University of Lancaster

Doctoral Thesis

Phase Equilibria from Molecular Simulation

Author:
Simon Boothroyd

Primary Supervisor:
Prof. Jamshed Anwar

Secondary Supervisor:
Dr. Andy Kerridge

Industrial Supervisor:
Dr. Anders Broo

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

in the

Department of Chemistry
Faculty of Science and Technology

May 2018

UNIVERSITY OF LANCASTER

Abstract

Department of Chemistry
Faculty of Science and Technology

Doctor of Philosophy

Phase Equilibria from Molecular Simulation

by Simon Boothroyd

Phase equilibria are at the heart of many properties of substances, such as their solubility, manufacturability, and stability. They are of significant industrial and commercial interest, perhaps most importantly to the pharmaceutical industry where drug stability and solubility are two of the largest challenges of drug development. The focus of this thesis then was to develop a molecular level understanding of phase equilibria, and produce tools and models to predict phase stability. An emphasis was given to exploring solid-solid and solid-liquid equilibria and stability. Specifically, the work presented here aimed to elucidate what drives the formation of multicomponent crystals, improve available models for exploring phase equilibria phenomena and explore solubility prediction from first principles as a potentially more powerful alternative to correlation based methods. These three fundamental areas were explored by employing molecular simulation in combination with the machinery of statistical mechanics, utilising advanced sampling methods and free energy calculations. This approach has led to the development of a foundation for understanding multicomponent crystal formation in terms of molecular affinities and packing, the characterisation of a set of soft coarse-grained potentials for use in phase equilibria studies, which overcome the main limitations of the most widely used potential, and finally, the development of a novel method for solubility prediction from first principles. Here, this novel method was successfully applied to an ionic (aqueous sodium chloride) and small molecular (urea in methanol and aqueous urea) system. In the future, these results are expected to lead to a set of guidelines for predicting (and perhaps prohibiting) multicomponent crystal formation, the development of a higher class of coarse-grained transferable force field with utility in studying phase equilibria, and powerful approach for predicting solubility of even large, flexible molecules (such as pharmaceuticals).

Declaration of Authorship

I, Simon Boothroyd, declare that this thesis titled, 'Phase Equilibria from Molecular Simulation' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

Acknowledgements

I would like to express my deepest gratitude to all those who have helped and supported me over the duration my PhD.

I would first like to thank my thesis advisor Prof. Jamshed Anwar for his constant support, guidance and always driving me to be my best. I would also like to thank Dr. Andy Kerridge for his invaluable feedback and discussions, and whose ability to simplify any problem was always helpful.

I would also like to thank all of the individuals with whom I spent time with while on site at AstaZeneca. In particular, I would like to thank Dr. Anders Broo and Dr. David Buttar for organising my visits there, for their insights into the industry and their valuable feedback on my work. Further, I would like to think Dr. Jim McCabe for allowing me into his lab, and letting me loose on some 'real' chemistry!

I would also like to express my sincerest gratitude to my friends and to my family. Without their support, I'm sure I would not have made it this far. And finally, I must thank my amazing partner Ana. Thank you for putting up with all the long nights and weekends spent in the office, for listening to me talk about my work, and for always making me happy - no matter how many simulations had failed.

List of Publications

This thesis is based on the following published and publishable manuscripts:

I. **Why Do Some Molecules Form Hydrates or Solvates?**
Simon Boothroyd, Andy Kerridge, Anders Broo, David Buttar, Jamshed Anwar
Cryst. Growth Des., 2018, 18, pp 1903–1908

S.B. and J.A. designed the research with help from A.K.; S.B. developed the methods and carried out the work; S.B. analysed the data; S.B. and J.A. wrote the manuscript with the help of all co-authors; all co-authors contributed to a critical discussion of the results and conclusions.

II. **Towards Realistic and Transferable Coarse-Grained Models: Phase Diagrams of Soft van der Waals Potentials**
Simon Boothroyd, Andy Kerridge, Jamshed Anwar
Manuscript in preparation

S.B. and J.A. designed the research; S.B. developed the methods and carried out the work; S.B. analysed the data; S.B. and J.A. wrote the manuscript with the help of all co-authors; all co-authors contributed to a critical discussion of the results and conclusions.

III. **Solubility prediction from first principles: A density of states approach**
Simon Boothroyd, Andy Kerridge, Anders Broo, David Buttar, Jamshed Anwar
In review in PCCP

S.B. and J.A. designed the research; S.B. developed the methods and carried out the work; S.B. analysed the data; S.B. and J.A. wrote the manuscript with the help of all co-authors; all co-authors contributed to a critical discussion of the results and conclusions.

IV. **Solubility prediction via chemical potentials from density of states**
Simon Boothroyd, Jamshed Anwar
Manuscript in preparation

S.B. and J.A. designed the research; S.B. developed the methods and carried out the work; S.B. analysed the data; S.B. and J.A. wrote the manuscript; all co-authors contributed to a critical discussion of the results and conclusions.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Scope of the Thesis

The primary aim of this thesis is to advance our understanding of (and develop methodologies to predict) phase equilibria. A molecular level insight into why, and under which conditions particular phases are favoured over others is key to predicting many properties of substances, such as their stability, solubility or even manufacturability. A deep understanding of phase equilibria is of significant commercial interest, with applications ranging from material development[1–3], gas storage and carbon capture[4–7], toxicology, food formulation and pharmaceutical development.

Of particular importance are phenomena arising from solid-solid and solid-liquid phase equilibria and stability. The issues arising from the relative stabilities of different solid forms are many - from the interconversion of polymorphs in pharmaceuticals to the formation of multicomponent crystals with degraded (or in cases enhanced) performance (as is discussed in Section 1.4). Similarly, the stability of a solid in solution (i.e its solubility) or lack thereof is a major issue for the pharmaceutical industry. This issue of solubility is two-fold - there is the challenge of initially dissolving a solid into solution, and then there is the potential issue of a new solid form with lower solubility recrystallising out. The second case can compromise the bioavailability of a pharmaceutical. A molecular level insight into what drives phase stability is thus critical. It would not only be key for developing relevant interventions (or exploiting potentially beneficial applications), but would also lay the foundation for predictive science. To this end, the aim of this thesis is to

- explore the phase stability of multicomponent crystals - the formation of which can be both exploited or highly problematic (see Sections 1.4.2 and 1.4.3).
- contribute to molecular simulation methodology employed to explore phase equilibria, with the intention of overcoming a number of the existing limitations explored in Section 2.3.
- explore methods for solubility prediction from first principles as a potentially powerful alternative to correlation based approaches.

Phase equilibria are investigated in the thesis using molecular simulation, which not only offers molecular level insight into the mechanism of phase transitions, but also when coupled with the machinery of statistical mechanics gives access to a kinetic and thermodynamic description of phase stability. Phase transitions are particularly challenging to simulate however, as transitioning between phases is a stochastic process that can occur over timescales much larger (from seconds all the way up to years) than are accessible from typical simulations (on the order of microseconds). Free energy calculations and biased simulations are employed here to overcome these challenges. Many of the simulations presented in this work were performed using PhaseMC - a bespoke Monte Carlo code that I developed for the purpose of exploring phase equilibria (see Appendix A).

This introductory chapter begins with a broad introduction into the kinetics and thermodynamics of phase equilibria. This is followed by an introduction to the different forms of the solid state, with an overview of their associated challenges and applications arising from phase stability. The second chapter provides an introduction to molecular simulation and statistical mechanics and their application to calculating phase equilibria is given. The remainder of this thesis is comprised of four significant components of original research (one of which has been published, and another of which is in review) resulting directly from this work, and a concluding chapter with a future perspective.

1.2 Nucleation

Nucleation is the first step of a phase transition. It involves atoms or molecules from the old phase self-assembling into small clusters of the new phase, known as nuclei. Nuclei whose size are below a certain threshold, the critical size r_c, are unstable and are likely to disassemble back into the old phase. Those that are above the critical size

however continue to grow until the new phase is fully formed[8]. The driving force behind nucleation from the melt is the degree of supercooling, from vapour it is the vapour pressure and from solution it is the degree of supersaturation. The magnitude of this driving force can be generalised as

$$\Delta\mu = \mu_{new} - \mu_{old} \tag{1.1}$$

where μ_{new} and μ_{old} are the chemical potentials of the new and old phases respectively[9].

Nucleation can be classified into two categories depending on how the nuclei form. The first, homogeneous nucleation, involves the formation of nuclei in a pure bulk medium without the presence of impurities or heterogeneous surfaces. The second, heterogeneous nucleation, involves the formation of nuclei at heterogeneous surfaces such as the walls of a beaker or on the surface of small impurities.

1.2.1 Homogeneous Nucleation

The process of homogeneous nucleation is described by the classical nucleation theory (CNT). When a nucleus of size r forms, there is a free energy penalty $G_{surface}$ associated with creating an interface between the old and the new phases. Conversely, provided that the new phase has a lower chemical potential, there is a competing favourable term G_{bulk} associated with particles transferring from the old to the new phase. The free energy change associated with a phase transition is thus given by

$$\Delta G = G_{surface} + G_{bulk} \tag{1.2}$$

Assuming a spherical nucleus (although this can be generalised to other shapes), these two terms are expressed as

$$G_{surface} = 4\pi r^2\gamma \tag{1.3}$$

$$G_{bulk} = \frac{4}{3}\pi r^3\Delta\mu \tag{1.4}$$

where γ is the surface energy density. Clearly as $G_{surface}$ depends on the surface area of the nucleus, it grows proportionally to r^2. Similarly, as G_{bulk} is dependent on the volume of the nucleus, it grows as r^3. At some size then, the critical size, the r^3 term begins to dominate and a maximum in the free energy is produced (Figure 1.1). The magnitude of this maxima ΔG_c (defined in Equation 1.5) is the height of the free energy barrier that must be overcome for a phase transition to occur.

$$\left.\frac{dG(r)}{dr}\right|_{r_c} = 0, \;\; \Delta G_c = \Delta G(r_c) \tag{1.5}$$

The rate constant of nucleation depends on ΔG_c, and is given by

$$k = A \exp\left(-\beta \Delta G_c\right) \tag{1.6}$$

where A is a kinetic prefactor, $\beta = \dfrac{1}{k_B T}$, k_B is the Boltzmann constant and T is temperature.

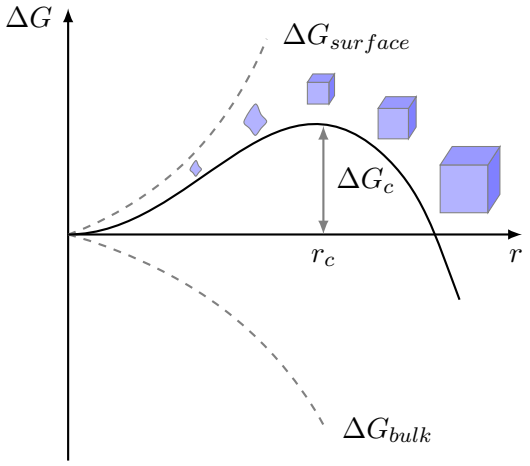


Figure 1.1: The competing contributions to the free energy of homogeneous nucleation as predicted by classic nucleation theory.

Although CNT gives a solid conceptual framework for understanding nucleation, and generally agrees qualitatively with experimental observations, it is often unable to match quantitative results. This is attributed to the assumption that bulk macroscopic properties, such as surface tension, can be used to describe microscopic nuclei, and that the nuclei have a sharp, rather than diffuse interface[10]. Further, CNT makes the assumption that nucleation is a single step process. There is evidence that nucleation can instead proceed via a multistep route[11].

1.2.2 Heterogeneous Nucleation

In almost all experiments, the system undergoing a phase transition is not homogeneous in composition and often contains many impurities such as dust particles. Further, the system will be in contact with a number of heterogeneous surfaces such as beaker walls or the fluid-air interface. As such, homogeneous nucleation is rarely observed in practice. Rather, heterogeneous nucleation is more common - nuclei form on the available heterogeneous surfaces.

Heterogeneous nucleation is more rapid than homogeneous nucleation as the surface will lower the unfavourable interfacial free energy barrier. The degree to which the surface will speed up nucleation depends on how greatly it mimics the structure of the final phase[12]. For crystallisation, an important type of heterogeneous nucleation is secondary nucleation that occurs on the surface of a pre-formed crystal. This seed crystal already matches the structure of the desired crystal, thus aiding nucleation. Molecular simulation has given insight into how crystals forming from solution can act as a further nucleation sites, thus catalysing nucleation even further[13].

1.3 Crystal Growth

Once nucleation has yielded nuclei of sufficient size, it becomes favourable for the nucleus to grow. Each new atom or molecule that adds to the nucleus acts a growth unit. If the new phase is crystalline, it is convenient to picture these growth units as simple cubes that assemble to form the larger crystal. Each growth unit will either be neighbours with other growth units, or a unit of the fluid phase. These model crystals are named Kossel crystals (Figure 1.2).

The surface of the Kossel crystal has three potential binding sites: facial sites that can form a single interaction, step sites that can form two, and kink sites that can form three (Figure 1.2). As facial sites only offer a single binding opportunity, they have the lowest binding energy while kink sites have the highest. As such, growth units will preferentially adsorb onto kink and step sites over facial ones[9]. Once the majority of kink and step sites have been occupied, 2-dimensional nucleation of the remaining flat surfaces of mainly facial sites must begin.

Given that facial sites only offer a single interaction with incoming growth units, adsorption is weak. There is a contest between adsorption and de-adsorption back to bulk fluid

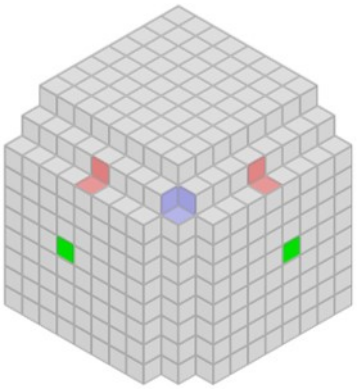


Figure 1.2: A three-dimensional Kossel crystal with examples of facial (green), step (red) and kink (blue) sites highlighted.

medium. Growth units that do adsorb to the surface will create new step sites. If other growth units adsorb into these new step sites, small islands begin to form thus creating new multiple binding sites where once there were only facial ones. These islands then continue to grow outwards into full layers (Figure 1.3).

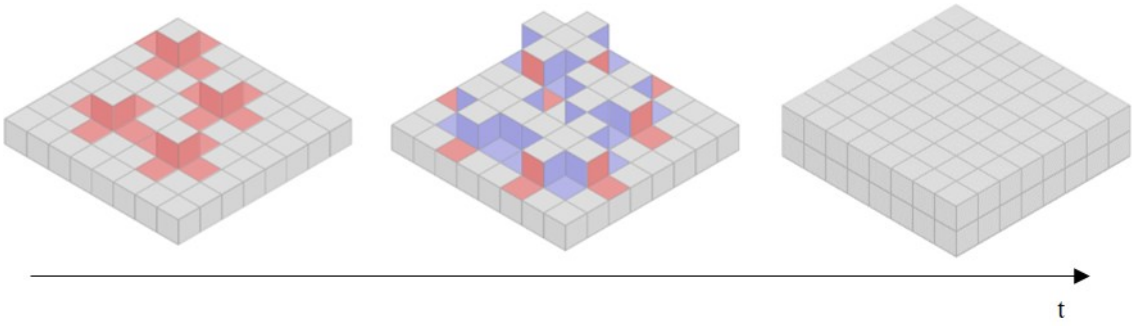


Figure 1.3: Two-dimensional surface nucleation growth with newly formed step and kink sites highlighted in red and blue respectively.

Although surface nucleation growth offers a good conceptual mechanism for growth on a surface, it makes the assumption that the growth surface is perfectly smooth. In reality, crystals contain many defects and dislocations[14]. In particular, screw dislocations (Figure 1.4) are responsible for spiral growth; screw dislocations produce continuous step-binding sites on the growth faces of crystals that enables new growth units to easily attach. By ignoring such dislocations, surface nucleation growth cannot fully account for the growth rates measured experimentally[9].

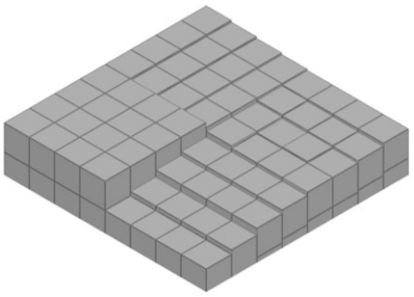


Figure 1.4: A screw dislocation.

1.4 The Solid State

The solid state is an integral aspect of chemistry. The properties of a solid are not solely reliant on the traits of a single molecule. Rather, bulk properties emerge from the collective behaviour and are not entirely predictable from the traits of the single molecule[15]. Phase stability of the solid state is one such bulk property. An understanding of phase stability is critical to many topical applications, as will be explored in this section.

The structure of a solid is classified as being either amorphous or crystalline. An amorphous solid is a structure that may have a short range, but no long range order. Its constituent molecules are locked into an almost completely random array. In contrast, crystals have a very well defined structure, constructed from tiled arrays of identical building blocks of atoms and molecules. The shape, size and composition of the individual blocks and how these are arranged in terms of orientations and translations on the lattice fully determine the structure of the crystal.

Amorphous solids tend to be less stable than their crystalline counterparts due to their disordered nature, and thus have a higher solubility, dissolution rate and can exhibit higher bioavailability when permeation across the gut membrane is not the rate limiting step[16]. This property is particularly desirable where solubility is an issue, as is often the case with pharmaceutical molecules. The lowered stability can be problematic however, as commercially viable products need to remain stable for their lifespan (typically 3-5 years). As such, there is a growing interest in stabilising amorphous materials by introducing various additives, for example polymers to act as stabilisers[16]. While currently a number of amorphous pharmaceutical products have made it to market[16], crystalline products tend to be more commonplace.

Crystalline solids can be further organised into three categories: polymorphs, solvates / hydrates and co-crystals[15] (Figure 1.5). These categories are not mutually exclusive, with some crystalline solids fitting into all such categories.

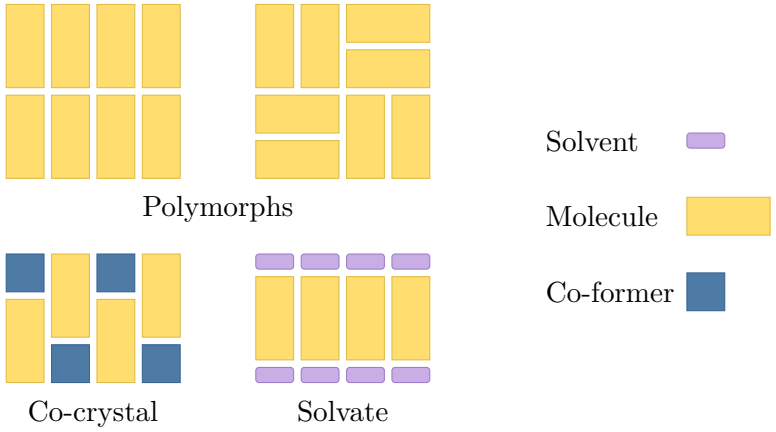


Figure 1.5: The three general categories of crystals.

1.4.1 Polymorphs

Polymorphs are crystals that are identical in composition but have a different arrangement or packing of their components within the crystal lattice. Differences in packing may vary from only slight rearrangements on a lattice to complete reorganisation. Further, polymorphs may arise from molecules adopting different conformations. Hence, polymorphism is especially common for larger, more flexible molecules - such as pharmaceutical molecules.

The physicochemical and mechanical properties (melting point, solubility, dissolution rate, crystal morphology, tensile strength to name a few) of different polymorphs tend to vary markedly[15,17,18]. This can be problematic. A single molecule may exhibit a large number of polymorphs, yet only one of the forms may have the required properties. There is no guarantee that the required polymorph will be the most stable however. Only one polymorphic form will be the most thermodynamically stable while the others will be metastable. This is not to say the metastable polymorphs cannot be used in applications. The metastable forms can be kinetically stable, taking months or even years to convert to the most stable form. Still, caution is required.

A notorious case of a metastable polymorph converting to a more stable form is Ritonavir[18]. While Ritonavir was produced as a polymorph with good solubility, over time it transformed into a previously unknown, more stable polymorph. The new form had a

lower solubility, which resulted in a reduced bioavailability. This led to a major product recall.

A comprehensive understanding of a molecule's polymorph landscape, and knowledge of their relative stabilities, is critical. While experimental screens are routinely employed to map out polymorph landscapes[18], there is no guarantee however that the most stable form will be found (as with Ritonavir). Computational screening offers a potential route to complement experimental screens, yet crystal structure prediction (CSP) is still very much an ongoing challenge. CSP is still mainly only successful for small, rigid molecules, although large strides are continually being made[19].

1.4.2 Solvates and Hydrates

Solvates and hydrates are one of the most prevalent types of multicomponent crystals. When crystals form from solution, the crystallisation solvent can become incorporated into the lattice. Crystals for which this phenomena is observed are known as solvates; hydrates are a special case where the incorporated solvent is water[15]. The formation of solvates and hydrates is common. It is predicted that roughly a third of drug molecules are able to form a hydrate[20].

How the solvent is incorporated into the lattice depends on the relationship between solute and solvent. The solvent can be directly integrated into the lattice, thus helping to stabilize the structure. As a general trend, the resulting solvate will be less soluble than the anhydrous form as the solvent has already interacted with the solute. Alternatively, the solvent can occupy channels that form between the solute lattice - these structures are known as channel solvates. The amount of solvent that occupies these channels is dependent on the vapour pressure[21].

As with polymorphs, the inclusion of solvent into the lattice results in a solid with markedly different properties to the pure form. The propensity for a solvate to have a decreased solubility can be detrimental to pharmaceuticals, as this can give rise to decreased bioavailability where dissolution is the limiting step. An example of this was the recall of generic carbamazepine pharmaceutical, due to the formation of a dihydrate[18] with compromised bioavailability resulting from a reduced solubility.

Even given the significance of solvates and hydrates, the fundamental question of why do some pairs of solute and solvents form solvates, while others only form anhydrous crystals, remains very much open. This issue is explored in depth in Chapter 3.

1.4.3 Co-crystals

Co-crystals, much like solvates, are another common form of multicomponent crystal. They are broadly defined as crystals that contain two or more components[22], although a precise definition is still under debate. Other definitions further stipulate that each component is solid[23]. Again, the inclusion of a second compound in the lattice can result in a solid whose properties vary markedly from those of the pure forms. Co-crystals are usually engineered to exploit this phenomenon.

The ability to modulate a compound's properties without the need to make covalent changes to the molecule's structure is highly desirable. Hence, co-crystals offer an attractive solution[22] where particular bulk properties of a system need to be enhanced[24,25]. Prediction of the final properties of a co-crystal from the properties of the individual components remains a challenge[26]. This, combined with an ability to predict co-former compatibility would open the possibility of designing co-crystals[27].

Forming a co-crystal is not always as simple as just mixing the active pharmaceutical ingredient (API) with the co-former. The temperature, concentrations and even solvent used during crystallisation all play an important role in what products will be formed[28]. Hence, the phase diagram of the systems often lies at the heart of co-crystal design.

Although the binary phase diagram offers insight into the compatibility of the API for the co-former[29,30], it does not offer a complete picture. Even though it predicts the formation of a co-crystal under certain conditions, changing the crystallisation solvent can result in no co-crystal being formed. This phenomena can be rationalised by studying a ternary phase diagram with axis of API, co-former and solvent concentration[28]. It can be inferred that in order to truly be able to predict the conditions needed for co-crystal formation, one needs to consider the solvent as much as the co-former[28,31,32].

A key consideration for solvent selection is solubility of the co-former and API. This will determine not only solvent compatibility, but also the crystallisation approach that must be taken to ensure the co-crystal is formed[31]. Hence knowledge of the co-former and API's solubility in a range of solvents, and at a range of conditions is necessary. Molecular simulations offer a potentially robust and efficient route to solubility prediction. A novel, robust and efficient method for which is presented in Chapter 5.

1.5 Thesis Outline

The primary aim of this thesis is to explore and predict phase equilibria using the molecular simulation techniques introduced in Chapter 2.

Chapter 3 aims to address the fundamental question of 'why do solvates and hydrates form' using molecular simulations of a simple model system. The intent is to elucidate the core principles that facilitate the formation of hydrates. Phase stability is explored as a function of saturation, pressure, solute-solvent affinity and solute-solvent size ratio. The work also provides a foundation for answering the question of 'why do only some pairs of molecules form co-crystals'.

While the simple, coarse grained model employed in Chapter 3 enabled solvate formation to be studied, it suffered from a number of limitations that prevented certain aspects of solvation to be fully explored. Namely the model has only a limited boiling point range. This forced us to study solvate formation for only a very small range of temperatures. Further, the model suffered from kinetic trapping - the solid form that crystallised out was, under certain conditions, the metastable one (i.e. the solvate occasionally formed when the anhydrous phase should have been more stable). The main source of both these issues was that the potential used to model the coarse-grained interactions was too hard as the repulsive wall is too steep. Hence, the aim of the work described in Chapter 4 is to characterise a 'softer' coarse grained potential model. This will enable the design of a better class of coarse grained models for use in a wide range of applications.

In Chapter 5, a novel, robust and efficient method is proposed for solubility prediction from first principles. The method is in principle able to calculate the solubility for even large, drug-like molecules for a large range of temperatures and pressures. The method should thus be capable of predicting the solubility gain / loss by forming a solvate / co-crystal as opposed to a single component crystal, which in turn could have large utility in the drug development process. In Chapter 6, this method is further extended to the calculation of chemical potential of fluid phases as a potentially more efficient route to solubility calculation, and applied to molecular systems.

Finally, the results and their significance are summarised in a concluding discussion chapter.

Chapter 2

Theory

2.1 Statistical Mechanics

Statistical mechanics at its heart is the bridge between the microscopic and macroscopic - relating atomic information such as position and momentum to more familiar and practically useful properties such as temperature, pressure, and chemical potential.

From a classical point of view, the microscopic state of a system of N atoms is described by $6N$ coordinates: $3N$ give the atomic positions $\boldsymbol{r}^N$, while the other $3N$ give the momenta $\boldsymbol{p}^N$. The full $6N$-dimensional coordinate space is defined as phase space. At the microscopic level, atomic coordinates evolve along deterministic trajectories though phase space, giving rise to fluctuations in macroscopic properties. Most often it is the average of these properties that are of interest, rather than the individual motions of the atoms themselves. This averaging forms the basis of statistical mechanics.

Statistical mechanics provides two approaches for calculating macroscopic averages. The first approach is to follow the microscopic trajectory of a system (by performing a molecular dynamics simulation for example), and take a *time average* of the observable of interest A, such that

$$\langle A \rangle = \lim_{\tau \to \infty} \frac{1}{\tau} \int_{t=0}^{\tau} A\left(\boldsymbol{r}^N(t), \boldsymbol{p}^N(t)\right) \mathrm{d}t \tag{2.1}$$

where τ is the length of time over which the average is taken. The second approach would conceptually involve constructing an ensemble of many replicas of the system,

identical in nature, but occupying different phase space coordinates. The time average could then be replaced by an *ensemble average* over the ensemble of configurations

$$\langle A \rangle = \int P\left(\boldsymbol{r}^N, \boldsymbol{p}^N\right) A\left(\boldsymbol{r}^N, \boldsymbol{p}^N\right) \mathrm{d}\boldsymbol{r}^N \mathrm{d}\boldsymbol{p}^N \tag{2.2}$$

where P is the ensemble's probability density. The ergodic hypothesis states that these two approaches are equivalent.

The probability density in Equation 2.2 arises from the imposition of macroscopic constraints on the microscopic trajectories. A closed and isolated system, for example, would naturally exist at a fixed volume, number of particles and energy. As such, only those elements of phase space that satisfy those conditions would have a non-zero P. An ensemble of configurations subject to these conditions is named the microcanonical ensemble (see 2.1.1). Of course, we are not limited to averaging in the microcanonical ensemble. By coupling the system of interest to a thermal bath, averages may be taken over an ensemble of configurations existing at a fixed temperature, rather than energy (see 2.1.2). Similarly, coupling the system to a barostat, or a permeable membrane connected to an infinite particle reservoir, allows averages at constant pressure or chemical potential may be calculated (see 2.1.3). These common ensembles are detailed in the following subsections.

2.1.1 The Microcanonical Ensemble

The microcanonical (NVE) ensemble is the simplest of the thermodynamic ensembles. As described above, it is one in which the energy E, volume V and number of particles N is fixed - i.e it is representative of a closed and isolated system.

The fundamental *a priori* probability postulate of statistical mechanics states that, for an isolated system, all microstates with an equal energy are equally probable. Hence, the probability density for this ensemble is given by

$$P\left(\boldsymbol{r}^N, \boldsymbol{p}^N\right) = \frac{1}{\Omega} \delta\left(\mathcal{H}\left(\boldsymbol{r}^N, \boldsymbol{p}^N\right) - E\right) \tag{2.3}$$

where Ω is the density of states, $\mathcal{H}$ is the Hamiltonian of the system and δ is the Dirac delta function. The density of states is the total number of microstates that have a non-zero probability for the given energy level. It is essentially a degeneracy, and is directly related to the system's entropy S by

$$S\left(N,V,E\right) = k_B \ln \Omega\left(N,V,E\right) \tag{2.4}$$

where k_B is the Boltzmann constant.

2.1.2 The Canonical Ensemble

The canonical (NVT) ensemble is one in which the temperature T, volume and number of particles is fixed. The probability of finding a given microstate subject to these macroscopic constraints is given by the Boltzmann distribution

$$P\left(\boldsymbol{r}^N, \boldsymbol{p}^N\right) = \frac{1}{Q\left(N,V,T\right)} \exp\left[-\beta \mathcal{H}\left(\boldsymbol{r}^N, \boldsymbol{p}^N\right)\right] \tag{2.5}$$

where $\beta = \dfrac{1}{k_B T}$, and Q is the canonical partition function, defined by

$$Q\left(N,V,T\right) = \frac{1}{h^{3N} N!} \int \exp\left[-\beta \mathcal{H}\left(\boldsymbol{r}^N, \boldsymbol{p}^N\right)\right] \mathrm{d}\boldsymbol{r}^N \mathrm{d}\boldsymbol{p}^N \tag{2.6}$$

Here the Planck constant h is introduced so that the classical partition function matches the quantum mechanics description of a particle in a box, and the $N!$ factor accounts for the indistinguishability of particles.

