These may have been derived from a previous computer simulation of the material, or could have been derived from the corresponding measured partial structure factors of the material. If the latter is the case it will be important that the correlation functions do not carry the truncation and other errors that are often present when Fourier transforming diffraction data. Methods for achieving the Fourier transform reliably have been discussed elsewhere (see for example Ref. [9]).

The starting point for the development is the *potential of mean force*, $\psi_{\alpha\beta}(r)$, between atoms α and β in the fluid [10], where

$$\psi_{\alpha\beta}(r) = -kT \ln\left[g_{\alpha\beta}(r) \right], \tag{2}$$

and $g_{\alpha\beta}(r)$ is the site–site pair correlation function between α and β. For the "data" there is an equivalent potential of mean force: $\psi^{D}_{\alpha\beta}(r) = -kT \ln[g^{D}_{\alpha\beta}(r)]$. These potentials are by definition purely pairwise potentials, but cannot be used in a computer simulation because they already contain the many-body cooperative effects arising from the packing of particles in the material. However if a computer simulation is performed with an assumed initial potential, then the potential of mean force can

be used to indicate where that potential needs to be modified if it is to reproduce the measured site–site pair correlation functions accurately.

The procedure is to set up a model fluid with the correct density and temperature of the material in question, using an assumed potential, the reference potential, $U_{\alpha\beta}^{\mathrm{ref}}(r)$ between sites α and β. Typically the reference potential will incorporate known hard-core limitations on atomic and molecular overlap. In addition if a material contains molecules the atoms are constrained so that the molecular structure is conserved as the simulation proceeds. The simulation with the reference potential gives rise to an initial estimate of the site–site pair correlation functions. Setting the initial potential, $U_{\alpha\beta}^{\mathrm{O}}(r)$, and the site–site correlation functions, $g_{\alpha\beta}(r)$, equal to their initial (reference) values, $U_{\alpha\beta}^{\mathrm{ref}}(r)$ and $g_{\alpha\beta}^{\mathrm{ref}}(r)$ respectively, the potential of mean force (1) is used to generate a new potential energy function, $U_{\alpha\beta}^{\mathrm{N}}(r)$, as a perturbation to the initial potential:

$$\begin{aligned} U_{\alpha\beta}^{\mathrm{N}}(r) &= U_{\alpha\beta}^{\mathrm{O}}(r) + \left[\psi_{\alpha\beta}^{\mathrm{D}}(r) - \psi_{\alpha\beta}(r)\right] \\ &= U_{\alpha\beta}^{\mathrm{O}}(r) + kT\left\{\ln\left[g_{\alpha\beta}(r)/g_{\alpha\beta}^{\mathrm{D}}(r)\right]\right\}. \end{aligned} \tag{3}$$

The new potential energy function $U_{\alpha\beta}^{\mathrm{N}}(r)$ now replaces $U_{\alpha\beta}^{\mathrm{O}}(r)$ in the simulation, which proceeds as before but with a revised potential.

Subsequently, after a sufficient number of particle moves to bring the simulation into equilibrium with the new potential, the simulation is stopped again and a further perturbation to the empirical potential is estimated, based on the current site–site pair correlation functions. This process is repeated until such point that, assuming the process described here is convergent,

$$U_{\alpha\beta}^{\mathrm{O}}(r) \approx U_{\alpha\beta}^{\mathrm{N}}(r) = U_{\alpha\beta}(r) \tag{4}$$

and hence

$$g_{\alpha\beta}(r) \approx g_{\alpha\beta}^{\mathrm{D}}(r) \tag{5}$$

for all r and all atom pairs. At that point a set of empirical potential energy functions has been derived which are able to reproduce the observed site–site pair correlation functions of the material in question. The sequence of steps described here is the basis for the algorithm developed in the present work.

Before proceeding to describe the practicalities of this approach, it is important to discuss the convergence of such a process, and how close this method will lead to a true empirical potential for the material in question. The convergence of the algorithm hinges on one basic assumption, i.e. that a set of site–site pairwise additive potentials exist which can generate the observed site–site pair correlation functions. If that is not possible then obviously no solution will be found, since all calculations have to be performed in the pair approximation. In practice, as will seen below, it is likely the method will find only one of a set of possible pair potentials, depending on the constraints imposed on the simulation, especially those related to the maximum distance to which the potential is to be defined.

