PCCP

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2018, **20**, 20981

Solubility prediction from first principles: a density of states approach

Simon Boothroyd,[a] Andy Kerridge,[a] Anders Broo,[b] David Buttar[c] and Jamshed Anwar *[a]

Solubility is a fundamental property of widespread significance. Despite its importance, its efficient and accurate prediction from first principles remains a major challenge. Here we propose a novel method to predict the solubility of molecules using a density of states (DOS) approach from classical molecular simulation. The method offers a potential route to solubility prediction for large (including drug-like) molecules over a range of temperatures and pressures, all from a modest number of simulations. The method was employed to predict the solubility of sodium chloride in water at ambient conditions, yielding a value of 3.77(5) mol kg^{-1}. This is in close agreement with other approaches based on molecular simulation, the consensus literature value being 3.71(25) mol kg^{-1}. The predicted solubility is about half of the experimental value, the disparity being attributed to the known limitation of the Joung–Cheatham force field model employed for NaCl. The proposed method also accurately predicted the NaCl model's solubility over the temperature range 298–373 K directly from the density of states data used to predict the ambient solubility.

Received 19th March 2018,
Accepted 19th July 2018

DOI: 10.1039/c8cp01786g

rsc.li/pccp

1. Introduction

When dissolving a substance in solution, there comes a point when no more will dissolve. The concentration at which this occurs is the solubility limit (the solubility) and depends on the properties of both the solute and solvent. Being a fundamental property, the solubility is of interest across a spectrum of application domains that include chemical toxicity, formulation of foods and development of chemical and pharmaceutical products,[1] weathering of the terrestrial and built environments, and formation and dynamics of ecological environments such as soil including fate of pollutants. The solubility is also an important factor in many disease states which include cholesterol deposition in atherosclerosis, formation of gall and kidney stones, and formation of amyloid plaques in disease such as Alzheimer's.[2] Another notable example is the interest in the solubility of carbon in the Earth's upper mantle, the latter represents the largest reservoir of carbon on Earth.[3] For each of these, considerations of solubility are important for devising relevant interventions. For some of these *e.g.* pharmaceuticals, being able to accurately predict the solubility from the molecular structure would be a 'game-changer'.[4,5]

There are three main approaches to solubility prediction: empirical, correlation-based methods,[6] quantum mechanical (QM) continuum solvation models such as COSMO-RS,[7] and molecular simulation.[8] Correlation methods include quantitative structure property relationships (QSPR) based on molecular descriptors, with the parameters being optimised against a dataset of molecular structures with known solubilities. Such models are limited in their usage, breaking down when predicting solubility for molecules that are distinct from the training set. Furthermore, the solubility can only be predicted at the conditions (*e.g.* temperature and pressure) at which the training set data were collected. The continuum solvation approaches neglect sampling of the solvent degrees of freedom and involve parameterisation, in particular requiring a fitted value for the free energy of fusion for the prediction of solubility of solids.

Molecular simulation offers potentially the more powerful approach to solubility prediction, with the solubility being accessed *via* statistical mechanics. There are two distinct approaches: *via* calculation of the chemical potentials[9] (summarised below), or direct (brute force) simulation of the dissolution of the solid in a solvent towards equilibrium.[10] The latter requires large system sizes to minimise finite-size effects and very long simulations to attain the essential near equilibrium conditions.

At the solubility limit x_s, the (undissolved) solid phase coexists with its solution. As the two are in equilibrium, the chemical potential of the solute in the solid phase (μ_{solid}) and that in solution (μ_{soln}) are identical at the given temperature T and pressure p, $\mu_{solid}(T,p) = \mu_{soln}(T,p)$. Prediction of the

[a] *Chemical Theory and Computation, Department of Chemistry, Lancaster University, Lancaster LA1 4YB, UK. E-mail: j.anwar@lancaster.ac.uk*
[b] *Pharmaceutical Science IMED Biotech unit, AstraZeneca, Pepparedsleden 1, Mölndal, 431 83, Sweden*
[c] *Pharmaceutical Science IMED Biotech unit, AstraZeneca, Silk Road Business Park, Macclesfield, SK10 2NA, UK*

solubility therefore requires in general the calculation of the chemical potential of the solute in solution for a series of concentrations, and then interpolation to find where it intersects the chemical potential of the solid (which is calculated separately). Both of these chemical potentials are accessible by molecular simulation. The chemical potential of the solid phase can be calculated *via* thermodynamic integration of an Einstein crystal[11,12] or by quasi-harmonic lattice dynamics. Calculation of the chemical potential of the solute in solution is more demanding, though the methods are well established and include thermodynamic integration,[13,14] the so-called perturbation approach,[15–17] expanded ensembles,[18,19] and variations on these.[20] These methods involve 'growing' the solute molecule from its reference state reversibly in the solvent. While both thermodynamic integration (TI) and perturbation techniques are robust and effective (particularly when coupled with soft-core[21] and dampening potentials[22]), large drug-like molecules are still challenging, and these methods are computationally very demanding. Each chemical potential determination requires at least a dozen or so separate simulations, which need to be repeated for any other temperature and pressure conditions of interest. To date there are only a few studies that have attempted to predict solubilities from molecular simulation *via* chemical potential calculations.[19,23–28] Much of the focus of these studies has been on the alkali halides with NaCl becoming a model test case.