While Q primarily acts as a normalisation constant, it is in fact one of the most fundamental quantities in thermodynamics. Although it equates to just a single number, all thermodynamic properties of a system at equilibrium can be determined from it, including the average system energy

$$\langle E \rangle = \frac{\partial \ln Q\left(N,V,T\right)}{\partial \beta} \tag{2.7}$$

and from it heat capacity

$$C_V = \frac{\langle E \rangle}{\partial T} \tag{2.8}$$

the entropy

$$S = \frac{\partial}{\partial T} \left(k_B T \ln Q \left(N, V, T \right) \right) \tag{2.9}$$

and perhaps most importantly, the Helmholtz free energy

$$F(N, V, T) = -k_B T \ln Q \left(N, V, T \right) \tag{2.10}$$

Evaluating the high dimensional integrals in Equation 2.6 analytically is unfeasible for all but the simplest systems however. Instead, numerical approaches must be taken - these often take the form of the molecular simulation techniques described in Section 2.2

The form of the partition function may be simplified slightly by analytically integrating out the momentum terms. The Hamiltonian can be written as the sum of the system's kinetic (K) and potential (U) energies

$$\mathcal{H} \left(\boldsymbol{r}^N, \boldsymbol{p}^N \right) = K \left(\boldsymbol{p}^N \right) + U \left(\boldsymbol{r}^N \right) = \sum_{i=1}^{N} \frac{\left| \boldsymbol{p}_i^2 \right|}{2m_i} + U \left(\boldsymbol{r}^N \right) \tag{2.11}$$

where the potential energy is only dependent on the positions, and the kinetic energy depends only on the momenta. Here m_i is the mass of atom i. Inserting this definition into Equation 2.6 yields

$$Q \left(N, V, T \right) = \frac{1}{h^{3N} N!} \int \mathrm{d}\boldsymbol{p}^N \exp \left[-\beta \sum_{i=1}^{N} \frac{\left| \boldsymbol{p}_i^2 \right|}{2m_i} \right] \int \mathrm{d}\boldsymbol{r}^N \exp \left[-\beta U \left(\boldsymbol{r}^N \right) \right] \tag{2.12}$$

where now the integral has been separated into one over all momenta, and another all positions. The integral over all momenta evaluates analytically to

$$\int \mathrm{d}\boldsymbol{p}^N \exp \left[-\beta \sum_{i=1}^{N} \frac{\left| \boldsymbol{p}_i^2 \right|}{2m_i} \right] = \frac{h^{3N}}{\Lambda^{3N}} \tag{2.13}$$

where $\Lambda = h/\sqrt{2\pi m k_B T}$ is the de Broglie wavelength of a classical particle. The final form of the partition function then becomes

$$Q\left(N,V,T\right) = \frac{Z\left(N,V,T\right)}{N!\Lambda^{3N}} \tag{2.14}$$

where

$$Z\left(N,V,T\right) = \int \mathrm{d}\boldsymbol{r}^N \exp\left[-\beta U\left(\boldsymbol{r}^N\right)\right] \tag{2.15}$$

is the configurational partition function. Given that the momentum has been integrated out, a corresponding probability distribution over only atomic positions is expressed as

$$P\left(\boldsymbol{r}^N\right) = \frac{1}{Z\left(N,V,T\right)} \exp\left[-\beta U\left(\boldsymbol{r}^N\right)\right] \tag{2.16}$$

This expression is fundamental to the Monte Carlo simulations described in Section 2.2.2

2.1.3 The Isothermal-Isobaric, and Grand Canonical Ensembles

Two useful extensions of the canonical ensemble are the isothermal-isobaric (NpT) and grand canonical (μVT) ensembles.

In the isothermal-isobaric ensemble, the number of particles, pressure (p) and temperature is fixed. Much like the canonical ensemble, a probability distribution

$$P\left(\boldsymbol{r}^N,\boldsymbol{p}^N,V\right) = \frac{1}{Q\left(N,p,T\right)} \exp\left[-\beta\left(\mathcal{H}\left(\boldsymbol{r}^N,\boldsymbol{p}^N\right)+pV\right)\right] \tag{2.17}$$

a corresponding partition function

$$Q\left(N,p,T\right) = \int \exp\left[-\beta pV\right] Q\left(N,V,T\right) \mathrm{d}V \tag{2.18}$$

and a free energy expression (in this case the Gibbs free energy)

$$G(N,p,T) = -k_B T \ln Q\left(N,p,T\right) \tag{2.19}$$

are defined for the isothermal-isobaric ensemble. It is often the preferred ensemble to work in as it is gives a close representation of experimental conditions.

The NpT ensemble is somewhat limited, however, by the condition of a fixed number of particles. Several applications require the number of particles in the system to fluctuate. These include studying the adsorption of particles into materials, or more importantly for this work, calculating the solubility of a system (see Chapter 5). For such applications, the grand canonical ensemble is employed - in the grand canonical ensemble the chemical potential (μ), volume and temperature of the system is fixed. Its partition function Ξ takes the form

$$\Xi\left(\mu, V, T\right) = \sum_{N=0}^{\infty} \exp\left[-\beta\mu N\right] Q\left(N, V, T\right) \tag{2.20}$$

Conceptually, this ensemble is representative of coupling the system of interest to an infinite reservoir of particles. The two subsystems would be separated by a permeable membrane that allows particles to transfer between the two subsystems.

2.2 Molecular Simulation

Molecular simulation attempts to simulate the microscopic world, offering an atomic resolution not accessible to experiment. Not only does it offer a powerful tool to study the dynamic behaviour of systems, thermodynamic and structural properties can be extracted using the machinery of statistical mechanics. The two most commonly employed molecular simulation techniques, molecular dynamics and Monte Carlo simulation, are described in the following subsections.

2.2.1 Molecular Dynamics Simulation

Molecular dynamics simulations employ Newtonian mechanics to evolve the state of a system over time. They give direct access to the dynamics of a system, as well as a route to measuring thermodynamic and structural properties.

At each step of a molecular dynamics simulation, the force acting on each atom in the system

$$\boldsymbol{f}_i = -\frac{dU\left(\boldsymbol{r}^N\right)}{d\boldsymbol{r}_i} \tag{2.21}$$

is calculated from a potential energy function U (discussed in greater detail in section 2.2.4), where $\boldsymbol{f}_i$ is the force acting on atom i and $\boldsymbol{r}_i$ is the position of atom i.

The acceleration for each atom is then derived using Newton's laws of motion

$$m_i \ddot{\boldsymbol{r}}_i = \boldsymbol{f}_i \tag{2.22}$$

Integrating the acceleration numerically yields a new position and velocity for each atom at a short time interval ahead. This procedure of generating forces, deriving accelerations and from these, generating new positions and velocities is repeated typically for millions of steps. In doing so, the system is evolved through time and a molecular trajectory is constructed.

The choice of the time interval is critical. If it is too small, then the computational time required to adequately explore phase space will be large. Conversely, if the time step is too large, the numerical integration will be unstable due to atoms grossly impacting and overlapping into each other resulting in high energy states. The best choice of value depends on the dynamics of the system being simulated. Simulations of largely flexible molecules require a much shorter timestep (typically around 1 fs) than simulations of rigid bodies, for example, due to the increased mobility / faster motion of the atoms.

The configurations generated by this procedure would be consistent with the micro-canonical ensemble - i.e the number of particles, the volume and the total energy of the system is conserved. For most simulations, however, it common to employ either NVT or NpT ensembles.

Sampling is performed in the NVT ensemble by coupling the sytem to a thermal bath in order fix the temperature of the system. The instantaneous temperature of a system at time t is related to the atomic velocities by

$$\frac{3}{2} k_B T(t) = \sum_i^N \frac{1}{2} m_i \left|\boldsymbol{v}_i\right|^2 \tag{2.23}$$

where $\boldsymbol{v}_i$ is the velocity of atom i. Temperature then can be controlled by modulation of the velocities of each atom. This is generally accomplished in practice by either

stochastically scaling the velocities directly, or by introducing a fictitious dynamical variable that modulates the velocities by scaling the simulation time itself[33].

Molecular dynamics can further be extended to the NpT ensemble by introducing a barostat. During an NpT simulation, the volume of the system is adjusted in order to maintain the system's pressure. During a simulation, the instantaneous pressure can be calculated according to

$$PV = Nk_BT + \mathcal{W} \tag{2.24}$$

where

$$\mathcal{W} = \frac{1}{3}\sum_{i=1}^{N} \boldsymbol{r}_i \cdot \boldsymbol{f}_i \tag{2.25}$$

is the system's virial.

2.2.2 Monte Carlo Simulations

Monte Carlo simulations are a broad class of numerical simulation, with applications that include the accurate evaluation of multidimensional integrals and generating states according to probability distributions. They are thus ideal for calculating the thermodynamic properties of molecular systems using the machinery of statistical mechanics. As was discussed in Section 2.1, the thermodynamic properties of a system are accessible through the partition function. For simplicity the technique will be introduced in the context of the NVT partition function, however it is easily extended to the many others.

Evaluation of the partition function for general systems is impossible both analytically and by numerical integration. A naive approach to evaluate Equation 2.15 would be to use one of the many quadrature methods, such as Simpson's rule, whereby the atomic coordinates for the N particles would be located on a uniformly distributed grid. The number of grid points that would be required to properly capture the curvature of the potential energy surface of a molecular system (even for relatively small systems) would be enormous however.

An alternative approach would be to generate configurations at random, and weight any observables of interest according to the Boltzmann distribution

$$\langle A \rangle = \frac{\sum_{i=1}^{N_{confs}} A\left(\boldsymbol{r}^N\right) \exp\left[-\beta U\left(\boldsymbol{r}^N\right)\right]}{\sum_{i=1}^{N_{confs}} \exp\left[-\beta U\left(\boldsymbol{r}^N\right)\right]} \tag{2.26}$$

where N_{confs} is the number of random configurations generated. This approach would be incredibly inefficient. The majority of sampled configurations would have a very high energy due to a large number of overlapping particles. Their Boltzmann factors would thus essentially be zero and hence they would barely contribute to Equation 2.26.

A better approach would be only to consider those configurations that do have a significant Boltzmann weight. In other words, we wish to generate configurations with a probability $\propto \exp\left[-\beta U\left(\boldsymbol{r}^N\right)\right]$. This is the basis of the Metropolis scheme[34].

The Metropolis scheme proceeds by performing trial moves that transform the old state of a system o to some new state n. These may include particle translations or rotations, moves that scale the volume of the box or even particle insertion / deletion moves depending on the desired ensemble (see Figure 2.1).

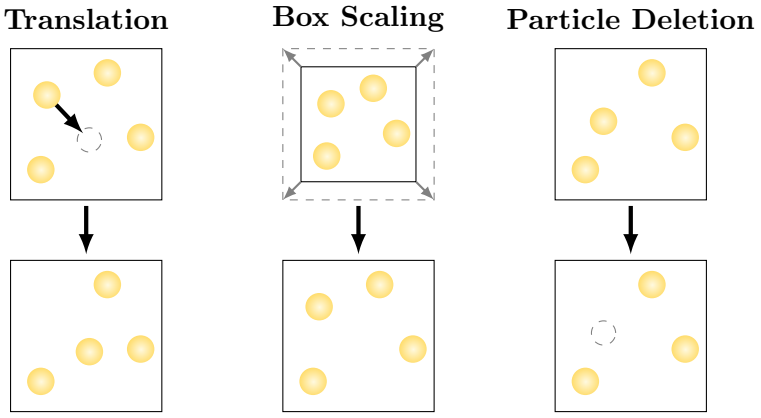


Figure 2.1: The state of the system is most commonly evolved in Monte Carlo simulations by performing particle translations, volume scaling (in the NpT ensemble) and particle insertions / deletions (in the μVT ensemble)

The moves are chosen such that the condition of detailed balance

$$P_o \pi_{o \to n} = P_n \pi_{n \to o} \tag{2.27}$$

is obeyed, where P_o, P_n are the desired equilibrium probabilities of being in states o and n respectively (in the NVT ensemble they take the form of Equation 2.16), and $\pi_{o \to n}$, $\pi_{n \to o}$ are the probabilities of transitioning from o to n and n to o respectively. Enforcing this rather strict condition guarantees the correct probability distribution is sampled.

The probability of transitioning between states can be split into two components

$$\pi_{o\to n} = g_{o\to n}\alpha_{o\to n} \tag{2.28}$$

where $g_{o\to n}$ is the probability of **proposing** the new state from the old state, and $\alpha_{o\to n}$ is the probability of actually **accepting** the proposed transition. $P_n\pi_{n\to o}$ is similarly defined.

Inserting Equation 2.28 into 2.27 and rearranging yields

$$P_o g_{o\to n}\alpha_{o\to n} = P_n g_{n\to o}\alpha_{n\to o} \implies \frac{\alpha_{o\to n}}{\alpha_{n\to o}} = \frac{P_n g_{n\to o}}{P_o g_{o\to n}} \tag{2.29}$$

While many choices of $\alpha_{o\to n}$ ensure this equation is satisfied, the choice of

$$\alpha_{o\to n} = \min\left[1, \frac{P_n g_{n\to o}}{P_o g_{o\to n}}\right] \tag{2.30}$$

is most commonly employed[33]. This acceptance criteria is the heart of a Monte Carlo simulation - it defines whether configurations generated by the trial moves should be either accepted, or rejected. It should be noted that this acceptance criteria is incredibly general - provided that the probability of generating a new configuration from an old one can be determined, virtually any trial move (even those that are unphysical) can be performed.

The most common trial moves employed in a Monte Carlo simulation are particle translations. A particle in the system is selected at random and displaced by a random amount between $-\delta_{max}$ and δ_{max} in all three dimensions. The probabilities for generating the forward and reverse moves are thus given by

$$g_{o\to n} = g_{n\to o} = \frac{1}{N}\frac{1}{(2\delta_{max})^3} \tag{2.31}$$

Combining these with Equations 2.16 and 2.30 yields an acceptance criteria of

$$\begin{aligned}\alpha\left(o \rightarrow n\right) = & \left\{\exp\left[U\left(\boldsymbol{r}_n^N\right)\right] \frac{1}{N} \frac{1}{\left(2\delta_{max}\right)^3}\right\} \\ & \times \left\{\exp\left[U\left(\boldsymbol{r}_o^N\right)\right] \frac{1}{N} \frac{1}{\left(2\delta_{max}\right)^3}\right\}^{-1} \\ = & \exp\left[U\left(\boldsymbol{r}_n^N\right) - U\left(\boldsymbol{r}_o^N\right)\right] \end{aligned} \tag{2.32}$$

Remarkably, the partition function has completely cancelled out. Thus in a Monte Carlo simulation we can directly sample and calculate properties from an ensemble without explicitly calculating its partition function.

In practice, a Metropolis Monte Carlo simulation in the NVT ensemble will be performed in the following way:

1. Select a particle at random and displace it by some random amount in the range $-\delta_{max}$ to δ_{max}.
2. Calculate the change in energy between the new and old states and from this the acceptance criteria in Equation 2.32.
3. Generate a random number between 0.0 and 1.0 and compare it with the acceptance criteria:
   i) if it less than the criteria, the move is accepted.
   ii) otherwise, the move is rejected and the system is returned to its prior state.

Given that each configuration is generated with the correct probability by this method, the average value of an observable can simply be calculated as the average over the stochastic trajectory generated by the successive trial Monte Carlo moves.

Within the framework of Metropolis, extension to other ensembles (or in fact to any arbitrary probability distribution) is trivial, and often only requires two alterations to the canonical example. First, the probability distribution in Equation 2.30 must be swapped with the distribution of interest. Secondly, new moves must be introduced to ensure that all of the external variables of an ensemble are explored. In the isothermal-isobaric ensemble, this means that moves that explore the accessible volume range of a system must be introduced. These are generally employed as moves that scale the size and coordinates of the simulation box. Similarly, particle insertion / deletion moves must be performed when sampling in the grand canonical ensemble.

2.2.3 Error Estimation on Averaged Quantities

The average value of any observable calculated by molecular simulation (whether that be Monte Carlo or molecular dynamics) will be subject to a statistical uncertainty, arising from the finite length of the trajectory over which the average was taken. It is often important to quantify this error, and how it propagates through calculations and into subsequent properties of interest.

The average value of a property A measured during a simulation is calculated according to

$$\bar{A} = \frac{1}{\tau_{run}} \sum_{i=1}^{\tau_{run}} A_i \tag{2.33}$$

where τ_{run} is the number of samples taken during the simulation, and A_i is the i'th sample of A taken. Here the bar notation is employed to distinguish between the simulation average, and the true ensemble average (Equation 2.2), which would be equivalent in the limit of infinite samples and provided sampling is ergodic. If the collected samples are uncorrelated, the estimated error in this average would be

$$\sigma_{\bar{A}} = \frac{1}{\tau_{run}} \sqrt{\sum_{i=1}^{\tau_{run}} \left(A_i - \bar{A}\right)^2} \tag{2.34}$$

Data is often sampled so frequently during a simulation however, that successive data points are heavily correlated. The most common approach to overcome this is to employ block averaging[38]. The sampled data set is split into a number of blocks (n_b) of length τ_b, so that $\tau_{run} = \tau_b * n_b$. The average from each block $\bar{A}_b$ is then calculated by

$$\bar{A}_b = \frac{1}{\tau_b} \sum_{i=1}^{\tau_b} A_i \tag{2.35}$$

As the block size is increased the block averages themselves become uncorrelated, so that the total error in the average may be estimated by

$$\sigma_{\bar{A}} = \frac{1}{n_b} \sqrt{\sum_{b=1}^{n_b} \left(\bar{A}_b - \bar{A}\right)^2} \tag{2.36}$$

The value which τ_b should take is often unknown *a priori*, and may be found in practice by plotting $\sigma_{\bar{A}}$ as a function of τ_b, and identifying the value at which the plot plateaus.

The uncertainties calculated in properties according to Equation 2.36 must be propagated through any calculations by the standard expressions in order to obtain uncertainties in the quantity of interest, such as free energies calculated by the thermodynamic integration method introduced in Section 2.3.3.2.

2.2.4 Modelling Molecular Interactions

In a classical simulation, electrons and protons are not simulated explicitly. While this greatly reduces the complexity and time needed to run a simulation, the many inter- and intramolecular interactions need to be approximated. This is accomplished using the potential energy function that approximately describes the nature of the interactions between the atoms.

Every intramolecular interaction (such as bond stretching, angle bending or torsional rotations) and every intermolecular interaction (both Coulombic and Van der Waals) will have an associated potential energy. Each of these can be approximated by an empirical function and an associated set of parameters - the energy of a bond stretching can be approximated by a harmonic potential for example, that is parameterised by a bond length and a bond stiffness.

The potential energy function is the combined sum of all such functions (Figure 2.2), and can be directly employed in the molecular simulations described in sections 2.2.1 and 2.2.2.

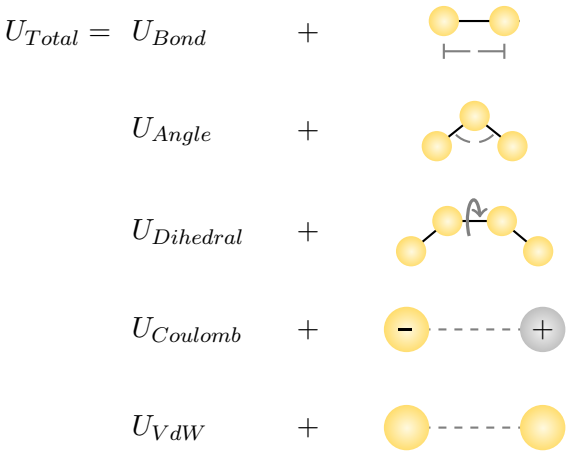


Figure 2.2: An example potential energy function as a sum of its individual components.

Each of the potential functions (and the parameters that describe them) must be chosen with great care, as these are what will determine the accuracy of the molecular models, and hence the results of any simulations that use them. They must represent the physical molecular interactions as closely as possible. The challenge, however, is that there is no unique way of making these choices. Although there are many ways to computationally model a classical molecule[35], there are two general approaches: the all-atomistic approach and the coarse grained approach.

The all-atomistic approach treats each atom in a molecule as a distinct particle. The parameters that describe each of the interactions shown in Figure 2.2, such as the bond length and stiffness in the case of a harmonic bond potential, are normally derived either by quantum mechanics or by empirically fitting to experimental data. Non-bonded intramolecular interactions (Van der Waals and Coulombic) between atoms separated by one or two bonds (referred to as 1-2 and 1-3 interactions respectively) are excluded from the energy function, as these are encoded within the bond and angle parameters respectively. In some cases, the non-bonded interactions between atoms separated by three bonds are also excluded (as they may already be accounted for by the dihedral parameters), although it is more common to include them, but to scale them by some constant.

While the all-atomistic approach can be used to closely match most of the physical molecular interactions well, and hence reproduce the bulk properties of the physical system, it rapidly becomes more expensive as the number of atoms in the system increases. This increasing cost limits both the size and the length of simulation can be run. Clearly then when simulating phenomena that occur over very large timescales (e.g protein folding), or require many particles (e.g studying crystal defects) a different approach must be taken.

The alternate approach, coarse graining, is to consider groups of atoms or even whole molecules as single particles. This has two main benefits. The first is that the number of particles that must be simulated (and hence the number of calculations that must be made) is dramatically reduced. The second is that the potential energy surface of the system becomes significantly smoother, allowing the phase space to be traversed much more rapidly. Combined, this means that larger systems can be efficiently simulated for much longer durations than would be possible for an atomic system, albeit at the cost of some accuracy.

The coarse grained approach has been used extensively in the modelling of crystallisation processes[13,36,37] and many other events that require simulating large systems over large time scales such a protein aggregation. It is discussed in greater detail in Chapter 4.

2.2.5 Periodic Boundary Conditions

While real systems may contain billions of atoms (a single mole alone contains 6.022×10^{23} atoms), this presents an issue for current hardware which limits us to at best only simulate a few million atoms for mere nanoseconds. An intuitive approach would be to only examine a chunk of the bulk system thus simulating fewer atoms. An isolated block of atoms would be entirely surrounding by vacuum however, and thus would be exposed to large surface effects - more than likely the block would instantly vapourise.

Periodic boundary conditions offer a solution to this. A small fraction of atoms from the bulk system are isolated in a volume known as the primary cell. Conceptually, this cell is then surrounded by replica images of itself (Figure 2.3). In doing this, atoms in the primary cell are then able to interact with the infinite array of replicas, so that the primary cell is now surrounded again by the bulk phase, rather than vacuum. In practice, this is achieved by simply translating any atom or molecule that leaves the simulation box so that it re-enters on the opposite side. Further when calculating any distances between atoms, the minimum image convention is applied - the smallest possible distance between an atom in the primary cell and one of its neighbours is used. Care must be taken that the size of the primary cell is not chosen to be too small, or else the system will essentially become periodic and almost crystalline in nature.

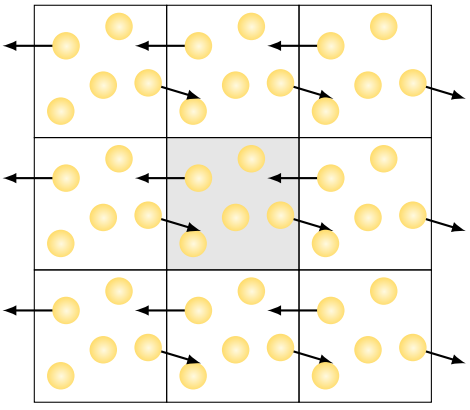


Figure 2.3: An example primary cell (grey) surrounded by eight of its images (white).

As the number of atoms in the system increases, it is typical to also employ neighbour lists to further increase the efficiency of the simulation[38].

2.2.6 Long-Range Electrostatic Interactions

The Coulomb potential is used to the compute the electrostatic interactions between pairs of atoms

$$U_{Coulomb} = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{4\pi\epsilon_0}\frac{q_i q_j}{|\boldsymbol{r}_{ij}|} \tag{2.37}$$

where $\boldsymbol{r}_{ij}$ is the separation between atoms i and j, q_i and q_j are their respective charges and ϵ_0 is the permittivity of free space. Problematically, the Coulomb potential decays slowly as a function of $|\boldsymbol{r}_{ij}|$, acting over ranges much larger than the typical size of a simulation box. Simply truncating the potential can result in large artefacts, especially for systems containing ionic species[38].

Although the summation can be rewritten as one that is between atoms in the primary box and the periodic images surrounding it

$$U_{Coulomb} = \frac{1}{2}\sum_{\boldsymbol{n}}{}'\left[\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{4\pi\epsilon_0}\frac{q_i q_j}{|\boldsymbol{r}_{ij}+\boldsymbol{n}|}\right] \tag{2.38}$$

this summation is only conditionally convergent. Here the sum over $\boldsymbol{n}$ is over lattice vectors between the primary and image cells, and the prime on the summation indicates that the interaction between atoms i and j is discarded when $\boldsymbol{n} = \boldsymbol{0}$. The Ewald summation is employed to overcome this issue of convergence.

In the Ewald approach, each atom in the system is surrounded by a neutralising Gaussian charge distribution of opposite sign. The sum of the atomic charges with the opposing distribution converges rapidly as a function of $|\boldsymbol{r}_{ij}|$. To recover the 'true' atomic interactions, the effects of the neutralising distribution needs to be removed. This is achieved by introducing a second set of Gaussian distributions with the same charge as the atom they are centred on (see Figure 2.4).

Provided the width of the Gaussian distributions is large enough, the screened atomic charges will only interact with the other screened charges within the primary cell, and hence can be computed directly in real space. The second set of distributions, on the other hand, will be located on a periodic lattice surrounding and including the primary cell. As such, their interactions with the charges can be represented by an also rapidly converging Fourier series calculated in reciprocal space. A final correction must be

applied to account for the self interaction between the atoms and the compensating charge distributions.

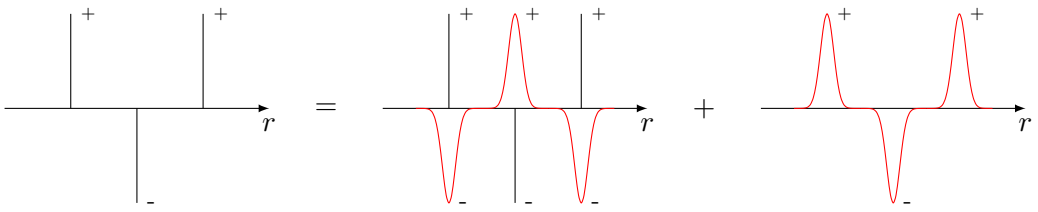


Figure 2.4: The atomic point charges are surrounded by a neutralising Gaussian distribution. This distribution is neutralised by an equal but opposite sum of distributions.

2.3 Phase Coexistence Methods

One of the great challenges that still plagues molecular simulation is calculating phase coexistence[11]. Nucleation of a new phase (whether the new phase is a liquid, a solid or a gas) is an entirely stochastic event that occurs rarely, as was discussed in Section 1.2. Even in real systems where the number of atoms is on the order of 10^{23}, a nucleation event may take seconds, minutes or even longer to be observed. The rarity of these events is only worsened in simulations. The volume element of the real system studied is very small and hence the number of atoms simulated is many orders of magnitudes smaller than experiment. Further, only microsecond timescales are accessible by simulation.

While phase transitions may be simulated by a brute force approach, they are generally inaccessible for all but the simplest of systems. The homogeneous freezing of ice, for example, took months of simulation time to be observed[39]. A more sophisticated approach is thus required. A number of these are described in the following subsections.

2.3.1 Direct Coexistence Approach

The direct coexistence approach overcomes the issue of simulating nucleation events by bypassing them completely. Simulations are run on a system containing the coexisting phases of interest within the same simulation box (Figure 2.5).

Over the course of the simulation the position of the interface between the two phases is monitored. If the interface remains stable, the two phases are coexisting. If not the simulation conditions need to be adjusted depending on the direction the interface moves until the right coexistence conditions are found (i.e, until the interface no longer moves).

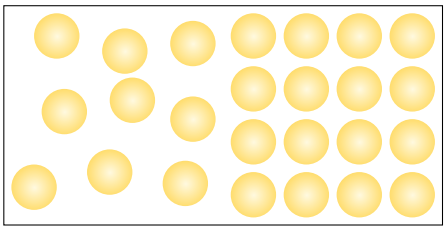


Figure 2.5: A 2-dimensional schematic of a liquid and solid phase coexisting in the same box.

While conceptually quite elegant, these types of simulation are far from ideal. Direct coexistence simulations can be very sensitive to the size and shape of the interface. Further, the time required to reach equilibration can be substantial (on the order of microseconds)[40].

2.3.2 Gibbs Ensemble Monte Carlo

Gibbs ensemble Monte Carlo (GEMC) simulations offer a possible solution to the interfacial issues experienced by direct coexistence simulations.

In a GEMC simulation, the different phases of interest are each simulated simultaneously in separate boxes (thus eliminating the interface between them). These boxes are coupled together by a series of Monte Carlo moves that aim to ensure that the temperature, pressure, and chemical potential in each box is identical, i.e that the phases in each box are coexisting.

In the simplest case, two boxes are constructed: one box with a volume V_{I} that contains N_{I} particles of phase I, and another of volume V_{II} that contains N_{II} particles of phase II. The total volume $V = V_{\mathrm{I}} + V_{\mathrm{II}}$, number of particles $N = N_{\mathrm{I}} + N_{\mathrm{II}}$ and temperature T remain fixed throughout the simulation. In this way, the multiple boxes being simulated can actually be thought of a single united system being kept in the NVT ensemble.

As would be expected of an NVT simulation, particles in each box are independently translated and rotated using the standard Monte Carlo moves. This ensures that each box is sampling the correct temperature distribution. Driving the pressure of each box to be equal is slightly more tricky. Periodically throughout the simulation, the volume of one of the boxes (say box I for example) is varied by some amount $V_{\mathrm{I}} = V_{\mathrm{I}} + \Delta V$. At the same time, V_{II} is changed by an equal but opposite amount $V_{\mathrm{II}} = V_{\mathrm{II}} - \Delta V$. In this way the total volume of the system is conserved, yet each box is simulated as if it were in the $N_{\mathrm{I}}pT$ / $N_{\mathrm{II}}pT$ ensemble. Here p is assumed to be equal in each box and is actually the

coexistence pressure. A similar approach is employed to ensure the equality of chemical potentials between the boxes. In this case however particle insertion / deletion moves are employed, rather than volume scaling ones. As a particle is deleted from one box, it is inserted into the other box and vice versa such that the total number of particles is again conserved.

The pressure and chemical potential (see the single-step Widom's method in Section 2.3.3.1) of each box are tracked over the course of the simulation to check when equilibration is achieved. Their values at equilibrium will be the coexistence pressure and chemical potential respectively.

While GEMC has been used extensively to study vapour-liquid phase coexistence (see also Chapter 4), it struggles to simulate solid phase coexistence. As is the case for all Monte Carlo simulations where insertion / deletion moves are involved, inserting particles into a dense system, such as a solid, will almost exclusively result in particle overlap and thus such moves would constantly be rejected. The inability to perform insertion moves means there is no way to couple the chemical potentials of each box, and hence, coexistence between boxes cannot be guaranteed. Alternate methods are thus required when considering dense phases.

2.3.3 Free Energy Methods

Free energy calculations offer a robust route for calculating phase coexistence. The free energy (or chemical potential in the case of multicomponent systems) of coexisting phases will be equal. Phase coexistence thus can be found by calculating the free energy of each phase as a function of some property (e.g. temperature, pressure or density), and determining where the free energy curves intersect.

Calculating absolute free energies by simulation however is challenging, as it requires the direct evaluation of a system's partition function. Instead, what is usually calculated is the free energy difference between the system of interest, and some reference state whose free energy is calculable analytically. The two main strategies to do this (although there are many variants of each) are thermodynamic integration and free energy perturbation.

2.3.3.1 Free Energy Perturbation

Consider the problem of determining the free energy difference between two states I and II

$$\Delta F\left(A \rightarrow B\right) = F_{\mathrm{II}} - F_{I} \tag{2.39}$$

that are each characterised by different Hamiltonians ($\mathcal{H}_{\mathrm{I}}$ and $\mathcal{H}_{\mathrm{II}}$ respectively), but sample largely the same configurations.