The convergence of the algorithm is established as a consequence of the uniqueness theorem for the site–site pair correlation function in pairwise additive systems (see Ref. [4], p. 178). This theorem states that for two pair potentials, $u_{\alpha\beta}^{(0)}(r)$ and $u_{\alpha\beta}^{(1)}(r)$, which differ by more than a constant, then it can be shown that

$$\int \mathrm{d}r \left[u_{\alpha\beta}^{(1)}(r) - u_{\alpha\beta}^{(0)}(r)\right]\left[g_{\alpha\beta}^{(1)}(r) - g_{\alpha\beta}^{(0)}(r)\right] < 0, \tag{6}$$

where $g_{\alpha\beta}^{(0)}(r)$, $g_{\alpha\beta}^{(1)}(r)$, are the site–site pair correlation functions which arise from to the two potentials. This inequality can only hold if $g_{\alpha\beta}^{(1)}(r) \neq g_{\alpha\beta}^{(0)}(r)$ when averaged over the *whole* of r-space, and hence a given set of $u_{\alpha\beta}(r)$ will give rise to a unique set of $g_{\alpha\beta}(r)$. Of course what is NOT addressed by this theorem is the *sensitivity* of the site–site pair correlation function to the pair potential: it is conceivable that a particular subtlety of the potential might not appear as a particular feature in the site–site pair correlation, simply due to the fact that any estimate of the latter quantity will have computational uncertainties associated with it. By the same token, if, as in the present case, the pair potential is available over only a limited range of radius values, and is assumed zero beyond that region, it may produce, over that r range and within computational limits, the same site–site pair correlation function as a different potential which is defined over all radius values. Hence the limits of the calculation mean that at the present time the derived potential may only be

unique for the particular set of constraints imposed on it.

It is however important to establish that the proposed algorithm will in fact find a solution. The convergence of the proposed algorithm stems from the fact that the left-hand side of (6) is *less* than zero. That is if we make a change to the pair potential the site–site pair correlation function will respond in a predictable direction, plus or minus, even if the magnitude of the change is not predictable. The fact that the left-hand side is always negative means than an increase in the potential function will always cause an overall decrease in the site–site pair correlation function, and vice versa. To see this, the site–site correlation function is written as a sum of two terms, i.e. the pair interaction term and the term arising from many-body effects:

$$g_{\alpha\beta}(r) = g_{\alpha\beta}^{(\mathrm{p})} + g_{\alpha\beta}^{(\mathrm{m})}(r), \tag{7}$$

where

$$g_{\alpha\beta}^{(\mathrm{p})}(r) = \exp\left[-\beta u_{\alpha\beta}(r)\right], \tag{8}$$

$$g_{\alpha\beta}^{(\mathrm{m})}(r) = g_{\alpha\beta}^{(\mathrm{p})}(r)\left\{\exp\left[\beta w_{\alpha\beta}(r)\right] - 1\right\}. \tag{9}$$

$\beta = 1/kT$ and $w_{\alpha\beta}(r)$ represents the difference between the pair potential and the potential of mean force. At zero density $w_{\alpha\beta}(r) = 0$ but at finite density the existence of the many-body terms prevents the algorithm from being a simple linear problem with exact solutions.

It will be apparent that any changes made to the site–site pair potentials, $\Delta u_{\alpha\beta}(r)$ will give rise, after a suitable equilibration period of the simulation, to changes in each of these terms, $\Delta g_{\alpha\beta}^{(\mathrm{p})}$, $\Delta g_{\alpha\beta}^{(\mathrm{m})}(r)$ respectively. In particular the form (8) dictates that $\Delta g_{\alpha\beta}^{(\mathrm{p})}(r)\Delta u_{\alpha\beta}(r) < 0$ for all r so that $-\int \mathrm{d}r\, \Delta g_{\alpha\beta}^{(\mathrm{p})}(r)\Delta u_{\alpha\beta}(r) > 0$. If this and (7) are substituted into (6) it is found that there is a *positive* upper bound on the integral of the many-body term:

$$\int \mathrm{d}r\, \Delta g_{\alpha\beta}^{(\mathrm{m})}(r)\Delta u_{\alpha\beta}(r) < -\int \mathrm{d}r\, \Delta g_{\alpha\beta}^{(\mathrm{p})}(r)\Delta u_{\alpha\beta}(r). \tag{10}$$

The kernel of the integral of the left-hand side of (10) is likely to have a variable sign as a function of r, since at some radius values the many-body term will enhance the changes to the pair term while at others it will tend to cancel them. The kernel of the integral on the right-hand side is however negative at all radius values. Therefore it is impossible that changes in the many-body term due to the change in the pair potential will completely annul the changes in the pair term when averaged over all r values. For this reason it is justifiable to use the potential of mean force to guide the choice of perturbation of the pair potential as in Eq. (4). The process of iterating the algorithm repeatedly will ensure that any overshoot or undershoot of the pair potential caused by the many-body interactions will eventually be cancelled. If a problem of undershoot or overshoot did become apparent it would always be possible to make the amplitude of the perturbation in (4) smaller, or else employ a more sophisticated refinement technique, such as Newton–Raphson.

