Solubility Prediction from First Principles

Author: James Cameron Carruthers MSci
Primary Supervisor: Prof. Jamshed Anwar
Secondary Supervisor: Dr. Andy Kerridge
Industrial Supervisor: Dr. Neil George

Thesis submitted in fulfilment of the requirements of the degree of Doctor of Philosophy

Department of Chemistry

June 2021


Chemistry
Lancaster
University

In memory of Charlotte.

Contents

List of Figures

List of Tables

Abstract

UNIVERSITY OF LANCASTER
Department of Chemistry
Doctor of Philosophy

Solubility Prediction from First Principles

by James Cameron Carruthers

Solubility is a phenomenon of critical importance in countless areas of nature and industry. Solubility drives geological evolution through sedimentation and erosion. The solubility of pharmaceuticals and agrochemicals determines their efficacy and how they have to be formulated for the best efficiency of resources. Solubility determines the fate of artificial chemicals in nature. There are many areas of science where recreating the system in a lab environment is physically impossible or prohibitively expensive so the ability to simulate these systems is a high priority.

This thesis is an exploration of methods to estimate solubilities from direct simulation of molecular systems and seeks to test their accuracy, precision and efficiency, and how they can be further improved.

The first study seeks to recreate the solubility of urea in water using two different thermodynamic cycles (molecular and atomic routes) and two different sets of force fields (Özpınar and TIP3P versus Hölzl and TIP4P/2005) of significantly different ages. This project is a test of simulation software to see if the thermodynamic cycles produce the same results and a test of the force fields to see if the newer force fields give a better estimate of the solubility of urea in water than the older force fields. Neither set of

force fields were actually tested in this way. The solubilities are also estimated using direct coexistence simulations to test the efficiency of this method with modern software and computing power. The newer Hölzl and TIP4P/2005 force fields are closer to reproducing the solubility of urea in water than the older Özpınar and TIP3P force fields according to direct coexistence method but the simulations take a very long time to equilibrate and a different solubility is obtained depending on whether the initial system is subsaturated or supersaturated. The Özpınar and TIP3P force fields failed to produce sensible chemical potential data. The chemical potentials derived from Hölzl and TIP4P/2005 through the molecular route agree with the direct coexistence results. The atomic route gives a too low estimate of the chemical potential difference between the crystal and solution.

The second study seeks to recreate the temperature/solubility phase diagram of butanol and water with direct coexistence simulations and free energy calculations to construct the curves of free energy of mixing using the GAFF and TIP3P force fields. The thermodynamics of mutual solvation are complicated by the competing solvation processes between phases and requires a more thorough analysis than for the solvation of solids which potentially means that direct coexistence simulations are competitive. The direct coexistence simulations were much more efficient than anticipated and gave statistically rigorous estimates of the solubilities of butanol and water. Numerically, they were not accurate estimates but reproduced the qualitative behaviour of the phase diagram and the critical temperature of miscibility was closely reproduced at just above 100°C. The free energy calculations failed to produce chemical potential data with the precision required to create the curves of free energy of mixing at any temperature but the excess chemical potential calculations showed the correct behaviour of electrostatic interactions being more favourable in water than butanol and dispersion interactions being more favourable in butanol than water.

The third study explores the phenomenon of polymorphism where a molecule can adopt multiple different crystal arrangements depending on temperature and pressure. The stability of polymorphs affects how soluble a molecule is in a particular solvent — higher stability means lower solubility. The drug molecule carbamazepine exists in four known polymorphs. The GAFF force field was tested on how well it can reproduce its polymorphs, judged on crystal unit cell parameters. The chemical potentials of

carbamazepine in its four polymorphs and in water were calculated to then see how the solubilities of the four polymorphs compare with experimental data. The GAFF force field closely reproduced three of the four polymorphs with one having issues on a single crystal axis. The stability hierarchy of the four polymorphs was reproduced but the experimental solubility of carbamazepine was over an order of magnitude lower than experimental data.

In conclusion, these studies show that there is still much progress required in general-use force field development in order for solubility estimation to go mainstream. In some applications, direct coexistence simulations will give faster solubility estimates than free energy calculations but they can't give the same thermodynamic insight. For free energy calculations, the thermodynamic cycle should be as simple as possible to avoid unnecessary errors — a thermodynamic adaptation of Occam's Razor. Finally, there needs to be development of dedicated software for setting up free energy simulations. There were thousands of simulations in these studies and much time was dedicated to writing input files by hand and troubleshooting errors in them. Dedicated software that automates the process will reduce errors and open up free energy simulations to wider use.

Acknowledgements

I express my deepest gratitude to Prof. Jamshed Anwar who went above and beyond to ensure my welfare in challenging circumstances. His enduring, compassionate support was critical in helping me maintain the mental fortitude required to complete this thesis. I started this PhD with minimal knowledge of statistical mechanics and the efficiency at which Prof. Anwar was able to assist me in learning all I needed to complete this research was amazing.

I am also deeply grateful to Prof Mauro Ferrario, who Prof. Anwar invited to assist in this difficult research. His wealth of experience and insight was of great help to me in understanding the intricacies of the scientific theory the whole project was built on.

I thank my friend Dr Simon Boothroyd for his expertise in programming greatly expediting the preparation for my research and providing valuable insight while he was also doing his PhD.

I thank Dr Neil George of Syngenta who personally invited me to a tour of the Jealott's Hill site which inspired a strong industrial perspective for my research and a vision of its potential. My gratitude is extended to Syngenta as a whole for twice inviting me to present at their Chemistry Collaborative Research Conference, teaching me valuable skills and broadening my horizons.

I am grateful for CCP5 and their Summer School which was hosted in Lancaster during my PhD. It gave me a great head start at the beginning of my PhD and I was invited to assist with the summer schools in the following two years which sharpened my skills and helped build a solid foundation of teaching skills if they are ever needed in future.

I could not have made it without the support of my colleagues and love of my friends, particularly through one of the darkest, loneliest times in all our lives.

Declaration

I, James Cameron Carruthers, declare that this thesis has been composed solely by myself and has not been submitted for the fulfilment of any other qualification. I confirm that the work contained in this thesis is my own, collaboration is explicitly indicated and all resources cited in support of my own work are properly attributed.

Chapter 1

Introduction

1.1 Introduction to Solubility

Solubility is one of the fundamental properties of multi-component chemical systems. Given a solute molecule and a solvent, the solubility of the solute is its maximum stable concentration in the solvent at a given temperature and pressure. Given this fundamental nature and interest, much effort is dedicated to collecting solubility data, particularly in water which is ubiquitous [1].

Solubility can range from infinite (fully miscible), such as ethanol in water, to poorly soluble for solids such as most metal phosphate salts. There is no commonly agreed threshold below which a compound is considered ”insoluble” and no compound has zero solubility due to entropic effects.

It is necessary to distinguish between solubility and the ability of a solvent to ”dissolve” a substance. A compound may react with the solvent to produce another compound which is then soluble. An example is insoluble zinc metal reacting with hydrochloric acid to produce soluble zinc chloride.

What determines solubility is the balance of the affinity for the solute in its own bulk state versus the affinity for the solution determined by a complex interplay of intermolecular interactions. In general the guideline is ”like dissolves like” — the familiar example being the poor solubility of oil and water. The structure of liquid water is a constantly fluctuating network of strong polar interactions. For something to be dissolved, a cavity in the water needs to be formed which incurs an energy penalty as the network is disrupted. The interactions of the solute with the solvent need to be strong enough to overcome this energy penalty otherwise it will be almost entirely excluded. Conversely, if the interactions between solute molecules are too strong, they will not be able to be pulled apart into solution.

The numerical definition of solubility is defined by the chemical potential (μ). The chemical potential of a solute N_X is the change in free energy upon the addition of a molecule to a system:

$$\mu(N_X) = \left(\frac{\partial G}{\partial N_X}\right)_{N_Y,\ldots,p,T}. \tag{1.1}$$

At the concentration of solubility, the chemical potential of the solute in solution is equal to its own bulk phase (Figure 1.2).

1.2 Applications of Solubility

Solvation is a crucial process in countless aspects of nature and technology from biochemistry to metallurgy. For some purposes solubilisation is a desired process such as for improving uptake of pharmaceuticals while for other purposes solubilisation has to be minimised, such as maintaining structural integrity of building materials.

1.2.1 Chemical Synthesis and Isolation

Exploiting the differing solubilities of reagents, intermediates and products in a variety of solvents is a convenient route to improve yields and purity in chemical synthesis.

Precipitation and Effervescence

One of the simplest applications of differential solubility is the reaction of soluble reagents to produce a product that falls out of solution as a solid (precipitation) or gas (effervescence). This can be an easy way to produce pure chemicals in high yields, particularly for gases [2]. The pH can also be manipulated to induce precipitation through displacement as unionised forms of chemicals are generally much less soluble in water than their salts.

Precipitation is a historically important method of inorganic analysis as ions can be differentiated by particular combinations of ions being insoluble. While largely superseded by modern analytical techniques, it still serves as a useful pedagogical process [3].

Liquid–liquid Extraction

Liquid–liquid extraction is used to isolate a particular component from a mixture by dissolving it into a solvent in which it has a much higher affinity. Typically, this is

demonstrated as separation of non-polar organic substances from a mixture using a non-polar solvent. Acidic and basic compounds can be pulled into an aqueous phase by converting them into their salt forms. It is a useful technique in a diverse range of applications [4, 5]. Metal ions can be selectively extracted into an organic solvent by using a selective chelating agent, such as the extraction of cobalt using trioctylamine into *m*-xylene [6]. DNA is isolated using a variety of extraction methods depending on cost and required purity [7]. Extraction of fragrances using solvent methods helps prevent degradation of delicate molecules that would otherwise occur through traditional distillation methods [8].

1.2.2 Biology and Pharmaceuticals

The human body is 60% water [9] and the knowledge that biological activity is tied with aqueous solubility has been known for over 150 years [10].

Differential aqueous solubility is naturally exploited as a driving force in homeostasis and creating biological structures. Lipids are used to create hydrophobic membranes that control the transport of molecules while specialised proteins are used to select what hydrophilic molecules are allowed to pass. Amino acid side groups have different affinities for water which is a large influence on the tertiary and quaternary structures of proteins [11].

Failure in homeostasis is the cause of many diseases. These can present in an acute or chronic manner. An acute example is decompression sickness, such as when a diver rises too quickly in deep water and blood gases can no longer be dissolved in the reduced pressure leading to bubbles forming, severely compromising blood circulation [12]. A chronic example is the formation of plaques of amyloid proteins through errors in folding. The misfolded protein influences the misfolding of adjacent proteins and they stick together in a useless mass. This process is suspected as a leading cause of Alzheimer's disease [13].

Drugs in the body will encounter a variety of environments and navigating these obstacles is key to getting to where it needs to be. Oral drugs, which comprise the vast majority of drugs consumed by humans, need to survive the highly acidic stomach and

then be able to pass through the various barriers of intestinal wall to enter the bloodstream.

Fick's first law of diffusion states that the rate of diffusion J between two different concentrations is proportional to the concentration gradient [14]. In a single dimension:

$$J = -D\frac{\partial \phi}{\partial x} \tag{1.2}$$

where D is the diffusion coefficient, ϕ is concentration and x is displacement. Therefore, if a drug cannot be provided in a high concentration or easily pass a barrier, the rate of absorption is severely limited.

Issues can also arise in the manufacture and storage of pharmaceuticals [15]. Poor solubility reduces the efficacy of candidate drugs as many are screened out before preclinical and Phase I studies, leading to major monetary losses [16, 17]. There has been an estimate of 39% of drug candidates failing due to poor pharmacokinetics (how well a drug is absorbed, carried to where it is needed, and excreted) [18]. Modern solubility prediction techniques are reducing this proportion as computing power has increased exponentially [19]. Despite this, it still remains difficult to develop these techniques as computational predictions can be significantly inaccurate [20]. Measuring solubility can also be time consuming if the chemical is relatively insoluble or prone to hydrolysis [21].

Combinatorial chemistry and high throughput screening (HTS) approaches to drug design provide fast results but can be biased towards high molecular weight and lipophilicity. These negatively correlate with aqueous solubility so, while a candidate may have high predicted bioactivity, its absorption may be slow [22]. This emphasises the role of accurate solubility data and its efficient measurement in determining the efficacy of new drugs [23]. Therefore, accurate solubility prediction from computational simulations is an important goal for pharmaceutical research in terms of both time and resources.