Substituting the canonical partition function into Equation 2.39, yields

$$\begin{aligned}
\beta\Delta F\left(N,V,T\right) &= -\ln\frac{Z_{\mathrm{II}}\left(N,V,T\right)}{Z_{\mathrm{I}}\left(N,V,T\right)} \\
&= -\ln\frac{\int\exp\left[-\beta\mathcal{H}_{\mathrm{II}}\left(\boldsymbol{r}^{N},\boldsymbol{p}^{N}\right)\right]\mathrm{d}\boldsymbol{r}^{N}\mathrm{d}\boldsymbol{p}^{N}}{\int\exp\left[-\beta\mathcal{H}_{\mathrm{I}}\left(\boldsymbol{r}^{N},\boldsymbol{p}^{N}\right)\right]\mathrm{d}\boldsymbol{r}^{N}\mathrm{d}\boldsymbol{p}^{N}} \\
&= -\ln\frac{\int\exp\left[-\beta\Delta\mathcal{H}\left(\boldsymbol{r}^{N},\boldsymbol{p}^{N}\right)\right]\exp\left[-\beta\mathcal{H}_{\mathrm{I}}\left(\boldsymbol{r}^{N},\boldsymbol{p}^{N}\right)\right]\mathrm{d}\boldsymbol{r}^{N}\mathrm{d}\boldsymbol{p}^{N}}{\int\exp\left[-\beta\mathcal{H}_{\mathrm{I}}\left(\boldsymbol{r}^{N},\boldsymbol{p}^{N}\right)\right]\mathrm{d}\boldsymbol{r}^{N}\mathrm{d}\boldsymbol{p}^{N}}
\end{aligned} \tag{2.40}$$

where $\Delta\mathcal{H} = \mathcal{H}_{\mathrm{II}} - \mathcal{H}_{\mathrm{I}}$. This final form is identical to taking an ensemble average (Equation 2.2) of $\exp\left[-\beta\Delta\mathcal{H}\left(\boldsymbol{r}^{N},\boldsymbol{p}^{N}\right)\right]$. Hence the free energy difference between states I and II becomes

$$\Delta F = -\frac{1}{\beta}\ln\left\langle\exp\left[-\beta\Delta\mathcal{H}\left(\boldsymbol{r}^{N},\boldsymbol{p}^{N}\right)\right]\right\rangle_{\mathrm{I}} \tag{2.41}$$

where the average is over an ensemble of configurations generated using the Hamiltonian of state I. Provided that the two states exist at the same temperature, and hence have an equal kinetic energy, the Zwangzig equation[41] is recovered

$$\Delta F = -\frac{1}{\beta}\ln\left\langle\exp\left[-\beta\Delta U\left(\boldsymbol{r}^{N},\boldsymbol{p}^{N}\right)\right]\right\rangle_{\mathrm{I}} \tag{2.42}$$

In practice, the average is calculated by running a simulation using the potential energy function of state I, but each time a new configuration is generated, the energy of the system is also calculated using the potential energy function of state II.

One of the most well known applications of free energy perturbation is the single-step Widoms insertion method. Consider a system of N fully interacting particles and a single ideal gas particle - i.e. a particle that does not interact with any of the other particles in the sytem. Compare this then to a system where the ideal particles interactions have now been turned on, so that the system now contains $N + 1$ interacting particles. The free energy difference between the two systems is actually the chemical potential.

The Widom method works in practice by simulating the system of N particles, and then periodically inserting a virtual particle that interacts but is not part of the system. The energy of interaction between this particle and the rest of the system is calculated, added to a running average, and then the particle is immediately removed again. The calculated average can then be inserted into Equation 2.42

As with all insertion schemes, the Widom method is only successful at calculating the chemical potentials in low density systems, such as the gas phase. Still it is a useful method, especially when combined with GEMC (section 2.3.2).

2.3.3.2 Thermodynamic Integration

Thermodynamic integration is another versatile method for calculating the free energy difference between states.

Unlike free energy perturbation, the states in thermodynamic integration do not have to share similar configurations. Instead, the state I is slowly transformed into state II via some reversible (but not always physical) pathway. The change in free energy is determined as a function of the progress along the path.

Let us start by more rigorously defining state I as being some state with a potential energy function U_{I}, and state II as having a potential energy function U_{II}. We also introduce a variable λ, a coupling parameter that measures the progress of the transition between states. At $\lambda = 0.0$ state I is recovered, and likewise at $\lambda = 1.0$ state II is recovered.

The change in free energy of the system as a function of λ can be easily derived directly from the partition function

$$\frac{\partial F}{\partial \lambda} = -\frac{1}{\beta}\frac{\partial}{d\lambda}\ln Z = -\frac{1}{\beta}\frac{1}{Z}\frac{\partial Z}{\partial \lambda} = -\frac{1}{\beta}\frac{1}{Z}\int \beta \frac{\partial U\left(\lambda, \boldsymbol{r}^N\right)}{\partial \lambda} \exp\left[-\beta U\left(\lambda, \boldsymbol{r}^N\right)\right] \mathrm{d}\boldsymbol{r}^N \quad (2.43)$$

where again the right hand side is simply just the ensemble average of $\partial U/\partial\lambda$. Here U is the potential energy function of the transitioning system. The change in free energy between states I and II can thus be obtained by integrating Equation 2.43

$$\Delta F = F_{\mathrm{II}} - F_{\mathrm{I}} = \int_0^1 \left\langle \frac{\partial U\left(\lambda, \boldsymbol{r}^N\right)}{\partial \lambda} \right\rangle \mathrm{d}\lambda \tag{2.44}$$

If the λ coupling is linear, i.e. $U\left(\lambda, \boldsymbol{r}^N\right) = \lambda U_{\mathrm{II}}\left(\boldsymbol{r}^N\right) + (1-\lambda)U_{\mathrm{I}}\left(\boldsymbol{r}^N\right)$, Equation 2.44 reduces to

$$\Delta F = F_{\mathrm{II}} - F_{\mathrm{I}} = \int_0^1 \left\langle U_{\mathrm{II}} - U_{\mathrm{I}} \right\rangle_\lambda \mathrm{d}\lambda \tag{2.45}$$

In a similar vein to the single-step Widom's method, the chemical potential of a system can be directly calculated from Equation 2.44. Here, state I would correspond to a system of N interacting and one non-interacting particles while state II would correspond to a system of $N+1$ interacting particles. The variable λ would act as a switching parameter. At $\lambda = 0$ the extra particles interactions would be fully switched off. As λ increases the interactions would be gradually turned on until $\lambda = 1$, at which point the extra particle would interact fully with the rest of the system.

Another widely employed application of thermodynamic integration is the Einstein crystal method[42], which calculates the free energies of solids. This method is discussed in more detail in Chapter 6.

2.3.4 Umbrella Sampling

Umbrella sampling is often employed to overcome large free energy barriers that hinder efficient sampling of phase space. A particularly good example of this is phase transitions, where the free energy barrier associated with nucleation (see Section 1.2) is large, and simulations often remain trapped in the metastable mother phase for large periods of time.

A bias is introduced that restrains the state of the system to various points along some reaction coordinate λ. The bias potential serves as an 'umbrella' bridging the end states. This may be the distance between two molecules, some torsion angle, or in the case of solid-liquid phase transitions, some measure of crystallinity such the Steinhardt Q6 order parameter[43].

Restraining the system along this path ensures that the full path between different states (including even unfavourable or metastable ones) is accurately sampled. The bias is typically introduced as a perturbation W to the system's potential energy, such that the biased probability of finding the system in some state is given by

$$\pi\left(\boldsymbol{r}^N\right) = \frac{\exp\left[-\beta\left(U\left(\boldsymbol{r}^N\right) + W\left(\lambda\right)\right)\right]}{\int \exp\left[-\beta\left(U\left(\boldsymbol{r}^N\right) + W\left(\lambda\right)\right)\right] \mathrm{d}\boldsymbol{r}^N} \tag{2.46}$$

Any observable calculated in this biased scheme would of course then be weighted in some way, as opposed to if it were calculated in an unbiased simulation. This weighting can be removed by applying the following

$$\left\langle A\left(\boldsymbol{r}^N\right)\right\rangle = \frac{\left\langle A\left(\boldsymbol{r}^N\right)\exp\left[\beta W\left(\lambda\right)\right]\right\rangle_W}{\left\langle \exp\left[\beta W\left(\lambda\right)\right]\right\rangle_W} \tag{2.47}$$

where the W subscript indicates the average is taken over configurations sampled according to Equation 2.46. A variant of the umbrella sampling approach is constraint molecular dynamics, where the system is constrained (rather than restrained as with umbrella sampling) at specific positions along the reaction coordinates.

Further to just enhancing sampling, Umbrella Sampling allows calculation of the free energy profile along the reaction coordinate of interest. A number of simulations are run, with each being restrained to different points along the reaction coordinate. For each simulation, a probability histogram is constructed, measuring the frequencies at which values of λ are visited. The histograms are unweighted and stitched together. The free energy profile is then given by

$$F\left(\lambda\right) = \frac{1}{\beta}\ln h\left(\lambda\right) - W\left(\lambda\right) + C \tag{2.48}$$

where C is an unknown additive constant that vanishes when computing free energy differences. Care must be taken to ensure that the histograms of neighbouring restrained simulations do indeed overlap by a large amount.

A major limitation of this method is the possibility to drive the system towards an unrealistic final configuration. Furthermore, if the collective variables are badly chosen, entire regions of the energy surface could be poorly sampled. The choice of collective variables are therefore crucial in ensuring a successful simulation[11].

2.3.5 Density of State Methods

Density of states simulations are closely related to, but are perhaps more robust than, umbrella sampling. They are discussed in detail and employed heavily in Chapters 4, 5 and 6.

Chapter 3

Why Do Some Molecules Form Hydrates or Solvates?

Abstract: *The discovery of solvates (crystal structures where the solvent is incorporated into the lattice) dates back to the dawn of chemistry. The phenomenon is ubiquitous, with important applications ranging from the development of pharmaceuticals to the potential capture of CO_2 from the atmosphere. Despite this interest, we still do not fully understand why some molecules form solvates. We have employed molecular simulations using simple models of solute and solvent molecules whose interaction parameters could be modulated at will to access a universe of molecules that do and do not form solvates. We investigated the phase behaviour of these model solute-solvent systems as a function of solute-solvent affinity, molecule size ratio, and solute concentration. The simulations demonstrate that the primary criterion for solvate formation is that the solute-solvent affinity must be sufficient to overwhelm the solute-solute and solvent-solvent affinities. Strong solute-solvent affinity in itself is not a sufficient condition for solvate formation: in the absence of such strong affinity, a solvate may still form provided that the self-affinities of the solute and the solvent are weaker in relative terms. We show that even solvent-phobic molecules can be induced to form solvates by virtue of a $p\Delta V$ potential arising either from a more efficient packing or because high pressure overcomes the energy penalty.*†*

*The manuscript presented in this chapter previously appeared in *Cryst. Growth Des.*[44], and is listed as Paper I in the list of publications.

†All spellings in this manuscript have been changed from the US (as were originally published) to the UK versions, so as to be consistent with the rest of the thesis.

3.1 Introduction

When a solute crystallizes from solution, it may do so either as a pure crystal or as a solvate, in which solvent molecules are incorporated into the lattice. When the incorporated solvent is water, the solvate crystals are termed hydrates. Solvate formation, in particular hydrate formation, is a common phenomenon.[45,46] About a third of all organic molecules are able to form hydrates and solvates,[47–49] an example exhibiting extreme promiscuity being the antibacterial sulfathiazole, for which over 100 solvates have been characterized.[50] Solvates can exhibit markedly different physicochemical properties relative to the corresponding anhydrous forms, including melting point, solubility, crystal habit, and mechanical properties. In the pharmaceutical industry, the choice of whether the form of the active substance is a solvate or anhydrous can affect its bioavailability and the ease (or otherwise) of manufacturing the product as well as its stability.[51] Hydrate formation is also an issue in the petroleum industry, where it can cause blockage of gas pipelines.[52] There are also other hugely beneficial potential applications ranging from hydrogen and natural gas storage to atmospheric carbon dioxide capture.[4–7]

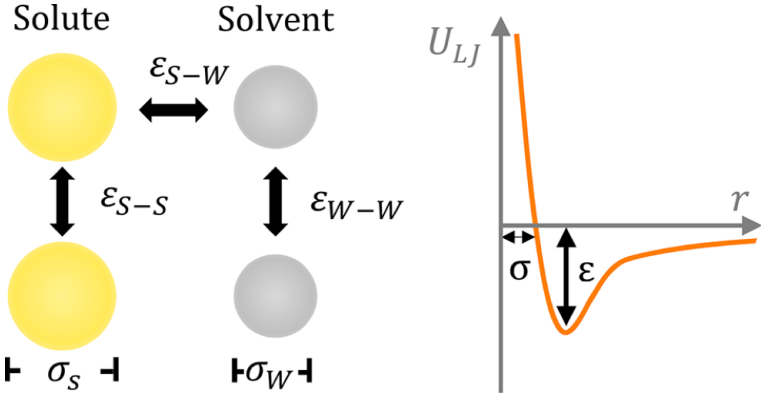
Despite this extensive interest, the fundamental question of why some molecules form solvates remains open. The thermodynamic perspective is that the solvated forms of these molecules have a lower free energy, but this is not insightful and begs the question of why they have a lower free energy. The thermodynamics approach is exemplified by studies comparing the potential energies (as approximations for free energies) of the various forms with a view to rationalizing why a particular molecule forms a hydrate while a related one does not.[53–55] While these methods offer some predictive capability, they inform us only about the system of interest rather than revealing broader insights. An alternative approach that addresses the posed question somewhat better has attempted to link molecular features to the propensity for hydrate formation. A series of surveys of the Cambridge Structural Database (CSD) revealed a strong correlation with the polar surface area and degree of branching within a molecule and with an increased number of polar functional groups (e.g., carbonyl (C=O), ether (C-O-C), hydroxyl (O-H), and primary amine (N-H)),[56–58] while no correlation was found with the ratio of hydrogen-bond donors to hydrogen-bond acceptors as previously suggested.[59,60] This suggests that a strong affinity for the solvent may be important, and yet there are many examples of substances with high solubility (i.e., those having a strong interaction with the solvent) that do not form solvates. Furthermore, how does one rationalize hydrates of hydrophobic molecules (e.g., gas hydrates)?[52]

At the heart of the question of why a particular molecule forms a solvate are the molecular interactions, specifically the interplay between the solute-solvent, solute-solute, and solvent-solvent interactions. Coupled to these interactions is the nature of the packing of the molecules in the potential anhydrous and solvate forms. Ideally, we need to explore and understand how the phase diagram of a solute/solvent system varies as a function of the strength of solute-solvent, solute-solute, and solvent-solvent interactions and molecular packing. How might one achieve this? A cursory review of the problem suggests that this is not feasible. To study the effect of variation of the intermolecular interactions on the phase behaviour requires the consideration of a series of solute and solvent molecules with a variety of molecular structures. The elucidation of the phase diagram for each of these solute-solvent pairs would be a major task in itself, independent of whether it is based on experiment or modelling. In addition to this, there is the difficulty of deconvoluting the effects of molecular packing from the strengths of the intermolecular interactions.

Here we access the phase behaviour of a universe of molecules that do and do not form solvates by means of molecular simulations using simple coarse-grained models of molecules. These simple models strip away the molecular complexity that otherwise obscures the core issue while enabling modulation of the intermolecular interactions by design. Thus, we investigate the crystallization behaviour of a series of solute-solvent systems as a function of the affinity and molecule size ratio (packing) between the solute and solvent. We show that solvate formation is promoted when the solute-solvent affinity overwhelms the solute and solvent self-affinities but that a strong solute-solvent affinity is not a sufficient condition in itself. Solvate formation can also occur for solutes with weak solvent affinity by virtue of the pΔV component of the Gibbs potential arising either from more efficient packing or because high applied pressure overcomes the energy penalty.

The phase behaviour of the solute-solvent systems was explored using molecular dynamics (MD) simulations. The solute and solvent molecules were represented by simple single-particle models based on Lennard-Jones (LJ) interactions. Such models are appropriate because solvate formation is a generic phenomenon, being observed in a wide class of materials. These models have been successfully employed by us earlier to probe crystal nucleation problems, including the identification of design rules for nucleation inhibitors[36,37] and for uncovering molecular processes in secondary nucleation.[13] The LJ model is characterized by two parameters (Figure 3.1): σ, the distance at which the interaction potential is zero, which serves as the effective molecule size, and ε, the potential energy well depth, which characterizes the affinity between the molecules. Our

choices of LJ parameters for the models were not arbitrary but based on the LJ phase diagram, which is known.[61] Thus, the chosen solvent parameters, σ_W = 0.47 nm and ε_{W-W} = 3.28 kJ mol^{-1}, define a liquid (the solvent) with a melting point of 273 K. The solute-phase packing parameters were in the range σ_S = 0.47-1.47 nm, while the solute self-affinity was fixed at ε_{S-S} = 5.00 kJ mol^{-1}. This chosen solute self-affinity for σ_S = 0.47 nm defines a solid with a melting point of approximately 421 K (about that of a typical organic solid).



$$U_{LJ} = 4\varepsilon\left(\frac{\sigma^{12}}{r^{12}} - \frac{\sigma^{6}}{r^{6}}\right)$$

Figure 3.1: Interactions between solute and solvent molecules (left) are characterized by the ε and σ parameters of the Lennard-Jones potential, shown plotted as a function of the separation distance r (right).

It should be noted that the large values of the affinity parameter employed here, up to ε = 8.0 kJ mol^{-1}, are well beyond the typical values characterizing van der Waals interactions. For comparison, the oxygen-oxygen van der Waals interaction for the TIP3P water model is characterized by ε = 0.6364 kJ mol^{-1}.[62] The implication is that the LJ model employed in the study serves as a molecular potential that encapsulates both the weak van der Waals and the stronger Coulombic interactions, albeit not strong formal charges. The LJ model as utilized here is used in the widely employed coarse-grained MARTINI force field[63] to represent molecular moieties containing up to four non-hydrogen atoms, e.g., -CH_2COOH, including water.

We investigated the crystallization behaviour of the solute for a universe of solute-solvent systems. The solute-solvent affinity was varied to encompass a range of systems: ε_{S-W} = 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, and 6.0 kJ mol^{-1}, where the larger values characterize systems with stronger affinities between the solute and the solvent. For each solute-solvent pair, we explored the crystallization behaviour of the solute from a series of

solutions with a range of solute concentrations: x_{solute} = 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100 mol %. The system size in all cases was 10 000 particles. The primary question for analysis was the following: which product crystallized out, the anhydrous form or the solvate?

3.2 Results and discussion

The first set of simulations explored the crystallization behaviour of solutes for a universe of solute-solvent systems with equal particle sizes ($\sigma_S = \sigma_W$ = 0.47 nm). The dependence of the crystallization product on the solute-solvent affinity is shown in the phase diagram in Figure 3.2. A weak solute-solvent affinity ε_{S-W} implies a low solubility. Consequently, at weak solute-solvent affinities, the solution becomes supersaturated at low concentrations, limiting the solution region (lower left region of the plot in Figure 3.2). At this weak solute-solvent affinity, the resulting product is the anhydrous structure. As the solute-solvent affinity increases (going up the y axis in Figure 3.2), the solubility increases, and the solution region becomes broader.

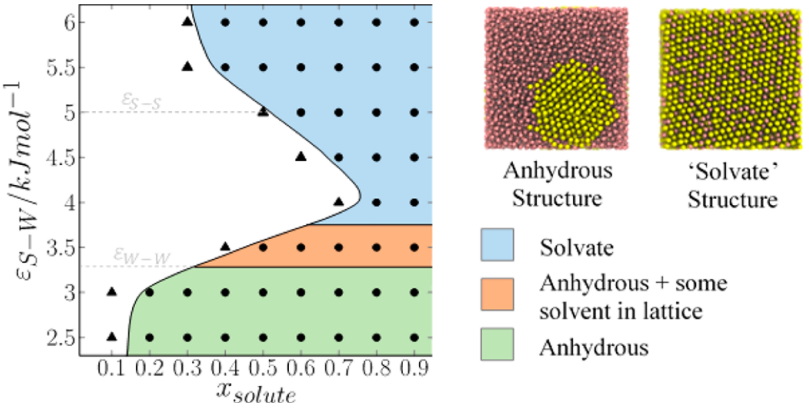


Figure 3.2: Phase diagram for equal-particle-size solute-solvent systems ($\sigma_W = \sigma_S$ = 0.47 nm) as a function of solute-solvent affinity ε_{S-W} and solute concentration x_{solute} at 283 K. The phase diagram exhibits four distinct regions: solution (white), solvate (blue), anhydrous (green), and anhydrous with some solvent inclusion (orange). Each data point on the plot represents a simulation result. Circles mark crystallization events (structures shown on the right), while triangles signify that the system remained a homogeneous solution. We note that the solvate structure is a lattice but is disordered with respect to occupation of the lattice sites. This is expected since close packing of two distinct but equal-sized particles cannot yield an interpenetrating lattice like that observed for NaCl.

At stronger solute-solvent affinities (ε_{S-W} > 3.28 kJ mol^{-1}), the solute-solvent affinity surpasses the solvent's affinity for itself, and each solute (solvent) particle shows a greater preference to have a solvent (solute) particle as a neighbour. At an affinity of ε_{S-W}

= 4.0 kJ mol^{-1} and above, the solute and solvent become fully integrated to yield a solvate lattice. At still stronger solute-solvent affinities, the solute (solvent) particles attract and order the solvent (solute) particles around themselves to such an extent that crystallization of the solvate is induced even at low concentrations. Consequently, the solution region in the phase diagram becomes more limited, with the saturation line tending toward lower concentrations (top left region of plot in Figure 3.2). These results suggest that the determining factor for solvate formation is the strength of the solute-solvent interactions relative to the solute-solute and solvent-solvent interactions.

In the above simulations, the solute and solvent particles were of equal size. We then considered the effects of packing, wherein we increased the solute particle size from σ_S = 0.47 nm first to σ_S = 1.18 nm and then to σ_S = 1.47 nm while keeping the solvent size fixed at σ_W = 0.47 nm. In both cases $\sigma_{S-W} = \frac{1}{2}(\sigma_S + \sigma_W)$. For the first of these systems, the particle sizes (σ_S = 1.18 nm and σ_W = 0.47 nm; solvent/solute radius ratio $\sigma_W/\sigma_S = 0.40$) were chosen to yield NaCl-type packing,[64] and indeed, this was the observed structure. In the second case, the solute molecules are substantially larger than those of the solvent (σ_S = 1.47 nm and σ_W = 0.47 nm; $\sigma_W/\sigma_S = 0.32$). These two systems show similar behaviour (Figure 3.3) that in broad terms is not too different from the behaviour of the equal-sized molecules. Strong solute-solvent affinities (compare the top left in Figures 3.2 and 3.3) yield the solvate phase while weaker solute-solvent affinities yield the anhydrous form. The second case, however, also shows an apparently unintuitive result that the solvate form is favoured even at the weakest solute-solvent affinities (the two points at $x_{solute} = 0.9$ and ε_{S-W} = 0.5 and 1.0 kJ mol^{-1}). We are unable to give a rigorous explanation for these results, despite carrying out repeat and additional simulations. At these data points the systems are 90% solute and 10% solvent, and as the solvent particle size is relatively very small, the solvent volume is miniscule. The entropy of solvent dispersion is probably more favourable, resulting in a solvate rather than the formation of a separate, small, subcritical condensed-phase cluster. The emergent solvates reveal a face-centered lattice for the solute molecules, with the solvent molecules either forming an interpenetrating face-centered lattice (the NaCl structure for $\sigma_W/\sigma_S = 0.40$) or filling the interstitial channels (for $\sigma_W/\sigma_S = 0.32$) (Figure 3.3). The latter structures are very similar to the class of nonstoichiometric channel solvates,[20,21,47,65] where the solvent molecules occupy channels formed within the solute lattice and can freely diffuse out depending on the relative vapour pressure of the solvent (relative humidity for a hydrate) in the environment. Indeed, the solvent particles in these simulated channel solvates exhibit significant diffusion (diffusion coefficient ~(3.5-7.5)×10^{-9} m^2 s^{-1}).

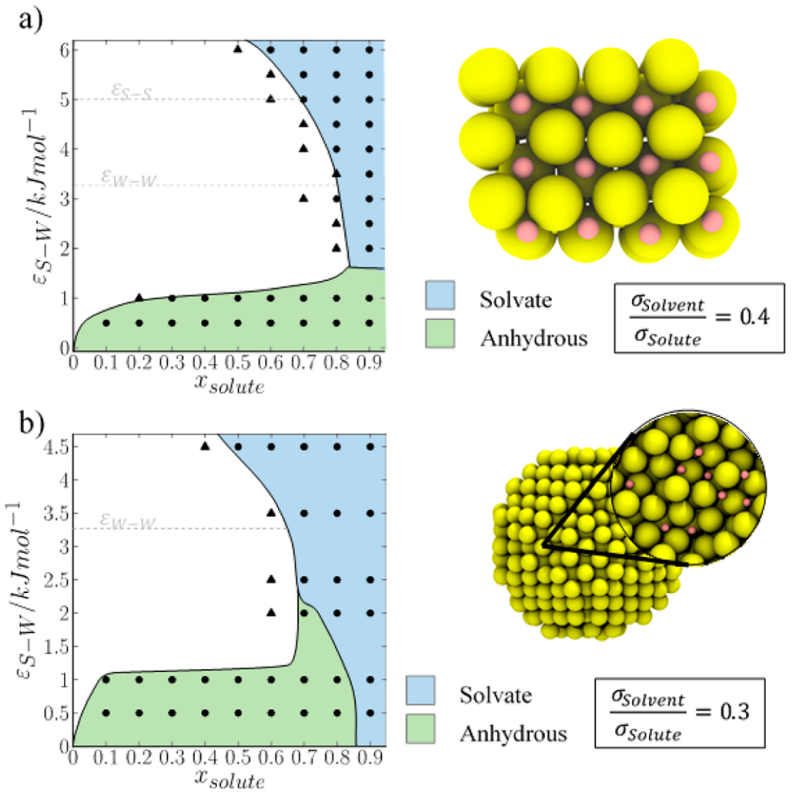


Figure 3.3: Phase diagrams for (a) NaCl-type and (b) channel-packing type solute-solvent systems as functions of solute-solvent affinity and solute concentration. The blue regions indicate solvate formation and the green regions the anhydrous form. Each data point on the graph represents a single simulation. Circles mark crystallization events (structure shown on the right), while triangles signify that the system remained a homogeneous solution.

For the system yielding the interstitial channels, we also looked closely at the extreme case of a solute with a very weak affinity for the solvent (ε_{S-W} = 0.3 kJ mol^{-1}), i.e., a solvent-phobic solute (see Figure 3.4). For this system, the solute-solute affinity was increased to ε_{S-S} = 8.0 kJ mol^{-1}, and we investigated the system at the low solute concentration of 1 mol %. (It should be noted that this system is quite different from the systems yielding the unintuitive data points at the bottom right in Figure 3.3b, as the solvent is in significant excess, rather than the solute). The strong solute-solute affinity and low molar concentration favoured the formation of a small solute crystallite in the bulk solvent, making it easier to observe whether the solvent was either included or excluded from the emergent structure. This system showed phase separation at (ambient) pressure p = 0.001 katm but yielded a solvate structure at a higher pressure of p = 10 katm. Thus, it is clear that even solvent-phobic solutes can form solvates when driven by the pΔV component of the Gibbs potential G. Indeed, the use of pressure to force the formation of hydrates experimentally has been noted earlier.[52,66,67]

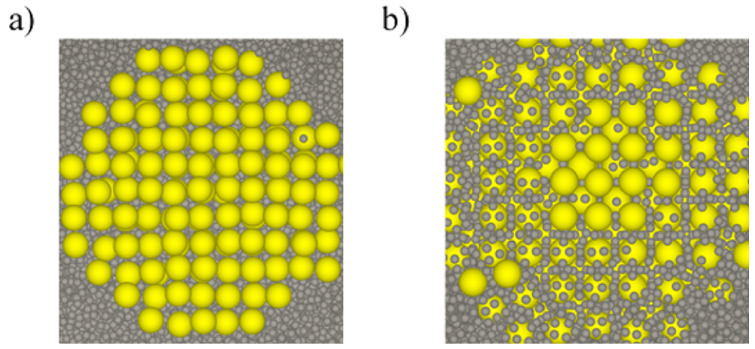


Figure 3.4: Slices taken from the final structures of the solvent-phobic (ε_{S-W} = 0.3 kJ mol^{-1}; σ_W / σ_S = 0.32) system. (a) At ambient pressure (p = 0.001 katm) the solvent was observed to be excluded from the solute structure, thus favoring the anhydrous form. (b) Increasing the pressure (p = 10 katm) caused the solvent to fill the interstitial channels between solute particles, similar to the behaviour observed in channel solvates.

The above results appear to show that a solvate is always formed when the solute-solvent affinity is strong but can also form when such affinity is lacking. Can we qualify this condition further? A limited number of additional simulations were carried out for the equal-particle-size system in which the solute-solute affinity was increased incrementally from the initially set value of ε_{S-S} = 5.0 to ε_{S-S} = 8.0 kJ mol^{-1} while the solute-solvent affinity was kept fixed at ε_{S-W} = 4.0 kJ mol^{-1}. This would be equivalent to a solute with a higher melting point but the same interaction with the solvent. One might expect that a such a system, given the strong (existing) solute-solvent affinity, would yield a solvate, reproducing the data points for ε_{S-W} = 4.0 kJ mol^{-1} in Figure 3.2. It did not. Instead, we observed that the anhydrous structure was the stable form. The inference is that a strong solute-solvent affinity in itself is not a sufficient condition for solvate formation. Rather, the solute-solvent affinity must be sufficient to overwhelm the solute and solvent self-affinities. These systems with strong solute-solute affinities (ε_{S-S} = 5.0-8.0 kJ mol^{-1}) tended to become kinetically trapped, and we had to resort to calculations of potential energy differences (as approximations for free energy differences) between the anhydrous and solvated forms to assess the stability.

The thermodynamic criterion for solvate formation (see Figure 3.5) is $\Delta G_{c,S\cdot nW} < (\Delta G_{v,S} + n\Delta G_{v,W})$, where $\Delta_{Gv,S}$ and $\Delta G_{v,W}$ are the molar free energy changes for vapourization of the solute crystal and the solvent fluid, respectively, $\Delta G_{c,S\cdot nW}$ is the molar free energy change associated with crystallization of the solvate from the vapour phase, and the integer n is the number of moles of solvent per mole of solute, as reflected in the stoichiometry for the reaction of the solute plus the solvent to form the solvate: $S + nW \rightarrow S \cdot nW$. For a 0 K (potential energy) approximation, the solvate formation criterion becomes $\Delta U_{c,S\cdot nW} < (\Delta U_{v,S} + n\Delta U_{v,W})$ where $\Delta U_{c,S\cdot nW}$ is the lattice energy

of the solvate form $S{\cdot}nW$, $\Delta U_{v,S}$ is the lattice energy of the anhydrous form, and $\Delta U_{v,W}$ is the lattice energy of the solvent crystal (as the solvent would be a solid at 0 K).