3. Practical considerations

The algorithm of the preceding section needs some refinements in order to be implemented in practice. Firstly any estimates of the $g_{\alpha\beta}(r)$ by computer simulation will have a degree of randomness, which arises from the fact that a computer simulation with typically only a few hundred or few thousand particles, necessarily samples only a tiny fraction of all possible particle configurations. This can in principle be avoided by running the simulation for a long period and taking the average of a very large number of configurations so that the $g_{\alpha\beta}(r)$ are smooth, but this would make any practical estimation of the effective pair potential extremely time consuming. If the estimated $g_{\alpha\beta}(r)$ are used in Eq. (4) then the noise present in the estimated correlation functions is transferred to the effective potential, and unless steps are taken to suppress this noise, it will grow in amplitude with each iteration of the algorithm and eventually render the estimated potential useless.

There are several possible ways around this difficulty. A simple solution that has been used in the present work is to measure the noise in the estimated potential function, using the noise function defined in Ref. [11]. If this noise becomes greater than a specified threshold value a smoother version of the potential is automatically generated by replacing each

value of the potential at a given radius r_j by the value

$$U^{N}_{\alpha\beta}(r_j) = \tfrac{1}{4}\left[U^{O}_{\alpha\beta}(r_{j-1}) + 2U^{O}_{\alpha\beta}(r_j) + U^{O}_{\alpha\beta}(r_{j+1})\right], \tag{11}$$

with the process repeated if necessary until the noise falls below the threshold value again. Of course such a smoothing process may modify the potential energy function and so enforce a limit on how well the derived potential energy can reproduce the supplied site–site pair correlation functions. The practice of smoothing the potential does not appear to cause undue distortions to the estimated correlation functions for the cases that have been dealt with to date.

Related to the question of smoothness is the fact that $g_{\alpha\beta}(r)$ can only be obtained at a discrete set of r values, which means the empirical potential will also be discrete. For the present simulations a simple linear interpolation between r values was found to give quite satisfactory results, although obviously more sophisticated interpolation schemes could be envisaged [12].

A third feature is that is there is no control of the form of the potential beyond the range of the supplied data, r_{max}. In conventional computer simulation the potential is truncated at an appropriate radius value, r_{cut} [12], and approximate corrections made for particles separated by distances greater than this cut-off distance. For the empirical potential there is no indication of the form of the potential beyond r_{max} so such corrections cannot be made.

In order to prevent discontinuities in the derivatives of the empirical potential occurring at r_{max} a damping function is used to modify the perturbation defined by Eq. (4):

$$U^{N}_{\alpha\beta}(r) = U^{O}_{\alpha\beta}(r) + kT\left\{\ln\left[g_{\alpha\beta}(r)/g^{D}_{\alpha\beta}(r)\right]\right\}D(r), \tag{12}$$

where

$$D(r) = \exp\left(-\frac{r^2}{2\kappa^2}\right),$$

so that the amplitude of the perturbation declines with increasing r. Here κ is a decay constant which was set to 4 Å for the examples used in the present work, which have a value of $r_{max} \approx 10$ Å. In practice the results of the simulation do not appear to depend strongly on the choice of damping function and other ones have been tested: this one was chosen because it is smoothly varying with r and has continuous derivatives.

Finally the formula for the potential of mean force is accurate provided $g_{\alpha\beta}(r) > 0$. As the site–site pair correlation functions approach zero at low r this formula becomes increasingly unreliable. To avoid the indeterminacies at low r the argument of the logarithm in (12) is replaced with $\{[g_{\alpha\beta}(r) + \delta]/[g^{D}_{\alpha\beta}(r) + \delta]\}$, with δ set to a small number such as $\delta = 0.0001$. Hence no adjustment to the empirical potential can be made when both $g_{\alpha\beta}(r)$ and $g^{D}_{\alpha\beta}(r)$ approach zero. However the assumed reference potential (which is not varied in the iteration procedure) will always be present to ensure that atomic overlap cannot occur. Because measured site–site pair correlation functions are not highly accurate at low r it is likely the empirical potential will also not be especially accurate at low r.