1.2.3 Environmental Science

Solubility is an important factor in how the Earth has developed in environments ranging from minerals deep in the mantle to nutrient distributions in soils [24, 25]. Computational methods are of particular importance in geology as the temperatures and pressures deep within Earth are practically impossible to recreate in a lab environment.

1.2.4 Agrochemicals

Although agrochemicals are dominated by organic chemistry just like pharmaceuticals and share some correlations, the motivations are very different [26]. There are major concerns with the environmental impact of agrochemical products and food safety [27]. Water scarcity is increasing and there have been millions of cases of people being poisoned through the abuse of pesticides [28, 29]. In addition, environmentally driven data collection faces similar time and cost difficulties to pharmaceuticals.

Solubility also determines how pollutants travel through the environment and uptake by organisms [30, 31]. The use of solubility in agricultural and environmental science is extended to measuring specialised data such as soil sorption coefficients [32].

1.2.5 Metallurgy

Alloys are solid solutions with a metal as the base solvent. Metallurgy is one of the oldest sciences in the world and the production of bronze, the alloy of tin in copper, is used as one of the major milestones in the development of civilisation [33]. Metals are typically alloyed to improve their strength and workability, though there are niche developments such as alloys with very low melting points used as coolants.

Iron and Steel

The smelting of iron is recognised as the next major technological milestone after bronze production. Iron is notable for its ability to dissolve carbon. The inclusion of carbon in iron improves its strength and this property is the etymology of the word ”steel”. Steel is ubiquitous in the modern world and many different varieties have been found and developed throughout history [34].

Amalgams

Amalgams are alloys of mercury, a few of which occur in nature. It is an easily smelted and vaporised metal and along with the softness of amalgams lead to it becoming a popular means of metal-plating materials. It was also a common method of extracting precious metals from ores. In modern times the recognised toxicity of mercury restricts its use but it still remains popular for a few purposes such as tooth fillings [35, 36, 37]. Its ability to dissolve and destroy metals means that it is forbidden to transport by air[38].

1.3 Experimental Determination of Solubility

Various methods for solubility measurement have been developed to satisfy the huge need and turnover required by industrial interests [39]. The methods fall under two broad categories based on the physical end point that is sought;

- Kinetic solubility — the concentration of solute when precipitate first starts to appear
- Thermodynamic solubility — the concentration after the amount of precipitate has stabilised in equilibrium with solution

1.3.1 Thermodynamic Methods

There have been a variety of thermodynamic methods developed to tackle different priorities [40, 41]:

- Shake-flask — hydrophilic compounds
- Column elution — solid compounds
- Potentiometric methods — ionisable compounds only
- Passive dosing — hydrophobic compounds
- Saturated vapour — volatile liquid compounds

Shake-flask Method

The shake-flask method of determining solubility is likely the most familiar to a layperson. An excess amount of solute is added to a solvent in a specialised vessel and is continuously stirred or agitated for a long time to ensure equilibration. The solution is then analysed to record the concentration. It is a simple process but is not suitable for high-throughput screening and requires a relatively large amount of the solute.

Column Elution

A chromatography column is loaded with a carrier coated with the compound of interest. The solvent is passed through the column and the eluate is collected to analyse. This method generally has a shorter saturation time than the shake-flask method but relies on the compound being stable on the carrier otherwise the solubility measurement will be inaccurate.

1.3.2 Kinetic Solubility Assay

The main method of kinetic solubility in high-throughput screening is to dissolve the test compound in DMSO and adding set amounts of this solution to the solvent of choice (usually water) until precipitation is first detected. Kinetic methods are quicker than thermodynamic methods but because there is an energy barrier to nucleation, kinetic methods may often overestimate solubility. This clash of speed and accuracy has been an issue for decades for industry as low solubility is a major source of attrition for target compounds [42].

1.4 Solubility Prediction

1.4.1 Qualitative Structure-Property Relations (QSPRs)

The current paradigm in the prediction of a particular property is to take a large set of molecules, called a training set, with known values for this property and fit a regression that relates the property of interest to molecular descriptors (such as counting a functional group) or other physical properties (such as the melting point). A typical quantitative structure-property relationship (QSPR) takes the form:

$$P = f\left(\sum^{N} g_i(p_i)\right) + E \tag{1.3}$$

where P is the core property of the study, f and g are functions to be determined on descriptor p, and E represents inherent model and observation error. This is then tested on a set of molecules outside the training set. Initially the focus was on biological activity and thus there also exists the term quantitative structure-activity relationship (QSAR) [43, 44]. A well-made QSPR can not only be used for predicting properties outside the training set, but also to fill data gaps for molecules in the training set.

An excellent broad background and applications of QSPR/QSAR are given in the perspective paper *QSAR Modeling: Where Have You Been? Where Are You Going To?* [45]. Here I focus on the application of QSPR to solubility.

Early Developments

The records of early QSAR studies are sporadic but go back remarkably far in scientific history for such a mathematically intensive technique [46]. The earliest known study into structure-activity relationships is that of Cros in 1863 [10]. QSPR specifically was initially developed in the late 19th century by looking at melting and boiling points in homologous series [47]. The first known solubility study was in 1924 when Fühner observed a reduction in solubility by a factor of 4–4.5 upon the addition of successive methylene groups to non-polar molecules [48]. There was eventually a consensus that these early solubility studies were all relating the property with molecular size [46]. In the latter half of the 20th century, solubility prediction studies expanded into the relationship with properties such as molar refraction [49] and melting point [50].

Modern Quantitative Structure Property Relationships

The modern form of QSPR is considered to have been developed by Hansch *et al.* in the 1960s through investigating the relationship between some of the properties of phenoxyacetic acids and their biological activity [51], though their predictive results were poor with an R^2 of 0.43 which they neglected to provide alongside their raw data. As computing power increased exponentially, the scope of these investigations grew from basic regression on small sets of organics to analysing thousands of molecules with the introduction of neural networks in 1991 and utilisation of more advanced analytical

techniques [52]. Despite the technique not being restricted to biological activity, the phrase "quantitative structure activity relationship" still dominates (fig. 1.1).

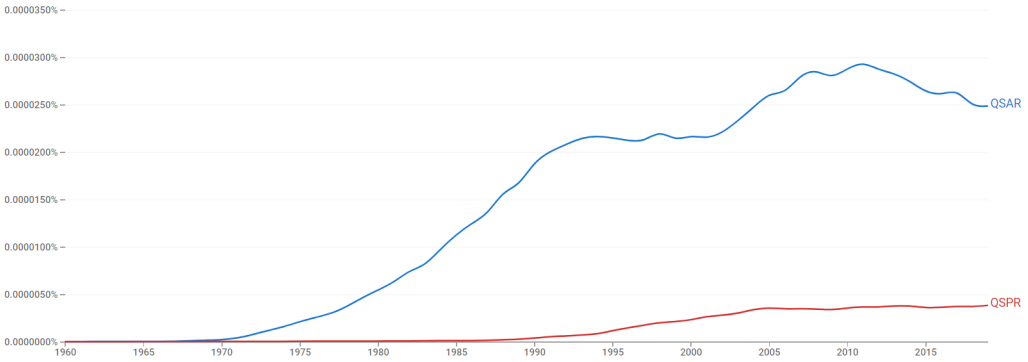


Figure 1.1: Ngram of QSPR and QSAR obtained from Google's Ngram Viewer.

For aqueous solubility, the molecular descriptors tended to be focused on the basic data like lipohilicity (log D), partition coefficients (log P) and the number of hydrogen bond donors and acceptors [53]. More recent studies have chosen larger, more sophisticated sets of descriptors [54], though concerns about overfitting have been around for a long time [55]. Notable studies and developments include:

- Sutter and Jurs [56] — A study early in the "QSPR rennaisance" which used 140 compounds and 53 descriptors refined from a set of 144. They report a suspiciously high R^2 of 0.987 for a nine-descriptor model yet they also go into much detail about the difficulties of QSPR development such as the quality of experimental data. They identified that errors are high for polychlorinated biphenyls and that functional groups need to be represented by multiple compounds to contribute well to the QSPR.

- Gao *et al.* [57] — An extensive QSPR study of a diverse training set of 930 compounds and 46 descriptors with a test set of 249. The study included agrochemicals and pollutants. They obtained an R^2 of 0.92 while the RMSE of $\log S$ was 0.53.

- DRAGON [58] — Software for calculating over 5,000 molecular descriptors — a useful utility for cheminformatics.

With QSPR becoming increasingly popular in research and industry, scientists have developed guidelines for the effective development and reporting of QSPR studies

[44, 45, 59]. However, in the context of solubility, there has been a persistent issue of accuracy, struggling to reduce RMSDs below 1 log unit regardless of the model [60]. This means that the predictive power of QSPR models are severely limited.

Challenges and Limitations of QSPR in Solubility Prediction

Creating an effective QSPR is a difficult endeavour that requires specialist knowledge and skills in statistical analysis. Over time there have been notable cases of naïve approaches to QSPR leading to fundamental errors in scientific analysis [61].

One of the greatest obstacles in QSPR development is the "QSPR paradox". One would expect that a small change in a molecule would lead to a correspondingly small change in the property of interest. However, this is often not the case, particularly in drug design [62].

1.4.2 Quantum Mechanics

The "purest" route to calculating molecular properties is to use quantum mechanics, where the electronic structure of molecules are calculated by solving the Schrödinger equation. However, quantum calculations are very intensive and system sizes are usually limited to tens of atoms. This means that using quantum methods to acquire thermodynamic data is a great challenge. As an approximation, implicit solvation can be used where the solvent is modelled as a continuum with a particular dielectric constant. There are a variety of methods that have been developed to model implicit solvation [63]. In a 2019 study, the solubilities of 51 drug molecules in various glycerides were predicted using the COSMO-RS method with an mean absolute error of 0.576 log units [64], which is respectable for a relatively unexplored solvation environment. The implicit solvation approach requires empirical data such as the melting point and lattice enthalpy and cannot account for local phenomena like hydrogen bonding and the hydrophobic effect.

1.4.3 Molecular Simulation

The ideal method of solubility prediction would only require molecular structures and a set of system conditions to generate a solubility value. A simple approach is to have

a mixed phase system of solute and solution, allow the system to equilibrate and measure the number density of the solute in solution. This is known as direct coexistence and can give good results with decent force fields. However, it is time intensive as it requires very large system sizes to minimise system size artifacts and simulation times on the order of microseconds to achieve equilibrium [65].

As an alternative, with smaller systems one can use techniques that calculate the chemical potential via free energy changes associated with adding a solute molecule in order to determine the concentration at solubility.

Chemical Potentials

Solubility can be determined by calculating chemical potentials which are a measure of how energetically favourable it is to add a molecule to a system. Given a solute and solvent, the basic approach is to calculate the concentration of solution which has the same chemical potential for the solute as the pure solute has for itself [66]. This works well for solids dissolving in fluids but more complex phase behaviours such as mutual solubility require a deeper considerations.

From the early years of molecular dynamics, the significance of solubility modelling was recognised. One of the earliest studies was in 1984 with Swope and Andersen modelling the solubilities of noble gases in water [67]. They determined their own estimates for the force field parameters from experimental data. Systems were limited to 64 water molecules and simulation times of up to 300 ps but despite the lack of computational resources they produced rigorous results and the team recognised the limitations that needed to be overcome.

Also in 1984, Frenkel published the first codification of the Einstein crystal method for determining the chemical potential of a crystal, using a simple hard-sphere model and Monte Carlo simulation [68].

By the 1990s, progress had been made with the Einstein crystal method, moving onto molecular crystals, codified in 1990 by Meijer and Frenkel with the calculation of the melting point of nitrogen [69].