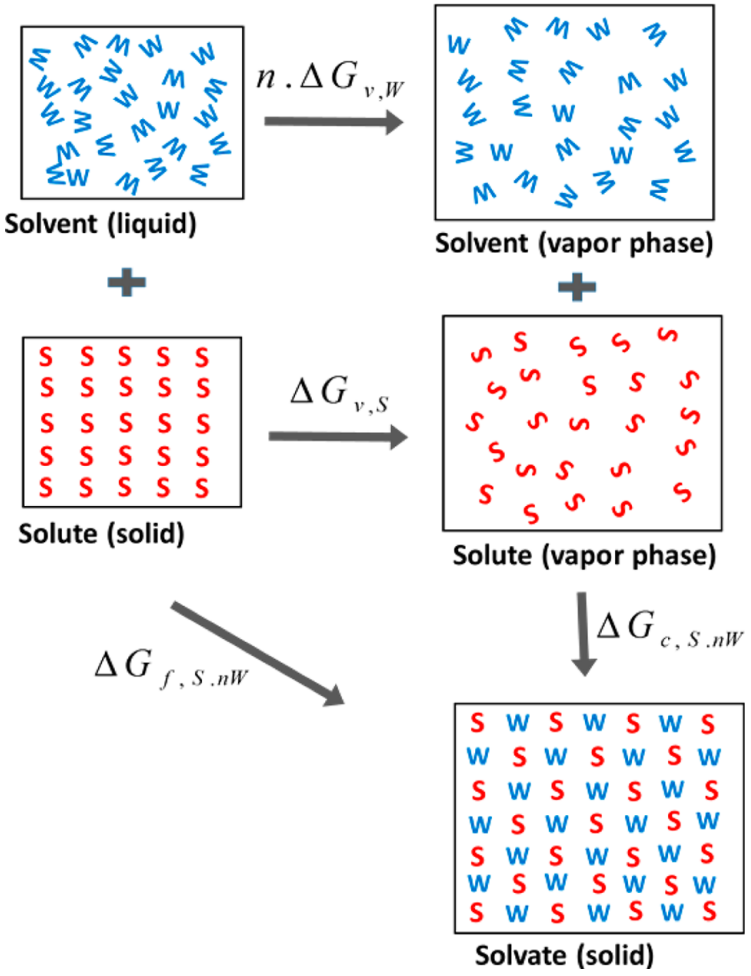


Figure 3.5: Thermodynamic cycle for the formation of a solvate from its components, the solute and solvent. $\Delta G_{f,S{\cdot}nW}$ is the molar free energy change for solvate formation, and $\Delta G_{v,S}$ and $\Delta G_{v,W}$ are the molar free energy changes for vapourization of the solute crystal and the solvent fluid, respectively. $\Delta G_{c,S{\cdot}nW}$ is the molar free energy change associated with crystallization of the solvate from the vapour phase, and the integer n reflects the stoichiometry S + nW $\rightarrow$ S $\cdot$ nW.

Within the spectrum of molecular interactions and packing ratios characterizing solvate formation, one can identify two limiting cases (Figure 3.6): (a) when there is strong solute-solvent affinity and (b) when the packing of the solute molecules is essentially independent of the solvent. Expressing the 0 K stability criterion, $U_{solvate} < (U_{solute} + U_{solvent})$, in terms of component atom-atom interactions yields $\left[\sum U_{S-S}(solvate) + \sum U_{W-W}(solvate) + \sum U_{S-W}(solvate)\right] < \left[\sum U_{S-S}(solute) + \sum U_{W-W}(solvent)\right]$. For the equal-molecule-size system with strong solute-solvent affinity, case (a), the dominating interactions within the solvate are those between the solute

and the solvent, as each solute (solvent) molecule is surrounded by solvent (solute) particles. The solute-solute and solvent-solvent interactions in the solvate are marginal. Consequently, for this case (to a first approximation), the stability criterion reduces to $\sum U_{S-W}(solvate) < \left[\sum U_{S-S}(solute) + \sum U_{W-W}(solvent)\right]$. For such a system, we can map the Lennard-Jones affinities onto the stability criterion by considering interactions between particles as pseudobonds.

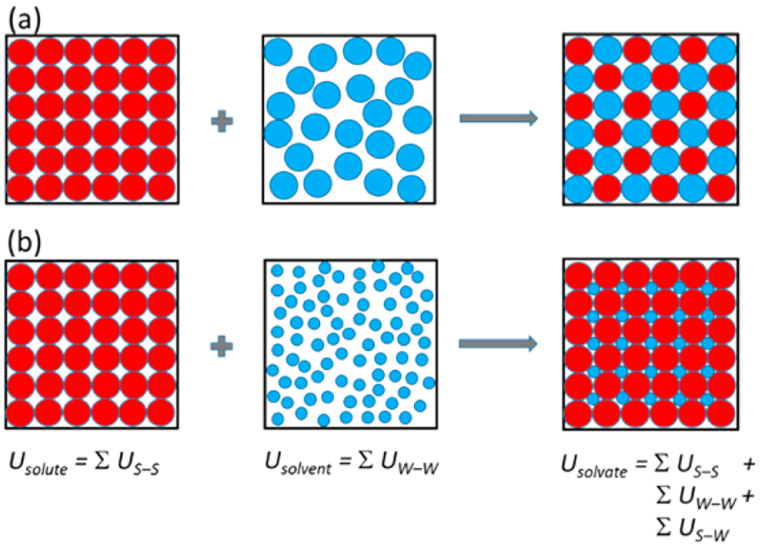


Figure 3.6: Two limiting cases of solvate formation, represented schematically: (a) equal-molecule-size system with strong solute-solvent affinity; (b) solvate formation where the solute packing is essentially the same in the anhydrous and solvate forms and independent of the solvent.

As a first-order approximation, we restrict the interactions to the first coordination sphere. For the solute in a face-centered-cubic lattice, there are 12 "bonds", and we approximate the strength of each by ε_{S-S}. Likewise, there are about 12 "bonds" for the liquid, for each of which we assume the strength ε_{W-W} (although the actual interaction is a little weaker since the particle separation distance is slightly greater in the liquid state). To form a solvate, the 12 solute-solute and 12 solvent-solvent "bonds" must be broken and replaced with 24 new solute-solvent "bonds", each with an approximate strength of ε_{S-W}. The approximate stability criterion for the Lennard-Jones system then becomes $24\varepsilon_{S-W} > 12\varepsilon_{S-S} + 12\varepsilon_{W-W}$, that is, $2\varepsilon_{S-W} > \varepsilon_{S-S} + \varepsilon_{W-W}$ (here the inequality operator has been switched from less than to greater than since ε is not the interaction energy but the energy well depth parameter). Substituting the self-affinity parameters utilized for the solute and solvent ($\varepsilon_{W-W} = 3.28$ kJ mol^{-1} and $\varepsilon_{S-S} = 5.00$ kJ mol^{-1}), the criterion indicates solvate stability above the solute-solvent affinity $\varepsilon_{S-W} = 4.1$ kJ mol^{-1}. This is entirely consistent with the switchover point for solvate formation observed in Figure 3.2, namely, about 4 kJ mol^{-1}.

For the limiting case (b) where the solute structure of the anhydrous form is essentially identical to that in the solvate (as in a nonstoichiometric channel solvate, e.g., the system shown in Figure 3.3b), $\sum U_{S-S}(solute) \approx \sum U_{S-S}(solvate)$, and the solvent-solvent interaction in the solvate is marginal, i.e., $\sum U_{W-W}(solvate) \to 0$. In this case, the stability criterion reduces to $\sum U_{S-W}(solvate) < \sum U_{W-W}(solvent)$, that is, the solute-solvent interaction must be stronger than the solvent-solvent interaction. This is intuitive, being akin to the interplay between the cohesive force of a fluid and the adhesive force that determines whether, for example, water will wet a nanopore (hydrophilic surface) or bridge it (hydrophobic surface exploited in high-tech wetwear that is waterproof and yet breathable). This issue is manifested by the system with weak solute-solvent affinity, where the solute is essentially solvent-phobic (Figure 3.4). At low pressures, the system phase-separates into the anhydrous form and the solvent. At the higher pressure of p = 10 katm, the pΔV component of the Gibbs potential overwhelms the solvent-solvent affinity, forcing the solvent into the lattice to form a solvate.

In conclusion, we have shown that the primary criterion for solvate formation is that the solute-solvent affinity must be sufficient to overwhelm the solute-solute and solvent-solvent affinities. A strong solute-solvent affinity in itself is not a sufficient condition. Solute molecules even with a weak affinity for a solvent can form solvates provided that the self-affinities of the solute and the solvent are weaker in relative terms. Indeed, as demonstrated, essentially solvent-phobic molecules can form solvates when driven by the pΔV term, i.e., under high pressure. In going forward, it would be insightful to carry out atomistic lattice or free energy calculations on solvate systems (using, e.g., Cambridge Crystallographic Data Centre data and tools), partitioning the energy into molecule-molecule (solute-solute, solute-solvent, and solvent-solvent) interactions to see how the insights ascertained here play out in realistic systems. Finally, we note that while the focus of this paper is solvate formation, the inferences are also applicable to cocrystal formation for binary systems,[22] where the second molecule in the lattice is not the solvent but another solute (solid-phase) molecule.

3.3 Methodology

Molecular dynamics simulations were carried out using the DL-POLY 4.06 software package[68] in the NPT ensemble using a Nosé-Hoover thermostat and barostat. All of the simulations were run at 283 K and a pressure of 1 atm unless otherwise indicated. The interactions (van der Waals) were truncated at $2.5\sigma_S$, and the standard long-range

corrections applied. All of the simulations were run for a minimum of 5 million steps using a 30 fs time step. The mass was set to 72 g mol^{-1} for all of the particles. The system size was 10 000 particles. Initial configurations comprised randomized coordinates.

Chapter 4

Towards Realistic and Transferable Coarse-Grained Models: Phase Diagrams of Soft van der Waals Potentials

Abstract: *Coarse-grained molecular simulations offer a robust route to simulating systems that would otherwise be too large, or require too long to simulate by a fully atomistic approach. Despite their numerous applications, the most commonly employed coarse-grained force fields utilise the Lennard—Jones (LJ) potential, which has proven to be too 'hard' to accurately reproduce molecular, rather than atomistic interactions. This inherent 'hardness' is identified as the source of the limitingly narrow temperature range over which models based on the LJ potential remain liquid. Here we characterise a set of 'softer', more representative potentials (the 9-6, 8-4 and 6-4 n-m potentials) by mapping their full phase diagrams. The mapped phase diagrams exhibit a broader liquid range than the more established LJ potential, thus enabling models based on these potentials to be employed in studies of most phases, and over a much wider range of conditions than would be previously accessible. Further, knowledge of these diagrams will enable the direct parameterisation of a set of transferable coarse-grained beads with a fundamentally physical grounding by employing the 'PhaseD' approach. This in turn will enable the construction of a more accurate, high class of coarse-grained force field.*[*]

[*]The manuscript presented in this chapter is listed as Paper II in the list of publications.

4.1 Introduction

Molecular simulations are a vital tool in exploring and explaining chemical phenomena at a molecular level, with widespread applications ranging from studying protein folding, membrane formation[69–72] and even crystallisation events[13,36,37]. The primary limitations are the accuracy of the force field parameters that characterise the molecular interactions, and the limited time- and length-scales that can be sampled. Currently, using high performance computing facilities, the accessible length scale is tens of nanometres (corresponding to an order of a 1 million particles) for simulation times of up to a few microseconds. The implication is that large systems (e.g. large biomolecular assemblies) and many phenomena (e.g. protein folding, phase transitions) remain inaccessible. There are two major approaches for dealing with length- and time-scale issues. For time-scale limited problems, one can resort to thermodynamic approaches, focusing on free energies calculations (free energy differences, chemical potentials, and free energy as a function of a reaction coordinate). For large systems, one can investigate the problem at a lower resolution – a coarse-grained perspective.

The coarse-grained approach treats groups of atoms, and potentially whole groups of molecules, as a single volume element. This element may be a spherical particle, an ellipsoid or some other variation.[73] This approach significantly reduces the number of particles that need to be simulated, thus enabling larger systems to be simulated at the cost of compromising atomistic resolution. There are also additional benefits. The interaction potential is softer enabling a larger timestep to be taken, from 0.002 ps to about 0.040 ps – a twenty fold advantage. Further, the coarse-graining softens the free energy surface, which enables faster equilibration of the system.

There are two distinct philosophical approaches to developing a coarse-grained representation of an atomistic system. In the chemical approach, the coarse-grained model is the best accurate representation of the atomistic model, encapsulating the full chemical specificity. Such parameterisation is typically carried out using Boltzmann inversion[74] or force-matching. In the physics-type approach, the philosophy is to develop the simplest generic model that encapsulates the essential physics of the chemical behaviour of interest. Coarse-grained models that encapsulate the full chemical specificity are tedious to develop, and are by design specific and hence, not transferable. Further, such models are only parameterised for conditions (e.g. temperature, pressure) at which they were derived[75]. In between the generic physics-type models and the coarse-grained, chemically-specific models are the transferable off-the-shelf coarse-grained models with

applicability to a large range of molecules. It is understood that the transferability implies loss of accuracy. Examples of such transferable coarse-grained models include MARTINI[63] and SDK[76,77]

Most transferable coarse grained force fields, including the popular MARTINI force field, use the Lennard-Jones (LJ) potential[71] to represent non-bonded interactions between the CG particles. MARTINI indeed lumps all electrostatics resulting from partial charges on atomistic sites including hydrogen bonding into the coarse-grained LJ parameters. The LJ energy parameter ε, for example, takes values of ε = 2.00-6.00 kJ mol^{-1} which is markedly higher than the typical value characterising an atomistic van der Waals interaction of ε =0.07-0.70 kJ mol^{-1}. The primary issue is that the LJ potential is too hard and does not represent well the softer interaction that characterises the non-bonded interaction between molecular moieties. This reveals itself in unphysical behaviour in such models, such as the over structuring of the fluid phase and fluid phases having a limited liquid-phase range[78]. The MARTINI water model, for example, freezes at ambient conditions, which must be circumvented by the inclusion of anti-freeze particles.

Chemically-specific coarse-grained models reveal that the non-bonded interaction between the coarse-grained particles are best described by softer *n-m* (Mie) potentials, relative to the 12-6 form of the Lennard Jones potential. Shelley et al, for instance, identified that for CG lipid models of dimyristoylphosphatidylcholine, the 9-6 potential form was the best description for the various CG lipid moieties, whilst the water model was described by a 6-4 potential[79].

Recently, we proposed a new approach to parameterising non-bonded interactions for off-the-shelf, transferable CG models or force fields, based on the phase diagram of the selected model potential[80]. The approach enables the design of CG particles whose melting points match that of the target molecule or moiety group. Specifically, values of ε and σ are directly identified from the phase diagram to give a CG particle with a particular melting point that corresponds to the melting point of the chemical moiety being represented. We term this the *PhaseD* approach, emphasizing the link with the phase (coexistence) diagram. This approach gives a good physical foundation for the CG particles, unlike ad-hoc but self-consistent parameterisations which can lead to non-realistic or unphysical behaviour. The procedure requires a knowledge of the phase diagram of the potential and involves fixing the mapping and the associated potential size parameter σ (which for example for MARTINI is σ = 0.47 nm for 4 atoms to 1 CG particle mapping), and then identifying the energy parameter ε that corresponds to the melting point T_{mp} from the phase diagram.

To exploit the more realistic softer *n*-*m* potentials for a transferable, and physically-founded CG force field as proposed by the *PhaseD* approach, we need to characterise the full phase diagram for these soft model potentials, which we do here. We characterise the phase behaviour of the 6-4, 8-4 and 9-6 *n*-*m* potentials, which could serve a softer alternative to the LJ potential:

$$U_{6-4}(r) = \frac{27}{4}\varepsilon\left[\left(\frac{\sigma}{r}\right)^6 - \left(\frac{\sigma}{r}\right)^4\right] \tag{4.1}$$

$$U_{8-4}(r) = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^8 - \left(\frac{\sigma}{r}\right)^4\right] \tag{4.2}$$

$$U_{9-6}(r) = \frac{27}{4}\varepsilon\left[\left(\frac{\sigma}{r}\right)^9 - \left(\frac{\sigma}{r}\right)^6\right] \tag{4.3}$$

where r is the distance between two particles, ε is the depth of the potential well and σ is the distance at which the potential is zero. Note that the forms given here differ slightly from the notation of some of the popular molecular simulation packages, such as DLPOLY[68], where the form of the potentials are

$$U_{6-4}(r) = \frac{1}{2}\varepsilon\left[4\left(\frac{r_0}{r}\right)^6 - 6\left(\frac{r_0}{r}\right)^4\right] \tag{4.4}$$

$$U_{8-4}(r) = \frac{1}{4}\varepsilon\left[4\left(\frac{r_0}{r}\right)^8 - 8\left(\frac{r_0}{r}\right)^4\right] \tag{4.5}$$

$$U_{9-6}(r) = \frac{1}{3}\varepsilon\left[6\left(\frac{r_0}{r}\right)^9 - 9\left(\frac{r_0}{r}\right)^6\right] \tag{4.6}$$

The two can be converted between using the relation $r_0 = \sigma(\frac{n}{m})^{\frac{1}{n-m}}$.

From theoretical considerations, one may expect the 6-exponent dispersive term to carry over for CG particles given its physical basis for atomic systems. However, chemically specific coarse graining suggests that for some chemical moieties a different dispersive exponent may be a better description. Several coarse-grained models based on the *n*-*m* potentials, with dispersive exponents ranging from 4 to 8.8, have already been shown to well reproduce vapour phase properties of chain molecules, as well as being used for a number of polymer simulations[79,81,82].

Elucidating phase diagrams from molecular simulations is still challenging, even for small molecular systems. For the simple *n*-*m* potentials, there are three components to

the phase diagram: the solid-liquid, solid-vapour and liquid-vapour coexistence curves. Molecular simulation offers a variety of methods for predicting such phase diagrams, some being general whilst others are more specific, finding utility only for a particular co-existence branch. The brute force approach for finding phase coexistence involves setting up a simulation box with the two phases separated by an interface, evolving the system using either molecular dynamics (MD) or Monte Carlo (MC) at a chosen temperature and pressure, and monitoring which way the interface between the phases moves. The temperature (for a fixed pressure) or pressure (for a fixed temperature) would then be varied until the conditions are found such that the location of the interface remains constant. Such direct coexistence approaches are tedious, requiring many simulations to narrow down to the coexistence condition. The thermodynamic approach for finding phase coexistence is to calculate the absolute free energies (or chemical potentials) of the individual phases as a function of temperature and pressure and searching for the conditions at which they are equal. This is typically achieved using either a thermodynamic integration[83,84] or perturbation approach[41,85,86]. Again, these methods can be tedious to employ in practice, requiring a number of simulations (usually as at least a dozen or more) to calculate even a single free energy. It is not necessary to access the full phase diagram via such free energy calculations, however. Once a single phase-coexistence point has been determined, the Gibbs-Duhem integration procedure[87] can be used to trace the rest of the coexistence curves from a much more modest number of simulations. Caution however is required when using this method. If the initial condition is far away from the true coexistence curve, the path traced by the Gibbs-Duhem integration can diverge due to cumulative errors from successive integrations. The method is best used in conjunction with additional coexistence points that serve as constraints.

An alternative and perhaps more elegant method is to employ a density of states approach, which offers an efficient and robust route to calculating phase coexistence for a wide range of conditions, all from a somewhat limited number of simulations. They enable calculation of a system's partition function (to within an unknown constant), from which most thermodynamic properties, including phase coexistence, can be determined. The isothermal-isobaric partition function is given by the weighted summation over all the microstates accessible to a system

$$Q\left(N,p,T\right)=\sum_{i}^{states}e^{-\beta(E_i+pV_i)} \tag{4.7}$$

where N is the number of particles in the system, p is pressure, T temperature, $\beta = 1/k_BT$ is the Boltzmann factor and E_i, V_i are the energy and volume of microstate i respectively. Given that a number of these microstates will be degenerate, the partition function may be rewritten as

$$Q\left(N,p,T\right) = \sum_{E}\sum_{V}\Omega\left(V,E\right)e^{-\beta(E+pV)} \tag{4.8}$$

where Ω is the the density of states (which is independent of both temperature and pressure), and the first summation (over E) is now over all possible energy levels. The corresponding probability distribution for the system is given by

$$P\left(V,E\right) = \frac{1}{Q\left(N,p,T\right)}\Omega\left(V,E\right)e^{-\beta(E+pV)} \tag{4.9}$$

The density of states itself can be determined using already well established Wang-Landau Monte Carlo (WLMC) simulations[88,89]. Given an estimated density of states, phase coexistence is found by fixing the temperature and varying pressure in Equation 4.9 (or vice versa) until the probability distribution exhibits two peaks of equal area, i.e. the system is equally likely to exist in two unique phases. This temperature and pressure pair is a single coexistence point. As the density of states is independent of temperature and pressure, many coexistence points can be determined from a single density of states calculation[88].

Here we predominantly employ the DOS approach[90] complemented with the Gibbs-Duhem method to map out the full, and largely unexplored phase diagrams of the softer 6-4, 8-4 and 9-6 potentials. The methodology is first validated on the 12-6 potential whose phase diagram is already well characterised[61]. The phase diagram of the softer van der Waals potentials will enable the development of a transferable, higher quality, coarse-grained force field that can better reproduce the interactions and properties of groups of atoms being represented by the coarse-grained model.

4.2 Methodology

The full phase diagram for each potential was constructed in a piecewise fashion: First, the solid-liquid and vapour-liquid coexistence lines were calculated using density of states calculations. The critical and triple points where determined directly from these two

curves. The solid-vapour coexistence curve was then calculated by Gibbs Duhem integration.

4.2.1 Vapour-liquid coexistence

The density of states for the liquid and vapour phases for each potential were estimated using Wang-Landau Monte Carlo simulation as outlined by Shell et al[88]. The determined density of states was inserted into Equation 4.9 and reweighted for a range of temperatures and pressure to map out a large region of the coexistence curve. In addition, Gibbs ensemble Monte Carlo and Gibbs-Duhem integration were used to predict the same regions of the coexistence curves to give further confidence in the results. The Gibbs-Duhem integration was run using one of the coexistence points generated by the density of states approach as an initial condition. Provided that the initial condition is in fact a coexistence point, the integration should trace out the liquid-vapour coexistence curve that passes through each of the other points produced by both the DOS and Gibbs ensemble simulations.

With the liquid-vapour coexistence curves determined, the critical temperature and density for each potential was estimated by fitting the coexistence densities and temperatures from the density of states calculations to the laws of rectilinear diameters and scaling[33]. The critical pressure was found by fitting the liquid-vapour curve to a function of the form

$$\ln P^* = a_0 T^{*-1} + a_1 \tag{4.10}$$

(where $P* = P\sigma/\varepsilon$ is the reduced unit coexistence pressure, $T* = k_B T/\varepsilon$ is the reduced unit coexistence temperature, k_B is the Boltzmann constant, and a_0 and a_1 are constants that were determined by least-squares fitting) and substituting in the critical point temperature.

4.2.2 Solid-liquid coexistence

Whilst the vapour-liquid coexistence curves were readily accessible using the combination of DOS and Gibbs-Duhem, the solid-liquid coexistence curves proved to be more challenging. The system in the DOS simulations became trapped for large periods of time in one of the two phases, meaning that the density of states was sampled much more

in one phase than the other. Thus, the estimated density of states was somewhat biased towards one phase over the other and hence could not accurately define phase coexistence. In view of this, we resorted to a free energy approach. First, the density of states of the liquid and the solid phases were sampled individually. The free energy of each phase can be determined from these independent density of states surfaces according to

$$F_{WL,solid} = -k_B T \ln \sum_E \sum_V \Omega_{solid}(V,E) e^{-\beta(E+pV)} + C_{solid} \tag{4.11}$$

$$F_{WL,liquid} = -k_B T \ln \sum_E \sum_V \Omega_{liquid}(V,E) e^{-\beta(E+pV)} + C_{liquid} \tag{4.12}$$

where C_{solid} and C_{liquid} are unknown constants arising from the WL algorithm only calculating the density of states to within some constant. As the density of states of each phase is sampled independently, $C_{liquid} \neq C_{solid}$. Thus, to compare the free energies produced by Equations 4.11 and 4.12, and hence find phase coexistence, the values of C_{liquid} and C_{solid} must be determined. Their values can be found provided a single absolute free energy for the liquid and solid phases is known (the free energy is calculated from the density of states at the conditions at which the known free energy was calculated, then the constant is chosen so that the two become equal). The Einstein molecule method was used to calculate the absolute free energy of the solid phase for a temperature and pressure close to a coexistence condition (as determined from the initial Wang-Landau calculations)[91], while a variation on the Wang-Landau algorithm as developed by us and described elsewhere[92] was employed to calculate the absolute free energy of the liquid phase for the same condition. These values for the free energies were compared to those calculated using Equations 4.11 and 4.12, and used to determine the values of C_{liquid} and C_{solid}. Once these two constants were determined, absolute free energies of the two phases, and thus phase coexistence, was calculated for a wide number of conditions by directly reweighting Equations 4.11 and 4.12, without the need to repeat the relatively tedious absolute free energy calculations. The Gibbs-Duhem integration was run using one of the coexistence points generated by this approach as an initial condition.

4.2.3 Triple point and solid-vapour coexistence

The triple point was determined from the point of intersection of the solid-liquid and vapour-liquid coexistence curves. Towards the triple point, the solid-liquid curve becomes vertical, thus effectively fixing the triple point temperature. All that remains is to determine the pressure at which the curves intersect. This was found by substituting the triple point temperature into Equation 4.10. The solid-vapour coexistence curve was then traced by Gibbs-Duhem integration, using the triple point as the initial condition.

4.2.4 Technical details

All simulations were run using an in-house Monte Carlo code with the exception of the brute force molecular dynamics simulations of direct coexistence, and the solid-liquid, liquid-vapour Gibbs-Duhem integration simulations which were run using DLPOLY 4.07[68]. The energy and length scales were defined by ε and σ, which were fixed at 1 kJ mol^{-1} and 1 $\mathring{A}$ respectively for all simulations. The 12-6 and 9-6 potentials were truncated after 3σ, the 8-4 and 6-4 after 4.5σ and the usual energy and virial corrections applied. For the low-exponent dispersion term potentials i.e. the 8-4 and 6-4 potentials, the interaction decays much more slowly with separation distance. Consequently, there is need to employ a larger cutoff, which is the basis for the larger cutoff of 4.5σ employed for these potentials. System size was 500 particles for simulations involving the 12-6 and 9-6 potentials, and 1372 particles for the 8-4 and 6-4 potentials. The larger system size was needed for the 8-4 and 6-4 potentials to accommodate the increased cut-off radius, as the box size needs to be twice the cutoff to eliminate the possibility of double inclusion of interactions resulting from periodic boundaries. For the direct coexistence simulations 24000 particles were used to minimise finite size effects. Wang-Landau Monte Carlo simulations were run until the modification factor was reduced to below 10^{-6} for the standard DOS calculations and to below 10^{-7} for the free energy calculations, with reductions taking place after the minimum histogram value was no less than 80% of the average value.

4.3 Results and discussion

The coexistence approach and methodology was first tested on the LJ potential, whose phase diagram has been determined. The calculated solid-liquid, vapour-liquid curves

on a reduced pressure ($P^* = p\sigma^3/\varepsilon$) – reduced temperature ($T^* = k_BT/\varepsilon$) plane were found to be in close agreement with those presented by Agrawal and Kofke[61] as shown in Figure 4.1. The LJ potential was found to have a triple temperature and pressure of $T^*_{tp} = 0.688$, $P^*_{tp} = 0.0012$ and a critical temperature and pressure of $T^*_c = 1.297$, $P^*_c = 0.120$ which are in good agreement with other literature estimates ($T^*_{tp} = 0.694$, $P^*_{tp} = 0.0013$, $T^*_c = 1.299$, $P^*_c = 0.123$)[93,94]. The slight difference in values are most likely attributed to the differences in cut-off length, and system size. The accurate reproduction of the LJ phase diagram gives confidence in the approach and methodology (and in particular our Monte Carlo code).

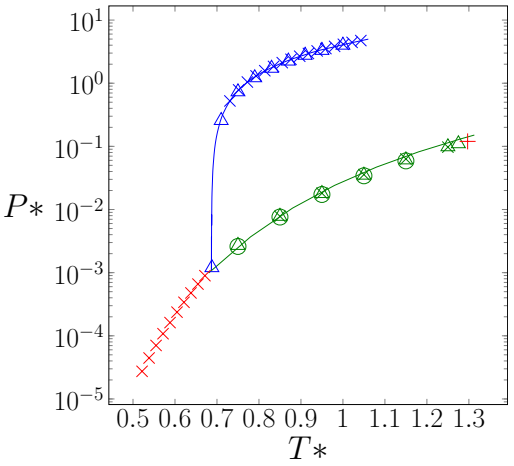


Figure 4.1: The melting (blue) and boiling curves (green) of the 12-6 potential as calculated by Agrawal and Kofke (solid line) and this study. Each circle (○) represents a coexistence point calculated by Gibbs-Ensemble Monte Carlo, each triangle (△) a point by Wang-Landau Monte Carlo and each cross cross (×) a point by Gibbs–Duhem integration.

Given this confidence, we proceeded to map the largely unknown phase diagrams of the 6–4 , 8–4 and 9–6 potentials which are presented in Figures 4.2 and 4.3 (tabulated data is available in Supplementary Information). The density of states approach was successfully used to determine the liquid-vapour coexistence curves of the 6–4, 8–4 and 9–6 potentials. The calculated curves were in excellent agreement with both the Gibbs-ensemble and Gibbs-Duhem integration techniques, which were employed for confirmation. The coexistence curve traced by Gibbs-Duhem integration passed straight through those generated by the other two methods without diverging, which is a good indication that each curve has indeed been calculated accurately. Similarly, for the solid-liquid coexistence, the combined free energy and DOS approach were in good agreement with Gibbs-Duhem integration for the 8–4 and 9–6 potentials. The results of the free energy calculations are presented in Table 4.1. However, the combined approach was

unable to yield consistent results for the 6–4 potential. The curve traced by Gibbs-Duhem integration using one of the Wang Landau Monte Carlo coexistence points as the initial condition diverged rapidly. We are unable to offer a rigorous explanation as to why the method was successfully used for the 8–4, 9–6 and 12–6 potential, and yet was unsuccessful for the 6–4 potential. We believe that the broader potential well of the 6-4 potential perhaps hindered sampling of the liquid phase during the WL sampling. Consequently, for the solid-liquid coexistence of the 6–4 potential, we resorted to direct coexistence simulations using molecular dynamics simulations to identify a number of points on the solid-liquid existence curve. One of these values was then used as the initial condition for the Gibbs-Duhem integration. The Gibbs-Duhem was able to trace the full curve being in excellent agreement with the discrete points determined by direct coexistence simulations.

Table 4.1: The results of the absolute free energy calculations for the liquid and solid phases for each of the potential models.