Here we present a novel method to calculate the solubility directly from the density of states of a system. Density of states (DOS) calculations are well established, being particularly effective and efficient for determining phase co-existence.[29–31] The application of DOS methods however has been largely restricted to single, pure component systems. We utilise the DOS framework for multicomponent systems to access phase coexistence of a solid in equilibrium with its solution, and hence the solubility. The method in principle is able to predict solubility for a range of temperatures, pressures and solid forms using a single, density of states. It is more efficient than thermodynamic integration and the perturbation approach. We have successfully applied the methodology to predict the aqueous solubility of sodium chloride.

2. Solubility from density of states

We start by considering a pure system to illustrate how phase coexistence can be determined *via* a density of states approach, before considering its application to more complicated multicomponent systems.

The isothermal–isobaric (*NpT*) partition function is given by

$$Q(N,p,T) = \sum_{i=1}^{\text{microstates}} \mathrm{e}^{-\beta(E_i+pV_i)} \tag{1}$$

where the summation is over all microstates and E_i and V_i are the energy and volume of microstate i respectively. Given that distinct states may have identical energies *i.e.* are degenerate, $Q(N,p,T)$ may be expressed in the form

$$Q(N,p,T) = \sum_E \sum_V \Omega(V,E)\mathrm{e}^{-\beta(E+pV)} \tag{2}$$

where $\Omega(V,E)$ is the density of states of the system[32] and the first summation is now over energy levels. The corresponding probability distribution is

$$P(V,E) = \frac{1}{Q(N,p,T)}\mathrm{e}^{\ln\Omega(V,E)-\beta(E+pV)} \tag{3}$$

If the density of states is known, the phase coexistence condition can be determined by exploring the probability distribution at a given pressure whilst scanning in temperature, or *vice versa*. The probability distribution of a single component at coexistence exhibits two peaks of equal area, indicating that both phases are equally likely under these conditions. A key feature of the DOS approach is that the density of states $\Omega(V,E)$ is independent of T and p. This means that, in principle, coexistence conditions can be determined for a range of temperatures and pressures all from a single density of states.[29]

We now consider a multicomponent system composed of a number of different molecular species i, j, k,…. Within this system, we allow the number of molecules of one component N_i to fluctuate, while the populations of the other components N_j,N_k,…, are kept fixed.

For such a system, the partition function is given by

$$\Xi(\mu_i,p,T,)_{N_j,N_k,\ldots} = \sum_E \sum_V \sum_{N_i} \Omega(N_i,V,E)_{N_j,N_k,\ldots}\mathrm{e}^{-\beta(E+pV-\mu_iN_i)} \tag{4}$$

where μ_i is the chemical potential of component i. The corresponding probability distribution is

$$P(N_i,V,E)_{N_j,N_k,\ldots} = \frac{1}{\Xi(\mu_i,p,T)_{N_j,N_k,\ldots}}\mathrm{e}^{\ln\Omega(N_i,V,E)_{N_j,N_k,\ldots}-\beta(E+pV-\mu_iN_i)} \tag{5}$$

As before, if $\Omega(N_i,V,E)_{N_j,N_k,\ldots}$ is known, exploration of the above probability distribution would enable coexistence conditions to be identified – including the sought-after coexistence point at which the solid phase of component i would be in equilibrium with its solution phase *i.e.* the solubility. Thus for a given temperature and pressure, tweaking the chemical potential for component i would yield a bimodal probability distribution as a function of number of particles N_i in the N_j,N_k,… mixture system at the solubility limit, from which the solubility concentration can be ascertained. The two coexistence states would be the 100% solute (solid) phase, and its saturated solution (Fig. 1).