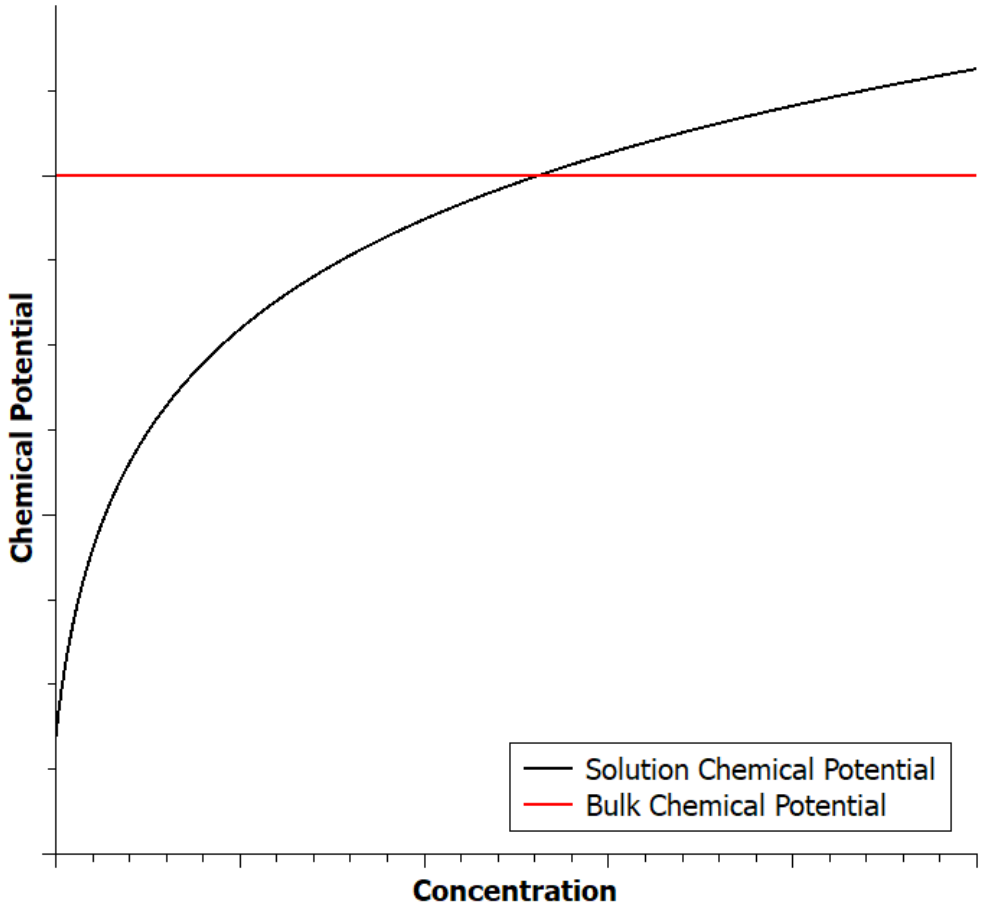


Figure 1.2: Chemical potential of a solute in a solvent versus the chemical potential for the pure bulk solvent.

In 2002, Ferrario et al. published the first major study into estimating solubility, calculating the solubility of KF in water [70]. The study demonstrated the robustness of the Einstein crystal method, with significant no numerical or statistical challenges, for this purpose while acknowledging the sensitivity of the results to the accuracy of the force fields.

Vega et al. then developed a new method of determining chemical potentials through an alternative reference state they called the "molecular Einstein crystal" [71]. The method uses fewer position restraints than the original Einstein crystal method and instead utilises a field restricting entire molecular rotations. While this method is effective, the orientation restriction field is not normally available for free energy calculations in common software.

In 2017, Li, Totton and Frenkel estimated the solubility of hydrophobic organic molecules using a modification of Vega's molecular Einstein crystal [66]. Instead of an orientation field, molecules were simply held in a specific position and rotation using three position restraints. This method proved effective but the study was limited by using a rigid molecule with no partial charges and the central position restraint for the molecular Einstein crystal was implemented through a dummy atom.

In 2019, Belluci et al. developed a robust protocol for the molecular Einstein crystal method that required no dummy atoms or orientation field, instead only using three non-colinear atoms to restrain position and rotation of the molecule [72]. They used the more soluble paracetamol as an example and also measured the real solubility of paracetamol to compare their simulation results, which turned out to be within 0.03 molar ratio of the experimental solubility.

Density of States

A recently developed method estimates solubility by constructing a comprehensive density of states for a range of concentrations in a single Monte Carlo simulation [73]. In the simulation, the system volume and number of molecules are allowed to change and the energy. The Wang–Landau algorithm is applied to fully explore the energy landscape and the system passes through supercritical states to avoid phase transitions

and gas phase configurations to aid molecule insertions and removals. Eventually, a bimodal probability distribution with respect to solute molar fraction can be produced with the peaks at solubility and pure solute.

While the technique is theoretically more effective than molecular dynamics as Monte Carlo considers unphysical moves to better explore phase space, the phase space volume needed to calculate the density of states is still very large which makes exploring all of it difficult. To compensate, the technique has been limited to small systems.

1.5 Research Aims and Objectives

The purpose of this project is to explore and refine methods of predicting solubility from molecular simulations with the aim of establishing robust protocols and identifying improvements needed to be able to use molecular simulations for predicting solubilities of novel compounds. There are a plethora of methods to determine solubility so in the interest of time this project focused on explicit solvation and molecular mechanics methods which are easy to implement and have a wide range of applications.

1.5.1 Solubility of Urea

This study is a comparison of two thermodynamic routes to obtain chemical potentials and the direct coexistence method. One thermodynamic route preserves the molecular structure and the other breaks down the molecules into individual atoms [72, 74]. As the two routes have the same end states, they should give the same chemical potential difference of the crystal and solution. As most studies have focused on poorly soluble molecules, it was decided that a study into one of the most soluble molecules in water would be a useful addition to the literature. Urea was chosen as it is a small molecule with a simple crystal structure. This study compares the estimated solubility from two urea force fields optimised for opposite sides of the equilibrium. Özpınar et al. developed a force field to reproduce the crystal structure while et al. developed a force field to accurately reproduce aqueous urea solutions [75, 76]. The aim is to gain insight into which approaches are more effective for acquiring accurate solubility estimates.

1.5.2 Mutual Solubility Phase Diagram of Butanol and Water

The thermodynamics of the solvation of liquids are complicated by them dissolving into each other, requiring much more free energy data than for the solvation of solids. Butanol and water have an interesting phase diagram where, as temperature increases, the solubility of butanol in water decreases and then increases until the two are miscible close to the boiling point of water. The aim here was to reconstruct the solubility-temperature phase diagram of butanol and water through both free energy and direct coexistence simulations. With the increased diffusion in liquids and diffusion taking place in two directions, direct coexistence simulations were anticipated to have good efficacy and could be potentially shown to be of comparable value to free energy based methods with modern computing power.

1.5.3 Aqueous Solubilities of Polymorphs of Carbamazepine

Carbamazepine is a drug molecule that has four stable polymorphs at standard conditions with very different solubilities. These differences caused major issues for its formulation. Here the aim is to recreate the stability hierarchy of the four polymorphs and determine the solubilities with the intention of carrying the technique forward to new candidates to detect similar issues much earlier in the development cycle of pharmaceuticals, agrochemicals etc.

Chapter 2

Methods

2.1 Introduction

In this chapter the theory and methodology of the studies in this project are explained: a review of the statistical mechanics that form the basis of molecular simulation and thermodynamics, the implementation of molecular dynamics and how energy calculations are efficiently carried out and how free energy and chemical potential data are extracted from molecular dynamics. Specific simulation details are later given in the study chapters.

2.2 Statistical Mechanics

2.2.1 Phase Space, Macrostates, Microstates and Thermodynamic Ensembles

Phase space is the space defined by the positions and momenta of all particles in a system [77]. A single particle is characterised in terms of its three dimensions of position and three dimensions of momenta. For a system of N particles, phase space is described by $6N$ coordinates. For a system of indistinguishable particles, each microstate is a set of $N!$ points corresponding to the permutations of the particles, i.e. swapping the positions and momenta of two identical particles results in the same microstate.

A macrostate is a set of microstates that meet a particular set of macroscopic thermodynamic constraints (e.g. a particular number of particles, system volume and temperature).

A thermodynamic ensemble is a region in phase space which can be reasonably be expected to be visited according to a set of thermodynamic conditions with a system at thermodynamic equilibrium. From a purely mathematical perspective, it is the probability distribution function of phase space according to those thermodynamic conditions. The most common thermodynamic ensembles are:

- Microcanonical Ensemble (NVE) — Constant particle number, system volume and system energy

- Canonical Ensemble (NVT) — Constant particle number, system volume and temperature
- Isothermal–isobaric Ensemble (NPT) — Constant particle number, system pressure and temperature
- Grand Canonical Ensemble (μVT) — Constant chemical potential, system volume and temperature

The NVE ensemble has the simplest statistical mechanics and is the only one to be ergodic (the entirety of valid phase space is accessible) but it is not experimentally realistic as it does not provide a complete link to thermodynamics. The NVT ensemble is usually used for equilibrating systems in preparation for NPT simulations. NPT simulations represent experimental conditions for a closed system and is most common for predicting real system properties. The μVT ensemble is useful for those wishing to explore properties of dilute or non-interacting gases but is usually infeasible for complex chemical systems as particle insertion carries too high a barrier for dense systems.

Ensemble Average

One of the main purposes of molecular simulation is to collect the average of a particular property of interest such as pressure or volume. The mean of a system property A is given by:

$$\langle A \rangle = \frac{1}{N} \sum^{N} A(\boldsymbol{q}) \tag{2.1}$$

with N being the number of samples and $\boldsymbol{q}$ is a point in phase space.

2.2.2 Partition Function and Thermodynamic Potential

The partition function is a dimensionless quantity which describes the number of distinct states available to a thermodynamic ensemble. One can derive almost all thermodynamic properties of a system from the partition function. The internal partition function of a molecule is commonly divided into separate contributions from translation, rotation, vibrational normal modes and electronic states:

$$q = q_{tr}\, q_{rot}\, q_{vib}\, q_{el} \tag{2.2}$$

with the following canonical partition function for N indistinguishable particles:

$$Z = \frac{q^N}{N!}. \tag{2.3}$$

The $N!$ factor comes about as swapping two identical particles does not result in a new microstate. Therefore the $N!$ is needed to avoid overcounting of microstates.

The partition function is associated with a thermodynamic potential defined as the logarithm of the partition function. The definition of the partition function depends on the ensemble.

Finally, the derivative of the thermodynamic potential with respect to particle number is the chemical potential, which determines reaction and phase equilibria — the desired data of this thesis.

Microcanonical Ensemble

The partition function is not generally definable for the microcanonical ensemble but the thermodynamic potential is simply the entropy. The ideal gas is a special case for which the microcanonical partition function can be defined [78]:

$$\Phi(E, V, N) = V^N \frac{(2\pi m E)^{3N/2}}{N! h^{3N} \Gamma(3N/2 + 1)} \tag{2.4}$$

Where Γ is the extension of the factorial function from positive integers to real numbers.

Canonical Ensemble

The generic partition function for a canonical ensemble of N identical particles is:

$$Z(N, V, T) = \frac{1}{h^{3N} N!} \int e^{-\beta H(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_N, \boldsymbol{q}_1, \ldots, \boldsymbol{q}_N)} d^{3N}\boldsymbol{p}\, d^{3N}\boldsymbol{q} \tag{2.5}$$

with h being Planck's constant to maintain a dimensionless Z and H is the Hamiltonian (sum of kinetic and potential energies). In general, this needs to be estimated numerically but there are reference systems for which the integration can be analytically calculated such as the ideal gas, which has zero potential energy. The associated thermodynamic potential is called the Helmholtz energy (F for "free energy" or A for Ger-

man *Arbeit* meaning "work") which is the maximum available thermodynamic work at a constant temperature:

$$F = -kT \ln Z \tag{2.6}$$

And the chemical potential for a substance in a mixed system is:

$$\mu_A(N_A, \ldots, V, T) = \left(\frac{\partial F}{\partial N_A}\right)_{N_B,\ldots,V,T} \tag{2.7}$$

Isothermal–Isobaric Ensemble

The partition function for the NPT ensemble is similar to the NVT ensemble:

$$\mathcal{Z}(N, P, T) = \frac{1}{h^{3N} N!} \int e^{-\beta p V} e^{-\beta H(\boldsymbol{p}_1,\ldots,\boldsymbol{p}_N,\boldsymbol{q}_1,\ldots,\boldsymbol{q}_N)} d^{3N}\boldsymbol{p}\, d^{3N}\boldsymbol{q} dV \tag{2.8}$$

$$= \int Z e^{-\beta p V} dV \tag{2.9}$$

Similarly, the Gibbs energy is related to the Helmholtz energy through the simple relationship:

$$G = F + pV. \tag{2.10}$$

2.2.3 Chemical Potential

The chemical potential of a molecule is the derivative of free energy with respect to the number of molecules in the system (as defined in Equation 1.1) [79].
In the NPT ensemble, for a single component system of N particles the chemical potential is defined as:

$$\mu_G = \frac{G}{N} = \frac{F}{N} + p\frac{V}{N} \tag{2.11}$$

In practice, $p\frac{V}{N}$ is a very small value. Given a pressure of 100 kPa and volume change of 0.1 nm^3, the associated energy is 6 J/mol — much smaller than common error sizes. Chemical potential simulations do not tend to even approach that amount of volume change. As a result, one could use Helmholtz and Gibbs chemical potentials interchangeably with no noticeable error.