The simulation itself proceeds along standard lines, (see for example Ref. [12]). The main difference is that instead of the potential being fixed at the outset, it varies iteratively along the lines described in this and the preceding sections. Once the initial configuration of particles has been equilibrated with the reference potential, it has not been found necessary to re-equilibrate the distribution at each iteration of the algorithm because the process of continually modifying the potential by small amounts ensures that the distribution is eventually equilibrated. It is worth noting that because the simulation is driven by a potential, it is found that individual distributions can deviate quite substantially from the data, even though the ensemble distribution averaged over many particle configurations follows the data closely. Hence it is likely that the particle configurations generated by empirical potential Monte Carlo simulation will sample a large range of phase space.

4. Application to liquid water

In order to test the method described in the previous sections, and also to check that the simulation

program was operating correctly, a set of site–site correlation functions for water were first derived using the SPCE water potential [6]. The simulation was performed with 256 water molecules in a cubic box of side 19.72 Å, giving a number density of 0.0334 molecules per Å^3. The OH and HH intramolecular distances were set to 1.0 and 1.62 Å respectively, as in the SPCE potential. The short range Lennard-Jones potential and the charge interactions were treated within the minimum image convention [12] extending throughout the box. However in order to save computing time no corrections for longer range interactions were included, so that the site–site pair correlation functions obtained differ slightly from those obtained with full Ewald summation of long range interactions [13]. The longer range interactions can be ignored in the present instance because the main purpose of the simulation here is not to describe the structure of SPCE water per se, but to establish whether the empirical potential Monte Carlo method is viable. As was seen in the previous section, the empirical potential simulation can make no estimate of the potential outside the maximum range of the $g(r)$ data, which for a computer simulation is half the dimension of the box in which the simulation is performed. In fact as shown in Fig. 1 the OO, OH and HH site–site pair correlation functions obtained in this work compare well with those obtained previously [13].

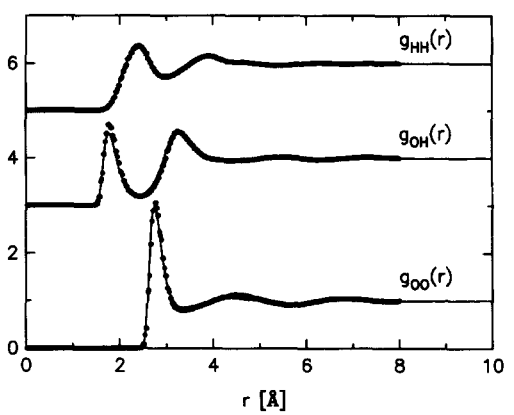

Fig. 1. Comparison of the site–site pair correlation functions for water obtained from the present Monte Carlo simulation program and using the true SPCE potential [6] (line), with the results given by Kusalik [13] (circles). This figure establishes the correctness of the simulation algorithm being used here.

Table 1
Lennard-Jones parameters for the empirical potential simulation of SPCE and experimental water

Atomic species	σ_α (Å)	ϵ_α (kJ per mol)
O	2.5	1.00
H	1.4	1.00

The reference potentials assumed in the empirical potential simulation work were in the form of site–site Lennard-Jones potentials,

$$U_{\alpha\beta}^{\mathrm{LJ}}(r) = 3\epsilon_{\alpha\beta}\left[\left(\frac{\sigma_{\alpha\beta}}{r}\right)^{12} - \left(\frac{\sigma_{\alpha\beta}}{r}\right)^{6}\right]T(r), \qquad (13)$$

with the hard core and well depth determined from the usual Lorentz–Berthelot mixing conditions:

$$\sigma_{\alpha\beta} = \tfrac{1}{2}(\sigma_\alpha + \sigma_\beta), \quad \epsilon_{\alpha\beta} = (\epsilon_\alpha \epsilon_\beta)^{1/2}.$$

Here $T(r)$ is a truncation function so that the Lennard-Jones potential goes smoothly to zero at $r = r_{\max}$. As with the damping function, $D(r)$, the results do not appear to depend significantly on the precise form of this truncation function. The form used here was that fifth-order polynomial in r which is continuous and has continuous first and second derivatives at r_1 and $r_{\max}$, with $r_1 = 4$ Å for the present examples [14].

Table 1 lists the values of these parameters which were used in the current analysis for water. It should be emphasized that these values are there primarily to prevent atomic and molecular overlap: the ϵ values were chosen to be large enough to prevent all possible atomic overlap in the event that the modifications to the reference potentials became large, and the σ values were determined from the lowest r values for which $g(r)$ data are non-zero. The same set of values were used for empirical potential simulations of both SPCE and experimental water.

