On the accuracy of one- and two-particle solvation entropies

Benedict W. J. Irwin[1] and David J. Huggins[1,2,a)]

[1]*Theory of Condensed Matter Group, Cavendish Laboratory, University of Cambridge, 19 J J Thomson Avenue, Cambridge CB3 0HE, United Kingdom*
[2]*Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom*

Evaluating solvation entropies directly and combining with direct energy calculations is one way of calculating free energies of solvation and is used by Inhomogeneous Fluid Solvation Theory (IFST). The configurational entropy of a fluid is a function of the interatomic correlations and can thus be expressed in terms of correlation functions. The entropies in this work are directly calculated from a truncated series of integrals over these correlation functions. Many studies truncate all terms higher than the solvent-solute correlations. This study includes an additional solvent-solvent correlation term and assesses the associated free energy when IFST is applied to a fixed Lennard-Jones particle solvated in neon. The strength of the central potential is varied to imitate larger solutes. Average free energy estimates with both levels of IFST are able to reproduce the estimate made using the Free energy Perturbation (FEP) to within 0.16 kcal/mol. We find that the signal from the solvent-solvent correlations is very weak. Our conclusion is that for monatomic fluids simulated by pairwise classical potentials the correction term is relatively small in magnitude. This study shows it is possible to reproduce the free energy from a path based method like FEP, by only considering the endpoints of the path. This method can be directly applied to more complex solutes which break the spherical symmetry of this study. © *2017 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).* [http://dx.doi.org/10.1063/1.4983654]

I. INTRODUCTION

The ability to estimate the free energy difference between two defined states is a useful tool in computational chemistry. Direct applications are predicting the free energy of a solvation process, whether it may be testing a small molecule in a solvent to see if the pair is likely to be miscible[1] or a larger system, for example, a peptide, protein,[2] or protein-ligand complex.[3] The latter has a direct implication to *in silico* drug designs. Being able to quantitatively measure such a change then allows the relative comparison of ligands for a given protein.[4] For the protein-ligand complex, if the free energy of the bound state is less than the unbound state then the equilibrium will favour the bound state.

There exist a number of methods of estimating changes in the free energy with computational simulations. These include the Free Energy Perturbation (FEP)[5] and Thermodynamic Integration (TI).[6] Both these methods rely on a well defined path from the reference state to the new state, but are very generally applicable and work with different levels of theory for the description of the physical system.[7] Another method which has seen growing success is Inhomogeneous Fluid Solvation Theory (IFST),[8–10] which does not need a well-defined path between the reference state and the new state; however, IFST can only calculate changes in the free energy in the

context of a solvation process. IFST performs this by using a direct computation of the change in entropy due to solvation and combining this with a direct energy measurement[11,12] to calculate the free energy. There are numerous examples of calculating and estimating such a change in free energy.[13–18]

This work attempts to expand on one method of measuring the solvation entropy change directly, namely the Mutual Information Expansion (MIE),[19] and uses a k-Nearest Neighbours (KNN)[20,21] estimator to evaluate an approximation to the change in the solvation entropy. This work is different from the majority of previous works, in that we truncate the MIE at a higher order, including an additional term. This additional term represents correlations between two solvent molecules in the presence of a solute, and such terms have been measured before in the context of water in protein binding pockets[22] using Grid Inhomogeneous Solvation Theory (GIST).[23,24] We seek to find quantitatively, the change in free energy associated with the extra information, and whether it is necessary to include this term in the MIE. The system under study in this work is fundamental and simple, a Lennard-Jones neon solvent with a fixed Lennard-Jones atom in the centre of the simulation cell which represents the solute in a solvation process. The changes in free energy will be compared to an equivalent FEP simulation. The comparison of FEP and IFST in this work is independent of the forcefield used.

First, we will discuss the theory used in all calculations, then review the computational methods used, and input simulation parameters. The results will then be discussed and then analysed.

---

a)Electronic mail: djh210@cam.ac.uk. URL: http://huggins-lab.tcm.phy.cam.ac.uk/.

146, 194111-1

II. THEORY

Our calculations of the solvation free energy associated with fixing a Lennard-Jones atom at the centre of a neon simulation box (see Fig. 1) are based on the following:

Step 1. A change in free energy in the canonical ensemble is given by a change in the Helmholtz free energy,

$$\Delta A = \Delta U - T\Delta S, \qquad (1)$$

where ΔU is the change in the internal energy of the system, T is the temperature of the system, and ΔS is the change in the entropy of the system. The changes here are the differences between the solute-liquid system and the bulk liquid system.

Step 2. By using a molecular dynamics (MD) simulation with a parameterized forcefield, it is possible to estimate ΔU as

$$\Delta U = \bar{U}_2 - \bar{U}_1, \qquad (2)$$

where $\bar{U}_1$ is the equilibrium expectation energy of a system of N neon atoms in a periodic cell of volume V at temperature T, and $\bar{U}_2$ is the equilibrium expectation energy of a system of N neon atoms in the presence of a fixed Lennard-Jones atom at volume V and temperature T.

Step 3. It is possible to write the total change in solvation entropy, ΔS in Eq. (1), as an expansion over correlations of one-body, two-body, three-body, and so on, as discussed by Baranyai and Evans[25] for the homogeneous fluid and by Lazaridis[8,9] for the inhomogeneous fluid. The theory for homogeneous fluids was extensively studied by Kirkwood,[6,26] Nettleton,[27] Raveche,[28] and Wallace.[29–31]

For the system of N ideal atoms with coordinates $\{\mathbf{r}\}_N = \{\mathbf{r}_1, \ldots, \mathbf{r}_N\}$ and momenta $\{\mathbf{p}\}_N = \{\mathbf{p}_1, \ldots, \mathbf{p}_N\}$, we have the N body distribution in positions and momenta

$$f_N = f_N(\{\mathbf{r}\}_N, \{\mathbf{p}\}_N) = g_N(\{\mathbf{r}\}_N) \prod_{k=1}^{N} f_k(\mathbf{p}_k), \qquad (3)$$

and the total entropy of a bulk fluid is given by

$$S_{\text{liquid}} = -\frac{Rh^{3N}}{N!} \int f_N \ln f_N \, d\{\mathbf{r}\}_N d\{\mathbf{p}\}_N, \qquad (4)$$

with R the gas constant and h the Planck constant. Then, the separability of the momentum can be exploited to give

$$S_{\text{liquid}} = S_{\text{momentum}} + S_{\text{configuration}}, \qquad (5)$$
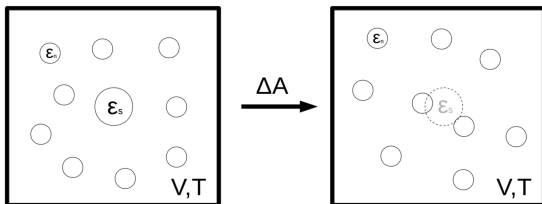