We do not, however, need to determine the density of states for the whole spectrum of mole fraction values from $x_i = 0$ (pure solvent) to $x_i = 1$ (pure solute) as implied, though we could. Given that at the solubility limit, $\mu_{\text{solid}}(T,p) = \mu_{\text{soln}}(T,p)$, one could substitute the chemical potential of the solid, if it were known, into the probability distribution (eqn (5)). This would guarantee that a peak is observed at $x_i = 1$. A second peak would then be expected at some lower mole fraction, which would correspond to the solubility (Fig. 1). Thus, we can calculate the chemical potential of the solid phase separately, and therefore focus on a limited mole fraction range where the solute remains in solution; the solubility condition will reveal

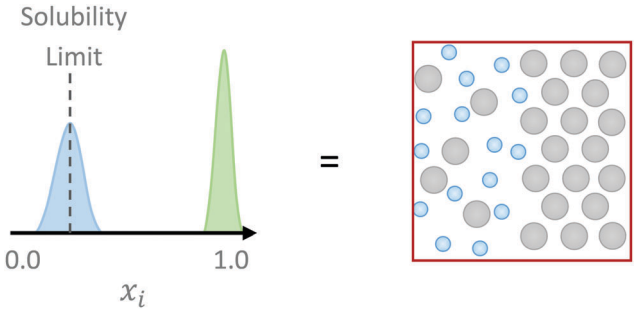

Fig. 1 A schematic probability distribution for a system of solute (grey particles) and solvent (blue particles) as a function of solute fraction. At the solubility limit, the solute particles will have an equal probability of being in both the solid phase (the green peak at x_i = 1.0) and the solution phase (the blue peak). The location (mole fraction) of the solution phase peak is the solubility limit.

itself as a single peak in the probability distribution located at the corresponding concentration.

The primary challenge therefore is to access the density of states $\Omega(N_i,V,E)_{N_j,N_k,\ldots}$, techniques for which are now well established.[33] Here we employ a 3-dimensional variant of the efficient Monte Carlo scheme originally developed by Wang and Landau.[32] Configurations are generated according to probability

$$P(N_i, V, E)_{N_j,N_k,\ldots} \propto \frac{1}{\Omega(N_i, V, E)_{N_j,N_k,\ldots}} \quad (6)$$

with $\Omega(N_i,V,E)_{N_j,N_k,\ldots}$ being developed and improved on-the-fly as the simulation proceeds in a self consistent manner. Every-time a particular point in $\Omega(N_i,V,E)_{N_j,N_k,\ldots}$ space is visited, its value is incremented according to $\ln \Omega(N_i,V,E)_{N_j,N_k,\ldots\text{new}} = \ln f + \ln \Omega(N_i,V,E)_{N_j,N_k,\ldots\text{old}}$, where $\ln f$ is an arbitrary modification factor. When $\Omega(N_i,V,E)_{N_j,N_k,\ldots}$ has converged to its true value, all possible states in the system would be visited with an equal probability. This convergence is tracked by means of a separate histogram of visits to particular states $h(N_i,V,E)$. The density of states is said to have converged when the histogram becomes 'sufficiently' flat.

The density of states is evolved over a number of iterations, beginning with a (gross) value of $\ln f = 1$. When the histrogram of visits $h(N_i,V,E)$ is sufficiently flat (in our case, when the minimum value is greater than 80% of the average), the value of $\ln f$ is reduced to $\ln f_{\text{new}} = \frac{1}{2}\ln f_{\text{old}}$, and the histogram of visits is reset to zero for the next iteration.

To explore the (N_i,V,E) space associated with $\Omega(N_i,V,E)_{N_j,N_k,\ldots}$, we employed Monte Carlo simulations involving particle translation, volume scaling, and solute insertion/deletion moves. The respective moves were accepted or rejected in accordance with the following criteria,[27] which are valid provided that the volume is sampled logarithmically:

$$P_{\text{translation}}(\text{A} \rightarrow \text{B}) = \min\left(1, \frac{\Omega(\text{A})}{\Omega(\text{B})}\right)$$

$$P_{\text{volume}}(\text{A} \rightarrow \text{B}) = \min\left(1, \frac{\Omega(\text{A})}{\Omega(\text{B})}\frac{V_{\text{B}}^{N_i+1}}{V_{\text{A}}^{N_i+1}}\right)$$

$$P_{\text{insertion}}(\text{A} \rightarrow \text{B}) = \min\left(1, \frac{\Omega(\text{A})}{\Omega(\text{B})}\frac{V}{N_{i,\text{B}}}\right)$$

$$P_{\text{deletion}}(\text{A} \rightarrow \text{B}) = \min\left(1, \frac{\Omega(\text{A})}{\Omega(\text{B})}\frac{N_{i,\text{A}}}{V}\right) \quad (7)$$

As is well known, insertion/deletion moves present a particular challenge for dense systems and large solute molecules. Insertions of such molecules in dense systems are invariably rejected due to overlaps, while deletion of species with a high affinity for each other *e.g.* ion pairs, will often be unfavourable. Here we have devised a creative solution wherein we extend the sampled volume space for the liquid (solution) state to the gas phase for each of the N_i systems, and then proceed to carry out the particle insertion/deletion there (see Fig. 2).