It should be noted that the molecules in this simulation were not strictly rigid: in addition to the Lennard-Jones potential for inter-molecular interactions, the atoms within the molecules were assumed to interact via harmonic potentials, but the force constants for these potentials were held sufficiently large that typical fluctuations of the OH and HH intramolecular distances were on the order of 0.02 Å, since only the intermolecular site–site pair correla-

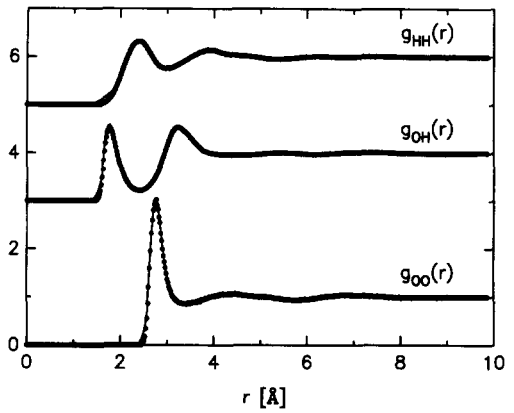

Fig. 2. Comparison of the site–site pair correlation functions for water obtained from simulation with the true SPCE potential (circles) with those obtained with the empirical SPCE potential (line).

tion functions are available from the SPCE potential. If intra-molecular peaks had been present in the site–site pair correlation functions then these force constants would be relaxed, and the intra-molecular potential would be refined in the same way that the intermolecular potential is refined here.

Fig. 2 shows the site–site pair correlation functions derived by empirical potential Monte Carlo (EPMC) simulation of the "data" of Fig. 1: it is found that the two sets of correlation functions are very close to one another. Fig. 3 shows the derived empirical site–site pair potentials and compares them with the true SPCE potentials. It is immediately

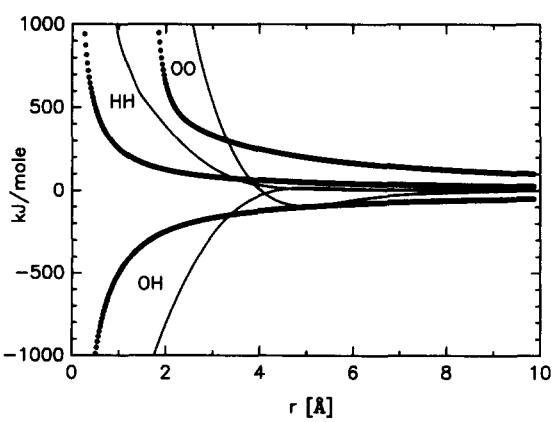

Fig. 3. Site–site empirical potentials as derived in the present work (line), compared to the true SPCE potential (circles).

apparent that the empirical potentials are rather different from the true SPCE potential, even though both potentials produce similar site–site pair correlation functions. This difference in the potentials is even more evident in the internal energies of each simulation. For the simulation with the true SPCE potential the excess internal energy was −47 kcal/mol, close to the published value of −41.4 kcal/mol for this potential [6]: the difference arises from the truncation imposed on the Lennard-Jones terms, Eq. (13), and the fact that long range corrections were neglected in this simulation. For the EPMC simulation of the SPCE pair correlation functions however the excess internal energy is a factor of 4 times more negative, at −161 kcal/mol. Such large differences in the simulated energies arise no doubt from the large differences in the site–site pair potentials as seen in Fig. 3. If a way could be found for constraining the energy of the EPMC simulation (and perhaps other thermodynamic quantities) it is very likely the discrepancies seen in this figure would be considerably reduced. Until such a constraint can be imposed it is not particularly meaningful to calculate the energy or other thermodynamic properties from the empirical potential simulation. The fact remains as shown in Fig. 2 that geometrical arrangements of molecules have been generated which closely reproduce the site–site pair correlation functions of the SPCE potential.

Nonetheless in spite of the obvious differences between true and empirical potentials, the *qualitative* trends are similar, i.e. the OO and HH potentials are mostly positive, with the HH potential roughly one quarter of the OO potential at low r values, as is the case of the true SPCE potential, and the OH potential is mostly negative, as expected.

The principal differences between empirical and true potentials occur at low r values where the empirical potentials diverge more rapidly than the true potential, and at large r values, where the empirical potential decays to zero much faster than the true potential. At least part of the discrepancy at low r arises from the different Lennard-Jones parameters used in the empirical potential simulation compared to the SPCE values [6]. The values for the empirical potential simulation were chosen primarily to prevent molecular overlap – it would have been perfectly feasible to run the simulation with the same

values as SPCE, but those values might not have been suitable for the simulation of the experimentally determined correlation functions of water. Therefore values were chosen to be compatible with the low r behaviour of the experimental site–site pair correlation functions.