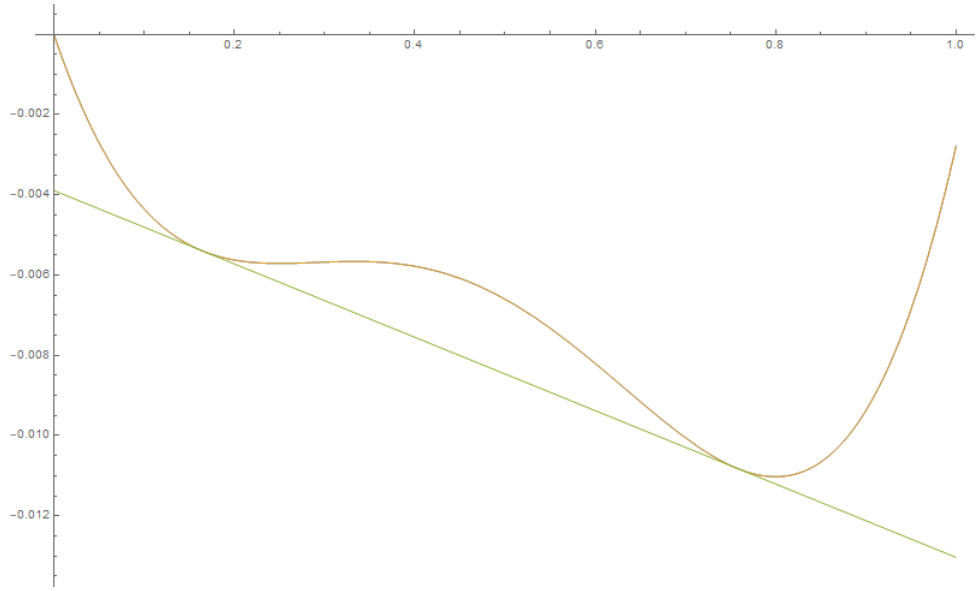


Figure 2.1: Demonstration of the common tangent for an arbitrary curve.

For solvation of a solid in a fluid where there is no solvation of the fluid in the solid, the concentration of solubility occurs where the chemical potential of the solution matches that of the solid.

If there is mutual solvation (typically two liquids), the calculation of the solubility equilibrium is more complicated. The free energy per molecule needs to be calculated as a function of molar fraction and the compositions of solubility are determined by the common tangent of the curve, as demonstrated in Figure 2.1, as this is where the chemical potentials are equal.
The chemical potential of the system as a function of composition is determined as the "free energy of mixing" (ΔG_{mix}) [80]. The definition of ΔG_{mix} for a binary system of components A and B is:

$$\Delta G_{mix}(x_A) = G_{sol}(x_A) - G_{pure}(x_A) + G_{ideal}(x_A) \tag{2.12}$$

$$\Delta G_{sol}(x_A) = x_A\mu_A(x_A) + (1 - x_A)\mu_B(x_A) \tag{2.13}$$

$$\Delta G_{pure}(x_A) = x_A\mu_A(1) + (1 - x_A)\mu_B(0) \tag{2.14}$$

$$\Delta G_{ideal}(x_A) = RT(x_A \ln(x_A) + (1 - x_A)\ln(1 - x_A)) \tag{2.15}$$

where $\mu_A(x_A)$ and $\mu_B(x_B)$ are the excess chemical potentials of components A and B at molar fraction x_A which need to be obtained through simulations. When a sufficient number of values for μ_A and μ_B have been obtained, the data can be interpolated using the second-order Redlich–Kister equation:

$$\Delta G_{mix}(x) = RT\, x\,(1-x)[A + B(2x-1) + C(2x-1)^2] + \Delta G_{ideal}(x). \tag{2.16}$$

2.3 Intermolecular Interactions

There are two main different ways of representing atomic interactions in simulations, quantum mechanics and molecular mechanics. Which to use depends on what properties are being sought and the computational resources available.

2.3.1 Quantum Chemistry

Quantum chemistry seeks to determine the forces on atoms in molecules by calculating the electronic structure of the molecule. The methods are split into two main approaches. *Ab initio* (Hartree–Fock etc.) methods seek to solve the many-electron Schrödinger equation given the nuclear positions. Density Functional Theory (DFT) seeks to calculate electron density with functionals (functions that take a function as an input and output a number) to determine the forces between nuclei. *Ab initio* methods are more accurate and do not require empirical data but come with great computational cost, while DFT methods trade ultimate accuracy in favour of much quicker calculations and can be expedited with empirical data.

Even with the efficiency of DFT, quantum methods for molecular dynamics are very intensive, limiting systems to very few atoms, so researchers tend to only resort to them in limited circumstances where they can use small systems or anticipate that the results are worth the wait [81].

2.3.2 Molecular Mechanics

Molecular mechanics ignores the electronic structure of molecules in favour of implementing what are called "force fields" to determine the forces between atoms. Force fields are a set of potential functions which define particular geometric aspects of the

molecule and intermolecular forces. A force field is typically the sum of the following forces:

- Covalent bond interactions
- Bond angle and dihedral interactions
- Dispersion forces / van der Waals interactions
- Coulombic/electrostatic forces.

Typical forms of each contribution are as follows:

$$U_{bond}(r) = k/2(r - r_0)^2 \tag{2.17}$$

$$U_{angle}(\theta) = k/2(\theta - \theta_0)^2 \tag{2.18}$$

$$U_{dihedral}(\phi) = k(1 + \cos(n\phi - \phi_0)) \tag{2.19}$$

$$U_{vdW}(r) = 4\epsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right] \tag{2.20}$$

$$U_{Coul}(r) = \frac{q_1 q_2}{4\pi\epsilon_0 r} \tag{2.21}$$

More specialist force fields can introduce extra potential functions to model interactions such as hydrogen bonding and polarisability.

Molecular mechanics is a very cheap method of modelling, easily supporting millions of atoms, but it relies entirely on empirical data and refining the force field takes many years of work. However, once the force field has been developed, they can be very good at generating and reproducing useful physical data such as heat capacity [82].

Molecular Mechanics Force Fields

The common force fields in use today are:

- AMBER (Assisted Model Building and Energy Refinement) — specialised for protein and DNA simulation [83]. Has a variant in GAFF (Generalised Amber Force Field) [84].
- CHARMM (Chemistry at HARvard Molecular Mechanics) —- specialised for biological molecules. CGenFF is the generalised version [85].

- GROMOS (GROningen MOlecular Simulation) — Initially specialised for condensed hydrocarbon systems with subsequent variants for different chemical systems [86].
- OPLS (Optimized Potential for Liquid Simulations) [87]

Much effort over decades has been dedicated to modelling water due to its ubiquity[88]. It has been a notoriously hard molecule to portray within the bounds of common force field parameters due to how anomalous its properties are such as its high density and phase transition temperatures[89]. The most commonly used explicit models are Transferable Intermolecular Potential n-Point (TIPnP, n from 3 to 5) and Simple Point Charge (SPC). In TIP3P, water is simply represented by the three atoms. In TIP4P, an extra charged particle with no mass is added towards the centre of mass, replacing the charge on the oxygen, to better represent the charge distribution. In TIP5P, there are massless charges added to the oxygen to represent the lone pairs. The force fields mentioned above are developed with the incorporation of the three point versions of either one of these. Although the three point versions are lacking accuracy in many aspects, they are considered "good enough" for representing solvation at ambient temperatures and are simple to implement.

2.4 Molecular Simulation Methods

2.4.1 Monte Carlo Simulation

Monte Carlo simulation is a method of simulating the evolution of a system through random perturbations of the positions of atoms or whole molecules. The most common implementation in the canonical ensemble is the Metropolis–Hastings algorithm [90]. When a new state is generated, the probability p of its acceptance is determined by the potential energies of the old and new states (E_{new} and E_{old}) and according to the temperature:

$$p = \min\left[1, \frac{\exp(-E_{new}/kT)}{\exp(-E_{old}/kT)}\right] \tag{2.22}$$

If the new energy is lower than the old energy, then the move is accepted. The probability of increasing the energy by a set amount increases as the temperature increases. If a move is rejected, the old state is counted again as an additional sample. This algorithm

is then used to sample the probability distribution of phase space. It should be noted that there is no time progression involved in Monte Carlo simulation so it cannot be used to simulate dynamic processes.

The main advantage of Monte Carlo is that it is more efficient at sampling high energy states compared to dynamic simulations. However, it is difficult to parallelise over multiple processors for faster sampling and sampling progress is sensitive to how large the perturbations are. Too small perturbation leads to inefficient sampling and too large perturbation will lead to a too high rejection rate for new states. With the recent progress in computing power, particularly with the advent of GPU coding, Monte Carlo has been sidelined in favour of dynamic simulation for most purposes as MC cannot provide dynamical information such as diffusion rates.

2.4.2 Molecular Dynamics

Molecular dynamics simulates the motions of atoms according to Newton's laws of motion. A useful property of phase space in MD is that trajectories in phase space never cross each other and therefore a trajectory is defined by a single Hamiltonian. This means that, except for niche systems like a harmonic oscillator, the system will never return to its original state. It is this property which allows us to effectively sample phase space in a dynamic way, assuming ergodicity. The mean of a system property in MD with continuous sampling is:

$$\langle A \rangle = \lim_{t \to \infty} \frac{1}{t} \int_{t_0}^{t_0+t} A(\tau) d\tau \tag{2.23}$$

where t is the sampling time. Since in MD we cannot actually sample continuously or indefinitely, we have to estimate the mean:

$$\langle A \rangle \approx \frac{1}{N} \sum^{N} A(\tau) \tag{2.24}$$

where N is the number of samples.
This is similar to the Monte Carlo process above but here extra caution must be taken as if the sampling frequency is too high, the samples will not be independent. However, sometimes a high frequency is desired if one is interested in the evolution of a property over time.

2.4.3 Integration Algorithms

Despite the core laws of motion being a small set of rules, integrating these in the necessarily discrete world of MD is not simple. Different integration algorithms have been developed over the years.

Leapfrog Integration

The leapfrog algorithm is so named because the positions $\boldsymbol{x}(t)$ and velocities $\boldsymbol{v}(t)$ are updated on alternate steps, "leapfrogging" each other in time. The method was codified by Störmer in 1907 [91]. The acceleration a_i is simply derived from the total force on the atom and its mass.

$$\boldsymbol{v}_{i+1/2} = \boldsymbol{v}_{i-1/2} + \boldsymbol{a}_i \Delta t \tag{2.25}$$

$$\boldsymbol{x}_{i+1} = \boldsymbol{x}_i + \boldsymbol{v}_{i+1/2} \Delta t \tag{2.26}$$

This scheme is time-reversible and inherently preserves the total energy of a dynamical system. However, calculating the Hamiltonian is more memory intensive than other schemes and initialising the half steps can be computationally tricky.

Verlet Integration

Verlet integration is an old integration scheme first developed in 1791 by Delambre but named after Loup Verlet when he implemented it into MD in the 1960s [92]. Like Leapfrog, it is time-reversible and preserves total energy but requires less memory. Given position $\boldsymbol{x}_n$ and acceleration $\boldsymbol{A}(\boldsymbol{x}_n)$, the algorithm works as follows:

$$\boldsymbol{x}_{i+1} = 2\boldsymbol{x}_i - \boldsymbol{x}_{i-1} + \boldsymbol{A}(\boldsymbol{x}_i)\Delta t^2 \tag{2.27}$$

Velocity Verlet

The Velocity Verlet algorithm is a variant of the Verlet algorithm that explicitly includes velocities [93]. It is similar to the Leapfrog algorithm but coordinates and velocities are calculated simultaneously. The basic algorithm is:

$$\boldsymbol{x}(t+\Delta t) = \boldsymbol{x}(t) + \boldsymbol{v}(t)\Delta t + \frac{1}{2}\boldsymbol{a}(t)\Delta t^2 \tag{2.28}$$

$$\boldsymbol{v}(t+\Delta t) = \boldsymbol{v}(t) + \frac{1}{2}\Delta t[\boldsymbol{a}(t) + \boldsymbol{a}(t+\Delta t)] \tag{2.29}$$

Stochastic Dynamics

Stochastic dynamics is an approach to MD that introduces an artificial randomness to the motions of atoms [94] ($\boldsymbol{F}_i(r)$ and "noise" ($\mathring{\boldsymbol{r}}_i$)) terms to the core differential equation:

$$m_i \frac{d^2\boldsymbol{r}_i}{dt^2} = -m_i\gamma_i\frac{d\boldsymbol{r}_i}{dt} + \boldsymbol{F}_i(r) + \mathring{\boldsymbol{r}}_i. \tag{2.30}$$

This approach does come at the cost of losing proper dynamic evolution of the system but it is advantageous for imposing temperature, equilibration and sampling phase space.