The procedure to predict the solubility, therefore, comprises two distinct stages:

(i) Determination of the 2-d density of states $\Omega(N_i,V,E)_{N_j,N_k,\ldots}$ for each solution concentration (…, $N_i - 1$, N_i, $N_i + 1$,…), calculated (independently) in the *NpT* ensemble. The energy and volume ranges are chosen so that both the liquid and gas states are sampled at each particular N_i.

(ii) Determination of the density of states in the gas phase of the full assembly of multiple concentration systems

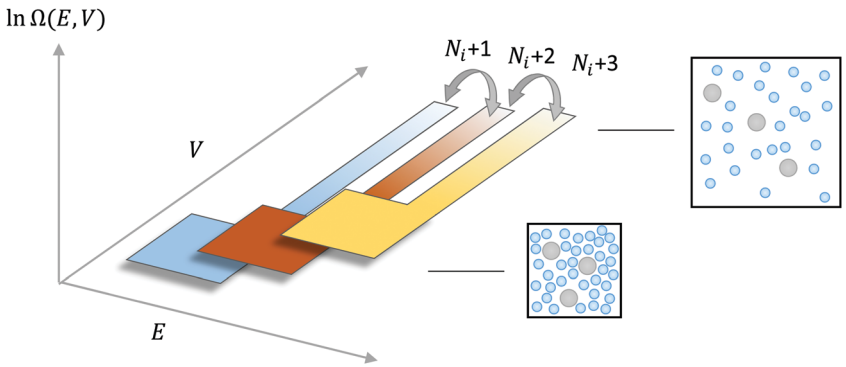

Fig. 2 The density of states is sampled independently for each concentration of interest in both in the liquid state and the gas states. Insertion/deletion moves between the different concentration windows are performed in the gas phase in order to connect the independent concentration windows.

$(\ldots, N_i - 1, N_i, N_i + 1, \ldots)$ in an μVT ensemble (involving particle insertions and deletions) over the entire chosen concentration range, where the volume is chosen such that the number density of the system is sufficiently low that insertion/deletion moves become feasible.

As the density of states for each window is calculated to within a multiplicative constant, the individual density of states windows must be combined using a fitting procedure. This requires finding a set of offsets C_m using least squares, which minimises the error function

$$e_{tot} = \sum_{m=1}^{M} \sum_{n} \left[\ln \Omega_{m,NpT}(n) + C_m - \ln \Omega_{\mu VT}(n)\right]^2 \qquad (8)$$

where M is the number of individual concentration windows, n is an index for all the overlapping points shared by the two windows,[29] $\Omega_{m,NpT}$ is the density of states of concentration window m and $\Omega_{\mu VT}$ is the density of states sampled in the μVT ensemble.

This approach has significant advantages. Firstly, the insertion/deletion moves are favourable even for large solute molecules – the minimum, system number-density (maximum volume) sampled can be increased arbitrarily to accommodate this. Secondly, exploring the volume and concentration dimensions independently greatly reduces the space that must be explored. Instead of having to sample the entire, combined 3-dimensional energy, volume and concentration space (E–V–N_i), one essentially samples the 2-dimensional E–V and E–N_i spaces. Finally, to study broader temperature and pressure ranges, only the solution (liquid) portion of the windows need to be expanded (so as to cover the energies and volumes accessible to the system over the range of conditions to be studied), the rest remains constant. This significantly reduces the number of simulations that must be run when exploring temperature and pressure.

3. Technical details

The above methodology was applied to predict the solubility of NaCl in water. The molecular system contained 200 water molecules and between 6 and 18 sodium chloride pairs, covering a concentration range of $\sim$1.67–5.00 mol kg^{-1}. The SPC/E model was used to represent the water molecules, while the sodium chloride ion pair were modelled by the Joung–Cheatham (JC/SPC/E) force field.[34] A short MC simulation in the NpT ensemble was run for each of the concentrations at T = 298 K and p = 1 atm and T = 373 K and p = 1 atm to determine the accessible energy and volume ranges for the liquid portions of each concentration window. The simulations were repeated in the NVT ensemble at the elevated temperature of 10 000 K to determine the maximum and minimum energies accessible for each concentration in the gas phase. The high temperature was necessary to ensure that NaCl ions did not cluster together into a single nucleus, the formation of which would hinder the particle removal moves. The volume for the gas phase was fixed at 28.38 nm^3 which, by trial and error, was found to be large enough to easily accommodate the solute insertion moves.

We explored two approaches for choosing the accessible volume and energy ranges for states between the liquid and gas regions, shown in Fig. 3. The first approach was to simply interpolate the accessible energies and volumes between the liquid and gas values. For the second approach, at low volumes (those accessible to the liquid) we allowed the system to explore energies ranging from the liquid values all the way to close to the gas values, essentially allowing the liquid to pass into a supercritical regime. At higher volumes, moving towards the gas volume, the system was restricted to exploring only the high energy states. This second pathway was found to give a much faster convergence of the density of states (possibly because the system navigates around the first-order gas to liquid transition), and hence was used in this study.