The deviation at large r is necessitated by the damping function, since the larger r behaviour of the potential is not available via the present empirical potential method. As discussed above it remains to be seen whether longer range information such as the experimental bonding energy per mol can be incorporated into the estimate of the empirical potential.

Another indication of the viability of the method can be obtained by comparing the size of the deviations of individual distributions about the average distribution. Fig. 4 shows the OO correlation for the two simulations with the standard deviation of individual distributions about the mean represented as the "error bars" in the figure. (All the simulations reported in this paper were accomplished with 256 water molecules.) It can be seen that these deviations are of a comparable magnitude in the two cases, implying that the EPMC simulation has sampled an equivalent range of molecular configurations as the true SPCE simulation. Note also that the noise fluctuations in the average curves are small compared to the error bars, which suggests both simulations have performed a true ensemble average over many configurations.

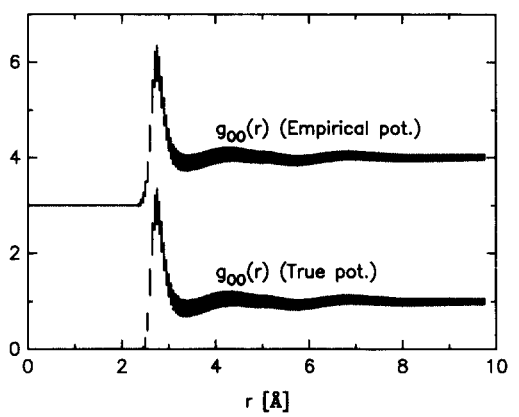

Fig. 4. Comparison of the standard deviation of individual distributions about the mean distribution for the oxygen–oxygen pair correlation. Empirical potential simulation is on top and the true potential simulation is on the bottom. The similarity in the height of the error bars for the two curves suggests that the empirical potential simulation is sampling a similar range of molecular configurations to those of the true potential simulation.

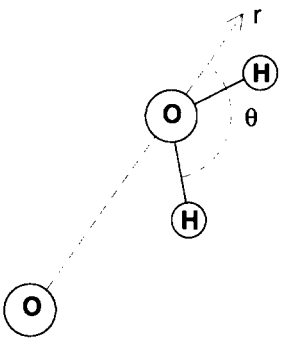

Fig. 5. Definition of the geometry used to define the O · · · O–H bond angle distribution. θ is the angle between the O–H bond and OO positive radial direction. Thus $\theta = 180$ ($\cos\theta = -1$) corresponds to one of the O–H bonds on the water molecule pointing directly towards a neighbouring oxygen.

As a further test of the procedure, the angle distribution O · · · H–O was calculated for both true SPCE simulation and for the empirical potential simulation. The angle between the O–H bond of one molecule and the O · · · O axis between neighbouring molecules is defined as θ, (see Fig. 5) and the bond angle distribution was evaluated in the range 2.5–3.5 Å, corresponding to the near-neighbour peak in $g_{OO}(r)$. The results of this comparison are shown in Fig. 6, where it is again evident that the empirical potential simulation has given an excellent representation of the true potential simulation.

Obviously many other tests could be devised, but it is already apparent that on the basis of the tests shown here the empirical potential Monte Carlo simulation is able to reproduce quantitatively the structure of the material in question. The fact that it works for a water-like potential, i.e. SPCE water, implies it will probably work in many other systems where the underlying pair potentials are probably as complicated, if not more so, and as long ranged as those for water.

5. Comparison with experimental water

Given the success of the method for looking at SPCE water, the experimental site–site pair correla-