	T^*	P^*	μ_{Solid}	μ_{Liquid}
12-6	1.0000	3.9400	0.825	0.819
9-6	1.2400	8.4176	4.747	4.740
8-4	1.9200	12.0735	-5.551	-5.494

The critical temperature T_c^* and density ρ_c^* of each potential were calculated by fitting the liquid-vapour curves to the scaling and rectilinear laws and are presented in Table 4.2. The curves were also fitted to Equation 4.10, yielding the coefficients that are also presented in Table 4.2. The critical pressures P_c^* were determined from these. For each potential, the calculated $\ln P^*$ varied linearly as a function of $1/T^*$, and hence the fitted curves were in excellent agreement with the calculated ones.

Table 4.2: Calculated critical points of the n-m potentials.

	T_c^*	P_c^*	ρ_c^*	a_0	a_1
12-6	1.2970	0.1199	0.314	-6.742	3.074
9-6	1.5918	0.1425	0.309	-7.765	2.924
8-4	4.9525	0.5276	0.356	-21.077	3.607
6-4	8.1626	1.0176	0.441	-33.555	4.111

The triple-point pressures were determined directly from the solid-liquid coexistence curves. Interestingly, the point comes naturally out of reweighting the solid-liquid density of states. Below the triple-point temperature, the weighted probability distribution only exhibits a single peak at an energy and density consistent with a solid, regardless of pressure. The triple-point temperature can thus be found by incrementing the temperature at some fixed low pressure until the liquid peak appears, and has an equal area

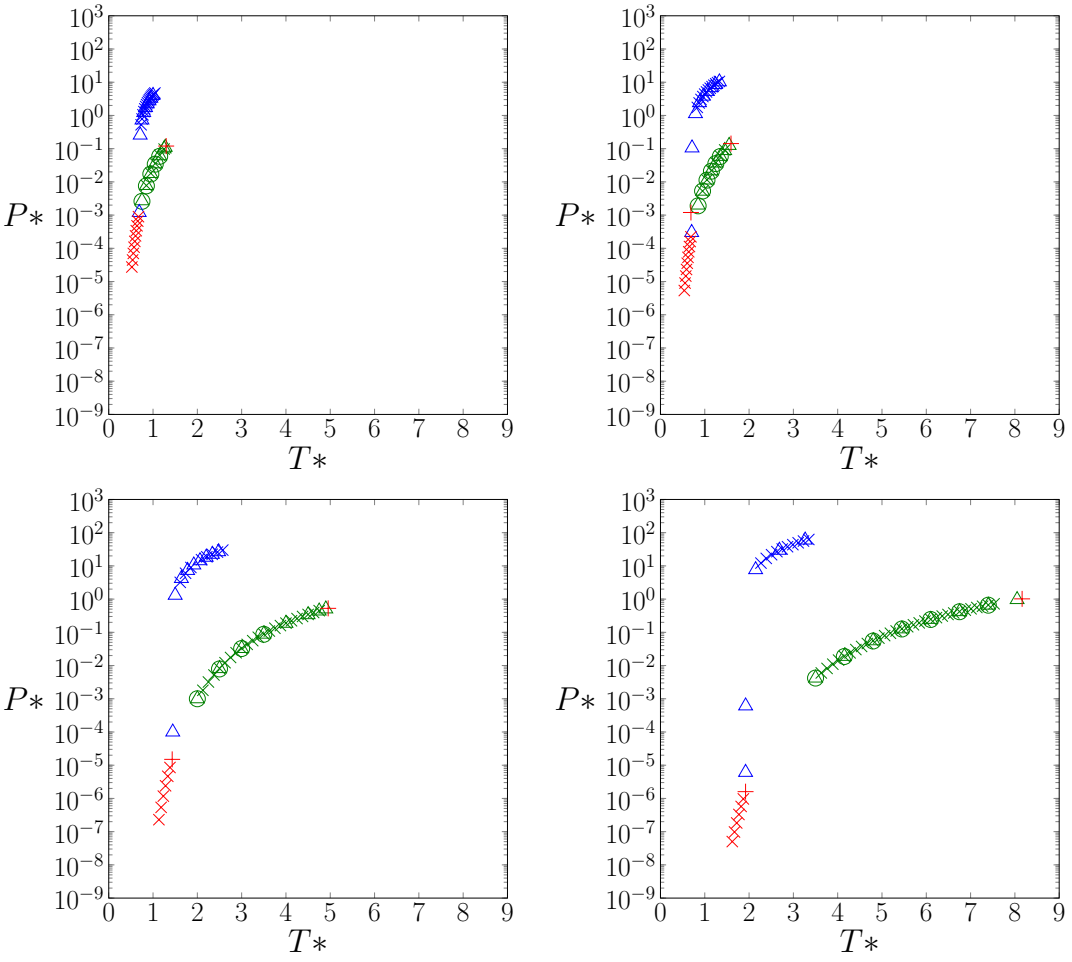


Figure 4.2: Calculated phase diagrams of the 12-6 (top left), 9-6 (top right), 8-4 (bottom left) and 6-4 (bottom right) potentials. The melting, boiling and sublimation curves are marked in blue, green and red respectively. Triangle (△), circle (○) and cross (×) symbols mark points calculated using Wang-Landau (or direct coexistence in the case of the 6-4 potential), Gibbs ensemble Monte Carlo and Gibbs-Duhem integration respectively.

to the solid peak. The triple pressure was determined by substituting the triple temperature into Equation 4.10, along with the coefficients in Table 4.2. The triple points for each potential are given in Table 4.3.

Table 4.3: Calculated triple points of the n-m potentials.

	T^*_{tp}	$P^*_{tp}(\times 10^{-4})$
12-6	0.688	11.990
9-6	0.705	3.038
8-4	1.432	0.150
6-4	1.921	0.016

While the critical and triple points are largely consistent with the literature, with some

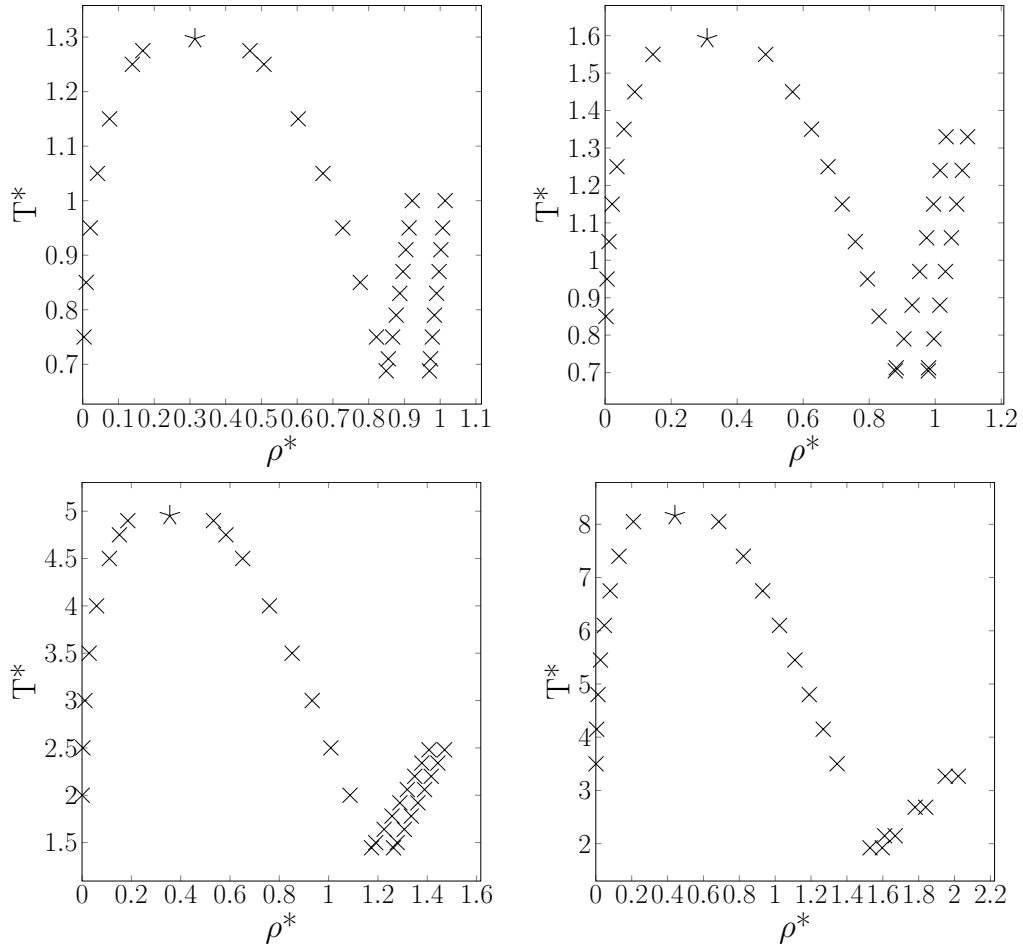


Figure 4.3: Coexistence densities determined from DOS calculations (and direct coexistence calculations in the case of the 6-4 potential) for the 12-6 (top left), 9-6 (top right) and 8-4 (bottom left), and 6-4 (bottom right) potentials. The critical point is shown by the star symbol.

minor deviations likely due to differences in cut-off length and system size (previous estimates for the 9-6 potential suggest triple and critical coexistence conditions as $T^*_{tp} = 0.720$, $P^*_{tp} = 0.00036$, $T^*_c = 1.616$, $P^*_c = 0.151$)[94–98], the values for the 6-4 and 8-4 differ significantly from those predicted by the equation of state (EOS) proposed for the n-m potential[99]. The deviations are most likely attributed to the EOS being fitted against mostly much harder potentials than are explored here, and hence the EOS is being extrapolated beyond its limits. The values of the critical and triple points presented here could perhaps be used to improve the EOS.

Whilst it has been identified that the lower exponent non-bonded interactions are better representations of the true interaction of a mapped groups of atoms, a clearer understanding as to why this so is lacking, other than that such potentials are softer. The calculated phase diagrams (presented in Figure 4.2) reveal the significant, actual impact

of the choice of a given *n-m* potential on the phase behaviour of the coarse-grained model. The most notable difference is the width of the liquid range; the range over which the model remains a liquid markedly increases as the potential is softened (as the exponents get lower). It is interesting to note that that the range over which the model remains a solid also expands, although not quite so dramatically.

The significance of the width of the liquid range for the *n-m* potentials, for example, can be illustrated by mapping a coarse-grained water model onto each of the potentials. We take the diameter of the coarse-grained water particle to be 0.47 nm, approximately the size of four clustered water molecules (although this value should likely be larger for the softer potentials), and the melting point is fixed to T_{mp} = 273 K, p_{mp}=1 atm. For these constraints one can read off the corresponding value of ε from the phase diagram. The sigma value of 0.47 nm and the identified value of ε, therefore, yields a water model that by design melts at 273 K. Linking the potential parameters to a melting point also fixes the boiling point (given approximately by Equation 4.13) of the model. The estimated boiling point for such a water model are summarised in Table 4.4. Particularly notable is the extremely limited liquid-phase range for the LJ water model spanning 273 – 286 K, a range of just 13 K. Thus, the origin of the unphysical freezing of the MARTINI water model becomes clear. In contrast, the softer, lower-exponent potentials clearly show a marked increase in the liquid range relative to the LJ potential. While the 8-4 potential gives the closest match to physical water (a liquid range of 100 K and boiling point of 373 K), the other soft potentials cover a wide range of phase behaviour, and hence could be effectively employed to represent of a spectrum of different molecules and moieties.

It is pertinent to note that liquid range identified for the water model depends on the choice of mapping. Should the coarse-grained mapping be say 3-1 (3 atoms being represented by 1 CG particle), the appropriate sigma value will be smaller, compared with the 4-1 mapping which gives a sigma value of 0.47 nm. A smaller sigma corresponds to a lower reduced pressure P*, and sampling the phase diagram at the lower reduced pressure yields a more limited liquid range. Likewise, a higher mapping (a more lower-resolution model) will mean a larger sigma, and hence higher reduced pressure, which on the phase yields a much broader liquid range. For example, a water model with a higher mapping of say approximately 8 water molecules that equating to a sigma of 9.65 nm would yield a liquid range of about 100 K.

We present here fitted equations that enable the identification of the melting (T_{mp}) and boiling (T_{bp}) points of each potential, and from these, an equation that yields the value of epsilon for a chemical moiety with a particular melting point for a given choice of

Table 4.4: Approximate parameters for a coarse-grained water model using the 12-6, 9-6, 8-4 and 6-4 potentials.

	ε / kJ mol^{-1}	T_{bp} / K	$(T_{bp} - T_{mp})$ / K
12-6	3.298	286	13
9-6	3.215	328	55
8-4	1.573	437	164
6-4	1.180	510	237

sigma (defined by the chosen mapping). The melting T^*_{mp} and boiling T^*_{bp} temperatures of each potential can be approximated to a good degree for pressures below the critical pressure by the following

$$T^*_{bp} = a_0 \left(\ln P^* - a_1\right)^{-1} \tag{4.13}$$

$$T^*_{mp} = b_0 + b_1 P^* + b_2 P^{*2} \tag{4.14}$$

where the coefficients b_0, b_1, b_2 are presented in Table 4.5. The coefficients were calculated from the WL coexistence data using least-squares fitting. Combining Equation 4.14 with the definition of the reduced LJ units and rearranging yields

$$\varepsilon = \frac{1}{2b_0}\left[k_B T_{mp} - b_1 p_{mp}\sigma^3 + \left(\left(b_1 p_{mp}\sigma^3 - k_B T_{mp}\right)^2 + 4b_0 b_2 p^2_{mp}\sigma^6\right)\frac{1}{2}\right] \tag{4.15}$$

Table 4.5: The coefficients derived by least square fitting used to approximate the melting point of the 6-4, 8-4, 9-6 and 12-6 potentials.

	b_0	b_1	b_2
12-6	0.6882	0.0855	-0.0019
9-6	0.7061	0.0738	-0.0013
8-4	1.4430	0.0483	-0.0004
6-4	1.9235	0.0292	-0.0001

4.4 Conclusion

To conclude, we have characterised the largely unknown phase diagrams of the softer 6-4, 8-4, 9-6 n-m potentials, using a combined methodology validated against the well characterised LJ phase diagram. These diagrams have given direct insight into the

nature and cause of the limited liquid range exhibited by the widely used LJ potential. The determined phase diagrams will form the basis for the development of new class of force field with strong physical basis, being linked to the melting points of the chemical moieties being represented. The universal nature of the approach and the diversity of the potentials will enable the parameterisation of a wide class of transferable coarse-grained beads, which will offer a more representative and robust representation of molecules. Further, the broadened liquid range inherent to the softer potentials will open the scope of systems and conditions at which simulations can be carried out, facilitating the study of a wide range of solid, liquid and vapour phenomena.

4.5 Supplementary Information

Table 4.6: Vapour-liquid coexistence points determined from Wang–Landau MC simulations for the 6-4 potential.

$T*$	$P*$	ρ_{vapour}	ρ_{liquid}
3.5000	0.0042	0.0012	1.3454
4.1500	0.0189	0.0047	1.2677
4.8000	0.0557	0.0123	1.1902
5.4500	0.1272	0.0259	1.1098
6.1000	0.2451	0.0476	1.0243
6.7500	0.4195	0.0802	0.9304
7.4000	0.6585	0.1293	0.8229
8.0500	0.9688	0.2094	0.6856

Table 4.7: Vapour-liquid coexistence points determined from Wang–Landau MC simulations for the 8-4 potential.

$T*$	$P*$	ρ_{vapour}	ρ_{liquid}
2.0000	0.0010	0.0005	1.0867
2.5000	0.0081	0.0033	1.0089
3.0000	0.0326	0.0116	0.9334
3.5000	0.0881	0.0287	0.8519
4.0000	0.1872	0.0591	0.7600
4.5000	0.3401	0.1109	0.6512
4.7500	0.4392	0.1516	0.5835
4.9000	0.5064	0.1857	0.5331

Table 4.8: Vapour-liquid coexistence points determined from Wang–Landau MC simulations for the 9-6 potential.

$T*$	$P*$	ρ_{vapour}	ρ_{liquid}
0.8500	0.0020	0.0024	0.8298
0.9500	0.0053	0.0058	0.7947
1.0500	0.0114	0.0118	0.7581
1.1500	0.0217	0.0214	0.7183
1.2500	0.0371	0.0359	0.6753
1.3500	0.0587	0.0573	0.6252
1.4500	0.0878	0.0898	0.5678
1.5500	0.1256	0.1448	0.4863

Table 4.9: Vapour-liquid coexistence points determined from Wang–Landau MC simulations for the 12-6 potential.

$T*$	$P*$	ρ_{vapour}	ρ_{liquid}
0.7500	0.0027	0.0037	0.8218
0.8500	0.0078	0.0098	0.7765
0.9500	0.0179	0.0214	0.7277
1.0500	0.0350	0.0413	0.6722
1.1500	0.0611	0.0751	0.6029
1.2500	0.0986	0.1389	0.5071
1.2750	0.1101	0.1681	0.4679

Table 4.10: Solid-liquid coexistence points determined from direct coexistence for the 6-4 potential.

$T*$	$P*$	ρ_{solid}	ρ_{liquid}
1.9209	0.0000061	1.597	1.530
1.9218	0.00061	1.597	1.529
2.1481	7.7058	1.669	1.610
2.6844	29.5825	1.839	1.780
3.2664	59.9791	2.021	1.948

Table 4.11: Solid-liquid coexistence points determined from Wang–Landau MC simulations for the 8-4 potential.

$T*$	$P*$	ρ_{solid}	ρ_{liquid}
1.44484	0.000015	1.263	1.174
1.5000	1.3079	1.278	1.191
1.6400	4.1840	1.307	1.225
1.7800	7.3038	1.335	1.257
1.9200	10.6780	1.362	1.289
2.0600	14.3113	1.389	1.319
2.2000	18.2021	1.416	1.349
2.3400	22.3405	1.443	1.378
2.4800	26.7172	1.470	1.407

Table 4.12: Solid-liquid coexistence points determined from Wang–Landau MC simulations for the 9-6 potential.

$T*$	$P*$	ρ_{solid}	ρ_{liquid}
0.7045	0.0003	0.979	0.879
0.7127	0.1053	0.980	0.882
0.7900	1.1347	0.996	0.905
0.8800	2.4250	1.014	0.930
0.9700	3.8137	1.031	0.953
1.0600	5.2950	1.049	0.974
1.1500	6.8666	1.065	0.995
1.2400	8.5267	1.082	1.015
1.3300	10.2720	1.098	1.033

Table 4.13: Solid-liquid coexistence points determined from Wang–Landau MC simulations for the 12-6 potential.

$T*$	$P*$	ρ_{solid}	ρ_{liquid}
0.6879	0.0012	0.970	0.849
0.7100	0.2568	0.973	0.855
0.7500	0.7316	0.978	0.867
0.7900	1.2210	0.984	0.877
0.8300	1.7253	0.990	0.887
0.8700	2.2428	0.997	0.896
0.9100	2.7712	1.002	0.904
0.9500	3.3117	1.007	0.913
1.0000	4.0052	1.014	0.922

Table 4.14: Vapour-solid coexistence points determined from Wang–Landau MC simulations for the 6-4 potential.

$T*$	$P*$	ρ_{solid}	ρ_{vapour}
1.6211	5.0244E-08	1.638	3.093E-08
1.6711	9.7761E-08	1.634	5.839E-08
1.7211	1.8281E-07	1.630	1.058E-07
1.7710	3.2913E-07	1.626	1.852E-07
1.8210	5.7302E-07	1.621	3.138E-07
1.8710	9.6707E-07	1.616	5.152E-07

Table 4.15: Vapour-solid coexistence points determined from Wang–Landau MC simulations for the 8-4 potential.

$T*$	$P*$	ρ_{solid}	ρ_{vapour}
1.1325	2.2944E-07	1.266	2.020E-07
1.1825	5.3811E-07	1.261	4.528E-07
1.2324	1.1737E-06	1.255	9.503E-07
1.2824	2.4017E-06	1.250	1.867E-06
1.3324	4.6432E-06	1.244	3.474E-06
1.3823	8.5415E-06	1.239	6.169E-06

Table 4.16: Vapour-solid coexistence points determined from Wang–Landau MC simulations for the 9-6 potential.

$T*$	$P*$	ρ_{solid}	ρ_{vapour}
0.5382	5.3673E-06	1.004	9.995E-06
0.5548	9.0038E-06	1.001	1.629E-05
0.5714	1.4634E-05	0.997	2.566E-05
0.5881	2.3126E-05	0.993	3.943E-05
0.6047	3.5579E-05	0.990	5.916E-05
0.6213	5.3476E-05	0.986	8.625E-05
0.6379	7.8616E-05	0.982	1.236E-04
0.6546	1.1323E-04	0.978	1.737E-04
0.6712	1.6009E-04	0.975	2.397E-04
0.6878	2.2235E-04	0.971	3.255E-04

Table 4.17: Vapour-solid coexistence points determined from Wang–Landau MC simulations for the 12-6 potential

$T*$	$P*$	ρ_{solid}	ρ_{vapour}
0.5216	2.7380E-05	1.003	5.258E-05
0.5383	4.4528E-05	0.999	8.280E-05
0.5549	7.0322E-05	0.996	1.272E-04
0.5715	1.0795E-04	0.992	1.897E-04
0.5882	1.6187E-04	0.988	2.773E-04
0.6048	2.3697E-04	0.984	3.951E-04
0.6214	3.3959E-04	0.979	5.525E-04
0.6380	4.7730E-04	0.976	7.581E-04
0.6547	6.5940E-04	0.971	1.022E-03
0.6713	8.9584E-04	0.967	1.354E-03

Chapter 5

Solubility prediction from first principles: A density of states approach

Abstract: *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.**

*The manuscript presented in this chapter is listed as Paper III in the list of publications.

5.1 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[100], 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[101]. 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[102]. 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'[103,104].

There are three main approaches to solubility prediction: empirical, correlation-based methods[105], quantum mechanical (QM) continuum solvation models such as COSMO-RS[106], and molecular simulation[107]. 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[108] (summarised below), or direct (brute force) simulation of the dissolution of the solid in a solvent towards equilibrium[40]. 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, the (undissolved) solid phase coexists with its solution. As the two are in equilibrium, the chemical potential of the solute in the solid phase and that

in solution are identical at the given temperature T and pressure p. Prediction of the 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[42,109] 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[83,84], the so-called perturbation approach[41,85,86], expanded ensembles[110,111], and variations on these[112]. These methods involve 'growing' the solute molecule from its reference state reversibly in the solvent. While both thermodynamic integration and perturbation techniques are robust and effective (particularly when coupled with soft-core[113] and dampening potentials[114]), 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, that 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[111,115–120]. 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[88,89,121]. 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 co-existence 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.

5.2 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

$$Q\left(N,p,T\right)=\sum_{i}^{states}\exp\left[-\beta\left(E_{i}+pV_{i}\right)\right] \tag{5.1}$$

where the first summation is over all states, with corresponding energy E_i and volume V_i. Given that distinct states may have identical energies i.e. are degenerate, $Q\left(N,p,T\right)$ may be expressed in the form

$$Q\left(N,p,T\right)=\sum_{E}\sum_{V}\Omega\left(V,E\right)\exp\left[-\beta\left(E+pV\right)\right] \tag{5.2}$$

where $\Omega\left(V,E\right)$ is the density of states of the system[90] and the summation over energy is now over energy levels. The corresponding probability distribution is then

$$P(V,E)=\frac{1}{Q\left(N,p,T\right)}\exp\left[\ln\Omega\left(V,E\right)-\beta\left(E+pV\right)\right] \tag{5.3}$$

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 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 $\Omega\left(V,E\right)$ 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[88].

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

For such a system, the partition function is given by

$$\Xi\left(\mu_{i},p,T\right)_{N_{j},N_{k},\ldots}=\sum_{E}\sum_{V}\sum_{N_{i}}\Omega\left(N_{i},V,E\right)_{N_{j},N_{k},\ldots}\exp\left[-\beta\left(E+pV-\mu_{i}N_{i}\right)\right] \tag{5.4}$$

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

$$P\left(N_i, V, E\right)_{N_j, N_k, \ldots} = \frac{\Omega\left(N_i, V, E\right)_{N_j, N_k, \ldots} \exp\left[-\beta\left(E + pV - \mu_i N_i\right)\right]}{\Xi\left(\mu_i, p, T\right)_{N_j, N_k, \ldots}} \tag{5.5}$$

As before, if $\Omega\left(N_i, V, E\right)_{N_j, N_k, \ldots}$ 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 $N_j, N_k, \ldots$ 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 (Figure 5.1).

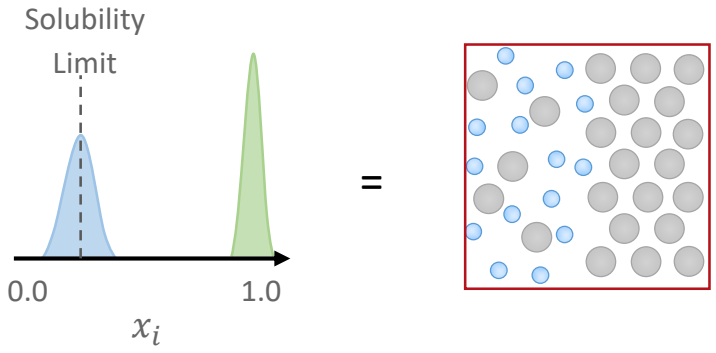


Figure 5.1: A schematic probability distribution for a system of solute (grey particles) and solvent (blue particles) as a function of solute fraction. At the solubility limit, the solute particles will have an equal probability of being in both the solid phase (the green peak at $x = 1.0$) and the solution phase (the blue peak). The location (mole fraction) of the solution phase peak is the solubility limit.

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, $\mu_{solid}\left(T, p\right) = \mu_{soln}\left(T, p\right)$, one could substitute the chemical potential of the solid, if it were known, into the probability distribution (Equation 5.5). 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 (Figure 5.1). 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 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 $\Omega\left(N_i, V, E\right)_{N_j,N_k,\ldots}$, techniques for which are now well established[122]. Here we employ a 3-dimensional variant of the efficient Monte Carlo scheme originally developed by Wang and Landau[90]. Configurations are generated according to probability

$$P\left(N_i, V, E\right)_{N_j,N_k,\ldots} \propto \frac{1}{\Omega\left(N_i, V, E\right)_{N_j,N_k,\ldots}} \tag{5.6}$$

with $\Omega\left(N_i, V, E\right)_{N_j,N_k,\ldots}$ being developed and improved on-the-fly as the simulation proceeds in a self consistent manner. Everytime a particular point in $\Omega\left(N_i, V, E\right)_{N_j,N_k,\ldots}$ space is visited, its value is incremented according to $\ln\Omega\left(N_i, V, E\right)_{N_j,N_k,\ldots,new} = \ln f + \ln\Omega\left(N_i, V, E\right)_{N_j,N_k,\ldots,old}$, where $\ln f$ is an arbitrary modification factor. When Ω 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 $h\left(N_i, V, E\right)$. 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 histogram of visits $h\left(N_i, V, E\right)$ 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 $\ln f_{new} = \frac{1}{2}\ln f_{old}$, the histogram of visits is reset to zero for the next iteration of the simulation.

To explore the (N_i, V, E) space associated with $\Omega\left(N_i, V, E\right)_{N_j,N_k,\ldots}$, 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[119], which are valid provided that the volume is sampled logarithmically:

$$\begin{aligned}
P_{translation}\left(A \rightarrow B\right) &= \min\left(1, \frac{\Omega\left(A\right)}{\Omega\left(B\right)}\right) \\
P_{volume}\left(A \rightarrow B\right) &= \min\left(1, \frac{\Omega\left(A\right)}{\Omega\left(B\right)}\frac{V_B^{N_i+1}}{V_A^{N_i+1}}\right) \\
P_{insertion}\left(A \rightarrow B\right) &= \min\left(1, \frac{\Omega\left(A\right)}{\Omega\left(B\right)}\frac{V}{N_{i,B}}\right) \\
P_{deletion}\left(A \rightarrow B\right) &= \min\left(1, \frac{\Omega\left(A\right)}{\Omega\left(B\right)}\frac{N_{i,A}}{V}\right)
\end{aligned} \tag{5.7}$$

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 Figure 5.2).

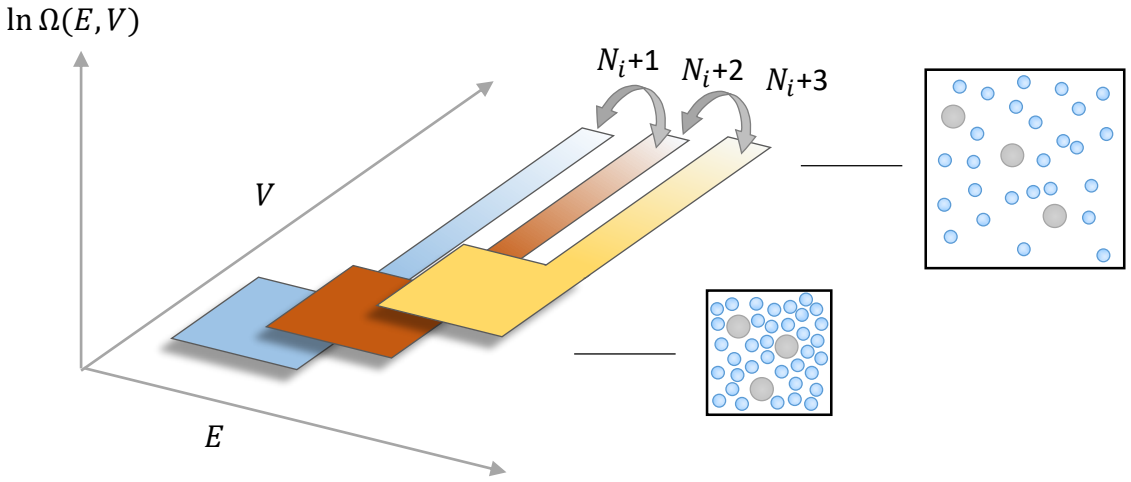


Figure 5.2: The density of states is sampled independently for each concentration of interest in both in the liquid state and the gas states. Insertion/deletion moves between the different concentration windows are performed in the gas phase in order to connect the independent concentration windows.

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

(i) Determination of the 2-d density of states $\Omega\left(N_i, V, E\right)_{N_j, N_k, \ldots}$ for each solution concentration $(\ldots, N_i-1, N_i, N_i+1, N_i+2, \ldots)$, 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 $(\ldots, N_i-1, N_i, N_i+1, N_i+2, \ldots)$ 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 using least squares, which minimises the error function

$$e_{tot} = \sum_{i=1}^{M} \sum_{k} \left[\ln \Omega_{i,NpT}(k) + C_i - \ln \Omega_{\mu VT}(k) \right]^2 \tag{5.8}$$

where M is the number of individual concentration windows, k is an index for all the overlapping points shared by the two windows[88], $\Omega_{i,NpT}$ is the density of states of concentration window 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.

5.3 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 ~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[123]. 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 Figure 5.3. 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.