FIG. 1. Schematic of the canonical ensemble simulation for both IFST and FEP. The N solvent neons with the Lennard-Jones parameter ε_n surround the solute with the Lennard-Jones parameter ε_s (left). The solute interactions are turned off, while the volume and temperature remain constant (right). Both methods used in this study are measuring the change in the Helmholtz free energy, ΔA, between these two states.

which is written explicitly as

$$S_{\text{liquid}} = \underbrace{-\frac{NR}{\rho} \int f_1(\mathbf{p}_1) \ln f_1(\mathbf{p}_1) \, d\mathbf{p}_1}_{\text{Momentum}}$$

$$\underbrace{-\frac{R\rho^N}{N!} \int g_N(\{\mathbf{r}\}_N) \ln g_N(\{\mathbf{r}\}_N) \, d\{\mathbf{r}\}_N}_{\text{Configuration}}. \qquad (6)$$

The momentum terms are the same as those of an ideal gas where the one-body distribution of momenta is given by

$$f_1(\mathbf{p}) = \rho(2\pi mkT)^{-3/2} \exp\left(\frac{\mathbf{p}^2}{2mkT}\right), \qquad (7)$$

where k is the Boltzmann constant, ρ is the number density of the equivalent ideal gas, and m is the mass of the atom. Then,

$$S_{\text{momentum}} = -\frac{NR}{\rho} \int f_1(\mathbf{p}_1) \ln f_1(\mathbf{p}_1) \, d\mathbf{p}_1$$

$$= \frac{3NR}{2} - NR \ln(\rho\lambda^3), \qquad (8)$$

with λ the thermal wavelength of an atom in the liquid.

For a vector of random variables $\mathbf{X} = (X_1, \ldots, X_n)$, we may write[32]

$$H(\mathbf{X}) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2)$$

$$+ \cdots + H(X_n|X_1, \cdots, X_{n-1}), \qquad (9)$$

where H is an information entropy and the notation $H(X|Y)$ is the conditional entropy of X given Y. In a similar fashion, we may write

$$\underbrace{-\frac{R\rho^N}{N!} \int g_N(\{\mathbf{r}\}_N) \ln g_N(\{\mathbf{r}\}_N) \, d\{\mathbf{r}\}_N = S_2 - I_3 + I_4 - \cdots}_{\text{Configuration}} \quad (10)$$

giving

$$S_{\text{liquid}} = S_{\text{momentum}} + S_{2\,\text{liquid}} - I_{3\,\text{liquid}} + I_{4\,\text{liquid}} - \cdots \quad (11)$$

This expansion was performed for a homogeneous fluid by Wallace. It applies to the canonical ensemble and is non-local, meaning that it is only exact when integrated over the entire system volume.[25,29] An ensemble invariant and a local form of the expansion are discussed by Baranyai and Evans.[25] In the conditions of this study, the distinguishing terms between the local and non-local forms cancel, so here the non-local form is used for simplicity. In this case, the expansion has the following terms:

$$S_{\text{momentum}} = -\frac{NR}{\rho} \int f_N^{(1)}(\mathbf{p}) \ln h^3 f_N^{(1)}(\mathbf{p}) \, d\mathbf{p}, \qquad (12)$$

$$S_{2\,\text{liquid}} = -\frac{R\rho^2}{2!} \iint g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2) \ln g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2) \, d\mathbf{r}_1 d\mathbf{r}_2, \quad (13)$$

$$I_{3\,\text{liquid}} = \frac{R\rho^3}{3!} \iiint g_N^{(3)}(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3) \ln \delta g_N^{(3)}$$

$$\times (\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3) \, d\mathbf{r}_1 d\mathbf{r}_2 d\mathbf{r}_3. \qquad (14)$$

We can write the part of the three-body correlation function which cannot be expressed multiplicatively by its marginal distributions as

$$\delta g_N^{(3)}(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3) = \frac{g_N^{(3)}(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3)}{g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2) g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_3) g_N^{(2)}(\mathbf{r}_2, \mathbf{r}_3)}, \qquad (15)$$

where each of these N particle correlation functions for k bodies can be written in terms of the k body density as[33]

$$g_N(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_k) = \frac{1}{\rho^k} \rho_N(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_k). \quad (16)$$

Finally, we may write the excess entropy of the liquid as

$$S_{\text{excess liquid}} = S_{\text{liquid}} - S_{\text{ideal}}, \quad (17)$$

we have

$$S_{\text{ideal}} = S_{\text{momentum}} + R = \frac{5NR}{2} - NR \ln(\rho \lambda^3). \quad (18)$$

Therefore,

$$S_{\text{excess liquid}} = S_{2 \text{ liquid}} - I_{3 \text{ liquid}} + I_{4 \text{ liquid}} - \cdots - R, \quad (19)$$

where terms with an S denote an entropy and terms with an I denote a mutual information term. The expansion with these terms is true for a homogeneous fluid. For the simulations used in this work, we include the presence of the central solute.[8] The central solute in our calculations is spherically symmetric, and our end goal is to generalise to any solute and solvent, so we must consider an inhomogeneous system. This was analytically performed by Lazaridis.[8,9]