2.4.4 Periodic Boundaries

To avoid finite size effects and energetic distortions from exposed surfaces, simulations are run with periodic boundaries. This means that atomic coordinates are treated according to modular arithmetic. Given a cubic box of edge length a, atomic coordinates are constrained to between 0 and a or $-a/2$ and $a/2$ depending on the software. If an atom moves beyond these bounds, a is added to or subtracted from the coordinates to restore its position to within the box. Atomic interactions are also calculated across the boundary on a nearest neighbour/minimum image basis. The maximum distance to calculate atomic interactions is constrained to below half the box length so that an atom does not interact with another atom twice in opposite directions.

2.4.5 Thermostats and Barostats

Thermostats control the momenta of the atoms in a system to maintain a particular temperature within reasonable fluctuations while taking care to maintain a correct distribution of momenta. The most common thermostats are:

- Andersen — Randomly assigns velocities to particles to produce a Maxwell–Boltzmann distribution. Inherently produces a correct ensemble but disrupts kinetics and transport phenomena [95].

- Berendsen — Controls the temperature through a time-dependent friction force acting on particle velocities. This method is simple but it fails to produce a correct ensemble [96].
- Nosé–Hoover — Similar to the Berendsen thermostat with an evolution of the friction coefficient and a tunable parameter, allowing the thermostat to produce a correct ensemble [97].
- Stochastic/Velocity Rescaling — A modification of the Nosé–Hoover thermostat with an added stochastic term to produce a correct ensemble [98].

The most famous and widely used strong coupling thermostat is the Nosé–Hoover thermostat [97]. It utilises a thermal bath and friction term into the equations of motion to control particle velocities.

Barostats manipulate the box axes to maintain an average pressure by monitoring the internal pressure calculated from what is known as the virial.

$$p = \frac{1}{V}\left(NkT + \frac{1}{3}\overline{\sum_{i<j} \boldsymbol{f}_{ij}\boldsymbol{r}_{ij}}\right) \tag{2.31}$$

The virial modifies the ideal gas pressure due to the pairwise forces between atoms with $\boldsymbol{f}_{ij}$ and $\boldsymbol{r}_{ij}$ being the force and displacement respectively between atoms i and j.

Like with thermostats, coupling can be strong or weak but volume fluctuations are much greater than temperature fluctuations, up to hundreds of bars for smaller systems. The common barostat used for equilibrating systems is the Berendsen barostat which applies instantaneous scaling factors to the box axes and atom positions [96]. Barostats such as Parrinello–Rahman treat the box axes like particles subject to the same equations of motion as the atoms and can have much stricter control on the volume [99]. They are not used for equilibration as they can induce large fluctuations which are not easily damped.

2.4.6 Constraints

There are two main forms of constraints in MD, molecular constraints and the centre-of-mass (COM) constraint. In Monte Carlo simulation, constraints are hard to implement

unless molecules are completely rigid and perturbations are applied only to centre-of-mass translation and rotation of the entire molecule.

Molecular Constraints

Molecular constraints enforce a desired geometry in a molecule to be constant. After each MD step, the new coordinates of the constrained set of atoms are reset within a set tolerance to satisfy the constraint while maintaining the COM of the constrained set.

An ubiquitous form of constraint is the bond constraint, which simply enforces a distance between two atoms. Maintaining a constant distance rather than having a harmonic potential allows a longer time-step and is a more accurate portrayal of a covalent bond, except for heavy atoms like chlorine, as bond vibrational modes are not excited at normal conditions. The common constraint algorithms are:

- SHAKE [100] — Can be applied to whole molecules but is an iterative process that can take many steps to be satisfied.
- SETTLE [101] — An evolution of SHAKE used for certain water models.
- LINCS [102] — A constraint algorithm specifically for bonds and isolated angles.

Centre-of-Mass Constraint

Normally, an MD simulation is set up so that the initial total momentum of the system is zero. However, incorrectly coded MD can build up errors that lead to the system gaining a total momentum. This runs the risk of producing a ”flying ice cube” where the total kinetic energy is largely absorbed into the COM motion and interatomic dynamics are suppressed. To avoid this, one can simply track the total momentum of the system and subtract it after each step when necessary. There is a free energy penalty associated with the removal of the degrees of freedom but with a large enough system, this penalty tends to be negligible and corrections can be implemented through the thermostat.

2.4.7 Restraints

Unlike constraints, restraints allow the geometry to change but with an energy penalty. They are easier to implement than constraints as they are just extra terms for the force field. Common forms are:

- Position restraints — an atom is restrained to a point in space.
- Angle restraints — a pair of atoms is restrained to a specified angle with another pair of atoms or a system axis
- Dihedral restraints — A dihedral angle is enforced between a set of four atoms
- Orientation restraints — The orientation of an entire molecule relative to a system axis is enforced. This is used for NMR simulations.

2.4.8 Efficient Calculation of Intermolecular Forces

Intermolecular forces are the major computational load in molecular simulation so the priority in software development is in making their calculation as efficient as possible. A variety of methods have been developed over the decades.

Interaction Cut-Off

The long-range behaviour of van der Waals forces decays very quickly (r^{-7} for the usual Lennard–Jones potential form) so with a reasonably large system these forces are negligible at long distances. With this in mind, a force cut-off is set beyond which van der Waals forces are ignored. Current practice has the cut-off set at 1.2 or 1.4 nm. To avoid artifacts from the potential form jumping to zero at the cut-off, the potential is modified so that it smoothly changes to zero at the cut-off.

Electrostatic forces are more important at long range as they only decay as a function of r^{-2}. However, they are still subject to a cut-off limited to half the system box length to avoid double interactions with atoms. Beyond this range, alternative methods have to be used to account for the electrostatic contribution to the potential energy.

Neighbour Lists

The neighbour list is the data structure that tracks which atoms are within the interaction cut-off distance of each other by detecting when atoms enter or leave a buffer region around each atom. It is updated at a lower frequency than the simulation takes place depending on how dynamic the system is. This increases the efficiency of simulation as the machine does not waste time calculating distances between far-apart atoms during the force calculation stage.

Ewald Summation

Ewald summation is a method of calculating long-distance interactions between atoms in periodic systems, usually electrostatics [103]. The potential is divided into a sum of short range and long range potentials:

$$\phi(\boldsymbol{r}) = \phi_{sr}(\boldsymbol{r}) + \phi_{lr}(\boldsymbol{r}) \tag{2.32}$$

where $\phi_{sr}(\boldsymbol{r})$ is calculated as with van der Waals interactions but $\phi_{lr}(\boldsymbol{r})$ is treated in Fourier space. The Fourier transformation allows a much faster computation for long-range interactions and later development of the Particle–Mesh–Ewald (PME) method employing the Fast Fourier Transform further improved computational efficiency [104].

2.5 Free Energy Calculations

As free energy plays a fundamental role in thermodynamics, a large part of simulation research is dedicated to calculating free energies for various purposes such as phase equilibria and protein binding.

For solvation of a pure crystal, the crystal has a set chemical potential and the aim is to find the concentration of the solute molecule in the solvent that gives the same chemical potential. Solvation of fluids is more complicated as there is solvation in both directions with more complex chemical potential behaviour.

2.5.1 Free Energy Differences and Reference States

As in real life, in molecular dynamics simulation it is only possible to calculate free energy differences, not absolute free energies.

To obtain an absolute free energy, one needs to perform a free energy calculation starting from a system with an analytically known partition function and absolute free energy. There are two common reference states, the ideal gas for fluids and the Einstein crystal for solids.

The partition function of a uniform atomic ideal gas is:

$$Z = \frac{V^N}{N!}\left(\frac{2\pi kTm}{h^2}\right)^{3N/2} = \frac{1}{N!}\left(\frac{V}{\Lambda^3}\right)^N \tag{2.33}$$

with Λ being the de Broglie wavelength of the particle. This equation is only valid for $V \gg \Lambda^3$ otherwise quantum effects have to be accounted for. For a polyatomic ideal gas, the partition function is:

$$Z = \frac{1}{N_{mol}!}\left(\frac{V\,q}{\prod_{i=1}^{n_{mol}}\Lambda_i^3}\right)^{N_{mol}} \tag{2.34}$$

where N_{mol} is the number of molecules, q is the molecular partition function (covering bonds etc.) and n_{mol} is the number of atoms in the molecule. The molecular partition function has dimension $\mathsf{L}^{3(n_{mol}-1)}$.

The Einstein crystal is a lattice of particles bound to points in space by harmonic potentials. Since the particles cannot swap the sites they are bound to, they are in effect isolated systems (in other words *distinguishable particles*) so there is no factorial term to consider. The partition function for an Einstein crystal with a single particle type and restraint strength K_E is:

$$Z = \left(\frac{2\pi kT}{K_E\Lambda^2}\right)^{3N/2}. \tag{2.35}$$

If the ideal gas or Einstein crystal has multiple species, then the total partition function is simply the product of the partition functions as if the species were separated.

2.5.2 Free Energy Perturbation

In the free energy perturbation (FEP) scheme, the free energy difference between states A and B is given by:

$$\begin{aligned}\Delta F(A \to B) &= -kT(\ln Q'_B - \ln Q_A) \\ &= -kT \ln \left(\frac{Q'_B}{Q_A} \right) \\ &= -kT \ln \left\langle \exp \frac{H_A - H'_B}{kT} \right\rangle_A \end{aligned} \tag{2.36}$$

Here, a simulation is performed according to the force field of state A and periodically, the potential energy is calculated according to state B and the difference between the potential energies is recorded [105]. Standard practice is to run two simulations to obtain $\Delta F(A \to B)$ and $\Delta F(B \to A)$ then take the average of the two.

To be effective, there must be significant overlap between the distributions of potential energies in the two simulations. If the difference is too great, intermediate states must be established to bridge the gap with additional simulations.

2.5.3 Bennett Acceptance Ratio

The Bennett Acceptance Ratio (BAR) scheme takes the same information as FEP but with some statistical insight to improve convergence [106]. The main downside is that the approach is based on an implicit equation that needs to be solved numerically:

$$\begin{aligned}&\sum_{i=1}^{n_A} \frac{1}{1 + \exp(\ln(n_A/n_B) + (U_A - U_B)_i/kT - \Delta F/kT)} \\ -&\sum_{j=1}^{n_B} \frac{1}{1 + \exp(\ln(n_B/n_A) + (U_B - U_A)_j/kT - \Delta F/kT)} = 0 \end{aligned} \tag{2.37}$$

with n_A and n_B being the number of samples from simulations A and B. There is still the requirement for the potential energy distributions to overlap well so BAR also requires intermediate states where necessary.

It was later determined that the BAR was the *maximum likelihood method* with respect to the data it uses [107]. This means that it has the minimum variance in the free energy values derived from the data provided to it.

2.5.4 Thermodynamic Integration

Thermodynamic Integration (TI) takes a different approach to FE calculations than perturbation schemes. TI takes two systems A and B and a thermodynamic pathway between the two defined with a scaling parameter λ.

$$U(\lambda) = U_A + \lambda(U_B - U_A) \tag{2.38}$$

When λ is 0 the system is A and at 1 the system is B. Taking the definition of free energy (Equation 2.6) and the above equation, the free energy change between the two states is given by:

$$\begin{aligned}
\Delta F(A \to B) &= \int_0^1 \frac{\partial F(\lambda)}{\partial \lambda} d\lambda \\
&= -\int_0^1 \frac{kT}{Q}\frac{\partial Q}{\partial \lambda} d\lambda \\
&= \int_0^1 \frac{kT}{Q} \sum_s \frac{1}{kT} \exp[-U_s(\lambda)/kT] \frac{\partial U(\lambda)}{\partial \lambda} d\lambda \\
&= \int_0^1 \left\langle \frac{\partial U(\lambda)}{\partial \lambda} \right\rangle_\lambda d\lambda \\
&= \int_0^1 \langle U_B(\lambda) - U_A(\lambda) \rangle_\lambda \, d\lambda
\end{aligned} \tag{2.39}$$

where s is a configuration satisfying the ensemble, and $U_A(\lambda)$ and $U_B(\lambda)$ represent the potential energy of a configuration produced by intermediate state λ but with the end-state force fields applied.