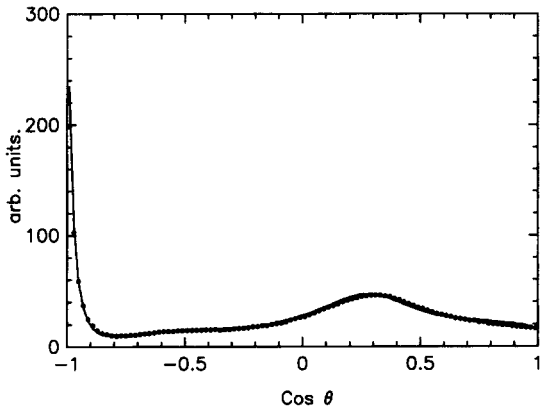

Fig. 6. O · · · O–H bond angle distribution for empirical potential (line) and true potential (circles) SPCE water. This distribution is evaluated in the radial range 2.5–3.5 Å, corresponding to the first coordination shell of a central water molecule. On the basis of the definitions in Fig. 5 the sharp peak at $\cos\theta = -1$ corresponds to the O–H bond on one water molecule pointing directly, on average, towards a neighbouring oxygen, such as will occur on hydrogen accepting sites of the central molecule. The broader hump at $\cos\theta = 0.3$ corresponds in part to the second hydrogen on the same molecule which will lie at roughly $\theta = 71°$ ($\cos\theta = 0.33$) for an SPCE water molecule if the first hydrogen is pointing towards a neighbouring oxygen. It also corresponds in part to hydrogens on neighbouring water molecules which are hydrogen bonded to the hydrogen donating sites of a central water molecule. Between these two features the density does not go to the zero that would be expected if the central molecule were only hydrogen bonded in tetrahedral positions, but instead there is a broad, low hump around $\cos\theta = -0.55$ ($\theta = 123°$) suggesting a significant degree of non-hydrogen bonded molecules in the first coordination shell.

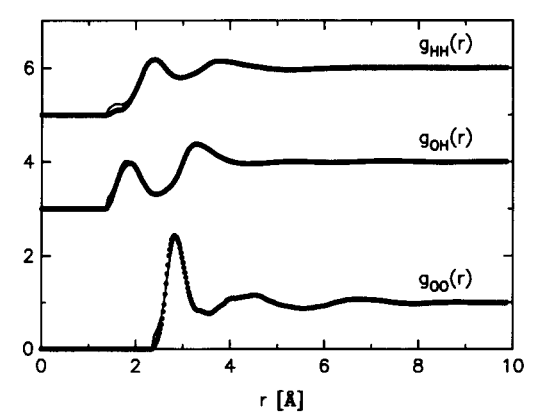

Fig. 7. Site–site pair correlation functions for experimental water, derived from the empirical potential simulation (line) and from the measured partial structure factors (circles).

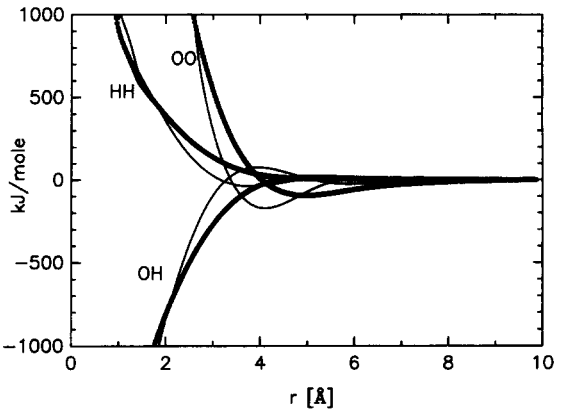

Fig. 8. The empirical potential used in the simulation of experimental water (line) compared to the SPCE empirical potential (circles).

tion functions derived previously [15] were also subject to the EPMC analysis: once again the excellent agreement with the "data" shown in Fig. 7. Fig. 8 shows the derived empirical potentials compared with those obtained from the EPMC analysis of SPCE water, while Fig. 9 compares the bond angle distributions in the two cases. It is evident from Fig. 8 that the experimental potentials are similar in form to the SPCE empirical potential, implying that the true SPCE potential may well be an excellent representation of the effective water potential, as has been

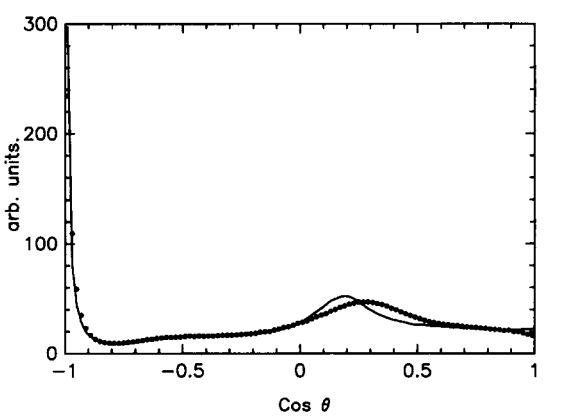

Fig. 9. O · · · O–H bond angle distribution for experimental water (line) compared to that for SPCE water (circles). Note the great similarity between these curves, suggesting the main geometry of hydrogen bonding is reproduced quite accurately by the SPCE potential. The difference in the position of the broad hump at $\cos\theta = 0.3$ for SPCE water and $\cos\theta = 0.2$ for experimental water arises primarily because of the different geometries of the two molecules.

established from other evidence [7]. The bond angle distributions in Fig. 9 show remarkable agreement between experimental and SPCE water, more so in fact than would be expected by comparing the site–site correlations, which do show some discrepancies. The only real difference occurs in the position of the hump at $\cos\theta = 0.2$ in experimental water compared to 0.3 in SPCE water. This difference arises simply from the slightly different geometry of the SPCE water molecule (OH distance 1.0 Å, HH distance 1.62 Å) compared to real water molecule (OH distance 0.98 Å, HH distance 1.55 Å). The close agreement between the curves suggests that SPCE models the degree of orientational correlation between hydrogen bonded molecules quite accurately.