Eqs. (12)–(14) are the relevant terms for the liquid, and for a system with a solute we may write

$$S_{1 \text{ solute}} = -R\rho \int g_N^{(1)}(\mathbf{r}_1|s) \ln g_N^{(1)}(\mathbf{r}_1|s) \, d\mathbf{r}_1, \quad (20)$$

$$I_{2 \text{ solute}} = \frac{R\rho^2}{2!} \iint g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2|s) \ln \delta g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2|s) \, d\mathbf{r}_1 d\mathbf{r}_2, \quad (21)$$

$$I_{3 \text{ solute}} = -\frac{R\rho^3}{3!} \iiint g_N^{(3)}(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3|s) \ln \delta g_N^{(3)} \\ \times (\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3|s) \, d\mathbf{r}_1 d\mathbf{r}_2 d\mathbf{r}_3, \quad (22)$$

where the $|s$ argument indicates the presence of a solute. We can write the part of the two- and three-body correlation functions which cannot be expressed by its marginal distributions as

$$\delta g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2|s) = \frac{g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2|s)}{g_N^{(1)}(\mathbf{r}_1|s) g_N^{(1)}(\mathbf{r}_2|s)}, \quad (23)$$

$$\delta g_N^{(3)}(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3|s) \\ = \frac{g_N^{(3)}(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3|s) g_N^{(1)}(\mathbf{r}_1|s) g_N^{(1)}(\mathbf{r}_2|s) g_N^{(1)}(\mathbf{r}_3|s)}{g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2|s) g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_3|s) g_N^{(2)}(\mathbf{r}_2, \mathbf{r}_3|s)}. \quad (24)$$

Some slight differences exist between the homogeneous and inhomogeneous formulations, $S_{2 \text{ liquid}}$ is not a mutual information term (denoted with an I rather than an S), as the marginal distributions vanish. We may then write

$$S_{\text{solute}} = S_{\text{momentum}} + S_{1 \text{ solute}} - I_{2 \text{ solute}} + I_{3 \text{ solute}} - \cdots \quad (25)$$

and subtracting the ideal gas terms, Eq. (18),

$$S_{\text{excess solute}} = S_{1 \text{ solute}} - I_{2 \text{ solute}} + I_{3 \text{ solute}} - \cdots - R. \quad (26)$$

To find the change in entropy from the addition of the solute, we calculate the *excess entropy of solution*, $\Delta S_{\text{exc soln}}$. This is then the excess entropy of the solute system, Eq. (26), minus the excess entropy of the neat liquid, Eq. (19),

$$\Delta S_{\text{exc soln}} = S_{\text{excess solute}} - S_{\text{excess liquid}}, \quad (27)$$

$$\Delta S_{\text{exc soln}} = [S_{1 \text{ solute}} - I_{2 \text{ solute}} + I_{3 \text{ solute}} - \cdots] \\ - [S_{2 \text{ liquid}} - I_{3 \text{ liquid}} + I_{4 \text{ liquid}} - \cdots], \quad (28)$$

where we see the factors of $-R$ from $S_{\text{excess solute}}$ and $S_{\text{excess liquid}}$ cancel. It is at this point we choose a truncation of $\Delta S_{\text{exc soln}}$. The most severe truncation generates what we will call the *conditional one particle entropy* (C1PE),

$$\Delta S_{1|s} = S_{1 \text{ solute}}. \quad (29)$$

This is achieved by removing all integrals with a subscript greater than 1. If we include the two-body terms, then we have the *conditional two particle entropy* (C2PE),

$$\Delta S_{2|s} = S_{1 \text{ solute}} - I_{2 \text{ solute}} - S_{2 \text{ liquid}}. \quad (30)$$

It is $\Delta S_{2|s}$ that this work is concerned with calculating.

Step 4. Instead of directly integrating numerically the terms in Eqs. (29) and (30), we can instead convert the integral into a sum which converges to the integral asymptotically in the limit of infinite data. We can then extrapolate toward that infinite limit by measuring the sum for finite quantities of data and fitting an appropriate extrapolating function. The estimators we use in this study are KNN estimators and are formulated as follows.

Inserting the explicit integral terms into Eq. (30) gives

$$\Delta S_{2|s} = -R\rho \int g_N^{(1)}(\mathbf{r}_1|s) \ln g_N^{(1)}(\mathbf{r}_1|s) \, d\mathbf{r}_1 \\ - \frac{R\rho^2}{2} \iint g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2|s) \ln \frac{g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2|s)}{g_N^{(1)}(\mathbf{r}_1|s) g_N^{(1)}(\mathbf{r}_2|s)} \, d\mathbf{r}_1 d\mathbf{r}_2 \\ + \frac{R\rho^2}{2} \iint g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2) \ln g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2) \, d\mathbf{r}_1 d\mathbf{r}_2, \quad (31)$$

we may separate out the denominators of the logarithm in the second term and then swap the coordinates $\mathbf{r}_1$ and $\mathbf{r}_2$ in one of those new integrals to give

$$\Delta S_{2|s} = -R\rho \int g_N^{(1)}(\mathbf{r}_1|s) \ln g_N^{(1)}(\mathbf{r}_1|s) \, d\mathbf{r}_1 \\ - \frac{R\rho^2}{2} \iint g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2|s) \ln g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2|s) \, d\mathbf{r}_1 d\mathbf{r}_2 \\ + R\rho^2 \iint g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2|s) \ln g_N^{(1)}(\mathbf{r}_1|s) \, d\mathbf{r}_1 d\mathbf{r}_2 \\ + \frac{R\rho^2}{2} \iint g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2) \ln g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2) \, d\mathbf{r}_1 d\mathbf{r}_2, \quad (32)$$

and it is then possible to integrate over $\mathbf{r}_2$ in the third term giving

$$\Delta S_{2|s} = R(N-2)\rho \int g_N^{(1)}(\mathbf{r}_1|s) \ln g_N^{(1)}(\mathbf{r}_1|s) \, d\mathbf{r}_1 \\ - \frac{R\rho^2}{2} \iint g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2|s) \ln g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2|s) \, d\mathbf{r}_1 d\mathbf{r}_2 \\ + \frac{R\rho^2}{2} \iint g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2) \ln g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2) \, d\mathbf{r}_1 d\mathbf{r}_2. \quad (33)$$

It is convenient to construct estimators to measure three distinct quantities

$$H_{1,s} = \rho \int g_N^{(1)}(\mathbf{r}_1|s) \ln g_N^{(1)}(\mathbf{r}_1|s) \, d\mathbf{r}_1, \quad (34)$$

$$H_{2,s} = \rho^2 \iint g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2 | s) \ln g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2 | s)\, d\mathbf{r}_1 d\mathbf{r}_2, \quad (35)$$

$$H_{2,l} = \rho^2 \iint g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2) \ln g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2)\, d\mathbf{r}_1 d\mathbf{r}_2, \quad (36)$$

giving an expression for the conditional one- and two-particle entropies in terms of these estimators

$$\Delta S_{1|s} = R H_{1,s}, \quad (37)$$

$$\Delta S_{2|s} = R(N-2) H_{1,s} + \frac{R}{2}\left(H_{2,l} - H_{2,s}\right). \quad (38)$$

In general, we may state the information entropy of a probability density function over p random variables as the p-dimensional integral

$$H[\rho] = -\int \rho(\{\mathbf{x}\}_p) \ln \rho(\{\mathbf{x}\}_p)\, d\{\mathbf{x}\}_p \quad (39)$$

as the expectation

$$H[\rho] = \mathbb{E}[-\ln \rho(\{\mathbf{x}\}_p)]. \quad (40)$$

However, in Eqs. (34)–(36), we have correlation functions in the logarithm rather than densities. We may then instead calculate

$$H[\rho] + \int \rho(\{\mathbf{x}\}_p) \ln \rho\, d\{\mathbf{x}\}_p. \quad (41)$$

This quantity can then be approximated by a finite sum which is performed in Sec. II A.