For example, in an FE calculation for Coulombic charges on a molecule, an intermediate simulation would calculate the potential energies for the current configuration with zero charge and full charge on the molecule.

TI is simpler to implement than BAR but it is sensitive to large variations in $\frac{\partial U(\lambda)}{\partial \lambda}$. There need to be enough intermediate states with the right λ values to accurately portray the

curve to integrate.

The chemical potential of a molecule in a system is the free energy change of turning on its intermolecular forces combined with the ideal chemical potential.

2.5.5 Soft-Core Potentials

In free energy calculations, when λ is close to zero the van der Waals potential gains a very harsh singularity that introduces significant statistical issues due to energy spikes. To avoid this, the force field for intermediate λ states is modified with a soft-core function that has the singularity removed. The soft-core potential used in Gromacs is [108]:

$$V_{sc}(r) = (1-\lambda)V_A(r_A) + \lambda V_B(r_B) \tag{2.40}$$

$$r_A = (\alpha\sigma_A^6\lambda^p + r^6)^{\frac{1}{6}} \tag{2.41}$$

$$r_B = (\alpha\sigma_B^6(1-\lambda)^p + r^6)^{\frac{1}{6}} \tag{2.42}$$

Where $V_A(r_A)$ and $V_B(r_B)$ are the potentials of state A and B, and α, σ and p are tunable parameters of the soft-core modification.

Chapter 3

A Comparison of Thermodynamic Routes and Force Fields in Modelling the Aqueous Solubility of Urea

3.1 Introduction

The purpose of this study is to attempt to reproduce the solubility of urea in water using different thermodynamic pathways that should give the same results. The success lies in a thorough and robust understanding of molecular interactions and, from this benchmark, research can move on to novel compounds.

The unique properties of urea solutions have fascinated scientists, especially in regards to its use as a protein denaturant. It improves the solubility of hydrophobic molecules and there is much controversy about its aggregation behaviour in water [109, 110]. Accurately modelling urea in water is challenging due to the strong and complex hydrogen bonding interactions. Like water, urea forms a relatively open crystal structure and easily forms porous frameworks. This can be exploited to form clathrates and co-crystals [111, 112].

In this study two urea models, one optimised for the crystal and the other for aqueous solutions, and two thermodynamic pathways shall be implemented and their results compared. One thermodynamic pathway maintains the molecular structure of the solute and the other breaks down the molecule into its constituent atoms. Another part of the study is investigating how the choice of restraint strength of the Einstein crystal affects the calculated crystal chemical potential.

As computing power and code efficiency have massively improved in the past decade, it is now considered feasible to run direct coexistence simulations for crystals and solvents in a reasonable time frame. It is still much slower than free energy methods but it should give the most accurate estimated range of solubility with a robust simulation setup.

3.2 Methods

The molecular simulation software used in this project is Gromacs [113]. It is free, open source and has an active community for development and support with comprehensive

Table 3.1: Partial charges on the urea molecule according to the two models used in this study. Cis and trans hydrogens are in relation to the oxygen. The unit e_c is the charge of the electron.

Atom	Özpınar Partial Charge / e_c	Hölzl Partial Charge / e_c
C	0.884	0.6068
N	–0.888	-0.8400
O	–0.660	-0.6162
H_{cis}	0.388	0.4026
H_{trans}	0.388	0.4421

Table 3.2: Lennard-Jones parameters for urea atoms according to the two models used in this study.

Atom	Özpınar σ / nm	Hölzl σ / nm	Özpınar ϵ / kJ/mol	Hölzl ϵ / kJ/mol
C	0.339967	0.36039	0.359824	0.35982
N	0.325000	0.34452	0.711280	0.51114
O	0.295992	0.31377	0.878640	0.59432
H	0.106908	0.11333	0.0656888	0.065689

documentation. The free energy data was analysed using the "alchemical-analysis" suite by MobleyLab [114].

3.2.1 System Data

The urea force fields used in this study are a refined GAFF force field for reproducing the crystal [75] and a custom force field designed to reproduce urea aqueous solutions [76] — their non-bonded parameters shown in Tables 3.1 and 3.2. The TIP3P water model was used for the Özpınar model and TIP4P/2005 was used for the Hölzl model. The crystal used for this study is Crystallography Open Database entry 1008776 based on neutron diffraction data [115].

Free Energy Calculation Systems

The crystal unit cell was tiled to 5×a, 5×b and 6×c to enable the use of a standard 1.2 nm cut-off for non-bonded forces giving a system size of 300 molecules. To generate

the reference sites for the Einstein crystal position restraints, the crystal was simulated with the constant-stress (NST) regime to test the stability and accuracy of the crystal parameters within the 5% threshold. The average configuration parameters were taken from an equilibrated NST run and then an NVT simulation was carried out. The average atomic positions from the NVT run were used as the reference positions.

Seven solution systems were made from 1000 molecules ranging from 0 to 300 urea molecules in 50 molecule intervals. This goes up to roughly the solubility of real urea in water at standard conditions [1].

The molecular dynamics integrator was stochastic dynamics for improved phase space sampling. This means less accurate kinetics on the approach to equilibrium but it is an acceptable compromise for faster acquisition of thermodynamic data. The timesteps used were 2 fs for systems with completely constrained bonds and 0.5 fs for systems with flexible bonds. Energy data were collected every 1000 steps. Non-bonded energy potentials were the potential-switch function for van der Waals interactions and Particle-Mesh Ewald for electrostatics with a cut-off distance of 1.2/nm and long range dispersion corrections were applied for energy and pressure. The barostat used for solution simulations was Parrinello-Rahman with a compressibility of 4.5×10^{-5} bar (matching that of water at standard conditions). Free energy calculations for van der Waals and electrostatics use 26 and 11 equidistant λ-states respectively. Calculations for bonds and position restraints unfortunately have to be tuned on an ad hoc basis with test runs. Soft-core parameters are left as Gromacs defaults.

Direct Coexistence

There are two systems that are used for determining direct coexistence solubility. The crystal can either be placed in contact with a pure solvent to give a lower bound (Figure 3.1) or in contact with a supersaturated solution for an upper bound (Figure 3.2). Care must be taken to ensure the regions away from the phase boundaries are large enough to be statistically reliable. The crystal used was produced by tiling the FE calculation crystal $2 \times 2 \times 2$ for a total of 2400 molecules and equilibrated.
For the pure water system, the box size was simply extended in the z direction and the space was filled with *gmx solvate*. The system was then checked to remove any

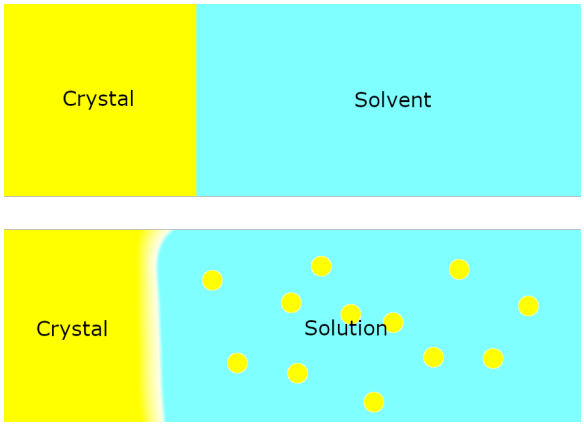


Figure 3.1: Schematic for a crystal dissolving into an initially pure solvent.

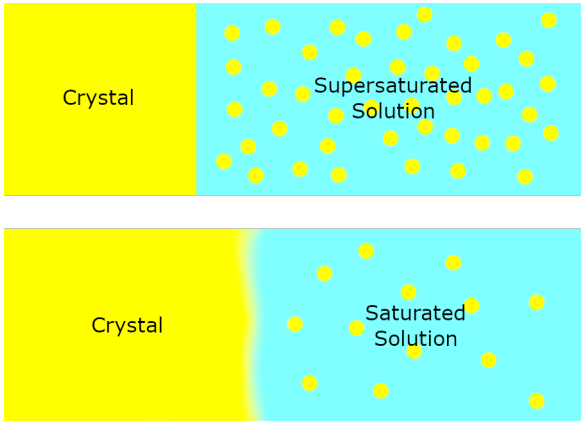


Figure 3.2: Schematic for a supersaturated solution depositing excess solute onto a crystal.

water molecules erroneously placed within the crystal (urea has a rather open crystal structure). The number added cannot be strictly controlled when trying to fill the entire space — in this case 15697 water molecules were present after removing the water molecules in the crystal. The system used anisotropic pressure scaling. The simulation was performed until the concentration of urea in the solvent had stabilised for a long enough time to get a good average.

The supersaturated system requires greater care as there is great uncertainty as to guessing the required concentration. The system was constructed similarly to the pure water system but the empty space had urea molecules added, too. The initial molar

ratio was chosen as 0.5 but this was far too high and the solvent part of the system became too small for good statistics. A new system was created with a roughly 0.2 molar ratio of urea (1980 urea to 7920 water) and this worked for statistics.

3.2.2 Thermodynamic Pathways

Two distinct approaches to calculating crystal free energies have been established in the literature. In the atomic route, the reference state is a mixed Einstein crystal with each atom attached to its average position in the real crystal. This is only possible when flexible bonds have defined parameters. In the molecular approach, the reference state is a "molecular" Einstein crystal where a single atom of the molecule (preferably close to its centre of mass) is bound to its average position. The molecule is free to rotate. If the software allows, one can maintain the internal non-bonded interactions to reduce free energy changes and their errors. A schematic for the two routes is given in Figure 3.3.

Atomic Approach

The thermodynamic pathway for the atomic approach from reference to real crystal is as follows:

1. Turn on bonded interactions within molecules (bonds, angles etc.)
2. Turn on vdW interactions. One can actually get away with not using a soft-core potential as the atoms are restrained from overlapping.
3. Turn on electrostatic charges.
4. Remove position restraints.

The thermodynamic pathway for the atomic approach to the solvation free energy of the probe molecule is as follows:

1. Calculate internal partition function of probe molecule (bonds, angles etc.)
2. Turn on vdW interactions.
3. Turn on electrostatic charges.