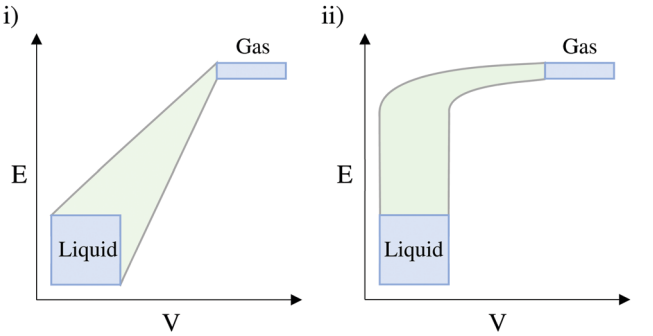

Fig. 3 The two choices explored for the accessible energies and volumes between the liquid and gas states: (i) direct interpolation between the liquid and gas states and (ii) transformation of the liquid to dense, high energy states before expanding to the gas state.

The energy range (for the whole system *i.e.* un-normalised by the number of molecules in the system) was discretised into bins of width 10 000 kJ mol^{-1} while the logged volume range was discretised into bins of width 0.008. These values where chosen so that the curvature of the peaks in the probability distributions was sufficiently captured, which is also a good indicator that the curvature of the density of states has been sufficiently captured.

The initial value of the Wang–Landau convergence factor was set to 1.0 and was allowed to decrease to 2.384×10^{-7}. By this point the relative change in the logged density of states between the current and previous iterations was sufficiently low, indicating that the density of states was converged. Further, both the chemical potentials and the probability distributions had also reached convergence by this point. It is crucial for this method that the density of states has indeed converged as small errors in the density of states can lead to large errors in the probability distribution.

The Monte Carlo code was parallelised using the scheme proposed by Vogel *et al.*[35] to expedite convergence and precision. The (N_i,E,V) space was partitioned into small overlapping chunks with multiple walkers being assigned to each chunk. Three walkers were found to be optimal for each liquid–gas window chunk and four walkers for each gas window chunk.

4. Results and discussion

The probability distribution for the JC/SPC/E model of sodium chloride at 298 K and 1 atm, calculated directly from the density of states by reweighting according to eqn (5), is shown

in Fig. 4. The NaCl solid chemical potential was taken as $\mu_i = -770.92$ kJ mol^{-1} as reported by Benavides *et al.*[8] Their choice of a de Broglie wavelength of 1.0 Å was adopted in this study. This choice does not affect the phase coexistence as the same value is used for the solution and solid phase calculations.[36]

The probability distribution reveals a dominant peak at about 13 NaCl pairs. Taking an ensemble average

$$\langle N_{\mathrm{NaCl}} \rangle_{T,p,N_{\mathrm{H_2O}}} = \sum_{E} \sum_{V} \sum_{N_{\mathrm{NaCl}}} N_{\mathrm{NaCl}} \times P(N_{\mathrm{NaCl}}, V, E)_{T,p,N_{\mathrm{H_2O}}} \quad (9)$$

gives an average of 13.57(18) sodium chloride pairs, and hence a solubility of 3.77(5) mol kg^{-1}, where P is the probability distribution given in eqn (5). Uncertainties in these values were calculated by averaging the results obtained from five independent DOS calculations. The calculated solubility is in close agreement with the values found in the literature for the Joung–Cheatham model (force field) for NaCl, the consensus literature value being 3.71(25) mol kg^{-1}.

This value is actually roughly half of the experimental solubility of 6.14 mol kg^{-1}. The disparity between the calculated and experimental solubility is due to the model itself (which is currently the best available).[8] In relative terms the solubility prediction is decent given that aqueous solubilities predicted by continuum solvation methods are at best within 4-fold of experimental data and often worse. The handful of solubility predictions from molecular simulation that have been reported (including the current study) reveal the critical nature of the force field parameters. Coexistence points are known to challenge force fields but for the same reason serve as essential data points for developing and optimising force field parameter sets.