A. k-Nearest Neighbours (KNN) estimators

To efficiently calculate the integrals in Eqs. (29) and (30), we used the k-nearest neighbours method which has shown previous success in calculating first order solvation entropies,[10,13,23] dihedral entropies in small molecules,[34] and entropies of water in protein binding pockets.[14,22] The general estimator for N_p, p-dimensional objects across F frames of data is given by

$$H^{(k)} \cong \frac{1}{N_p F} \sum_{i=1}^{N} \sum_{j=1}^{F} \ln\left(\frac{N_p (F-1) \pi^{p/2} d_{ij,k}^p}{\Gamma\left(\frac{p}{2}+1\right)}\right) - \psi(k), \quad (42)$$

where $\Gamma(x)$ is the Euler gamma function, $\psi(k)$ is the Euler digamma function, and the symbol $\cong$ denotes an asymptotic equivalence in the limit of an infinite number of frames F. This estimator was initially worked on for the first nearest neighbour by Leonenko[20] and extended for the kth nearest neighbour by Singh[21] and has been used by various studies.[10,13,14,19,34,35] This estimator is a limiting case of the adaptive anisotropic elliptic kernel estimators.[36,37]

The estimator for an H_1 term, given F sufficiently uncorrelated MD frames containing N neon atoms, is given by the expression

$$H_1^{(k)} \cong \frac{1}{NF} \sum_{i=1}^{N} \sum_{j=1}^{F} \ln\left(\frac{4N(F-1)\pi d_{ij,k}^3}{3V}\right) - \psi(k), \quad (43)$$

which is used in Refs. 13 and 14. The next order expression is then

$$H_2^{(k)} \cong \frac{2}{N(N-1)F}$$
$$\times \sum_{i_1=1}^{N} \sum_{i_2>i_1}^{N} \sum_{j=1}^{F} \ln\left(\frac{N(N-1)(F-1)\pi^3 d_{i_1 i_2 j,k}^6}{6V^2}\right)$$
$$- \psi(k). \quad (44)$$

In previous work, Eqs. (43) and (44) have been shown to fit a power law for increasing values of F[13,14,19]

$$H_k(F) = a_k F^{b_k} + H_\infty \quad (45)$$

with a_k and b_k some constants for the kth neighbour selected, and H_∞ the asymptotic value of the entropy. b_k is a negative constant such that

$$\lim_{F \to \infty} H_k(F) = H_\infty. \quad (46)$$

To extract this value, a power law can be fitted to $H(F)$.

III. SIMULATION DETAILS

FEP and IFST were used to calculate the free energy change associated with adding a fixed Lennard-Jones particle to the centre of a neon simulation box in the canonical ensemble. The IFST free energy with only $\Delta S_{1|s}$ and with both $\Delta S_{1|s}$ and $\Delta S_{2|s}$ were calculated. All of the simulation data used in this study came from the same forcefield and MD simulations were all carried out in NAMD.[38] Thus, the resulting free energies are directly comparable.

A. Systems setup

The neon MD parameters were taken from the CHARMM27 forcefield.[39] There are no electrostatic interactions for this solvent-solute system, so all potential energy terms are in the form of a Lennard-Jones potential

$$V_{LJ}(\mathbf{r}_{ij}) = \sqrt{\varepsilon_i \varepsilon_j}\left(\left(\frac{R_i + R_j}{2|\mathbf{r}_{ij}|}\right)^{12} - 2\left(\frac{R_i + R_j}{2|\mathbf{r}_{ij}|}\right)^6\right), \quad (47)$$

where r_{ij} is the distance between atoms i and j, and R_k is the radius of minimum potential energy for atom k, $R_k = 2^{1/6}\sigma_k$. 4 different solutes were taken which had varying Lennard-Jones ε_s parameters as shown in Table I. All solutes had the same R_k parameter as the neon solvent. Each solute was initially solvated with 900 neon atoms in a cubic box of edge length 27.5 Å. Where reduced units are given for reference they are calculated as the temperature $T^* = k_B T / \varepsilon$, pressure $P^* = p\sigma^3/\varepsilon$, length $L^* = L/\sigma$, and number density $\rho^* = \rho\sigma^3$.

For the 12-6 LJ fluid, the triple point density of both the liquid ρ_{tl}^* and solid ρ_{ts}^* phases have been measured by previous studies to be in the ranges $\rho_{tl}^* = 0.818$–0.864 and

TABLE I. Forcefield parameters used for each of the solute atoms where ε_n is the value for bulk neon.

Solute name	ε_s [kcal/mol]	R_k [Å]	L	ρ [Ne Å^{-3}]	T (K)
SOL1	0.215	3.06	27.3760	0.043 82	25
SOL2	0.430	3.06	27.3710	0.043 84	25
SOL3	0.645	3.06	27.3707	0.043 84	25
SOL4	0.860	3.06	27.3685	0.043 85	25

Solute name	$\varepsilon_s/\varepsilon_n$	σ_s [Å]	L^*	ρ^*	T^*
SOL1	2.5	2.73	10.0420	0.8887	0.578
SOL2	5.0	2.73	10.0401	0.8892	0.578
SOL3	7.5	2.73	10.0400	0.8893	0.578
SOL4	10.0	2.73	10.0392	0.8895	0.578

$\rho_{ts}^* = 0.96$–0.978, and the triple point has approximate temperature and pressure $T^* = 0.661$ and $p^* = 0.0018$, respectively.[40] Thus, the simulated neon was in the liquid phase.