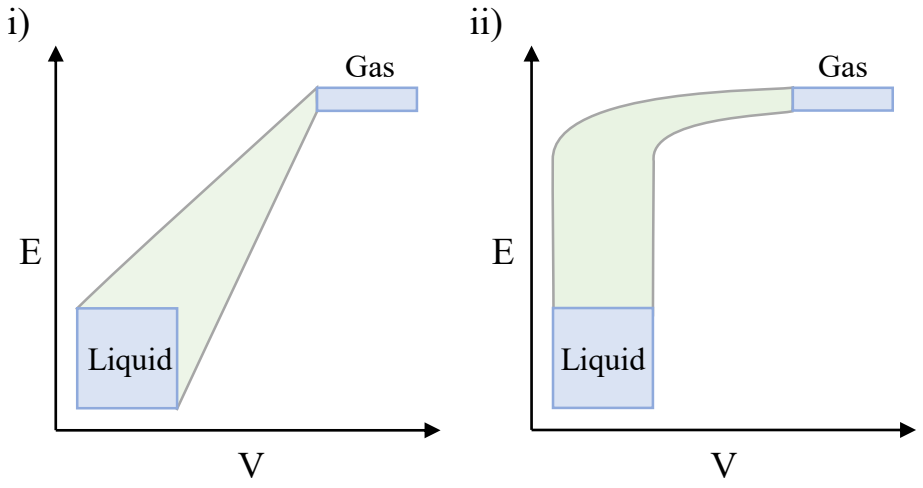


Figure 5.3: The two choices explored for the accessible energies and volumes between the liquid and gas states: i) direct interpolation between the liquid and gas states and ii) transformation of the liquid to dense, high energy states before expanding to the gas state, thereby avoiding a first order liquid-gas phase transition.

The energy range was discretised into bins of width 10,000 kJ mol^{-1} while the logged volume range was discretised 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 also.

For each of the simulations the initial value of the Wang-Landau convergence factor was set to 1.0, and was allowed to decrease until it was less than 2×10^{-7}. By this point the relative change in the logged density of states between the current and previous iterations

was 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[124] to expedite convergence and precision. Three walkers were found to be optimal for the liquid-gas windows and four walkers for the gas windows.

5.4 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 Equation 5.5, is shown in Figure 5.4. The NaCl solid chemical potential was taken as = -770.92 kJ mol^{-1} as reported by Benavides et al[107]. Their choice of a de Broglie wavelength of 1.0 Åwas 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[125].

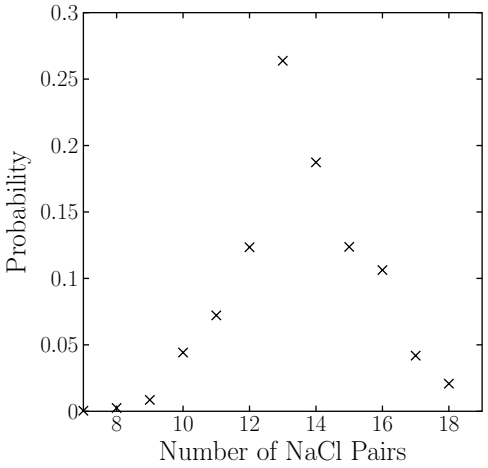


Figure 5.4: The probability distribution for the aqueous sodium chloride system at T=298 K and p=1 atm, averaged over five independant runs.

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

$$\langle N_{NaCl} \rangle_{T,p,N_{H_2O}} = \sum_{E} \sum_{V} \sum_{N_{NaCl}} N_{NaCl} \times P\left(N_{NaCl,V,E}\right)_{T,p,N_{H_2O}} \tag{5.9}$$

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 Equation 5.5. Uncertainties in these values were calculated by averaging the results obtained from five independant DOS calculations. The calculated solubility is in close agreement with the values found in the literature for the Joung-Cheatam 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}. This disparity between the calculated and experimental solubility is due to the model itself (which is currently the best available) [107]. 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 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 reasonably good agreement, the latter approach turned out to be more accurate - the data for which is presented in Figure 5.5 along with values presented in the literature for this model [107,118]. 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 potential of solutions.

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. 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. [116] and is presented in Table 5.1.

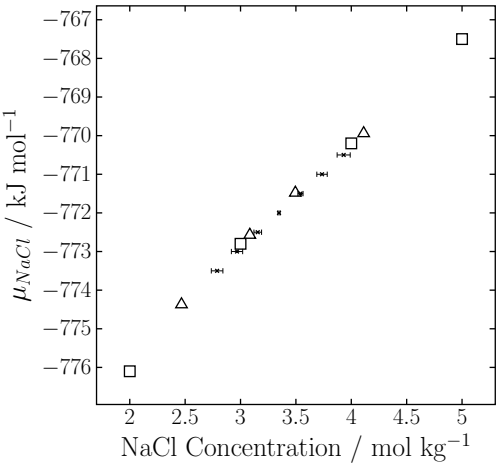


Figure 5.5: The chemical potential of the JC/SPC/E NaCl model as a function of concentration as calculated by this work (crosses), Vega *et al*[107] (triangles), Panagiotopoulos *et al*[118] (squares).

Table 5.1: 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)

These chemical potential values of the NaCl solid, along with the density of states, were inserted into Equation 5.5 in order to generate probability distributions for each temperature, from which the NaCl solubility was determined as before. The predicted solubility as a function of temperature is presented in Figure 5.6. Counter-intuitively, the solubility of the NaCl model actually decreases as the temperature increases. This unexpected behaviour has also been reported by others in the literature[118], again attributed as a limitation of the model itself.

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

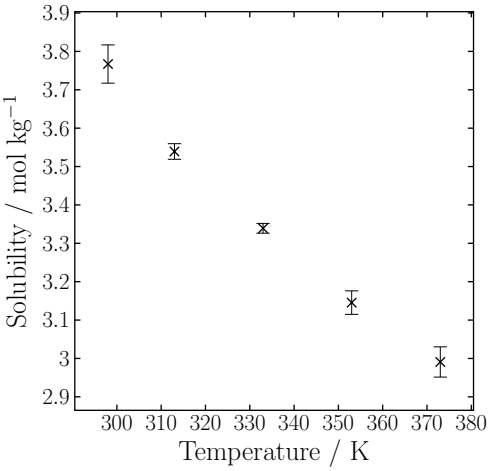


Figure 5.6: The solubility of the JC/SPC/E NaCl model as a function of temperature.

caused by the modification factor reduction scheme. The errors due to saturation and detailed balance have been discussed at depth in the literature[88,126], and are expected to be small relative to the sampling error. Notably, the overall estimated errors in the solubility and chemical potential calculations, being determined by performing five independant sets of simulation, are relatively small.

While the method has so far 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. 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.

Chapter 6

Solubility prediction via chemical potentials from density of states

Abstract: *The solubility of compounds is of fundamental significance to most fields, yet its prediction from first principles (starting from only knowledge of the solute and solvent's structure) remains a challenge. Recently we proposed a robust and efficient method to this end, employing classical molecular simulations to access the density of states (DOS) of a system of solute and solvent, and from this solubility. Here we improve the efficiency, and indeed the generality, of the method by extending it to calculate solution chemical potentials, from which solubility may be accessed. We employ this method to predict the chemical potential of urea in water and urea in methanol for a range of concentrations at ambient conditions. These were validated against values calculated by thermodynamic integration, and were found to be in excellent agreement. They were further used to obtain the solubility of urea in water (20.15 mol kg^{-1}) and in methanol (4.11 mol kg^{-1}) at ambient conditions, and for further temperatures up to 338 K.**

*The manuscript presented in this chapter is listed as Paper IV in the list of publications.

6.1 Introduction

Solubility is perhaps one of the most fundamental properties in chemistry, arising from the complex interactions between a solute and solvent. It is of marked importance for most fields, including material development, toxicology, food processing, the oil industry[127] and, pharmaceutical development[128] where many (if not most) drug molecules have poor solubility which in turn can hinder bioavailability. In each of these fields, the ability to accurately and efficiently predict the solubility would be a significant utility. Such methods will also give access to the solubility of compounds that would be a challenge to study in the laboratory and at conditions inaccessible to experiment e.g. high temperature and pressures. Molecular simulation offers a potentially powerful first principles route to this end.

For some applications, e.g. the development of pharmaceuticals, there is a need to predict the soubility of molecules that have yet to be synthesised, and hence for which the structure of the solid is unknown. In such cases, the polymorph landscape of the molecule would need to be first identified. This is becoming increasingly feasible with improving crystal structure prediction methods[129]. Molecular simulation would then be employed to gain access to the solubility of each possible form.

A route to predict solubility from simulation would be to employ a direct coexistence approach. Whilst this is promising, there are limitations, a key one being the time required to reach equilibrium can be unfeasable (on the order of microseconds)[40]. In contrast, the chemical potential route to solubility prediction is more robust and efficient. At the solubility limit, the chemical potentials of the solute in solution and in the solid phases are equal, such that

$$\mu^{solute}_{solid}(T,p) = \mu^{solute}_{solution}(T,p) \tag{6.1}$$

where T, p are the system's temperature and pressure respectively. While μ^{solute}_{solid} is readily calculated by employing the Einstein molecule[130] (or crystal[42]) method described below, calculating $\mu^{solute}_{solution}$ is typically more involved. In general, there are two approaches to do this: the first would be to determine the solution concentration that would exist at a given chemical potential (i.e at μ^{solute}_{solid}), while the second would invole calculating the chemical potential of the solution for a range of concentrations, and determining the concentration at which the solid and solution chemical potentials intersect[116]. Calculating

concentration for a given chemical potential is generally achieved by performing simulations in the semi grand-canonical ensemble[117], where molecules are grown or removed step-wise from the solution until equilibrium at the desired chemical potential is achieved. While this method works well for calculations involing ionic or atomic solutes[117,131,132], large solute molecules pose a challenge. The alternative approach, determining the chemical potential as a function of concentration, is more general and established; the methods include thermodynamic integration[83,84] (TI), perturbation[41,85,86] or expanded ensemble calculations[110,133]. TI in particular can require dozens of simulations to calculate even a single value for a single concentration for a single condition. These would then need to be repeated for each concentration / condition of interest. Further, these methods too are challenged by larger molecules, although employing soft-core potentials[113] or the recently employed cavitation method[120] go some way to overcome this.

Recently we proposed a novel method to calculate solubility directly from a systems density of states (DOS) that, in principle, is able to overcome both these limitations[134], demonstrating the method for predicting the solubility of NaCl. The DOS gives access to most properties of a system, including the probability of the system existing at different concentrations as a function of chemical potential from which solubility can be determined. The approach employes a variant of the Wang–Landau algorithm[88,90], where solution simulations are bridged to the vapour phase for the required insertion / deletion moves, so that insertion of even large molecules may be facilitated. Further, as the density of states is independent of temperature and pressure, the DOS gives access to solubility for wide range of conditions from a single DOS calculation.

Here, we extend the new DOS methodology for predicting solubility from focussing on the co-existence distributions to a more efficient approach of predicting solubility from chemical potentials calculated from DOS. In the original DOS solubility approach, one identifies the location of the probability distribution in the discrete solution concentration space (N, N+1, N+2, N+3 .. systems) at a particular chemical potential – the chemical potential of the solid phase. To accurately capture this distribution, the DOS must be determined for all concentrations that have a non-zero probability of existing at the given chemical potential. When the solubility limit is completely unknown *a priori*, it is then necessary to include a large spectrum of discrete solute concentrations within the DOS calculation as one does not know the location of the probability distribution in concentration space. Much of this information, however, is redundant, since the important concentrations are only those that contribute to the distribution peak that identifies the solubility concentration. In the original DOS-based solubility study, we

exploited prior knowledge of the solubility of the NaCl model, and employed 12 discrete concentrations with the DOS calculation. Here we reformulate the DOS-based solubility prediction method to calculate chemical potentials as a function of concentration, rather than the other way around. The free energy of solution of a given concentration is directly accessible (to within a given constant) from its density of states as calculated by our approach. If several such free energies are determined, an analytical function may be fitted to these as a function of solute concentration, then the chemical potential is simply its derivative. While a certain number of free energies (equating to the number of discrete concentrations) must be determined to produce an accurate fit, this in general will be much less than would be required for the distribution route. It should be noted that while this method is presented in the context of solubility calculation, it is in fact a general approach for calculating the chemical potential of fluid phases – the free energy change would be expected to vary linearly as a function of the number of molecules of interest, with the gradient being the chemical potential. We demonstrate and apply the method to predict the solubility of the organic molecule urea in both methanol and water for a range of temperatures. The chemical potentials calculated using DOS have been validated by thermodynamic integration.

6.2 Chemical potential of solution from DOS

The isothermal-isobaric partion function of a system of N_{solute} solute and $N_{solvent}$ solvent molecules is given by

$$Q(T,p,N_{solute},N_{solvent}) = \frac{q_{solute}^{N_{solute}} q_{solvent}^{N_{solvent}}}{N_{solute}!N_{solvent}!\Lambda_{solute}^{3N_{solute}}\Lambda_{solvent}^{3N_{solvent}}} \times \sum_{i}^{states} \exp\left[-\beta\left(E_i + pV_i\right)\right] \tag{6.2}$$

where q_{solute}, $q_{solvent}$ are the molecular partition functions (i.e rotational, vibrational and electronic) of the solute and solvent species respectively, Λ_{solute}, $\Lambda_{solvent}$ are their de Broglie wavelengths, $\beta = \frac{1}{k_B T}$ and k_B is the Boltzmann constant [120,135]. Here the first summation is over all possible microstates (with energy E_i and volume V_i) adopted by the system. Given that certain microstates are degenerate, Equation 6.2 can be rewritten as

$$Q(T,p,N_{solute},N_{solvent}) = \frac{q_{solute}^{N_{solute}} q_{solvent}^{N_{solvent}}}{N_{solute}!N_{solvent}!\Lambda_{solute}^{3N_{solute}}\Lambda_{solvent}^{3N_{solvent}}} \times \sum_E \sum_V \Omega_{conf}(V,E) \exp\left[-\beta\left(E+pV\right)\right] \tag{6.3}$$

where Ω_{conf} is the configurational density of states[88] and the summation is now over all energy levels. For convenience, the molecular partition functions and *de Broglie* wavelengths of the solute and solvent will be set to unity. This choice can only be made provided that these terms take the same values in both the solid and solution phases or provided the two system share a common reference state. In this work, the common reference state for both systems is an ideal gas of fully formed molecules, whose rotation is unrestrained. The free energy of this system is then

$$\begin{aligned} G(T,p,N_{solute},N_{solvent}) &= -\frac{1}{\beta}\ln Q(T,p,N_{solute},N_{solvent}) \\ &= -\frac{1}{\beta}\ln\left[\sum_E\sum_V \Omega_{conf}(V,E)\exp\left[-\beta\left(E+pV\right)\right]\right] \end{aligned} \tag{6.4}$$

Given that the density of states is independent of temperature and pressure, Equation 6.4 can in principle be used to determine the free energy for a range of temperatures and pressures, all from a single density of states calculation. Should the free energy be determined for a series of concentrations (enforcing the condition that number of solvent particles is fixed, and only the number of solute particles is allowed to vary), the solution chemical potential is found by fitting a polynomial as a function of N_{solute}, and analytically taking the derivative. As noted by Vega *et al*[116], a more accurate fit can be achieved by splitting the free energy into an ideal (G_{id}), and an excess (G_{ex}) component $G = G_{id} + G_{ex}$ where

$$\beta G_{id} = N_{solute}\ln\frac{N_{solute}}{V} - N_{solute} \tag{6.5}$$

and fitting the polynomial to the excess, rather than full free energy. The rationale behind this is that at low concentrations, the free energy profile is dominated by the log term of the ideal free energy, while the excess free energy varies more smoothly. In

fact, the excess free energy can be fitted to a good approximation to a second order polynomial, such that

$$G_{ex} = a_0 N_{solute}^2 + a_1 N_{solute} + a_2 \tag{6.6}$$

where a_0, a_1 and a_2 are coefficients to be determined by least squares fitting. The excess chemical potential is then

$$\mu_{ex} = 2a_0 N_{solute} + a_1 \tag{6.7}$$

and the full chemical potential is recovered by

$$\mu = \mu_{ex} + \frac{1}{\beta} \ln \frac{N_{solute}}{V} \tag{6.8}$$

where the righthand term is the ideal chemical potential.

The challenge then is to calculate the density of states of the system of solute and solvent for a range of concentrations. This can be accomplished by employing the method proposed previously by us[134]. In this approach, the DOS of the solute in solution is calculated for each concentration of interest. This DOS window must be large enough so as to encompass all possible and energies and volumes that would be available at each temperature / pressure at which the solubility will be calculated. A second set of DOS windows are then calculated, which extend the energy range sampled in the original windows to energies which would be accessible to the system at a temperature / pressure above the critical point. A third set of DOS windows are then calculated which extends the volume range sampled in the supercritical state to also cover the volume of some low density gas phase. In the gas phase, a DOS window spanning the entire concentration range of interest is calculated – employing solute insertion / deletion moves to transition between concentrations. The advantange of calculating the DOS in this way is two-fold: firstly, by first transitioning the system to a supercritical state, the system may then be transitioned to the gas phase without having to undergo a first order transition (which are known to challenge simulation); secondly, as the insertion / deletion moves are performed in the gas phase, there will be sufficient space to insert solute molecules into (the volume of the gas phase can simply be expanded further to facilitate larger solutes) without them overlapping with existing molecules in the system (a problem commonly encountered when employing such moves).

6.3 Chemical potential of solid

The chemical potential of solid urea was calculated using the Einstein molecule method. The reference state in this calculation is an ideal Einstein lattice of fully formed urea molcules, which are restrained to their lattice sites by harmonic potentials of the form

$$U_{pos} = \sum_{i}^{N_{solid}} K(r_{i,C} - r_{0,i,C})^2 \tag{6.9}$$

where N_{solid} is the number of molecules in the solid, K is the spring stiffness, and $r_{i,C}$, $r_{0,i,C}$ are the instantaneous and lattice positions of the carbon atom of urea molecule i respectively. Here the restraints are attached to the central carbon atom as a good approximation of the molecule's centre of mass. To prevent the diffusion of the centre of mass of the system (see Frenkel and Ladd[42] for the reasoning behind this), the position of one of the carbon atoms in the system is kept fixed. The free energy of this reference state is then

$$\frac{\beta}{N_{solid}} A_0 = \frac{3}{2}\left(1 - \frac{1}{N_{solid}}\right) \ln \frac{\beta K}{\pi} + \frac{1}{N_{solid}} \ln \frac{N_{solid}}{V_{solid}} \tag{6.10}$$

where V_{solid} is the volume of the solid[91]. To ensure parity with the solution phase calculations, the molecular partition functions and *de Broglie* wavelength terms are chosen here to be unity. This reference state is transformed into the full, unrestrained crystal by three successive steps, such that the total free energy of the crystal is given by

$$A_{solid} = A_0 + \Delta A_0 + \Delta A_1 + \Delta A_2 + A_{sym} \tag{6.11}$$

The first step is to introduce two extra tethers per urea molecule that effectively fix its orientation. The free energy change associated with this step is calculated by thermodynamic integration

$$\Delta A_0 = \int_{\ln c}^{\ln(K+c)} \left\langle \sum_{i}^{N_{or}} (r_i - r_{0,i})^2 \right\rangle_K (K+c) \mathrm{d}\ln(K+c) \tag{6.12}$$

where N_{or} is the total number of atoms that will be restrained by these new orientational tethers. In the case of urea, these restraints are attatched to each of the nitrogen

atoms, so that $N_{or} = 2N_{solid}$. The extra constant $c = \exp[3.5]$ is introduced to improve the accuracy of the integral[33]. The second step is to reintroduce the intermolecular interactions. The free energy difference between the ideal, non-interacting crystal and the fully interacting one (ΔA_1) is calculated by free energy perturbation

$$\beta\Delta A_1 = \beta U_{lattice} - \ln\langle\exp\left[-\beta\left(U_{solid} - U_{lattice}\right)\right]\rangle \tag{6.13}$$

where the average is evaluated over configurations sampled employing the ideal Hamiltonion (i.e one that only evaluates the tethered and intramolecular interactions), U_{solid} is the instantaneous energy of the solid evaluated using the full system hamiltonion and $U_{lattice}$ is the energy of the perfect lattice. The final step involves removing all restraints from the system, where the corresponding free energy change ΔA_2 is calculate by thermodynamic integration

$$\Delta A_2 = -\int_{\ln c}^{\ln(K+c)} \left\langle \sum_{i}^{N_{tethers}} \left(r_i - r_{0,i}\right)^2 \right\rangle_K (K+c)\mathrm{d}\ln(K+c) \tag{6.14}$$

where $N_{tethers}$ is the total number of restrained atoms (for urea $N_{tethers} = 3N_{solid}$). The final term in Equation 6.11, A_{sym}, accounts for the orientation field not having the same symmetry as the molecule of interest[130]. As urea has a point group of C_{2v}, $\beta\Delta A_{sym} = -N_{solid}\ln 2$

6.4 Technical details

The solubility of urea in methanol, and urea in water was explored as a function of temperature using the above methodology. 125 methanol and between 1 and 20 urea molecules were employed in the methanol solution calculations, spanning a concentration range of 0.25-5.00 mol kg^{-1}, while 216 water and between 1 and 9 urea molecules were employed in the aqueous calculations, spanning a concentration range of 0.26-2.31 mol kg^{-1}. The Amber GAFF force field[136] was used to model the urea and methanol interactions while the TIP3P water model was employed[62]. The urea and water molecules were treated as rigid bodies. For the density of states calculations, an energy bin size of 10 kJ mol^{-1} was used, while a logged volume bin size of 0.008 and 0.011 was used for the methanol and aqueous systems respectively. Gas phase volumes of 73617.7$\mathring{A}$ and 25592.7$\mathring{A}$ were used for the methanol and aqueous studies respectively. These values

were chosen by trial and error so that the employed grand canonical insertion / deletion moves were easily accommodated. The Wang-Landau modification factor was allowed reduced to 1×10^{-6}, at which point the results were well converged. For the thermodynamic integration calculations, the general scheme utilising soft core potentials proposed by Shirts and Pande[137] was followed. A 16 point Gaussian quadrature was employed to evaluate both the van der Waal and Coulomb integrals. All solution phase calculations were performed using our in house Monte Carlo simulation code.

The chemical potential of the solid phase was calculated at 298, 308, 318, 328 and 338 K. The structure of crystalline urea was taken from the Cambridge Structure Database[138] (reference code UREAXX29). From this a $4 \times 4 \times 4$ crystal was constructed. The cell vectors and angles were equilibrated at each temperature of interest by molecular dynamic simulations performed using DLPOLY 4.07[68]. Simulations were ran for 100000 steps with a timestep of 5 fs in the 'nst' ensemble (cell lengths / angles were allowed to vary anisotropically) using a Nose-Hoover thermostat and barostat. In each case the angles of the box, while permitted to change, remained orthogonal to a good degree. For each temperature, a perfect lattice with the equilibrated cell lengths was constructed. The remainder of the simulations were perfomed using our in house Monte Carlo code. The integrals in Equations 6.12 and 6.14 were evaluated using a 32- and 16- point Gauss-Legendre quadrature respectively.

6.5 Results and discussion

The density of states of urea in methanol, and of urea in water was calculated for a range of concentrations by employing the procedure outlined previously by us[134]. The free energies (to within an unknown constant arising from the DOS being also calculated to within an unknown multiplicative constant) of both systems, for each concentration studied, were obtained at 298 K by weighting these DOS surfaces according to Equation 6.4, and are given in Tables 6.1 and 6.2. The excess free energies were calculated from these by subtracting the ideal component (Equation 6.5), which were then fitted to a polynomial of the form of Equation 6.6 by the least squares method. The coefficients of the fit are given in Table 6.3. The excess solution chemical potential was calculated from the fitted coefficients according to Equation 6.7, from which the total chemical potential of solution (presented in Tables 6.1 and 6.2, and graphically in Figure 6.1) was determined. In addition, the chemical potentials of both systems were determined by thermodynamic integration, and are also shown in Figure 6.1.

Table 6.1: The solution free energies of urea in water calculated at 298 K to within an unknown (but identical for each concentration) additive constant, and the chemical potentials calculated from the polynomial fitted to this data.

N_{solute}	V / $\AA^3$	G / kJ mol^{-1}	G_{id} / kJ mol^{-1}	G_{ex} / kJ mol^{-1}	μ_{id} / kJ mol^{-1}	μ_{ex} / kJ mol^{-1}	μ / kJ mol^{-1}
1	6583.17	-1179.05	-24.26	-1154.78	-21.78	-52.96	-74.75
3	6719.78	-1325.22	-64.77	-1260.45	-19.11	-53.27	-72.39
5	6860.39	-1469.71	-101.88	-1367.83	-17.90	-53.59	-71.48
7	6993.73	-1612.55	-137.13	-1475.42	-17.11	-53.90	-71.01
9	7135.35	-1754.31	-171.15	-1583.16	-16.54	-54.21	-70.75

Table 6.2: The solution free energies of urea in methanol calculated at 298 K to within an unknown (but identical for each concentration) additive constant, and the chemical potentials calculated from the polynomial fitted to this data.

N_{solute}	V / $\AA^3$	G / kJ mol^{-1}	G_{id} / kJ mol^{-1}	G_{ex} / kJ mol^{-1}	μ_{id} / kJ mol^{-1}	μ_{ex} / kJ mol^{-1}	μ / kJ mol^{-1}
1	8235.32	-384.10	-24.82	-359.29	-22.34	-53.92	-76.26
3	8353.55	-531.07	-66.39	-464.68	-19.65	-54.12	-73.77
5	8461.38	-678.98	-104.48	-574.51	-18.42	-54.32	-72.74
7	8569.37	-824.89	-140.65	-684.24	-17.62	-54.52	-72.14
10	8748.58	-1040.59	-192.61	-847.98	-16.78	-54.82	-71.61
15	9066.09	-1397.90	-275.17	-1122.73	-15.87	-55.33	-71.19
20	9376.21	-1755.18	-354.30	-1400.88	-15.24	-55.83	-71.07

It can be seen that there is an excellent agreement between the two methods for both systems. This gives a good degree of confidence that the DOS approach is indeed able to accurately calculate the chemical potential of even molecular systems, in addition to simple ionic ones. Although urea is a relatively small molecule, the creative pathway employed when determining the density of states as a function of concentration appears to transfer well to molecules without modification. While standard grand-canonical simulations are challenged when performing insertion moves involving molecular species (due to the inserted molecules overlapping with existing molecules in the system leading to unfavourable high energy states), we have demonstrated that our employed pathway of first vapourising the solution, and then performs all solute insertion / deletion moves in the gas phase overcomes this limitation. It would seem then that the method should continue to scale well as larger molecules are considered.

For both systems the free energies were further calculated at 308 K, 318 K and 328 K (as well as 338 K for the aqueous system) by reweighting the DOS used in the 298 K calculations according to Equation 6.4. The excess portion of these were then fitted to polynomials of the form given in Equation 6.6 (the coefficients of which are given in

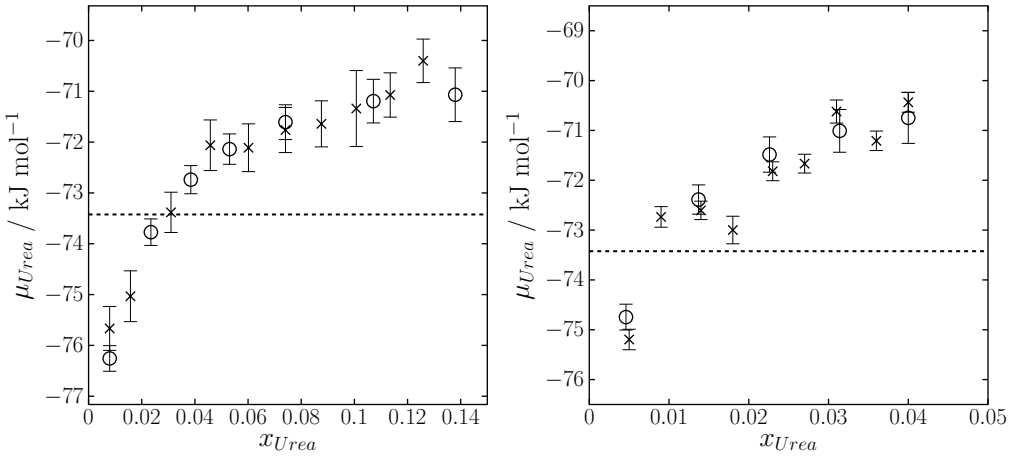


Figure 6.1: The total chemical potential of urea in methanol (left) and urea in water (right) as a function of molefraction of urea (x_{Urea}). Each cross represents a result calculated by thermodynamic integration, and each circle a result calculated by the DOS approach. The dashed horizontal line represents the chemical potential of the solid phase at 298 K, calculated by this work using the Einstein molecule method.

Table 6.3) yielding the chemical potentials shown in Figure 6.2. As would be expected, the chemical potential is seen to increase smoothly as a function of temperature for both systems.

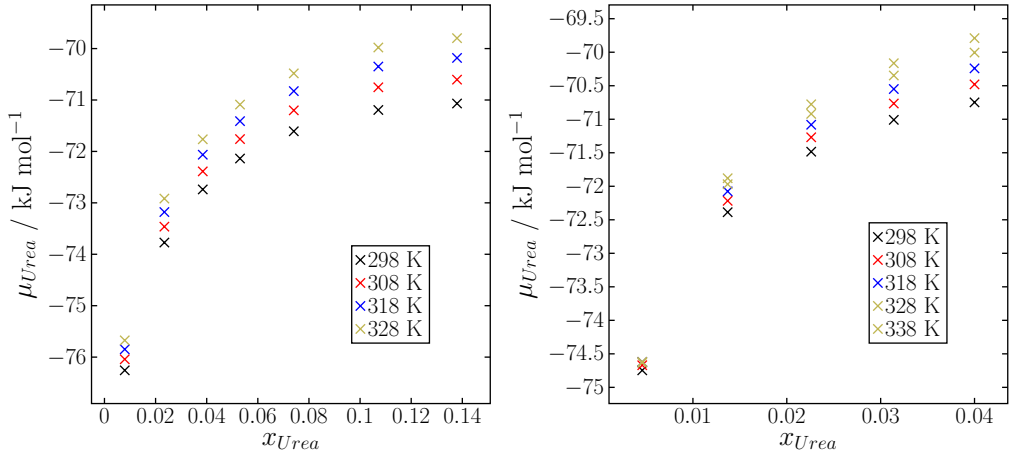


Figure 6.2: The chemical potentials of urea in methanol (left) and urea in water (right) for different temperatures.

In order to calculate the solubility of these systems, the chemical potential of the solid phase was calculated as a function of temperature using the Einstein molecule method. The results of these calculations are presented Table 6.4, and graphically in Figure 6.3. The solubility at each temperature was then determined by finding the point of intersection between the solid and solution curves. To a good approximation, the volume of the solution phase can be fitted to a second order polynomial of the form

Table 6.3: The coefficients calculated by fitting the excess free energies calculated by the DOS approach fit to Equation 6.6.

T	Methanol			Water		
	a_0	a_1	a_2	a_0	a_1	a_2
298	-0.05 (12)	-53.82 (25)	-304 (1)	-0.078 (25)	-52.81 (25)	-1101.72 (54)
308	-0.05 (11)	-52.82 (25)	-371 (1)	-0.077 (22)	-51.98 (22)	-1162.25 (48)
318	-0.051 (12)	-51.84 (25)	-437 (1)	-0.076 (21)	-51.18 (21)	-1222.61 (45)
328	-0.051 (12)	-50.87 (26)	-503 (1)	-0.072 (20)	-50.42 (21)	-1282.78 (44)
338	-	-	-	-0.069 (19)	-49.68 (20)	-1342.85 (42)

Table 6.4: The individual components of the solid phase free energies as calculated by the Einstein molecule method.