We then used the determined density of states to ascertain how the chemical potential of NaCl solutions varies as function of concentration, using two distinct approaches. Firstly, we

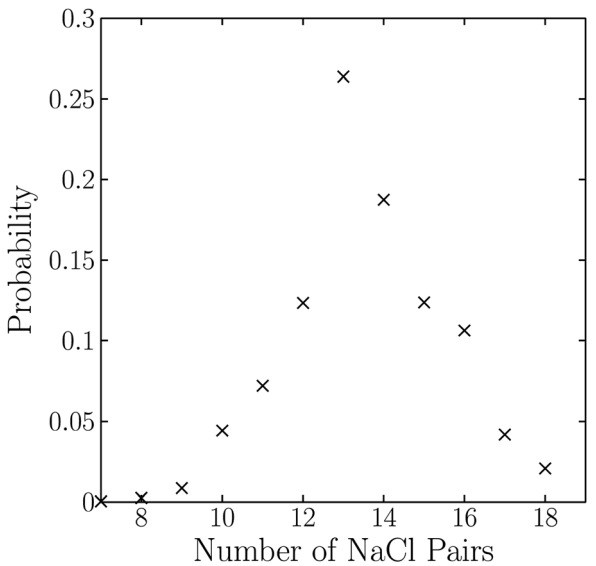

Fig. 4 The probability distribution for the aqueous sodium chloride system at T = 298 K and p = 1 atm, averaged over five independent runs.

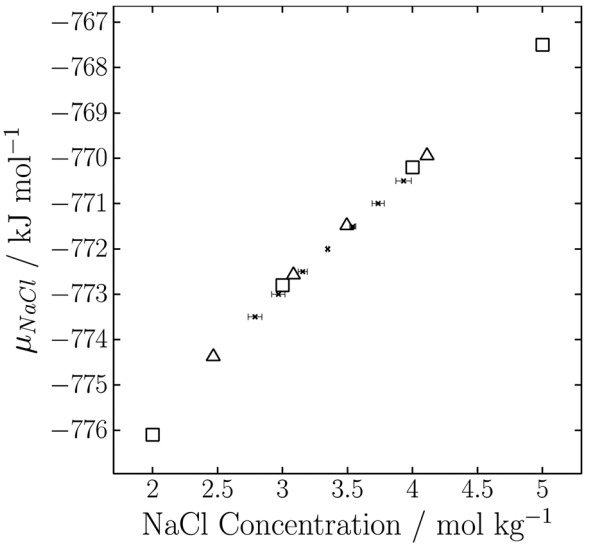

Fig. 5 The chemical potential of the JC/SPC/E NaCl model as a function of concentration as calculated by this work (crosses), Vega *et al.*[8] (triangles), Panagiotopoulos *et al.*[26] (squares).

calculated the chemical potential from the density of states for a series of NaCl concentrations by calculating the free energy as a function of concentration, to which a polynomial was fitted and then differentiated with respect to N_i. In the second approach we switched the independent–dependent variables and estimated the NaCl concentrations from probability distributions (as for NaCl solubility) corresponding to a series of chosen chemical potential values between −770.5 and −773.5 kJ mol^{-1}. While both approaches were in good agreement, the latter approach yielded values closer to those calculated by others – the data for which are presented in Fig. 5 along with values presented in the literature for this model.[8,26] As can be seen, the predicted values are in excellent agreement with the literature values, confirming that the presented DOS methodology not only offers a robust route to solubility prediction, but also enables the calculation of chemical potentials. In principle, these chemical potentials could be used to obtain the mean ionic activity coefficient of NaCl[37] as a function of temperature, provided the chemical potential of NaCl at infinite dilution is also estimated.

As a further validation of the method, the solubility of the JC/SPC/E NaCl model was calculated for a range of temperatures between 298 K and 374 K, from the same density of states surface as used for the calculation at 298 K. The predicted solubility as a function of temperature is presented in Fig. 6. For each of these calculations, the chemical potential of the NaCl crystal is required at the respective temperature, which was calculated following the procedure outlined by Argones *et al.*[38] and is presented in Table 1.

These chemical potential values of the NaCl solid, along with the density of states, were inserted into eqn (5) in order to generate probability distributions for each temperature, from which the NaCl solubility was determined as before. Counter-intuitively, the solubility of the NaCl model actually decreases

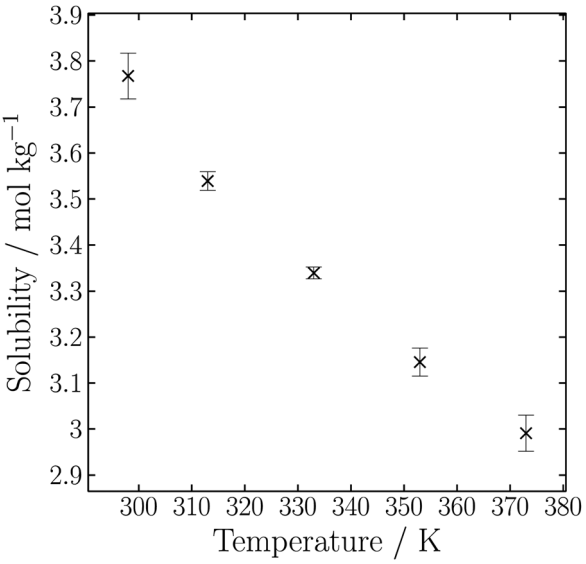

Fig. 6 The predicted solubility of the JC/SPC/E NaCl model as a function of temperature.