B. IFST

1. IFST equilibration

16 replicate equilibrations per solute type were performed in the NpT ensemble for 4 ns to find the natural densities of the simulations cells with each solute present. The resulting edge lengths are shown in Table I.

A 2 ns equilibration was then performed for each replicate in the NVT ensemble at $T = 25$ K ($T^* = 0.578$) and 1 atm ($p^* = 0.003\,44$) using the Langevin temperature control. An MD timestep of 2.0 fs was used. Electrostatic interactions were turned off as they were not present in the forcefield. van der Waals interactions were removed for separations over 10.5 Å, and switching was used between 9.5 Å and 10.5 Å. The simulations took place with cubic periodic boundary conditions. This process was repeated for each of the 4 solutes and one bulk system with no solute.

2. IFST production

The production simulations were carried out in the NVT ensemble for 60 ns at 25 K ($T^* = 0.578$) with the same MD parameters as the NVT equilibration. NAMD produces a trajectory (dcd) file, which is a set of system coordinate snapshots. A trajectory frame was saved every 500 fs such that neighbouring frames in the dcd file were not too strongly correlated; this is important for the KNN estimator when used later to extract entropies as commented in the previous work by Huggins.[13] The energies $\Delta U = \bar{U}_{solute} - \bar{U}_{bulk}$ were calculated by taking the average energy across all 16 repeats of the 60 ns production simulation, and these are shown in Table II along with the standard deviation across 16 repeats, $\sigma_{\Delta U}$.

3. IFST free energy calculations

To calculate the change in free energy for IFST simulations, we need to calculate each component on the right hand side of

$$\Delta A_{1|s} = \Delta U - T\Delta S_{1|s} \tag{48}$$

and

$$\Delta A_{2|s} = \Delta U - T\Delta S_{2|s}. \tag{49}$$

ΔU was calculated using Eq. (2), and therefore the quantities $\Delta S_{1|s}$ and $\Delta S_{2|s}$ must be calculated using Eqs. (37) and (38). This then requires collecting a set of nearest neighbour distances and nearest pair distances from the respective MD data for the entropy estimators.

C. K-Dimensional (K-D) trees

Previous studies using IFST or GIST have used a grid of voxels or cut-cell approach to find nearest neighbour distances.[13,14] For the pair nearest neighbour distances required for Eq. (44), this method was not practical. In order to increase the efficiency and allow the calculation of the pair terms, a K-Dimensional (K-D) tree method was used. This is covered in more detail in the supplementary material.

D. Free energy perturbation (FEP)

1. FEP protocol

The FEP simulation measures the change in free energy of removing a fixed solute from a box of neon. This simulation is parameterized by a variable λ such that when $\lambda = 0$, the solute is fully present and when $\lambda = 1$, the solute is fully annihilated.

In this study, each forward and backward FEP simulation had $N = 64$ "λ-windows," and each window has a different value of lambda which is labeled as *degree of annihilation* in the schedule displayed in Fig. 2. For the nth window out of a total of N windows, the value of λ in that window was generated with the expressions

$$\lambda_f(n) = \frac{1}{N}\sum_{k=1}^{N} 1 - \left(1 - \frac{n}{N}\right)^{\max(k-N+n,1)}, \tag{50}$$

$$\lambda_b(n) = \frac{1}{N}\sum_{k=1}^{N} 1 - \left(\frac{n}{N}\right)^{\max(k-n,1)}, \tag{51}$$

which satisfy $\lambda_f(0) = 0, \lambda_f(N) = 1, \lambda_b(0) = 1$, and $\lambda_b(N) = 0$, where $\lambda_f(n)$ and $\lambda_b(n)$ are the forward and backward schedules, respectively. These curves were picked to sample the endpoints of the FEP simulation more heavily to "avoid the end point catastrophe."[41] The endpoints are when the solute is close to full annihilation in the forward simulation ($n \approx N$) or when the solute is just being created in the backward simulation ($n \approx 0$). A van der Waals soft core potential parameter of

TABLE II. Average free energy results for each solute. $\sigma_{\Delta U}$ is the standard deviation of the 16 energy results for each solute. σ_{FEP} is the standard deviation across all 16 repeats of the FEP free energy change ΔA_{FEP}. $\sigma_{1|s}$ and $\sigma_{2|s}$ are the standard deviations across all 16 repeats of the IFST free energy changes $\Delta A_{1|s}$ and $\Delta A_{2|s}$, respectively. All units are in kcal/mol.

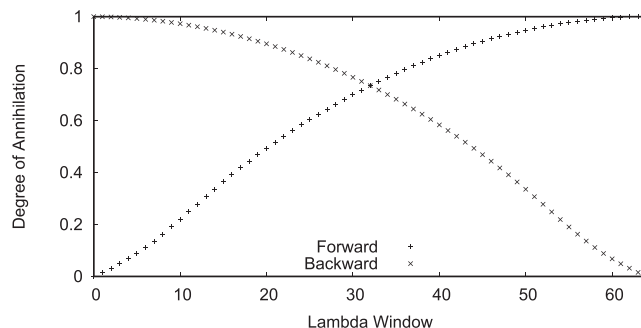|	ΔA_{FEP}	ΔU	$\Delta S_{1|s}$	$\Delta S_{2|s}$	$\Delta A_{1|s}$	$\Delta A_{2|s}$
SOL1	1.036	−1.717	−0.541	$<10^{-4}$	1.176	1.176
SOL2	1.820	−2.561	−0.589	$<10^{-4}$	1.972	1.972
SOL3	2.438	−3.257	−0.694	$<10^{-4}$	2.563	2.563
SOL4	2.975	−3.839	−0.773	$<10^{-4}$	3.067	3.067
	σ_{FEP}	$\sigma_{\Delta U}$	$\sigma_{\Delta S_{1|s}}$	$\sigma_{\Delta S_{2|s}}$	$\sigma_{1|s}$	$\sigma_{2|s}$
SOL1	0.004 95	0.0279	0.052	$<10^{-4}$	0.062	0.062
SOL2	0.003 82	0.0281	0.131	$<10^{-4}$	0.126	0.126
SOL3	0.005 85	0.0238	0.088	$<10^{-4}$	0.100	0.100
SOL4	0.006 24	0.0155	0.061	$<10^{-4}$	0.064	0.064

FIG. 2. The lambda schedule for the forward and backward FEP simulations across the 64 lambda windows as generated by Eqs. (50) and (51).