T	$\frac{\beta A_{sym}}{N_{solid}}$	$\frac{\beta A_0}{N_{solid}}$	$\frac{\beta \Delta A_0}{N_{solid}}$	$\frac{\beta \Delta A_1}{N_{solid}}$	$\frac{\beta \Delta A_2}{N_{solid}}$	$\frac{\beta A_{solid}}{N_{solid}}$
298	-0.693	11.639	18.280	-41.676	-17.185	-29.636
308	-0.693	11.639	18.281	-40.303	-17.303	-28.380
318	-0.693	11.639	18.281	-39.010	-17.427	-27.211
328	-0.693	11.639	18.281	-37.796	-17.546	-26.116
338	-0.693	11.639	18.281	-36.651	-17.666	-25.091

$$V_{solution} = b_0 N_{solute}^2 + b_1 N_{solute} + b_2 \tag{6.15}$$

where b_0, b_1 and b_2 are coefficients found a least squares fitting procedure (Table 6.5). Combining Equations 6.1, 6.6 and 6.15 yields

$$2a_0 N_{solute} + a_1 + \frac{1}{\beta} \ln \frac{N_{solute}}{b_0 N_{solute}^2 + b_1 N_{solute} + b_2} = \mu_{solid} \tag{6.16}$$

which can easily be solved for the solubility limit by applying the Newton—Raphson algorithm. Three iterations were required for the algorithm to convergence. The solubilities calculated by this approach are presented in Figure 6.4.

Table 6.5: The coefficients calculated by fitting the solution phase volumes calculated by the DOS approach fit to Equation 6.15.

T	Methanol			Water		
	b_0	b_1	b_2	b_0	b_1	b_2
298	0.288 (69)	54 (2)	8183 (6)	0	68.91 (32)	6514 (2)
308	0.315 (54)	54 (1)	8293 (5)	0	69.77 (21)	6569 (1)
318	0.301 (56)	54 (1)	8405 (5)	0	70.64 (11)	6628.99 (64)
328	0.203 (68)	56 (1)	8517 (6)	0	71.2 (14)	6695.69 (78)
338	-	-	-	0	71.53 (21)	6768 (1)

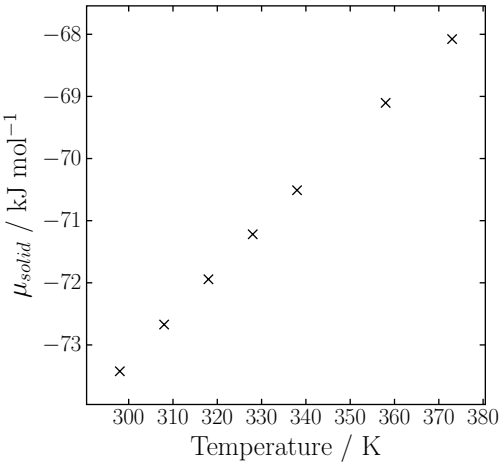


Figure 6.3: The chemical potential of solid urea as a function of temperature.

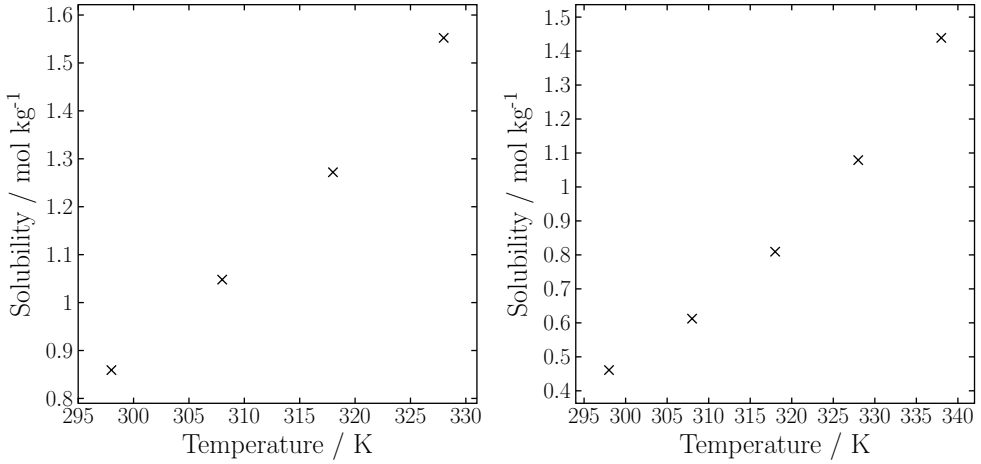


Figure 6.4: The solubility of urea in methanol (left) and water (right) as a function of temperature.

For both systems the calculated solubility of the model is markedly lower than experiment – at 298 K urea has an experimental solubility of 4.11 mol kg^{-1} in methanol and 20.15 mol kg^{-1} in water. While the calculated solubility in methanol (0.86 mol kg^{-1}) is roughly 4 fold lower than experiment, the aqueous solubility (0.46 mol kg^{-1}) is more significantly underestimated by roughly 40 fold the experimental value.

There are two potential sources of this departure from experiment: i) the method itself or ii) the computational model. We argue that the source of this discrepancy is most likely the latter, rather than the former. The solution phase chemical potentials of the two systems were calculated by both the proposed DOS approach, and well established TI calculations for a number of concentrations, and where found to be in excellent agreement (to within the statistical uncertainties of each method). If the DOS approach was flawed,

a strong departure from the TI values would be expected. The more likely explanation is that the model parameters employed to represent either urea or water were insufficient; it has been commonly observed now that existing force field parameters are challenged by solubility, even though they may be able to reproduce well other properties of molecules (such as their melting point)[107,120]. Further, it should be considered that only a single model, using a single charge set was considered for all species. It may be the case that one of the other available models (such as the model proposed by Duffy *et al*[139]) may better reproduce the solubility of urea. In future we aim to explore this further, and aim to identify whether it is the solid, solution or both phases which leads to the poor solubility.

These results further highlight how critical it is to have a robust and efficient method for calculating a model's solubility, as such a method will be invaluable in not only testing, but also helping to optimise existing force fields. A higher quality of force field is clearly required if solubility prediction from classical molecular simulations are to be routinely performed (or more importantly, trusted). This must be one of the major focuses of future work. To accommodate this, the method proposed here would in future would benefit from extension to larger, and more flexible systems. It is expected that application of the method to larger solute molecules will not be an issue, as the only change that is anticipated to be made is to increase the volume of the gas phase in the DOS calculations. Furthermore the method is expect to work well in combination with configurational bias Monte Carlo simulation moves[140], which would enhance sampling of the internal degrees of freedom of flexible molecules, and in turn, would enable the solubility of even flexible molecules to be determined with a good degree of accuracy.

To conclude, in this work we have shown how the chemical potential of a system may be determined from density of states calculations, demonstrating that the accuracy of the method is comparable to that of more established methods. The employed method, while preciously only applied to ionic systems, is shown to map well on molecular systems. Further, it was demonstrated that the method is able to calculate chemical potentials for a range of temperatures from knowledge of a single density of states surface. The ability to calculate which has led to the calculation of the solubility of urea in methanol, and urea in water as a function of temperature from a relatively modest number of simulations compared to what would be required for more traditional approaches.

Chapter 7

Concluding Remarks

The main aim of this thesis has been to gain molecular level insights into phase equilibria, and to produce methodologies to predict phase stability with an emphasis on solid-solid and solid-liquid equilibria. More specifically, the thesis aimed to address three fundamental issues. Firstly, it aimed to explore what drives multicomponent crystals to form. Secondly, it aimed to overcome a number of limitations present in current simulation methods used to explore phase coexistence. Finally, it aimed to develop methodology to calculate solubility from first principles as a potentially more powerful approach than correlation based methods. These aims have been accomplished by using a combined molecular simulation / statistical mechanics approach.

As a direct result of this work: a framework for understanding how the affinities between molecules and their packing complementarity facilitate solvate formation has been constructed; a set of coarse-grained interaction potentials with large application to studying phase equilibria and much more have been characterised; and a computationally robust, from first principles, methodology for solubility prediction of molecular and ionic systems was developed.

The solvates study presented in Chapter 3 has identified three key regimes in which solvates are expected to form. The solvate is expected to be favoured provided that i) the solute-solvent affinity is strong enough to overcome the combined solute-solute and solvent-solvent self affinities; ii) there is sufficient complementarity in solute-solvent packing such that voids large enough to incorporate solvent molecules are formed or iii) a strong $p\Delta V$ potential is required in cases where the solute-solvent affinity is weak relative to the solute-solute and solvent-solvent affinities. The identification of these regimes lays the foundation in future for designing guidelines for determining when solvates

may form. The existence of such guidelines would be fundamental in identifying where solvate formation may be an issue far in advance, allowing interventions or alternative approaches to be devised. Further, although initially derived in the context of solvates, the findings are expected to be applicable to co-crystals. An understanding of what may constitute a 'good' co-former could be a significant utility to drug development process, where co-crystallisation is becoming more popular. Furthermore, the observation that a strong $p\Delta V$ potential is able to facilitate solvate formation even in the absence of strong affinities offers an explanation as to why gas clathrates (where a strong affinity is lacking) are able to form. It thus may also offer a future pathway for forming solvates and co-crystals when traditional methods fail.

While the work in Chapter 3 provides a strong starting point for understanding the driving force behind solvate formation, it would in future benefit from the gradual re-introduction of molecular detail, such as the presence of specific molecular moieties. The challenge will be to identify how best to map the broad properties of affinity and size ratio (packing) onto a molecular language. The hydrogen bond descriptors or indicators of polarity that are commonly employed are unlikely to capture the formation of low affinity solvates, for example. The work would most likely proceed, at least in part, by a targeted survey of the Cambridge structural database (CSD). In addition, a number of limitations would be addressed. The first is to investigate solvate formation as a function of temperature. The Lennard–Jones model employed has a limited range over which it is liquid. As such, the influence of temperature could not be fully explored. Further, the produced phase diagrams contain a handful of anomalous points arising from kinetic trapping of unfavourable phases. Should a similar methodology be employed to study solid-solid phase phenomena, such as why do some molecules form co-crystals, this trapping would likely be emphasised. Finally, the solid phase only emerged at high concentrations. As the system was predominantly solute - there was always a mix of solvate and anhydrous forms in the simulation box. Ideally, a model with a lower solubility relative to its melting point would be employed, so that individual crystallites are observed. Not only would this make identifying the presence of solvates as opposed to pure crystals easier, but would also help overcome the issues of kinetic trapping.

The limited boiling range and the kinetic trapping observed in the solvate study can in part be attributed to the relative 'hardness' of the Lennard–Jones potential. To address this, this work aimed to characterise a set of softer potentials. To ensure the models employing these potentials have a strong physical grounding (the solute parameters were chosen in Chapter 3 to have properties similar to a typical organic molecules for example)

their phase diagrams must be determined. This was done in the work presented in Chapter 4. This work has developed a much clearer understanding of where the limited liquid range of the Lennard–Jones potential arises from - namely that the dispersive and attractive exponents are too high. The broad range of behaviour of the characterised potentials opens the possibility of constructing a better class of coarse-grained force field. A force field using these potentials would not only be able to reproduce a wide range of chemical specificity, but would have widespread application in studying the behaviour of most phases over a broad range of conditions. Future work will be focussed on constructing such a force field. There are two key challenges to doing this: the first will be determining which potential on average is best able to reproduce the interactions of a wide range molecular moieties. Once identified, homogeneous interaction parameters can be extracted directly from the calculated phase diagram of the potential. The second challenge will be to determine the heterogeneous interaction parameters - while the Lorentz–Berthelot mixing rules are applied in most situations to calculate these, a cursory investigation has revealed that they are in fact not suitable for use in coarse-graining. It was found that the affinity between beads predicted by these rules was too strong. It is anticipated that a new set of rules will need to be derived - this will most likely be achieved by an empirical study of the interactions of many types of chemical moieties, and performing some form of fitting procedure.

The final pieces of work presented in Chapters 5 and 6 were focused on developing a robust, accurate and efficient method for solubility prediction from first principles. The developed method is, in principle, capable of calculating the solubility of even large, flexible molecules for a large number of temperatures and pressures, from a reasonably modest number of simulations. The applicability of the method to even large molecules is possible due to the use of a creative sampling pathway, that overcomes the common issues associated with grand-canonical (and similar) simulations. Further, the methods ability to calculate solubilities over a wide range of temperatures and pressures significantly reduces the number of simulations that must be run. The method has been shown to be successful at not only predicting the solubilities of both a molecular and ionic system, but also to offer a robust and accurate route to calculating chemical potentials of fluid phases in general. Further, the method will serve as a useful tool for benchmarking the accuracy of existing force fields, and hence, would have future application in parameter development and optimisation. The next step in developing this method forward is to apply it to larger, more flexible molecules - the transition to which is expected to be smooth. The employed insertion moves should be able to accommodate even large

molecules, and the lack of any thermal barriers, especially when combined the configuration bias moves, should enable even largely flexbile molecules (such as pharmaceutical drug molecules) to be studied. Finally, in the future the method will be applied to systems containing multiple solute / solvent species. This would have applications in the development, and predicting the properties of co-crystals and even solvates, and hence has the potential to be a large utility in the drug development process and numerous other applications.

To summarise then, the work compiled in this thesis has advanced the calculation and understanding of solid-liquid phase equilibria. The thesis has offered insight into why molecules are able to form multicomponent crystals, providing a solid foundation for studying the phenomenon further, has characterised a set of more molecular like coarse-grained interaction potentials for future use in exploring phase equilibria, and finally, has presented a method for predicting solubility - arguably one of the most important properties of a system arising from phase stability.

Appendix A

Monte Carlo Simulation Code

All Monte Carlo simulations utilised by this thesis were performed using a bespoke Monte Carlo simulation code, nicknamed PhaseMC. It was written in C++ entirely by myself over a three year period, and consists of over 10,000 lines of code. Access to the source code will be provided upon request.

The core aim of PhaseMC is to facilitate the calculation of phase equilibria, and properties of phases. It supports simulating in the most common ensembles:

- NVT
- NpT
- μVT

and implements a number of advanced techniques such as

- Wang-Landau sampling,
- Gibbs-Ensemble simulations
- Einstein crystal calculations
- Thermodynamic integration calculations

The code is relatively flexible in terms of the type of system that can be studied. So far it has been applied to simple Lennard–Jones type coarse-grained systems, rigid molecules,

ionic species and a limited number of small flexible molecules. Electrostatic interactions are fully supported, and implemented as Ewald summations.

PhaseMC is currently limited to studying relatively small simulations (<1000 molecules). This is in part by design, as development time has predominantly been focused on feature development, and rigourous testing. At present there is no parallelisation for efficiency gain, and no neighbour lists are implemented.

The following sections aim to provide an overview of the key structures of the code, as well as a brief description of the input files required to run simulations with PhaseMC.

A.1 Coding Overview

PhaseMC is written in an object orientated style. The advantages of this are many. Object orientated code is significantly easier to maintain due to reduced code redundancy, is often simple to understand and navigate, and perhaps most importantly, is readily and rapidly extensible. A heavily simplified overview of the code structure is presented in Figure A.1

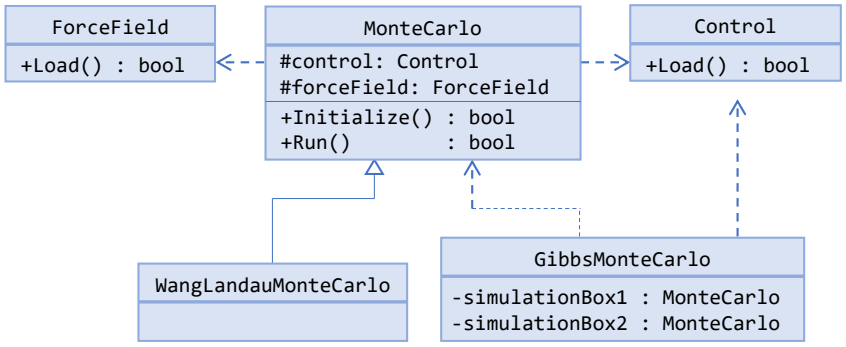


Figure A.1: A simplified overview of the PhaseMC code structure. Most methods / fields are omitted for clarity

There are three main kernel classes (`MonteCarlo`, `WangLandauMonteCarlo` and the `GibbsEnsembleMonteCarlo`) that implement the Monte Carlo simulation loops for the different sampling approaches.

These are supported by a `ForceField` and `Control` class. The `ForceField` class is responsible for loading and maintaining the systems force field. It contains helper methods to evaluate each term in the potential energy function. The `Control` class stores

and loads any parameters (such as temperature or number of steps) required by the simulation.

Given that the kernel classes form the heart of the code, they are discussed in more detail in the following subsections.

A.1.1 Metropolis Monte Carlo

The foundation of PhaseMC is the Metropolis Monte Carlo kernel, implemented in the `MonteCarlo` class. It's key methods have been highlighted in Figure A.2

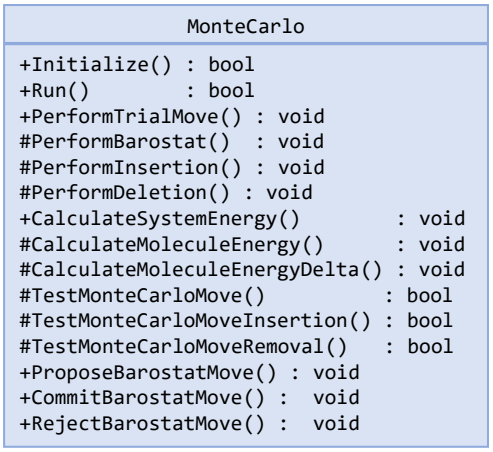


Figure A.2: An overview of the `MonteCarlo` class. Only a selection of key methods are presented.

The Initialize method is called before any simulation is started. It is responsible for reading and validating all input files. This includes loading the force field, the simulation control file and the atomic coordinate file. Further, it is responsible for constructing all atomic (and any other required) arrays.

Provided that `Initialize` is successful, the `Run` method is called. `Run`, as the name suggests, is responsible for maintaining the main simulation loop. It proceeds according to the flow chart shown in Figure A.3. Key here are the different `Perform...` methods.

Each of the `Perform...` methods is an implementation of a Monte Carlo trial move - `PerformTrialMove` handles molecule / atom displacements and rotations, `PerformBarostat` performs the box scaling barostat moves, and `PerformInsertion` / `PerformDeletion` are responsible for any μVT insertion / deletion moves. They are responsible for generating any new configurations, calculating the changes in energy, volume, particle number and then deciding whether to accept or discard the move.

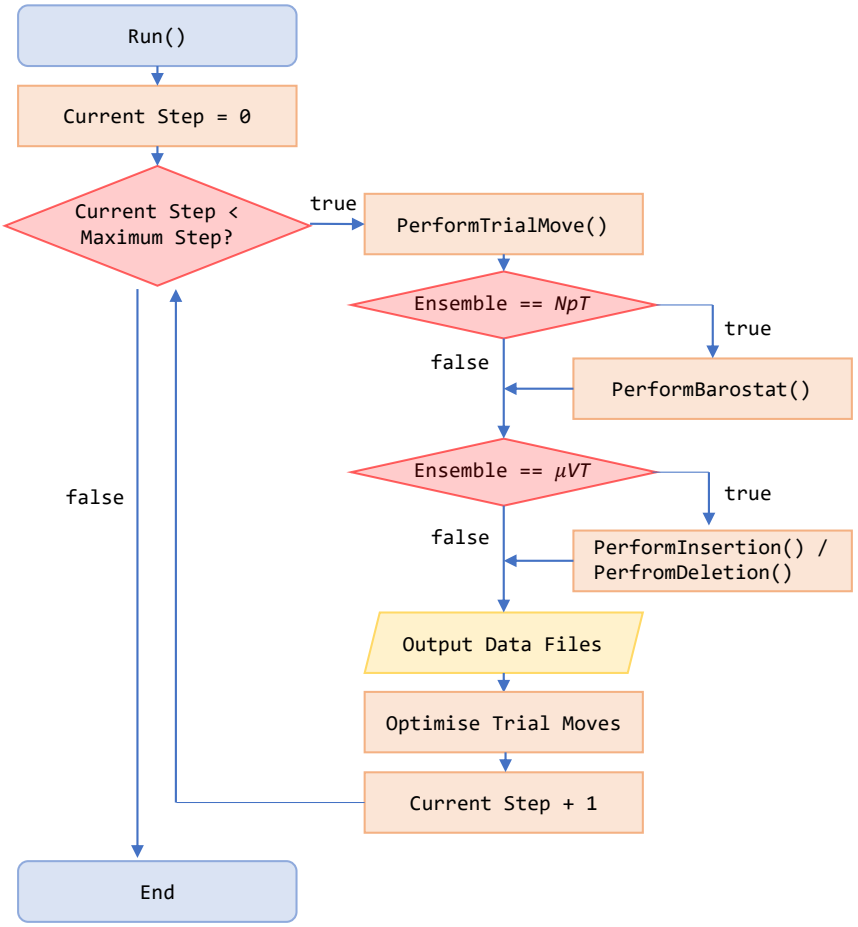


Figure A.3: A flow diagram of the main simulation loop implemented by the `MonteCarlo` class.

The coordinate generation is move specific, and not particularly interesting so will not be discussed here. More important is the energy calculation. Calculating energies is arguably the slowest, but one of the most critical part of any Monte Carlo code. In PhaseMC, three different methods to calculate energy are provided:

- one for calculating the energy of the entire system (`CalculateSystemEnergy`).
- one for calculating the change in energy of a single molecule (`CalculateMolecule-EnergyDelta`).
- one for calculating the absolute energy of single molecule (`CalculateMolecule-Energy`).

Combined, these methods enable almost all changes in the systems energy to be calculated in an optimised way regardless of the trial move. As such, implementing new

moves only requires a new method for generating or perturbing the systems coordinates, and one to implement the Metropolis acceptance / rejection criteria.

The changes in energy (and other state variable such as volume) are passed to an appropriate `Test...` method. These methods implement the Metropolis acceptance / rejection criteria, and ensure detailed balance is satisfied. They return a back a simple *true* or *false* boolean, which determine whether to accept or reject the move. They are implemented as virtual functions that can be overridden. This allows any child kernel class, such as `WangLandauMonteCarlo`, to implement their own variations, and hence sample from a probability distribution of their choosing without having to reimplement all of the trial move code. This significantly reduces the amount of redundant code, and makes extension to other ensembles simple.

The `Perform...` method takes the result of the acceptance or rejection criteria and then either returns the system back its orginal state before the move or copies the new state over the old one. Implementation wise, two sets of structures are maintained - one set that stores the current, accepted system state and another that stores any proposed changes. This avoids the need to performing any form of caching prior to the move being made, and further means that when a move is rejected, nothing needs to be done. Accepting a move simply involves copying the new state over the old.

Over the course of the simulation, a set of counters are maintained that tracks the frequency with which moves are either accepted or rejected. These are used to optimise the moves on the fly to acheive a roughly 50% acceptance ratio. In the case of particle translations, for example, the size of the displacement is optimised.

At the end of the simulation, the final system energy is recalculated from scratch and compared to the final energy reported by the main loop. Agreement shows that the various energy calculations (such as the change in molecular energy) are performing as expected. A deviation means there is a problem in the code and hence serves as a useful diagnostic.

A.1.2 Wang-Landau Monte Carlo

The Wang–Landau sampling method is implemented in the `WangLandauMonteCarlo` class, whose key structure is shown in Figure A.4.

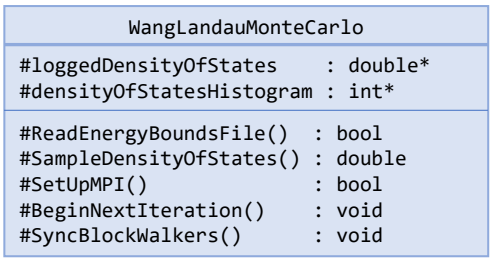


Figure A.4: An overview of the `WangLandauMonteCarlo` class. Only a selection of key methods are presented.

It inherits the `MonteCarlo` class. As was alluded to, the Wang–Landau method is mainly implemented by overriding the different `Test...` methods. The following additions are also required:

- A density of states array and a histogram array is defined and updated after each trial move.
- The `Perform...` methods are also overridden so that extra checks are performed that ensures the state of the system does not depart from the defined sampling window.

To run a Wang–Landau Monte Carlo simulation a sampling window must be defined. In the NVT ensemble this is simply and energy range, in the NpT and μPT ensembles however both an energy and volume range must be specified. These are stored in the `Control` file. In addition to defining a global window, the accessible energies for individual densities are defined using the `EnergyBounds.dat` input file, read by the `ReadEnergyBoundsFile` method.

A large extension made by the `WangLandauMonteCarlo` class is the introduction of MPI threading. It has two functions in the code. The first is to split the density of states window into blocks. Each block is assigned an MPI thread that runs a separate copy of the simulation, but samples in a different region. The second is to assign multiple walkers to sample the same region. This improves both the efficiency and precision of the algorithm. All MPI variables are initialised in `SetUpMPI`. At the end of each WL iteration, all walkers within a block are forced to wait until the others in that block have also reached the end of their iteration. The density of states from each walker is then sent to a dedicated 'lead' walker. This lead walker superimposes and averages the different DOS windows, and returns the combined one back to each walker ready for

the next iteration[124]. This is implemented in the `SyncBlockWalkers` method. The next iteration is started by the `BeginNextIteration` method.

A particularly important routine is `SampleDensityOfStates`. This method is used to sample the density of states for a set of coordinates (e.g for a given (E, V) pair). It employs either nearest neighbour, linear or bilinear sampling. While performing different sampling methods is trivial on rectangular surfaces, it is more complex for irregular shaped surfaces. At a high level, the method proceeds by discretising the systems state into a density of states bin. It then checks whether each of the neighbouring bins is within the accessible sampling window. Depending on the number of available neighbours, either bilinear, linear or nearest neighbour sampling is employed to sample the density of states.

PhaseMC is complemented by a separate toolkit of utilities, named the DOStoolkit. These include tools that

- stitch multiple density of states windows into a single one.
- reweight a density of states window to produce probability distributions at different T and P and μ.
- detect bimodal probability distributions and evaluating the difference in their probabilities.
- determine the free energy of phases / different concentrations.

A.1.3 Gibbs-Ensemble Monte Carlo

The Gibbs-Ensemble Monte Carlo sampling mode is implemented as a separate entity to the `MonteCarlo` class. It's key methods and properties are shown in Figure A.5.

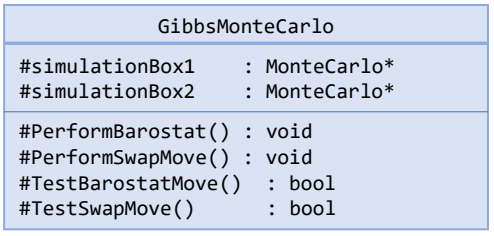


Figure A.5: An overview of the `WangLandauMonteCarlo` class. Only a selection of key methods are presented.

Within the class, two MonteCarlo simulation objects are created, and initialised using identical control and topology files, but different configurations. The `GibbsMonteCarlo`

class implements an entirely separate main loop, bypassing those of the individual `MonteCarlo Run` methods. It proceeds according to Figure A.6.

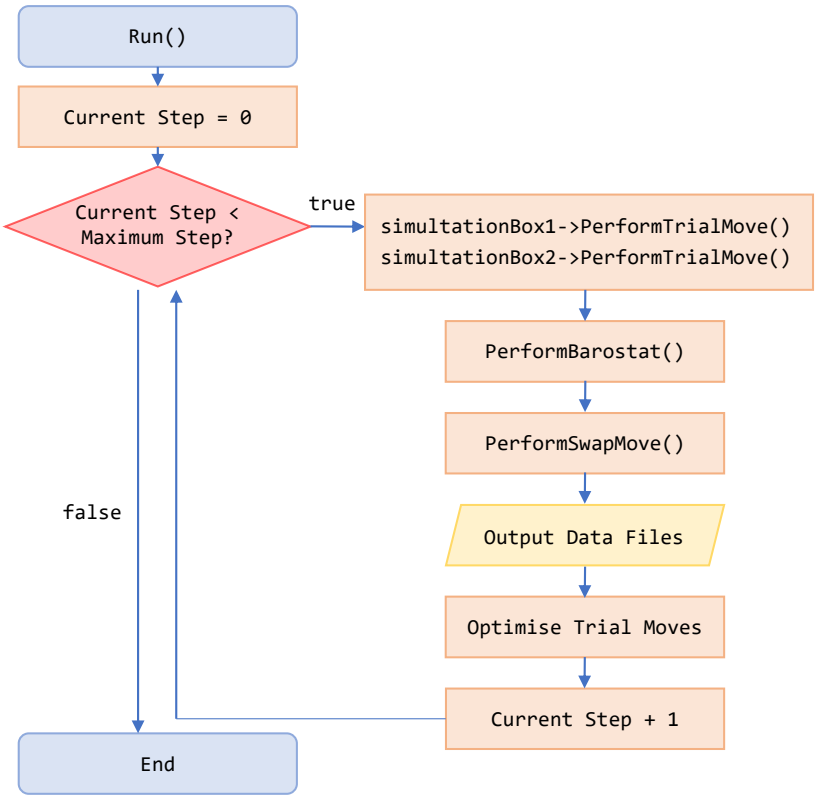


Figure A.6: A flow diagram of the main simulation loop implemented by the `GibbsMonteCarlo` class.

Within the loop, the individual `PerformTrialMove` methods of each object are called. This is followed by calls to the `GibbsMonteCarlo` implementation of the `PerformBarostat` and `PerformSwapMove` moves. These are almost idential to the same class of methods implemented by `MonteCarlo`. They do not however directly alter the coordinates of the two `MonteCarlo` objects. Rather, they call the objects `Propose...` methods. These essentially force each object to generate a new set of coordinates, pass back the changes in state, but not enforce the acceptance / rejection criteria. This is instead performed by the `GibbsMonteCarlo` implementation of the `Test...` methods. Acceptance / rejection of the move is then enforced by calling the individuals objects `Commit...` or `Reject...` methods. This setup results in a remarkably simple and clean `GibbsMonteCarlo` implementation, that is easily expanded to an arbitrarily number of coexisting boxes.

A.2 Input Files

Within this section, the required input files needed to run a simulation using PhaseMC are presented.

A.2.1 CONFIG File

The `CONFIG` file defines the dimensions of the simulation box, and contains the coordinates, and types of all atoms in the system. It is identical to the `CONFIG` file used in DLPOLY simualtions[68]. A full description of the file format can be found in the DLPOLY 4 user manual.