Table 1 Calculated chemical potential of the solid phase of JC/SPC/E NaCl model as a function of temperature

T/K	μ_{solid}/kJ mol^{-1}
313.00	−770.288(2)
333.00	−769.359(2)
353.00	−768.473(2)
373.15	−767.610(4)

as the temperature increases. This result has also been observed by Mester and Panagiotopouls,[26] who calculated the solubility of the Joung–Cheatham model of NaCl at 373.15 K to be 3.01(5) mol kg^{-1} – in close agreement with our value of 2.99(4) mol kg^{-1}. This unexpected behaviour is again attributed as a limitation of the model itself.[26]

A possible issue with the density of states approach for determining coexistence points is the potential for inadequate sampling of the coexistence states. The required nucleation step characterising first-order transitions (particularly the solid–liquid transition) is often suppressed as the creation of a surface involves an energy penalty. This is not an issue for the solubility prediction approach developed here. We are not sampling the dissolution of the solid nor its crystallisation but rather determining the density of states for the most part of the solution state, albeit around saturation.

There are three main sources of error within the methodology: errors associated with insufficient sampling, detailed balance not being satisfied, and the saturation of error caused by the modification factor reduction scheme. The errors due to saturation and detailed balance have been discussed at depth in the literature,[29,39] and are expected to be small relative to the sampling error. Notably, the overall estimated errors in the solubility and chemical potential calculations as determined from five independent sets of simulation were relatively small.

The efficiency gain of the DOS approach with respect to total computing requirement is probably not that marked for a single point solubility calculation relative to existing methods, such as TI or perturbation. The key gain arises from the DOS approach's inherent ability to predict solubilities over a range of temperature and pressure conditions from a single density of states. The calculation of the solid chemical potential is the same regardless whether the DOS approach or an existing methodology is employed, and so no efficiency is gained here. For the solution phase, a chemical potential calculation for a single concentration at a single fixed temperature by TI would involve around 21 simulations, one per lambda state.[8] Calculating the chemical potential of, say 10 concentrations, would then require 210 simulations. Should these then be repeated for the four additional temperatures considered by us, would result in a total of 1050 simulations. In contrast, for the DOS approach, we will probably need to investigate between 10–20 different concentrations to locate and fully capture the probability distribution that corresponds to the solution at saturation. For each of these concentrations we require three simulations – one to calculate the DOS of the solution phase, one to connect the solution and supercritical states, and one to connect the supercritical and gas states. For say 20 concentrations, this would amount to 60 simulations, plus one additional DOS calculation in the gas phase that connects the individual concentrations. Comparing the 61 DOS simulations with 1050 TI simulations yields a speed-up of about 17×.

While the DOS method has been applied only to a simple ionic system, we do not expect any significant challenges in extending the approach to larger solute molecules (including drug-like) in both aqueous and non-aqueous solvents. The switching from ions to molecules only requires a change in the density of the gas phase, avoiding the problematic creation and annihilation of particles in a condensed phase, which is a key benefit of the proposed method. Further, as the method samples only according to the density of states (*i.e.* entropy space), thermal barriers, such as those limiting dihedral rotations are expected to be less of an issue here than perhaps in other methods. For more challenging flexible molecules, the method could be coupled with established configurational-bias Monte Carlo moves to facilitate more efficient sampling of their molecular degrees of freedom.

In summary, we have developed and demonstrated a density of states approach to predicting solubility from molecular simulation. The method entails calculation of the density of states for a multicomponent solution, followed by exploration of the probability distribution as a function of number of solute particles in the system and the chemical potential of the solid, to identify coexistence conditions corresponding to the solubility. The density of states calculation is made possible by a unique pathway that avoids the problematic annihilation and/or creation of particles which is common to established methods. Consequently, the method is expected to perform well even for large, drug-like molecules. Further, it is able to yield, relatively efficiently, solubilities over a range of temperatures and pressures. The predicted solubility of the NaCl model at 298 K was found to be in close agreement with the literature.

Author contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

S. B. would like to thank the EPSRC, Lancaster University, and AstraZeneca for his PhD CASE Studentship (EP/L504804/1). We acknowledge the use of Lancaster University's high-performance computing (HPC) facility (HEC) and the N8 (consortium of the U.K.'s northern universities) HPC facility (Polaris).