5.0 was used. One energy reading was stored every 100 steps which equates to every 0.2 ps.

2. FEP equilibration

The end points of the NVT equilibrations for the IFST simulations were used as starting points for the FEP simulations. Each solute had 16 replicates. Each simulation underwent a further 0.2 ns equilibration at $T = 25$ K ($T^* = 0.578$) from this point to adjust to their individual λ values.

3. FEP simulation

For each λ-window, 0.469 ns of simulation was performed in the NVT ensemble at $T = 25$ K ($T^* = 0.578$). Then, considering the 64 windows in both directions the overall time was 60.1 ns, which is comparable to the 60 ns used for the IFST production run.

4. FEP calculations

The final change in the free energy from the FEP simulations is found by summing the changes in free energy between each pair of neighbouring λ windows. Then, the forward change is

$$\Delta A_{f,FEP} = \sum_{n=0}^{63} \Delta A_{\lambda(n),\lambda(n+1)} \tag{52}$$

and the backward change is

$$\Delta A_{b,FEP} = \sum_{n=1}^{64} \Delta A_{\lambda(n-1),\lambda(n)}. \tag{53}$$

To calculate the changes in free energy between the λ-windows, the Bennett Acceptance Ratio (BAR) method was used,[42] which is included in the ParseFEP Plugin[43] for Visual Molecular Dynamics (VMD)[44] in which all FEP results for this study were calculated. The BAR estimator provides a calculated statistical error, and these errors were less than 0.007 kcal/mol for all of the 64 simulations.

IV. RESULTS AND DISCUSSION

A. Results

Fig. 3 shows the solute-fluid radial distribution functions for all solutes and bulk neon.

Fig. 4 shows a comparison of the free energy estimates from both levels of the IFST results and from FEP.
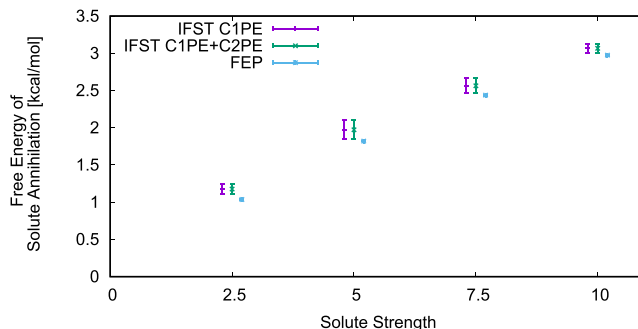
FIG. 4. The free energy estimates for the two IFST approximations and the FEP result at ε_s values of $2.5\varepsilon_n$, $5.0\varepsilon_n$, $7.5\varepsilon_n$, and $10.0\varepsilon_n$. Different levels of theory have been spaced for clarity. C1PE is the conditional one-particle entropy correction. C2PE is the conditional two-particle entropy correction. The error bars are the spread in the 16 repeats of each result.

Table II shows the comparison of the average free energies from FEP and IFST calculations, and the IFST calculations show the conditional one-particle and conditional two-particle entropy values.

The IFST free energies in Table II were calculated by Eqs. (48) and (49), the entropy terms were calculated by extrapolating Eq. (45) for the conditional one-particle entropies every 1000 frames from 1000 to 12 000, and Eq. (45) for conditional two-particle entropies, 16 repeats were taken at intervals of 1000, 2000, and 3000 frames and 2 repeats were extended to 1000, 2000, 3000, 4000, 5000, 7500, and 10 000 frames. Fig. 5 shows an example of the conditional two-particle entropy extrapolation for all solutes. The extrapolation routine used was from the gnuplot software, when fitting it weighted data points by the standard deviation of the 16 repeats of that point. For the conditional two-particle entropy fitting, where no standard deviation was available it was estimated to be $10/F$ kcal/mol, where F is the number of frames for that data point.

B. Discussion

Fig. 3 shows the radial distribution functions (RDFs) from the central solute to all solvent atoms. The troughs become deeper and the peaks become higher for increased solute potential ε_s parameter. Also plotted is the RDF for the bulk solvent, which has the shallowest troughs and lowest peaks. There is a slight distortion in the second solvation shell, which becomes more pronounced with higher ε_s. This is likely to
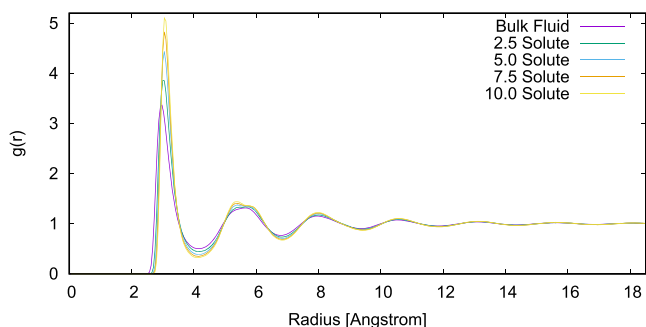
FIG. 3. The radial distribution functions for the solute to all fluid atoms. The fluid-fluid RDF is also plotted. For stronger central potentials, the peaks get higher and the troughs get deeper along with a consistent distortion in the second solvation shell.
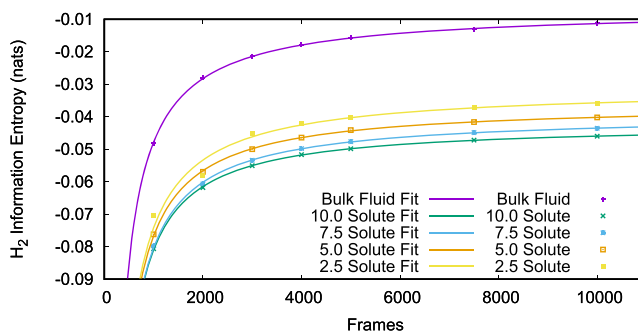
FIG. 5. A plot demonstrating how the asymptotic values for the $H_{s,2}$ and $H_{l,2}$ estimators were extracted. A power law is fitted to the three data points for each solute. The constant in the limit of infinite frames is used as in Eq. (45).

be a gentle crowding effect as the solvent molecules in the first solvation shell draw close to the solute, and the closest locations for second shell atoms correspond to the pits in the already tightly packed first solvation shell. The magnitude of the solvent-solvent correlation entropy was expected to be small in the system of monatomic species used in this study. It has been shown that in the Lennard-Jones system, the primary source of the solvent structure comes from packing,[45] and this will correspond to an entropy associated with the volume exclusion of the central solute which is captured fully by the C1PE term. For more complex solvents, namely water, evidence already exists for the solvent-solvent correlations.[46,47]