An example `CONFIG` file for a system of Lennard–Jones particle in a cubic simulation box is given below:

```
An example LJ system
0         1          5
6.8476934433e+00    0.0000000000e+00    0.0000000000e+00
0.0000000000e+00    6.8476934433e+00    0.0000000000e+00
0.0000000000e+00    0.0000000000e+00    6.8476934433e+00
S          1
0.070714            0.070714            0.070714
S          2
0.070714            0.070714            1.048956
S          3
0.070714            0.070714            2.027198
S          4
0.070714            0.070714            3.005440
S          5
0.070714            0.070714            3.983681
```

A.2.2 Control File

The Control file (`Control.inp`) contains all parameters for the simulation, such as the type of sampling mode to use, or the number of steps to run for. It is a free-formatted text file, consisting of a list of keywords, and values. An example file is given below:

```
mode                mc
ensemble            npt

temperature         269.42507
pressure            183.61707

cycles              100000

cutoff              3.0

trajectory          1000
statistics          100
```

A selection of the main keywords and a description of their role are presented below:

keyword	description
temperature *f*	The temperature of the system / K.
cutoff *f*	The cut-off radius for van der Waals interactions / Å.
cycles *n*	The number of cycles to run the simulation for.
ensemble	The ensemble to run in - one of nvt, npt, or muvt.
equilibration *n*	The number of equilibration steps.
ewald-cutoff *f*	The real space cut-off radius for electrostatic interactions / Å.
ewald-precision *f*	The precision of the ewald summation.
mode	The sampling mode to run in - one of mc, wl or gibb.
mu *f*	The chemical potential of the system (μVT) / kJmol^{-1}.
pressure *f*	The pressure of the system (NpT) / katm.
statistics *n*	The frequency with which to print the statistics about the system.
trajectory *n*	The frequency with which to print the simulation trajectory.

A.2.3 Topology File

The topology file (`Topol.top`) describes the chemical composition, and force field parameters that are to be used in the simulation. An example file is given below:

```
[ atom_types ]
#   name  charge
     OW  -0.8476
     HW   0.4238
[ end ]

[ molecule_type Water ]
    [ atoms 3 ]
    #  index  type
         1     OW
         2     HW
         3     HW
    [ end ]
    [ properties ]
        rigid
    [ end ]
[ end_type ]

[ intermolecular ]
# atomA atomB type param1 param2 param3 param4
# e.g.          lj    eps     sigma
# or            nm    eps     sigma      n      m
# or           bhm      A        B       C          D        sigma
     OW    OW   lj    650.000  3.166
     OW    HW   lj    000.000  0.000
     HW    HW   lj    000.000  0.000
[ end ]

[ system_molecules ]
    Water     200
[ end ]

[ reservoir_molecules ]
    Water     0
[ end ]
```

A.2.4 Energy Bounds File

The energy bounds file (`EnergyBounds.dat`) is an optional input file used when performing Wang–Landau sampling simulations. It is used to define the accessible energy ranges for individual densities. It consists of four columns: the first index column is the index of volume bin to set the energy range for. Similarly the second column is the index of particle bin. The third and fourth columns are minimum and maximum energies of the range respectively, given in J mol^{-1}. An example `EnergyBounds.dat` file is given below:

```
0    0    -3719469    -3598791
1    0    -3690450    -3570256
2    0    -3661836    -3542092
3    0    -3633616    -3514290
4    0    -3605784    -3486846
5    0    -3578333    -3459751
```

References

[1] Dhami, N. K.; Reddy, M. S.; Mukherjee, A. Biomineralization of calcium carbonates and their engineered applications: a review. *Frontiers in Microbiology* **2013**, *4*, 314–327.

[2] Soldatov, D. V. *Stimuli-Responsive Supramolecular Solids: Functional Porous and Inclusion Materials*; 2005; pp 214–231.

[3] Newnham, R. E. Phase Transformations in Smart Materials. *Acta Crystallographica Section A Foundations of Crystallography* **1998**, *54*, 729–737.

[4] Herslund, P. J.; Thomsen, K.; Abildskov, J.; von Solms, N. Modelling of tetrahydrofuran promoted gas hydrate systems for carbon dioxide capture processes. *Fluid Phase Equilibria* **2014**, *375*, 45–65.

[5] Yang, M.; Song, Y.; Jiang, L.; Zhao, Y.; Ruan, X.; Zhang, Y.; Wang, S. Hydrate-based technology for CO2 capture from fossil fuel power plants. *Applied Energy* **2014**, *116*, 26–40.

[6] Sugahara, T.; Haag, J. C.; Prasad, P. S. R.; Warntjes, A. A.; Sloan, E. D.; Sum, A. K.; Koh, C. A. Increasing hydrogen storage capacity using tetrahydrofuran. *Journal of the American Chemical Society* **2009**, *131*, 14616–14617.

[7] Sugahara, T.; Haag, J. C.; Warntjes, A. A.; Prasad, P. S. R.; Sloan, E. D.; Koh, C. A.; Sum, A. K. Large-Cage Occupancies of Hydrogen in Binary Clathrate Hydrates Dependent on Pressures and Guest Concentrations. *The Journal of Physical Chemistry C* **2010**, *114*, 15218–15222.

[8] Markov, I. V. *Crystal Growth For Beginners: Fundamentals of Nucleation, Crystal Growth and Epitaxy*; World Scientific Publishing: Singapore, 1995.

[9] Sunagawa, I. *Crystals: Growth, Morphology and Perfection*; Cambridge University Press: Cambridge, 2005.

[10] Karthika, S.; Radhakrishnan, T. K.; Kalaichelvi, P. A Review of Classical and Nonclassical Nucleation Theories. *Crystal Growth & Design* **2016**, *16*, 6663–6681.

[11] Anwar, J.; Zahn, D. Uncovering molecular processes in crystal nucleation and growth by using molecular simulation. *Angewandte Chemie (International ed. in English)* **2011**, *50*, 1996–2013.

[12] Davey, R.; Garside, J. *From Molecules to Crystallizers*; Oxford University Press, 2000.

[13] Anwar, J.; Khan, S.; Lindfors, L. Secondary Crystal Nucleation: Nuclei Breeding Factory Uncovered. *Angewandte Chemie International Edition* **2015**, *54*, 14681–14684.

[14] Frank, F. C. The influence of dislocations on crystal growth. *Discussions of the Faraday Society* **1949**, *5*, 48–54.

[15] Florence, A.; Attwood, D. *Journal of Pharmaceutical Sciences*, 4th ed.; Pharmaceutical Press: London, 2006; Vol. 72.

[16] Baghel, S.; Cathcart, H.; O'Reilly, N. J. Polymeric Amorphous Solid Dispersions: A Review of Amorphization, Crystallization, Stabilization, Solid-State Characterization, and Aqueous Solubilization of Biopharmaceutical Classification System Class II Drugs. *Journal of Pharmaceutical Sciences* **2016**, *105*, 2527–2544.

[17] Bernstein, J. *Polymorphism in Molecular Crystals*; Clarendon Press: Oxford, 2002.

[18] Byrn, S. R.; Zografi, G.; Chen, X. S. *Solid-State Properties of Pharmaceutical Materials*; 2017.

[19] Reilly, A. M. et al. Report on the sixth blind test of organic crystal structure prediction methods. *Acta Crystallographica Section B Structural Science, Crystal Engineering and Materials* **2016**, *72*, 439–459.

[20] Griesser, U. J. *Polymorphism*; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, FRG, 2006; pp 211–233.

[21] Ahlqvist, M. U.; Taylor, L. S. Water dynamics in channel hydrates investigated using H/D exchange. *International Journal of Pharmaceutics* **2002**, *241*, 253–261.

[22] Schultheiss, N.; Newman, A. Pharmaceutical Cocrystals and Their Physicochemical Properties. *Crystal growth & design* **2009**, *9*, 2950–2967.

[23] Grothe, E.; Meekes, H.; Vlieg, E.; ter Horst, J. H.; de Gelder, R. Solvates, Salts, and Cocrystals: A Proposal for a Feasible Classification System. *Crystal Growth & Design* **2016**, *16*, 3237–3243.

[24] Aakeröy, C. B.; Forbes, S.; Desper, J. Using cocrystals to systematically modulate aqueous solubility and melting behavior of an anticancer drug. *Journal of the American Chemical Society* **2009**, *131*, 17048–17049.

[25] Thakuria, R.; Delori, A.; Jones, W.; Lipert, M. P.; Roy, L.; Rodríguez-Hornedo, N. Pharmaceutical cocrystals and poorly soluble drugs. *International journal of pharmaceutics* **2013**, *453*, 101–125.

[26] Rama Krishna, G.; Ukrainczyk, M.; Zeglinski, J.; Rasmuson, Å. C. Prediction of Solid State Properties of Cocrystals Using Artificial Neural Network Modeling. *Crystal Growth & Design* **2018**, *18*, 133–144.

[27] Berry, D. J.; Steed, J. W. Pharmaceutical cocrystals, salts and multicomponent systems; intermolecular interactions and property based design. *Advanced Drug Delivery Reviews* **2017**, *117*, 3–24.

[28] Chiarella, R. A.; Davey, R. J.; Peterson, M. L. Making Co-Crystals - The Utility of Ternary Phase Diagrams. *Crystal Growth & Design* **2007**, *7*, 1223–1226.

[29] Yamashita, H.; Hirakura, Y.; Yuda, M.; Teramura, T.; Terada, K. Detection of cocrystal formation based on binary phase diagrams using thermal analysis. *Pharmaceutical research* **2013**, *30*, 70–80.

[30] Yamashita, H.; Hirakura, Y.; Yuda, M.; Terada, K. Coformer screening using thermal analysis based on binary phase diagrams. *Pharmaceutical research* **2014**, *31*, 1946–1957.

[31] Ainouz, A.; Authelin, J.-R.; Billot, P.; Lieberman, H. Modeling and prediction of cocrystal phase diagrams. *International journal of pharmaceutics* **2009**, *374*, 82–89.

[32] Holan, J.; Stěpánek, F.; Billot, P.; Ridvan, L. The construction, prediction and measurement of co-crystal ternary phase diagrams as a tool for solvent selection. *European journal of pharmaceutical sciences : official journal of the European Federation for Pharmaceutical Sciences* **2014**, *63*, 124–31.

[33] D. Frenkel,; B. Smit, *Understanding Molecular Simulation: from Algorithms to Applications*, 2nd ed.; Academic Press: London, 2002.

[34] Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics* **1953**, *21*, 1087–1092.

[35] Leach, A. *Molecular Modelling: Principles and Applications*, 2nd ed.; Pearson Education: Harlow, 1996.

[36] Anwar, J.; Boateng, P. K. Computer simulation of crystallization from solution. *Journal of the American Chemical Society* **1998**, *120*, 9600–9604.

[37] Anwar, J.; Boateng, P.; Tamaki, R.; Odedra, S. Mode of Action and Design Rules for Additives That Modulate Crystal Nucleation. *Angewandte Chemie International Edition* **2009**, *48*, 1596–1600.

[38] Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Clarendon Press: Oxford, 1989.

[39] Matsumoto, M.; Saito, S.; Ohmine, I. Molecular dynamics simulation of the ice nucleation and growth process leading to water freezing. *Nature* **2002**, *416*, 409–413.

[40] Espinosa, J. R.; Young, J. M.; Jiang, H.; Gupta, D.; Vega, C.; Sanz, E.; Debenedetti, P. G.; Panagiotopoulos, A. Z. On the calculation of solubilities via direct coexistence simulations: Investigation of NaCl aqueous solutions and Lennard-Jones binary mixtures. *The Journal of Chemical Physics* **2016**, *145*, 154111.

[41] Zwanzig, R. W. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *The Journal of Chemical Physics* **1954**, *22*, 1420–1426.

[42] Frenkel, D.; Ladd, A. J. C. New Monte Carlo method to compute the free energy of arbitrary solids. Application to the fcc and hcp phases of hard spheres. *The Journal of Chemical Physics* **1984**, *81*, 3188–3193.

[43] Steinhardt, P. J.; Nelson, D. R.; Ronchetti, M. Bond-orientational order in liquids and glasses. *Physical Review B* **1983**, *28*, 784–805.

[44] Boothroyd, S.; Kerridge, A.; Broo, A.; Buttar, D.; Anwar, J. Why Do Some Molecules Form Hydrates or Solvates? *Crystal Growth & Design* **2018**, *18*, 1903–1908.

[45] Nangia, A.; Desiraju, G. R. Pseudopolymorphism: occurrences of hydrogen bonding organic solvents in molecular crystals. *Chemical Communications* **1999**, *0*, 605–606.

[46] Görbitz, C. H.; Hersleth, H.-P. On the inclusion of solvent molecules in the crystal structures of organic compounds. *Acta Crystallographica Section B Structural Science* **2000**, *56*, 526–534.

[47] Brittain, H. G. *Polymorphism in pharmaceutical solids*; Informa Healthcare, 2009.

[48] Threlfall, T. L. Analysis of organic polymorphs. A review. *The Analyst* **1995**, *120*, 2435–2460.

[49] Stahly, G. P. Diversity in single- and multiple-component crystals. the search for and prevalence of polymorphs and cocrystals. *Crystal Growth and Design* **2007**, *7*, 1007–1026.

[50] Bingham, A. L.; Hughes, D. S.; Hursthouse, M. B.; Lancaster, R. W.; Tavener, S.; Threlfall, T. L. Over one hundred solvates of sulfathiazole. *Chemical Communications* **2001**, *0*, 603–604.

[51] (a) Khankari, R. K.; Grant, D. J. Pharmaceutical hydrates. *Thermochimica Acta* **1995**, *248*, 61–79; (b) Datta, S.; Grant, D. J. W. Crystal structures of drugs: advances in determination, prediction and engineering. *Nature Reviews Drug Discovery* **2004**, *3*, 42–57.

[52] Sloan, E.; Koh, C. *Clathrate Hydrates of Natural Gases*; CRC Press, 2006.

[53] Braun, D. E.; Karamertzanis, P. G.; Price, S. L. Which, if any, hydrates will crystallise? Predicting hydrate formation of two dihydroxybenzoic acids. *Chemical Communications* **2011**, *47*, 5443–5445.

[54] Hulme, A. T.; Price, S. L. Toward the prediction of organic hydrate crystal structures. *Journal of Chemical Theory and Computation* **2007**, *3*, 1597–1608.

[55] Cruz-Cabeza, A.; Day, G.; Jones, W. Towards Prediction of Stoichiometry in Crystalline Multicomponent Complexes. *Chemistry - A European Journal* **2008**, *14*, 8830–8836.

[56] Takieddin, K.; Khimyak, Y. Z.; Fábián, L. Prediction of Hydrate and Solvate Formation Using Statistical Models. *Crystal Growth & Design* **2016**, *16*, 70–81.

[57] Infantes, L.; Chisholm, J.; Motherwell, S. Extended motifs from water and chemical functional groups in organic molecular crystals. *CrystEngComm* **2003**, *5*, 480–486.

[58] Infantes, L.; Fábián, L.; Motherwell, W. D. S. Organic crystal hydrates: what are the important factors for formation. *CrystEngComm* **2007**, *9*, 65–71.

[59] Brychczynska, M.; Davey, R. J.; Pidcock, E. A study of methanol solvates using the Cambridge structural database. *New Journal of Chemistry* **2008**, *32*, 1754–1760.

[60] Brychczynska, M.; Davey, R. J.; Pidcock, E. A study of dimethylsulfoxide solvates using the Cambridge Structural Database (CSD). *CrystEngComm* **2012**, *14*, 1479–1484.

[61] Agrawal, R.; Kofke, D. A. Thermodynamic and structural properties of model systems at solid-fluid coexistence. *Molecular Physics* **1995**, *85*, 43–59.

[62] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **1983**, *79*, 926–935.

[63] Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; De Vries, A. H. The MARTINI force field: Coarse grained model for biomolecular simulations. *Journal of Physical Chemistry B* **2007**, *111*, 7812–7824.

[64] Filion, L.; Dijkstra, M. Prediction of binary hard-sphere crystal structures. *Physical Review E* **2009**, *79*, 046714.

[65] Braun, D. E.; Koztecki, L. H.; McMahon, J. A.; Price, S. L.; Reutzel-Edens, S. M. Navigating the Waters of Unconventional Crystalline Hydrates. *Molecular Pharmaceutics* **2015**, *12*, 3069–3088.

[66] Tomkowiak, H.; Olejniczak, A.; Katrusiak, A. Pressure-Dependent Formation and Decomposition of Thiourea Hydrates. *Crystal Growth & Design* **2013**, *13*, 121–125.

[67] Olejniczak, A.; Katrusiak, A. Pressure induced transformations of 1,4-diazabicyclo[2.2.2]octane (dabco) hydroiodide: diprotonation of dabco, its N-methylation and co-crystallization with methanol. *CrystEngComm* **2010**, *12*, 2528–2532.

[68] Todorov, I. T.; Smith, W.; Trachenko, K.; Dove, M. T. DL_POLY_3: new dimensions in molecular dynamics simulations via massive parallelism. *Journal of Materials Chemistry* **2006**, *16*, 1911–1918.

[69] Grime, J. M. A.; Dama, J. F.; Ganser-Pornillos, B. K.; Woodward, C. L.; Jensen, G. J.; Yeager, M.; Voth, G. A. Coarse-grained simulation reveals key features of HIV-1 capsid self-assembly. *Nature communications* **2016**, *7*, 11568.

[70] He, X.; Lin, M.; Sha, B.; Feng, S.; Shi, X.; Qu, Z.; Xu, F. Coarse-grained molecular dynamics studies of the translocation mechanism of polyarginines across asymmetric membrane under tension. *Scientific reports* **2015**, *5*, 12808.

[71] Marrink, S. J.; de Vries, A. H.; Mark, A. E. Coarse Grained Model for Semi-quantitative Lipid Simulations. *The Journal of Physical Chemistry B* **2004**, *108*, 750–760.

[72] Pluhackova, K.; Böckmann, R. A. Biomembranes in atomistic and coarse-grained simulations. *Journal of physics. Condensed matter : an Institute of Physics journal* **2015**, *27*, 323103.

[73] Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A. E.; Kolinski, A. Coarse-Grained Protein Models and Their Applications. *Chemical Reviews* **2016**, *116*, 7898–7936.

[74] Reith, D.; Pütz, M.; Müller-Plathe, F. Deriving effective mesoscale potentials from atomistic simulations. *Journal of Computational Chemistry* **2003**, *24*, 1624–1636.

[75] Moore, T. C.; Iacovella, C. R.; McCabe, C. Derivation of coarse-grained potentials via multistate iterative Boltzmann inversion. *The Journal of chemical physics* **2014**, *140*, 224104.

[76] Shinoda, W.; DeVane, R.; Klein, M. L. Multi-property fitting and parameterization of a coarse grained model for aqueous surfactants. *Molecular Simulation* **2007**, *33*, 27–36.

[77] Shinoda, W.; DeVane, R.; Klein, M. L. Zwitterionic Lipid Assemblies: Molecular Dynamics Studies of Monolayers, Bilayers, and Vesicles Using a New Coarse Grain Force Field. *The Journal of Physical Chemistry B* **2010**, *114*, 6836–6849.

[78] Marrink, S. J.; Tieleman, D. P. Perspective on the Martini model. *Chemical Society reviews* **2013**, *42*, 6801–22.

[79] Shelley, J. C.; Shelley, M. Y.; Reeder, R. C.; Sanjoy, B.; Klein, M. L. A Coarse Grain Model for Phospholipid Simulations. *J. Phys. Chem. B* **2001**, *105*, 4464–4470.

[80] Anwar, J. An approach for developing simple physics-type force field models for molecular simulation. **Manuscript in preparation**,

[81] Avendaño, C.; Lafitte, T.; Adjiman, C. S.; Galindo, A.; Müller, E. A.; Jackson, G. SAFT-γ Force Field for the Simulation of Molecular Fluids: 2. Coarse-Grained Models of Greenhouse Gases, Refrigerants, and Long Alkanes. *The Journal of Physical Chemistry B* **2013**, *117*, 2717–2733.

[82] Srinivas, G.; Shelley, J. C.; Nielsen, S. O.; Discher, D. E.; Klein, M. L. Simulation of Diblock Copolymer Self-Assembly, Using a Coarse-Grain Model. *The Journal of Physical Chemistry B* **2004**, *108*, 8153–8160.

[83] Straatsma, T. P.; Berendsen, H. J. C. Free energy of ionic hydration: Analysis of a thermodynamic integration technique to evaluate free energy differences by molecular dynamics simulations. *The Journal of Chemical Physics* **1988**, *89*, 5876–5886.

[84] Straatsma, T. P.; McCammon, J. A. Computational Alchemy. *Annual Review of Physical Chemistry* **1992**, *43*, 407–435.

[85] Chipot, C.; Pohorille, A. *Calculating Free Energy Differences Using Perturbation Theory*; Springer, Berlin, Heidelberg, 2007; pp 33–75.

[86] Kirkwood, J. G. Statistical Mechanics of Fluid Mixtures. *The Journal of Chemical Physics* **1935**, *3*, 300–313.

[87] Kofke, D. A. Direct evaluation of phase coexistence by molecular simulation via integration along the saturation line. *The Journal of Chemical Physics* **1993**, *98*, 4149–4162.

[88] Shell, M. S.; Debenedetti, P. G.; Panagiotopoulos, A. Z. Generalization of the Wang-Landau method for off-lattice simulations. *Physical Review E* **2002**, *66*, 056703–056709.

[89] Yan, Q.; Faller, R.; de Pablo, J. J. Density-of-states Monte Carlo method for simulation of fluids. *The Journal of Chemical Physics* **2002**, *116*, 8745–8749.

[90] Wang, F.; Landau, D. Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States. *Physical Review Letters* **2001**, *86*, 2050–2053.

[91] Aragones, J. L.; Valeriani, C.; Vega, C. Note: Free energy calculations for atomic solids through the Einstein crystal/molecule methodology using GROMACS and LAMMPS. *The Journal of Chemical Physics* **2012**, *137*, 146101.

[92] Boothroyd, S.; Anwar, J. Solubility prediction via chemical potentials from density of states. **Manuscript in preparation**,

[93] Panagiotopoulos, A. Z. Critical parameters of the restricted primitive model. *The Journal of Chemical Physics* **2002**, *116*, 3007–3011.

[94] Sousa, J. M. G.; Ferreira, A. L.; Barroso, M. A. Determination of the solid-fluid coexistence of the n - 6 Lennard-Jones system from free energy calculations. *The Journal of Chemical Physics* **2012**, *136*, 174502.

[95] Ahmed, A.; Sadus, R. J. Solid-liquid equilibria and triple points of n-6 Lennard-Jones fluids. *The Journal of Chemical Physics* **2009**, *131*, 174504.

[96] Charpentier, I.; Jakse, N. Phase diagram of complex fluids using an efficient integral equation method. *The Journal of Chemical Physics* **2005**, *123*, 204910.

[97] Khrapak, S. A.; Chaudhuri, M.; Morfill, G. E. Freezing of Lennard-Jones-type fluids. *The Journal of Chemical Physics* **2011**, *134*, 054120.

[98] Okumura, H.; Yonezawa, F. Liquid–vapor coexistence curves of several interatomic model potentials. *The Journal of Chemical Physics* **2000**, *113*.

[99] Ramrattan, N.; Avendaño, C.; Müller, E.; Galindo, A. A corresponding-states framework for the description of the Mie family of intermolecular potentials. *Molecular Physics* **2015**, *113*, 932–947.

[100] Pinho, S. P.; Macedo, E. A. *Developments and Applications in Solubility*; Royal Society of Chemistry: Cambridge, 2007; pp 305–322.

[101] Harper, J. D.; Lansbury, P. T. Models of Amyloid Seeding in Alzheimer's Disease and Scrapie: Mechanistic Truths and Physiological Consequences of the Time-Dependent Solubility of Amyloid Proteins. *Annual Review of Biochemistry* **1997**, *66*, 385–407.

[102] Dasgupta, R.; Walker, D. Carbon solubility in core melts in a shallow magma ocean environment and distribution of carbon between the Earth's core and the mantle. *Geochimica et Cosmochimica Acta* **2008**, *72*, 4627–4641.

[103] Gardner, C. R.; Walsh, C. T.; Almarsson, Ö. Drugs as materials: valuing physical form in drug discovery. *Nature Reviews Drug Discovery* **2004**, *3*, 926–934.

[104] Faller, B.; Ertl, P. Computational approaches to determine drug solubility. *Advanced Drug Delivery Reviews* **2007**, *59*, 533–545.

[105] Jorgensen, W. L.; Duffy, E. M. Prediction of drug solubility from structure. *Advanced Drug Delivery Reviews* **2002**, *54*, 355–366.

[106] Klamt, A. The COSMO and COSMO-RS solvation models. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2018**, *8*, e1338.

[107] Benavides, A. L.; Aragones, J. L.; Vega, C. Consensus on the solubility of NaCl in water from computer simulations using the chemical potential route. *The Journal of Chemical Physics* **2016**, *144*, 124504.

[108] Ferrario, M.; Ciccotti, G.; Spohr, E.; Cartailler, T.; Turq, P. Solubility of KF in water by molecular dynamics using the Kirkwood integration method. *The Journal of Chemical Physics* **2002**, *117*, 4947–4953.

[109] Vega, C.; Noya, E. G. Revisiting the Frenkel-Ladd method to compute the free energy of solids: The Einstein molecule approach. *The Journal of Chemical Physics* **2007**, *127*, 154113.

[110] Lyubartsev, A. P.; Martsinovski, A. A.; Shevkunov, S. V.; Vorontsov-Velyaminov, P. N. New approach to Monte Carlo calculation of the free energy: Method of expanded ensembles. *The Journal of Chemical Physics* **1992**, *96*, 1776–1783.

[111] Paluch, A. S.; Cryan, D. D.; Maginn, E. J. Predicting the Solubility of the Sparingly Soluble Solids 1,2,4,5-Tetramethylbenzene, Phenanthrene, and Fluorene in Various Organic Solvents by Molecular Simulation. *Journal of Chemical & Engineering Data* **2011**, *56*, 1587–1595.

[112] Moučka, F.; Nezbeda, I.; Smith, W. R. Molecular force fields for aqueous electrolytes: SPC/E-compatible charged LJ sphere models and their limitations. *The Journal of Chemical Physics* **2013**, *138*, 154102.

[113] Beutler, T. C.; Mark, A. E.; van Schaik, R. C.; Gerber, P. R.; van Gunsteren, W. F. Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chemical Physics Letters* **1994**, *222*, 529–539.

[114] Anwar, J.; Heyes, D. M. Robust and accurate method for free-energy calculation of charged molecular systems. *The Journal of Chemical Physics* **2005**, *122*, 224117.

[115] Barroso, M. A.; Ferreira, A. L. Solid–fluid coexistence of the Lennard-Jones system from absolute free energy calculations. *The Journal of Chemical Physics* **2002**, *116*, 7145–7150.

[116] Aragones, J. L.; Sanz, E.; Vega, C. Solubility of NaCl in water by molecular simulation revisited. *J. Chem. Phys.* **2012**, *136*, 244508.

[117] Lísal, M.; Smith, W. R.; Kolafa, J. J. Molecular simulations of aqueous electrolyte solubility: 1. The expanded-ensemble osmotic molecular dynamics method for the solution phase. *The journal of physical chemistry. B* **2005**, *109*, 12956–12965.

[118] Mester, Z.; Panagiotopoulos, A. Z. Temperature-dependent solubilities and mean ionic activity coefficients of alkali halides in water from molecular dynamics simulations. *The Journal of Chemical Physics* **2015**, *143*, 044505.

[119] Herdes, C.; Totton, T. S.; Müller, E. A. Coarse grained force field for the molecular simulation of natural gases and condensates. *Fluid Phase Equilibria* **2015**, *406*, 91–100.

[120] Li, L.; Totton, T.; Frenkel, D. Computational methodology for solubility prediction: Application to the sparingly soluble solutes. *The Journal of Chemical Physics* **2017**, *146*, 214110.

[121] Mastny, E. A.; de Pablo, J. J. Direct calculation of solid-liquid equilibria from density-of-states Monte Carlo simulations. *The Journal of chemical physics* **2005**, *122*, 124109.

[122] Singh, S.; Chopra, M.; de Pablo, J. J. Density of States–Based Molecular Simulations. *Annual Review of Chemical and Biomolecular Engineering* **2012**, *3*, 369–394.

[123] Joung, I. S.; Cheatham, T. E. Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *The Journal of Physical Chemistry B* **2008**, *112*, 9020–9041.

[124] Vogel, T.; Li, Y. W.; Wüst, T.; Landau, D. P. Generic, Hierarchical Framework for Massively Parallel Wang-Landau Sampling. *Physical Review Letters* **2013**, *110*, 210603.

[125] Vega, C.; Sanz, E.; Abascal, J. L. F.; Noya, E. G. Determination of phase diagrams via computer simulation: methodology and applications to water, electrolytes and proteins. *J. Phys.: Condens. Matter* **2008**, *20*, 153101.

[126] Schneider, S.; Mueller, M.; Janke, W. Convergence of Stochastic Approximation Monte Carlo and modified Wang–Landau algorithms: Tests for the Ising model. *Computer Physics Communications* **2017**, *216*, 1–7.

[127] Park, S. J.; Ali Mansoori, G. Aggregation and Deposition of Heavy Organics in Petroleum Crudes. *Energy Sources* **1988**, *10*, 109–125.

[128] Di, L.; Fish, P. V.; Mano, T. Bridging solubility between drug discovery and development. *Drug Discovery Today* **2012**, *17*, 486–495.

[129] Nyman, J.; Reutzel-Edens, S. Crystal structure prediction is changing from basic science to applied technology. *Faraday Discussions* **2018**,

[130] Aragones, J. L.; Noya, E. G.; Valeriani, C.; Vega, C. Free energy calculations for molecular solids using GROMACS. *The Journal of chemical physics* **2013**, *139*, 034104.

[131] Moučka, F.; Lísal, M.; Škvor, J.; Jirsák, J.; Nezbeda, I.; Smith, W. R. Molecular simulation of aqueous electrolyte solubility. 2. Osmotic ensemble Monte Carlo methodology for free energy and solubility calculations and application to Nacl. *Journal of Physical Chemistry B* **2011**, *115*, 7849–7861.

[132] Moučka, F.; Lísal, M.; Smith, W. R. Molecular simulation of aqueous electrolyte solubility. 3. Alkali-halide salts and their mixtures in water and in hydrochloric acid. *The journal of physical chemistry. B* **2012**, *116*, 5468–78.

[133] Lyubartsev, A. P.; Laaksonen, A.; Vorontsov-Velyaminov, P. N. Free energy calculations for Lennard-Jones systems and water using the expanded ensemble method A Monte Carlo and molecular dynamics simulation study. *Molecular Physics* **1994**, *82*, 455–471.

[134] Boothroyd, S.; Kerridge, A.; Broo, A.; Buttar, D.; Anwar, J. Solubility prediction from first principles: A density of states approach. **Manuscript in preparation**,

[135] Ben-Naim, A. Standard thermodynamics of transfer. Uses and misuses. *The Journal of Physical Chemistry* **1978**, *82*, 792–803.

[136] Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *Journal of Computational Chemistry* **2004**, *25*, 1157–1174.

[137] Shirts, M. R.; Pande, V. S. Solvation free energies of amino acid side chain analogs for common molecular mechanics water models. *The Journal of Chemical Physics* **2005**, *122*, 134508.

[138] Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C.; IUCr, The Cambridge Structural Database. *Acta Crystallographica Section B Structural Science, Crystal Engineering and Materials* **2016**, *72*, 171–179.

[139] Duffy, E. M.; Severance, D. L.; Jorgensen, W. L. Urea: Potential Functions, log P, and Free Energy of Hydration. *Israel Journal of Chemistry* **1993**, *33*, 323–330.

[140] Mooij, G. C. A. M.; Frenkel, D.; Smit, B. Direct simulation of phase equilibria of chain molecules. *Journal of Physics: Condensed Matter* **1992**, *4*, L255–L259.