References

1 S. P. Pinho and E. A. Macedo, *Dev. Appl. Solubility*, Royal Society of Chemistry, Cambridge, 2007, pp. 305–322.
2 J. D. Harper and P. T. Lansbury, *Annu. Rev. Biochem.*, 1997, **66**, 385.
3 R. Dasgupta and D. Walker, *Geochim. Cosmochim. Acta*, 2008, **72**, 4627.
4 C. R. Gardner, C. T. Walsh and Ö. Almarsson, *Nat. Rev. Drug Discovery*, 2004, **3**, 926.
5 B. Faller and P. Ertl, *Adv. Drug Delivery Rev.*, 2007, **59**, 533.
6 W. L. Jorgensen and E. M. Duffy, *Adv. Drug Delivery Rev.*, 2002, **54**, 355.
7 A. Klamt, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2018, **8**, e1338.
8 A. L. Benavides, J. L. Aragones and C. Vega, *J. Chem. Phys.*, 2016, **144**, 124504.
9 M. Ferrario, G. Ciccotti, E. Spohr, T. Cartailler and P. Turq, *J. Chem. Phys.*, 2002, **117**, 4947.
10 J. R. Espinosa, J. M. Young, H. Jiang, D. Gupta, C. Vega, E. Sanz, P. G. Debenedetti and A. Z. Panagiotopoulos, *J. Chem. Phys.*, 2016, **145**, 154111.
11 D. Frenkel and A. J. C. Ladd, *J. Chem. Phys.*, 1984, **81**, 3188.
12 C. Vega and E. G. Noya, *J. Chem. Phys.*, 2007, **127**, 154113.
13 T. P. Straatsma and H. J. C. Berendsen, *J. Chem. Phys.*, 1988, **89**, 5876.
14 T. P. Straatsma and J. A. McCammon, *Annu. Rev. Phys. Chem.*, 1992, **43**, 407.
15 J. G. Kirkwood, *J. Chem. Phys.*, 1935, **3**, 300.
16 R. W. Zwanzig, *J. Chem. Phys.*, 1954, **22**, 1420.
17 *Free Energy Calculations Theory and Applications in Chemistry and Biology*, ed. C. Chipot and A. Pohorille, Springer, Berlin, Heidelberg, 2007, pp. 33–75.
18 A. P. Lyubartsev, A. A. Martsinovski, S. V. Shevkunov and P. N. Vorontsov-Velyaminov, *J. Chem. Phys.*, 1992, **96**, 1776.
19 A. S. Paluch, D. D. Cryan and E. J. Maginn, *J. Chem. Eng. Data*, 2011, **56**, 1587.
20 F. Moučka, I. Nezbeda and W. R. Smith, *J. Chem. Phys.*, 2013, **138**, 154102.
21 T. C. Beutler, A. E. Mark, R. C. van Schaik, P. R. Gerber and W. F. van Gunsteren, *Chem. Phys. Lett.*, 1994, **222**, 529.
22 J. Anwar and D. M. Heyes, *J. Chem. Phys.*, 2005, **122**, 224117.
23 M. A. Barroso and A. L. Ferreira, *J. Chem. Phys.*, 2002, **116**, 7145.
24 J. L. Aragones, E. Sanz and C. Vega, *J. Chem. Phys.*, 2012, **136**, 244508.
25 M. Lísal, W. R. Smith and J. J. Kolafa, *J. Phys. Chem. B*, 2005, **109**, 12956.
26 Z. Mester and A. Z. Panagiotopoulos, *J. Chem. Phys.*, 2015, **143**, 44505.
27 C. Herdes, T. S. Totton and E. A. Müller, *Fluid Phase Equilib.*, 2015, **406**, 91.
28 L. Li, T. Totton and D. Frenkel, *J. Chem. Phys.*, 2017, **146**, 214110.
29 M. S. Shell, P. G. Debenedetti and A. Z. Panagiotopoulos, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2002, **66**, 56703.
30 E. A. Mastny and J. J. de Pablo, *J. Chem. Phys.*, 2005, **122**, 124109.
31 Q. Yan, R. Faller and J. J. de Pablo, *J. Chem. Phys.*, 2002, **116**, 8745.
32 F. Wang and D. P. Landau, *Phys. Rev. Lett.*, 2001, **86**, 2050.
33 S. Singh, M. Chopra and J. J. de Pablo, *Annu. Rev. Chem. Biomol. Eng.*, 2012, **3**, 369.
34 I. S. Joung and T. E. Cheatham, *J. Phys. Chem. B*, 2008, **112**, 9020.
35 T. Vogel, Y. W. Li, T. Wüst and D. P. Landau, *Phys. Rev. Lett.*, 2013, **110**, 210603.
36 C. Vega, E. Sanz, J. L. F. Abascal and E. G. Noya, *J. Phys.: Condens. Matter*, 2008, **20**, 153101.
37 Z. Mester and A. Z. Panagiotopoulos, *J. Chem. Phys.*, 2015, **142**, 44507.
38 J. L. Aragones, C. Valeriani and C. Vega, *J. Chem. Phys.*, 2012, **137**, 146101.
39 S. Schneider, M. Mueller and W. Janke, *Comput. Phys. Commun.*, 2017, **216**, 1.