Fig. 4 shows the calculations of the free energy of solute annihilation from all three methods of evaluation with the standard deviations of 16 repeats used as error bars. The average values of these estimates are displayed in Table II along with the standard deviations of the 16 repeats. The average values for all methods are in good agreement. The IFST results including the conditional two-particle correction have the same averages and standard deviations of the first order IFST. This demonstrates that the C2PE has no clear contribution to the free energy of this system. The FEP results have the lowest standard deviation of all results even though a comparable length of MD run was used. The increased uncertainty associated with the IFST results likely arises from fitting power laws to extract the asymptotic entropy. The standard deviation in the free energy is at least 3 times greater than that of the MD energies used, which indicates the added uncertainty is from the entropy. This could potentially be remedied by taking more data points during the extrapolation process at the expense of data processing time. Both the average IFST results are within 0.16 kcal/mol of the FEP results. These results indicate that it is possible to reconstruct a fairly accurate measurement of the free energy change of this kind of process by only using the start and end points of an equivalent FEP path. The results indicate that the entropy term contributes relatively less to the free energy of solute annihilation for solutes with larger ε.

V. CONCLUSIONS

The translational entropy associated with the solute-solvent correlation and the solvent-solvent correlation can be used to evaluate the free energy of solvation in the IFST framework for a system of neon atoms solvating a fixed central Lennard-Jones potential. Both IFST estimates reproduce the estimate by an equivalent FEP simulation to within around 0.16 kcal/mol and appear to consistently overestimate slightly. The IFST method has an advantage over FEP and can give a free energy estimate without having to define a path between the two systems. However, the accuracy associated with FEP estimates is greater, so it is a better tool for computation. We conclude that IFST in its current state is the better tool for physical interpretation as it avoids the nonphysical lambda states utilized by FEP. We note that FEP is an already well-developed method, and IFST may yet have future improvements to increase its optimization. The conditional two-particle entropy term did not contribute a noticeable

change in free energy for this system for any of the strengths of the solute potential tested.

The IFST framework can give the spatial contributions of configurational entropy when analysed with voxel based methods (as shown in Fig. 3 of the supplementary material). This allows areas of interest around the solute to be highlighted. Although the two-particle terms in the MIE do not appear to be significant in this simple system, they may be significant in a system with a liquid water solvent with hydrogen bonding and charges. The methods used in this work may be extended to such a system.

If the solvent-solvent interactions become very strong in a particular solvation process, it may be necessary to calculate the *conditional three-particle entropy*

$$\Delta S_{3|s} = S_{1\,\text{solute}} - I_{2\,\text{solute}} + I_{3\,\text{solute}} - S_{2\,\text{liquid}} + I_{3\,\text{liquid}} \quad (54)$$

and the methodology used in this work could be extended to such a calculation.

SUPPLEMENTARY MATERIAL

See supplementary material for a description of the implementation of K-D trees in this work, and for a spatial illustration of the C2PE contributions around a fixed solute. The data and codes used for calculations in this work will be available at https://doi.org/10.17863/CAM.9335.

ACKNOWLEDGMENTS

[1] T. Lazaridis, "Solvent reorganization energy and entropy in hydrophobic hydration," J. Phys. Chem. B **104**(20), 4964–4979 (2000).
[2] C. N. Nguyen, A. Cruz, M. K. Gilson, and T. Kurtzman, "Thermodynamics of water in an enzyme active site: Grid-based hydration analysis of coagulation factor xa," J. Chem. Theory Comput. **10**(7), 2769–2780 (2014).
[3] M. K. Gilson and H.-X. Zhou, "Calculation of protein-ligand binding affinities," Annu. Rev. Biophys. Biomol. Struct. **36**, 21–42 (2007).
[4] G. J. Rocklin, D. L. Mobley, and K. A. Dill, "Separated topologies—A method for relative binding free energy calculations using orientational restraints," J. Chem. Phys. **138**(8), 085104 (2013).
[5] R. W. Zwanzig, "High-temperature equation of state by a perturbation method. I. Nonpolar gases," J. Chem. Phys. **22**(8), 1420–1426 (1954).
[6] J. G. Kirkwood, "Statistical mechanics of fluid mixtures," J. Chem. Phys. **3**(5), 300–313 (1935).
[7] J. Kästner, H. M. Senn, S. Thiel, N. Otte, W. Thiel, and J. Ka, "QM/MM free-energy perturbation compared to thermodynamic integration and umbrella sampling: Application to an enzymatic reaction," J. Chem. Theory Comput. **2**, 452–461 (2006).
[8] T. Lazaridis, "Inhomogeneous fluid approach to solvation thermodynamics. 1. Theory," J. Phys. Chem. B **102**(18), 3531–3541 (1998).
[9] T. Lazaridis, "Inhomogeneous fluid approach to solvation thermodynamics. 2. Applications to simple fluids," J. Phys. Chem. B **102**(18), 3542–3550 (1998).

[10]D. J. Huggins and M. C. Payne, "Assessing the accuracy of inhomogeneous fluid solvation theory in predicting hydration free energies of simple solutes," J. Phys. Chem. B **117**(27), 8232–8244 (2013).

[11]N. Matubayasi, L. H. Reed, and R. M. Levy, "Thermodynamics of the hydration shell. 1. Excess energy of a hydrophobic solute," J. Phys. Chem. **98**(41), 10640–10649 (1994).

[12]N. Matubayasi and M. Nakahara, "Theory of solutions in the energetic representation. I. Formulation," J. Chem. Phys. **113**(15), 6070–6081 (2000).

[13]D. J. Huggins, "Estimating translational and orientational entropies using the k -nearest neighbors algorithm," J. Chem. Theory Comput. **10**, 3617 (2014).

[14]D. Huggins, "Quantifying the entropy of binding for water molecules in protein cavities by computing correlations," Biophys. J. **108**(4), 928–936 (2015).

[15]C. Velez-Vega, D. J. J. McKay, T. Kurtzman, V. Aravamuthan, R. A. Pearlstein, and J. S. Duca, "Estimation of solvation entropy and enthalpy via analysis of water oxygen-hydrogen correlations," J. Chem. Theory Comput. **11**(11), 5090–5102 (2015).