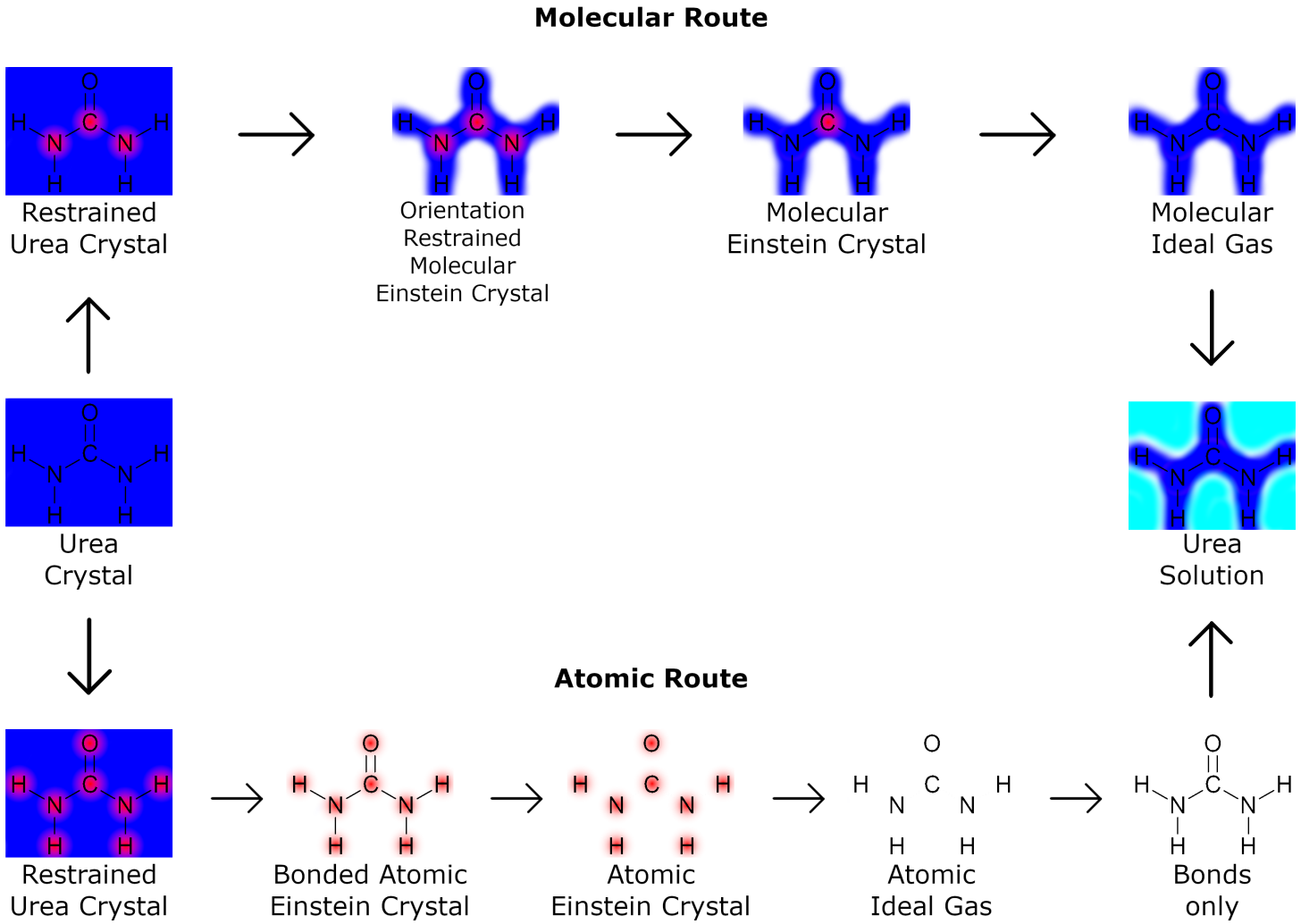


Figure 3.3: Schematic of the two thermodynamic routes used to calculate the solvation free energy of urea. Red represents harmonic restraints, blue represents urea non-bonded interactions and cyan represents water non-bonded interactions.

The partition function of a bond potential U in a cuboid box of dimensions a, b, c is:

$$Z_{bond} = \int_{-a/2}^{a/2} \int_{-b/2}^{b/2} \int_{-c/2}^{c/2} \exp(-U(x, y, z)/kT)\, dx\, dy\, dz \tag{3.1}$$

The partition function of the molecule before activating angle and dihedral potentials is simply the product of the partition functions of every bond. As angle and dihedral potentials are coupled, their free energy contributions have to be calculated numerically. The subsequent calculations for the vdW and Coulombic free energies are no more complicated than for the crystal.

Molecular Approach

The thermodynamic pathway for the molecular approach from reference to real crystal is as follows:

1. Add position restraints to molecules to enforce correct rotation
2. Turn on vdW interactions. As with the atomic approach, a soft-core potential may not be necessary.
3. Turn on partial charges.
4. Remove position restraints.

For the solvation free energy, one only needs to scale the intermolecular non-bonded forces (i.e. turn on vdW potentials then charges).

3.2.3 Einstein Crystal Restraint Strength Testing

Part of refining the free energy calculations is using the optimum strength for position restraints. If the strength is too low, atoms could still overlap and energy spikes decrease precision. If the strength is too high, the free energy calculation for their removal also involves huge energies and makes the integration difficult. Another question is whether the choice of restraint strength has a significant effect on the value of the final free energy estimate. To this end, the crystal free energy was calculated for the Özpınar model with restraint strength ranging from 10000 to 5000000 $\mathrm{kJ/mol\,nm^2}$ according to the 1-2-5 scheme.

3.3 Results

3.3.1 Özpınar Model

Direct Coexistence

The evolution of the molar fraction of urea in solution as a function of time for pure water and a supersaturated solution are given in figures 3.4 and 3.5. The evolution of the systems are surprisingly straightforward with distinct equilibria being shown. The rate of urea dissolution fluctuates as successive crystal layers are removed, a behaviour previously explored by Piana and Gale [116]. It can be seen that the solubility is between around 0.035 and 0.06 molar fraction. This is a small fraction of the real solubility of about 0.3 molar fraction.

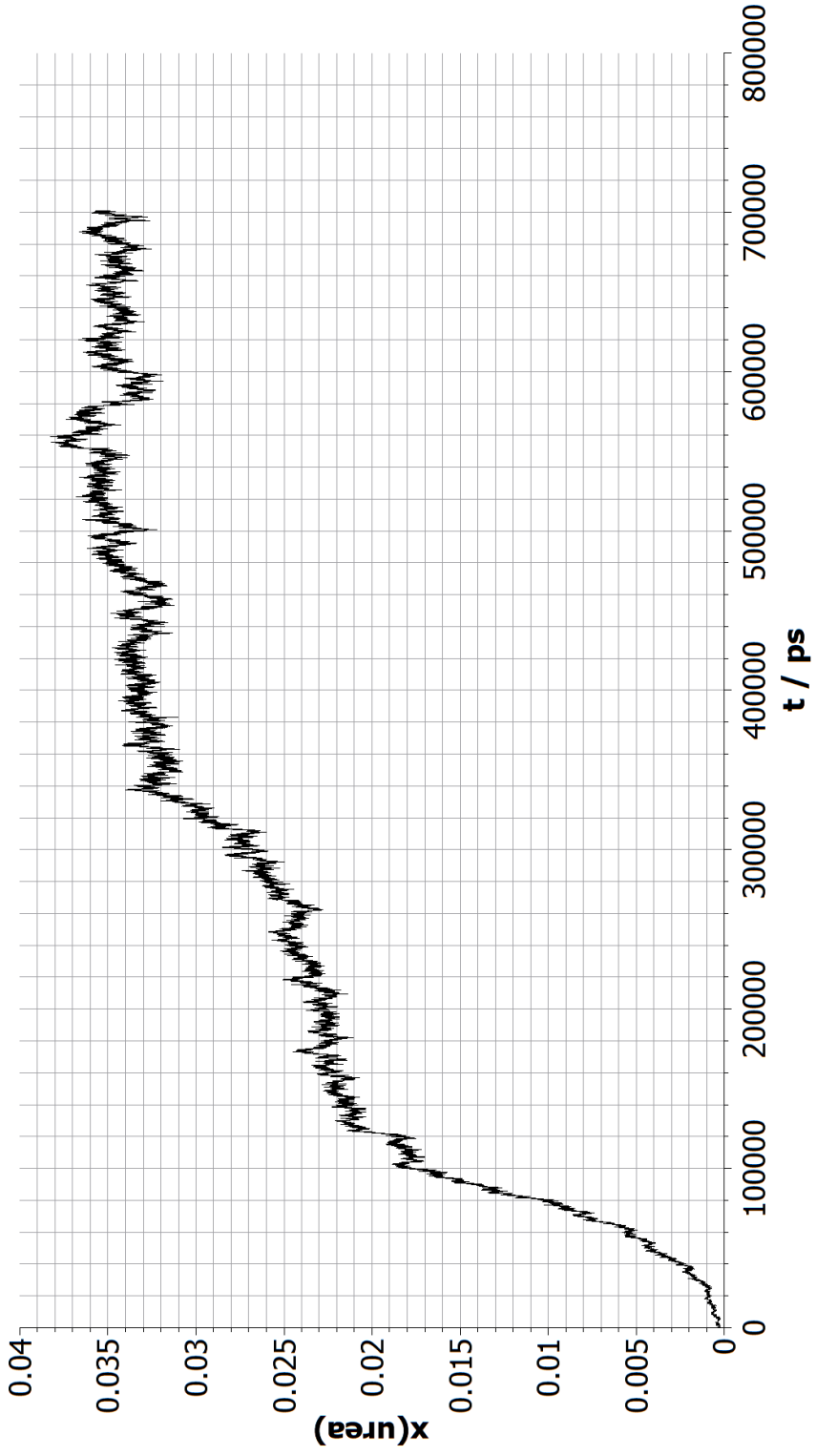


Figure 3.4: Evolution of molar fraction of urea in an initially pure system of water in direct coexistence with a urea crystal according to the Özpinar and TIP3P models. Solvation rate increases rapidly then stalls as crystal layers are depleted

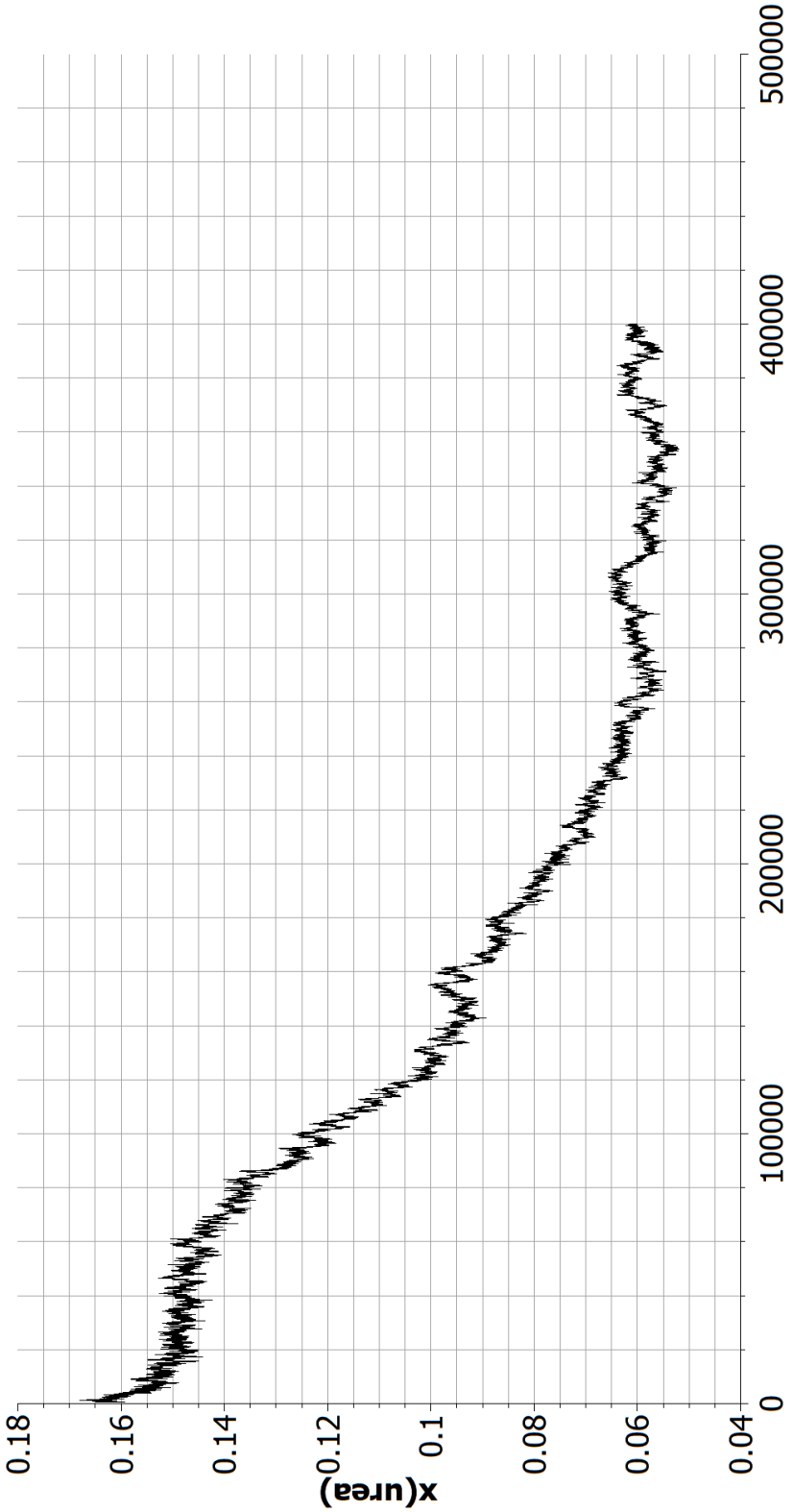


Figure 3.5: Evolution of molar fraction of urea in a supersaturated aqueous solution in contact with a urea crystal according to the Özpınar and TIP3P models. Fluctuation in the rate of crystallisation is not as extreme as the previous dissolution process.