6. Discussion and conclusion

In this paper I have shown how a simple bootstrap procedure can be used to derive an empirical potential energy function which when used in a computer simulation reproduces a set of measured or simulated site–site pair correlation functions. The method has been tested on SPCE and experimental water and found to give excellent results in both cases: it was used to highlight both the similarities and differences between the model potential and the empirical potential, but also revealed the great similarities between the hydrogen bonding in SPCE water compared to experimental water.

I would submit that an important advance has been made here. There are many instances where diffraction data for a liquid or solution are available, but only a subset of all the possible site–site pair correlation functions can be measured. For complex systems it will never be possible to obtain a complete set of site–site pair correlation functions by diffraction experiments. Furthermore for complex systems and complex-shaped molecules the use of alternative methods such as spherical harmonic analysis [15] is not always practical.

By combining a potential which incorporates known coordination and overlap restrictions between molecules, as well as known intra-molecular correlations, with the empirical potential technique, it should prove possible to simulate structures which are both physically reasonable *and* consistent with the available diffraction data. The choice of the starting potential is entirely in the hands of the user: it can in principle build in all known coordination effects such as molecular structure, torsional potentials, etc., if these are known a priori. It remains to be seen whether thermodynamic and other information can be used to constrain the empirical potential in the same way that the site–site pair correlation functions are used here.

The empirical potential derived here for water could in principle now be used in a predictive sense to calculate the structure of water under non-ambient conditions. Comparison of these predictions with measured correlation functions under the same conditions would provide an indication of how many-body interactions affect the potential. The present substantial disagreement between the simulated internal energy and other thermodynamic quantities for the true SPCE potential and those from a simulation with the EPMC SPCE potential indicates that the EPMC potential is not yet accurate enough to make that comparison worthwhile, and the next stage is to determine whether thermodynamic evidence can be used alongside the site–site pair correlation function data in deriving the empirical potential.

Acknowledgements

This work was performed with partial support from EEC Twinning Agreement #SC1-CT91-0714. My thanks to I. Svischev and P. Kusalik for supplying tables of the site–site pair correlation functions from their molecular dynamics simulation of SPCE water.

Note added in proof

Since this work was accepted the author's attention was drawn to a paper by A.P. Lyubartsev and A. Laaksonen, Phys. Rev. E 52 (1995) 3730, which appears to achieve essentially the same result, but with a different application in mind.

References

[1] R.L. McGreevy and L. Pusztai, Mol. Sim. 1 (1988) 359.
[2] R.L. McGreevy, Nucl. Instrum. Methods Phys. Res. A 354 (1995) 1.

[3] H.M. Rietveld, J. Appl. Cryst. 2 (1969) 65.
[4] C.G. Gray and K.E. Gubbins, Theory of molecular fluids, I (Oxford, 1984).
[5] R.A. Evans, Mol. Sim. 4 (1990) 409.
[6] H.J.C. Berendsen, J.R. Grigera and T.P. Straatsma, J. Phys. Chem. 91 (1987) 6269.
[7] Y. Guissani and B. Guillot, J. Chem. Phys. 98 (1993) 8221.
[8] P. Postorino, R.H. Tromp, M.-A. Ricci, A.K. Soper and G.W. Neilson, Nature 366 (1993) 668.
[9] A.K. Soper, Neutron Scattering Data Analysis 1990, ed. M.W. Johnson (IOP Conference Series No. 107, Bristol, 1990).
[10] J.P. Hansen and I.R. McDonald, Theory of simple liquids (Academic Press, London, 1986).
[11] A.K. Soper, C. Andreani and M. Nardone, Phys. Rev. E 47 (1993) 2598.
[12] M.P. Allen and D.J. Tildesley, Computer simulation of liquids (Oxford, 1987).
[13] I.M. Svishchev and P.G. Kusalik, J. Chem. Phys. 99 (1993) 3049.
[14] T.A. Andrea, W.C. Swope and H.C. Anderson, J. Chem. Phys. 79 (1983) 4576.
[15] A.K. Soper, J. Chem. Phys. 101 (1994) 6888.