[16]F. Reinhard and H. Grubmüller, "Estimation of absolute solvent and solvation shell entropies via permutation reduction," J. Chem. Phys. **126**(1), 014102 (2007).

[17]B. M. Dickson, H. Huang, and C. B. Post, "Unrestrained computation of free energy along a path," J. Phys. Chem. B **116**(36), 11046–11055 (2012).

[18]L. Young and C. B. Post, "Free energy calculations involving internal coordinate constraints to determine puckering of a six-membered ring molecule," J. Am. Chem. Soc. **115**(13), 1964–1970 (1993).

[19]V. Hnizdo, T. Jun, B. J. Killian, and M. K. Gilson, "Efficient calculation of configurational entropy from molecular simulations by combining the mutual-information expansion and nearest-neighbor methods," J. Comput. Chem. **29**(10), 1605–1614 (2008).

[20]L. F. Kozachenko, N. N. Leonenko, "Sample estimate of the entropy of a random vector," Probl. Peredachi Inf., **23**(2), 9–16 (1987).

[21]H. Singh, N. Misra, V. Hnizdo, A. Fedorowicz, and E. Demchuk, "Nearest neighbor estimates of entropy," Sciences **23**(3-4), 301–321 (2003).

[22]C. N. Nguyen, T. Kurtzman, and M. K. Gilson, "Spatial decomposition of translational water-water correlation entropy in binding pockets," J. Chem. Theory Comput. **12**, 414 (2015).

[23]C. N. Nguyen, T. K. Young, and M. K. Gilson, "Erratum: 'Grid inhomogeneous solvation theory: Hydration structure and thermodynamics of the miniature receptor cucurbit[7]uril' [J. Chem. Phys. 137, 044101 (2012)]," J. Chem. Phys. **137**(14), 149901 (2012).

[24]C. Nguyen, M. K. Gilson, and T. Young, "Grid inhomogeneous solvation theory: Hydration structure and thermodynamics of the miniature receptor cucurbit [7] uril" J. Chem. Phys. **137**(4), 044101 (2012); e-print arXiv:1108.4876.

[25]A. Baranyai and D. J. Evans, "Direct entropy calculation from computer simulation of liquids," Phys. Rev. A **40**(7), 3817–3822 (1989).

[26]J. G. Kirkwood and F. P. Buff, "The statistical mechanical theory of solutions. I," J. Chem. Phys. **19**(6), 774–777 (1951).

[27]R. E. Nettleton and M. S. Green, "Expression in terms of molecular distribution functions for the entropy density in an infinite system," J. Chem. Phys. **29**(6), 1365 (1958).

[28]H. J. Raveche, "Entropy and molecular correlation functions in open systems. I. Derivation," J. Chem. Phys. **55**(5), 2242 (1971).

[29]D. C. Wallace, "On the role of density fluctuations in the entropy of a fluid," J. Chem. Phys. **87**(4), 2282 (1987).

[30]D. C. Wallace, "Statistical entropy and a qualitative gas-liquid phase diagram," Phys. Rev. A **38**(1), 469–472 (1988).

[31]D. C. Wallace, "Statistical theory for the entropy of a liquid," Phys. Rev. A **39**(9), 4843–4847 (1989).

[32]E. Friedgut, "Hypergraphs, entropy, and inequalities," Am. Math. Mon. **111**(9), 749–760 (2004).

[33]H. Reiss, "Superposition approximations from a variation principle," J. Stat. Phys. **6**(1), 39–47 (1972).

[34]V. Hnizdo, E. V. A. Darian, A. Fedorowicz, E. Demchuk, S. Li, and H. Singh, "Nearest-neighbor nonparametric method for estimating the configurational entropy of complex molecules," J. Comput. Chem. **28**(3), 655–668 (2007).

[35]L. Wang, R. Abel, R. A. Friesner, and B. J. Berne, "Thermodynamic properties of liquid water: an application of a nonparametric approach to computing the entropy of a neat fluid" J. Chem. Theory Comput. **5**(6), 1462–1473 (2009).

[36]U. Hensen, H. Grubmüller, and O. F. Lange, "Adaptive anisotropic kernels for nonparametric estimation of absolute configurational entropies in high-dimensional configuration spaces," Phys. Rev. E **80**(1), 011913 (2009).

[37]U. Hensen, O. F. Lange, and H. Grubmüller, "Estimating absolute configurational entropies of macromolecules: The minimally coupled subspace approach," PLoS One **5**(2), e9179 (2010).

[38]J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten, "Scalable molecular dynamics with NAMD," J. Comput. Chem. **26**(16), 1781–1802 (2005).

[39]S. E. Feller, A. D. Mackerell, Jr., and A. D. J. Mackerell, "An improved empirical potential energy function for molecular simulations of phospholipids supporting 2," J. Phys. Chem. B **104**(31), 7510–7515 (2000).

[40]A. Ahmed and R. J. Sadus, "Solid-liquid equilibria and triple points of *n*-6 Lennard-Jones fluids," J. Chem. Phys. **131**(17), 174504 (2009).

[41]A. Pohorille, C. Jarzynski, and C. Chipot, "Good practices in free-energy calculations," J. Phys. Chem. B **114**(32), 10235–10253 (2010).

[42]C. H. Bennett, "Efficient estimation of free energy differences from Monte Carlo data," J. Comput. Phys. **22**(2), 245–268 (1976).

[43]P. Liu, F. Dehez, W. Cai, and C. Chipot, "A toolkit for the analysis of free-energy perturbation calculations," J. Chem. Theory Comput. **8**(8), 2606–2616 (2012).

[44]W. Humphrey, A. Dalke, and K. Schulten, "VMD: Visual molecular dynamics", J. Mol. Graphics **14**, 33 (1996).

[45]J. D. Weeks, D. Chandler, and H. C. Andersen, "Role of repulsive forces in determining the equilibrium structure of simple liquids," J. Chem. Phys. **54**, 5237–5247 (1971).

[46]T. Lazaridis and M. Karplus, "Orientational correlations and entropy in liquid water," J. Chem. Phys. **105**, 4294–4316 (1996).

[47]J. Zielkiewicz, "Structural properties of water: Comparison of the SPC, SPCE, TIP4P, and TIP5P models of water," J. Chem. Phys. **123**(10), 104501 (2005).