Crystal Chemical Potential

The TI curves for low, intermediate and high restraint strengths are compared in figures 3.6 to 3.13. It is apparent that there is a balancing act for the restraint strength as precision suffers if it is too low or too high. The chemical potential with 500000 kJ/mol nm^2 was used to compare with the solution chemical potentials.

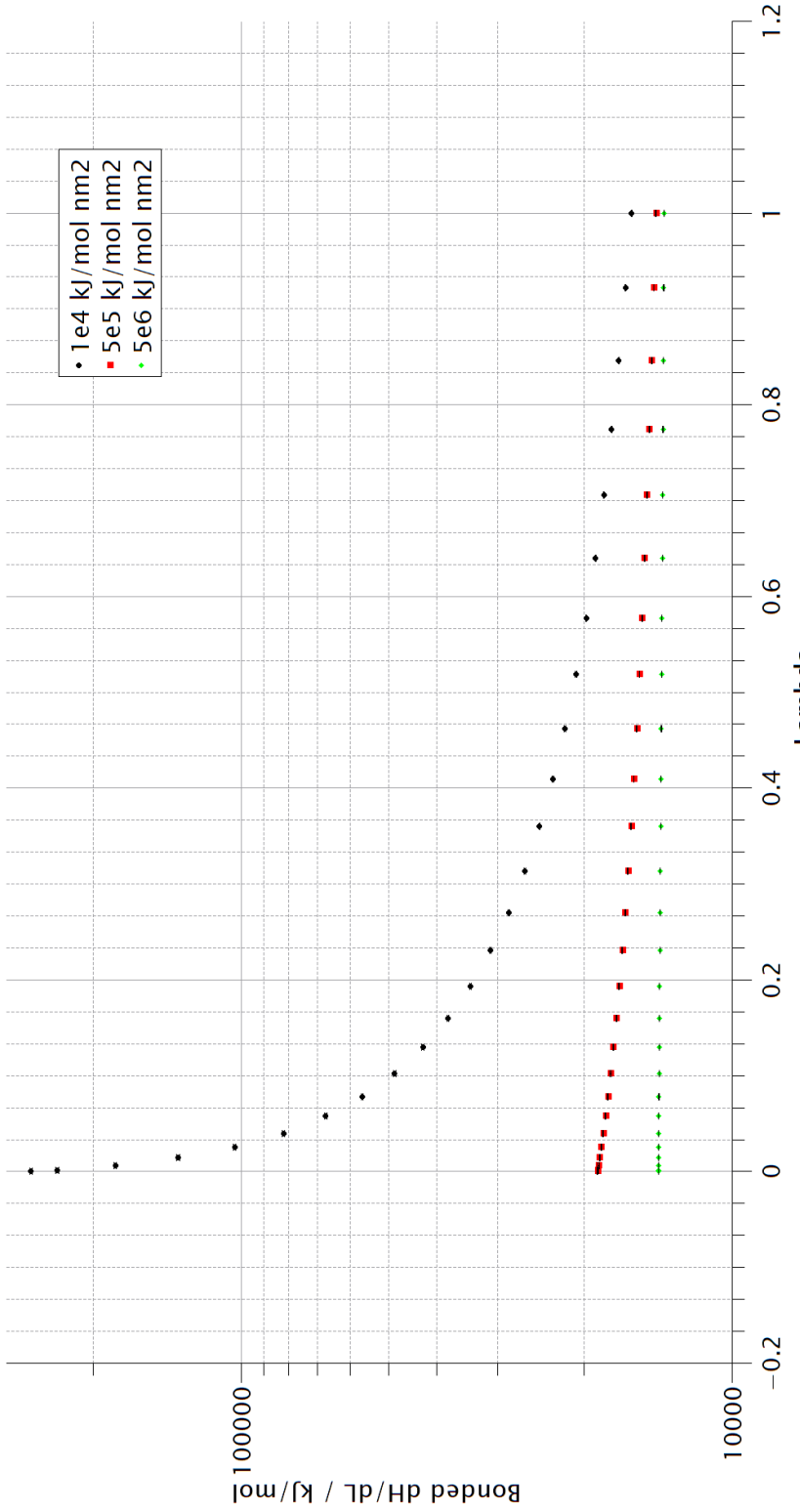


Figure 3.6: Plots of dH/dλ for turning on the bonded interactions in the urea molecule according to the Özpınar model. The curve flattens out with increasing restraint strength.

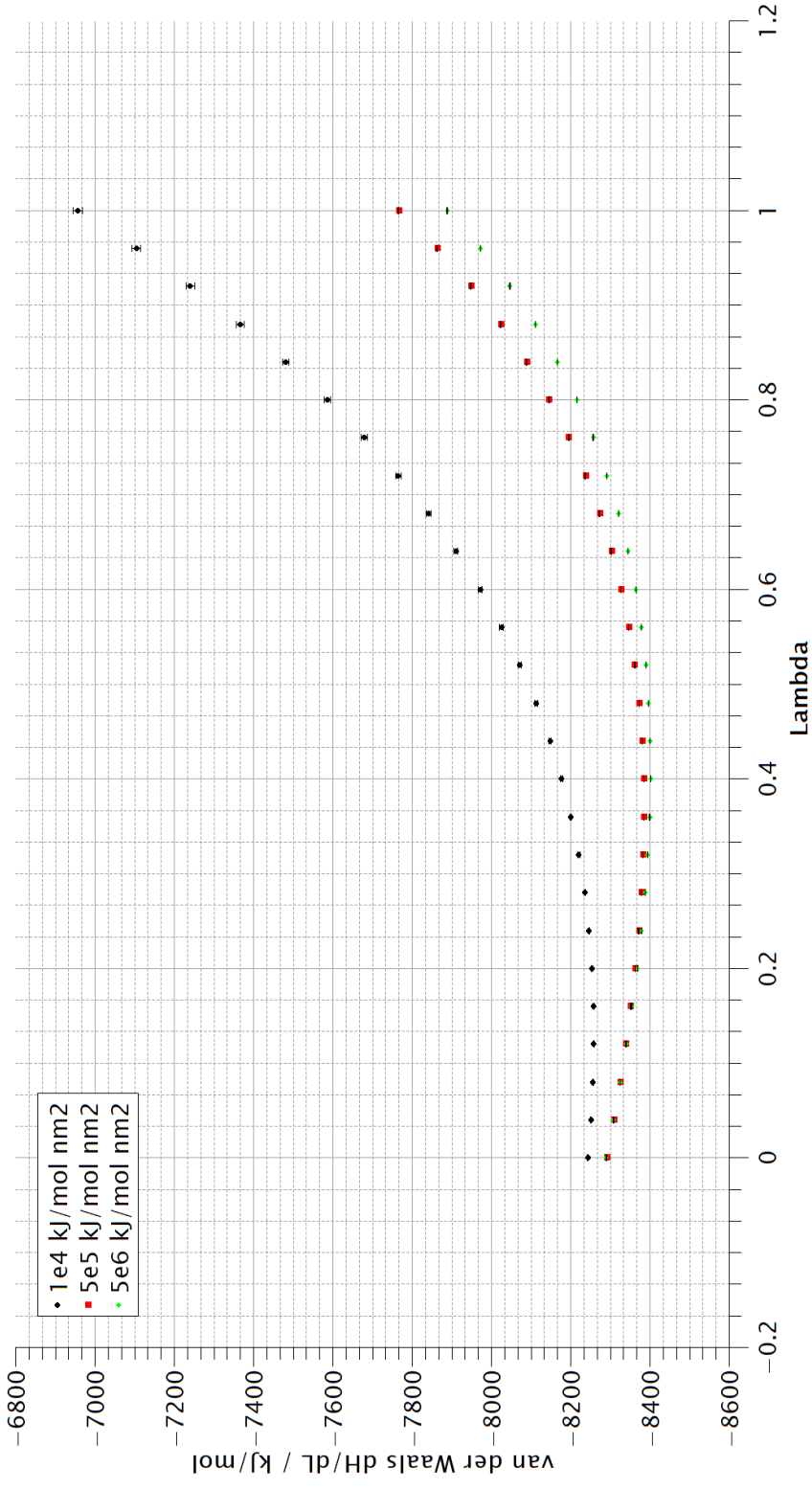


Figure 3.7: Plots of dH/dλ for turning on the van der Waals interactions in the urea crystal according to the Özpınar model. The free energy change decreases with increasing restraint strength as the atoms are prevented from getting close to each other by the restraints.

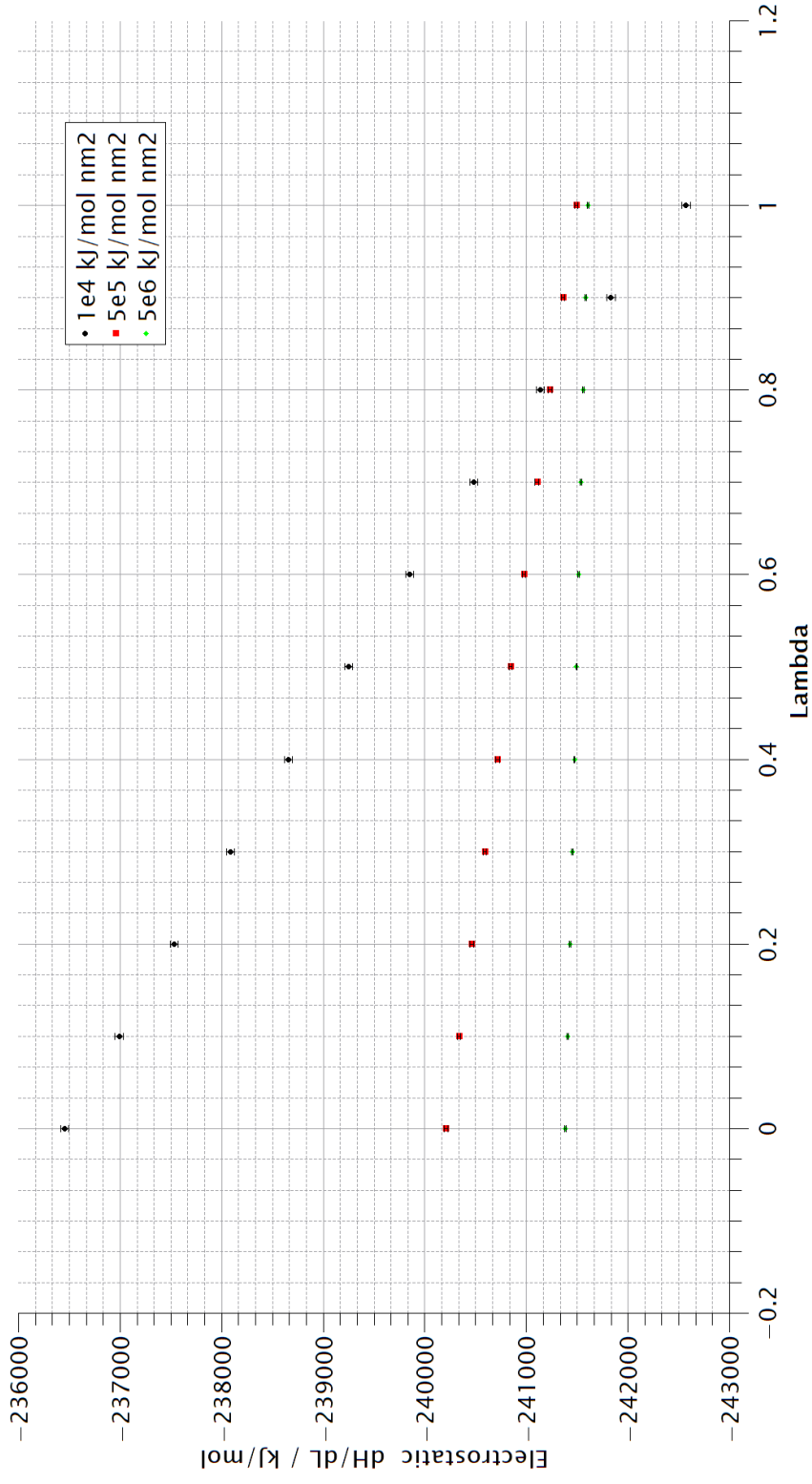


Figure 3.8: Plots of $dH/d\lambda$ for turning on the electrostatic interactions in the urea crystal according to the Özpınar model. As with the bonds, the curve flattens with increasing restraint strength but the curve becomes less negative near 1, possibly due to restraints restricting favourable interactions.

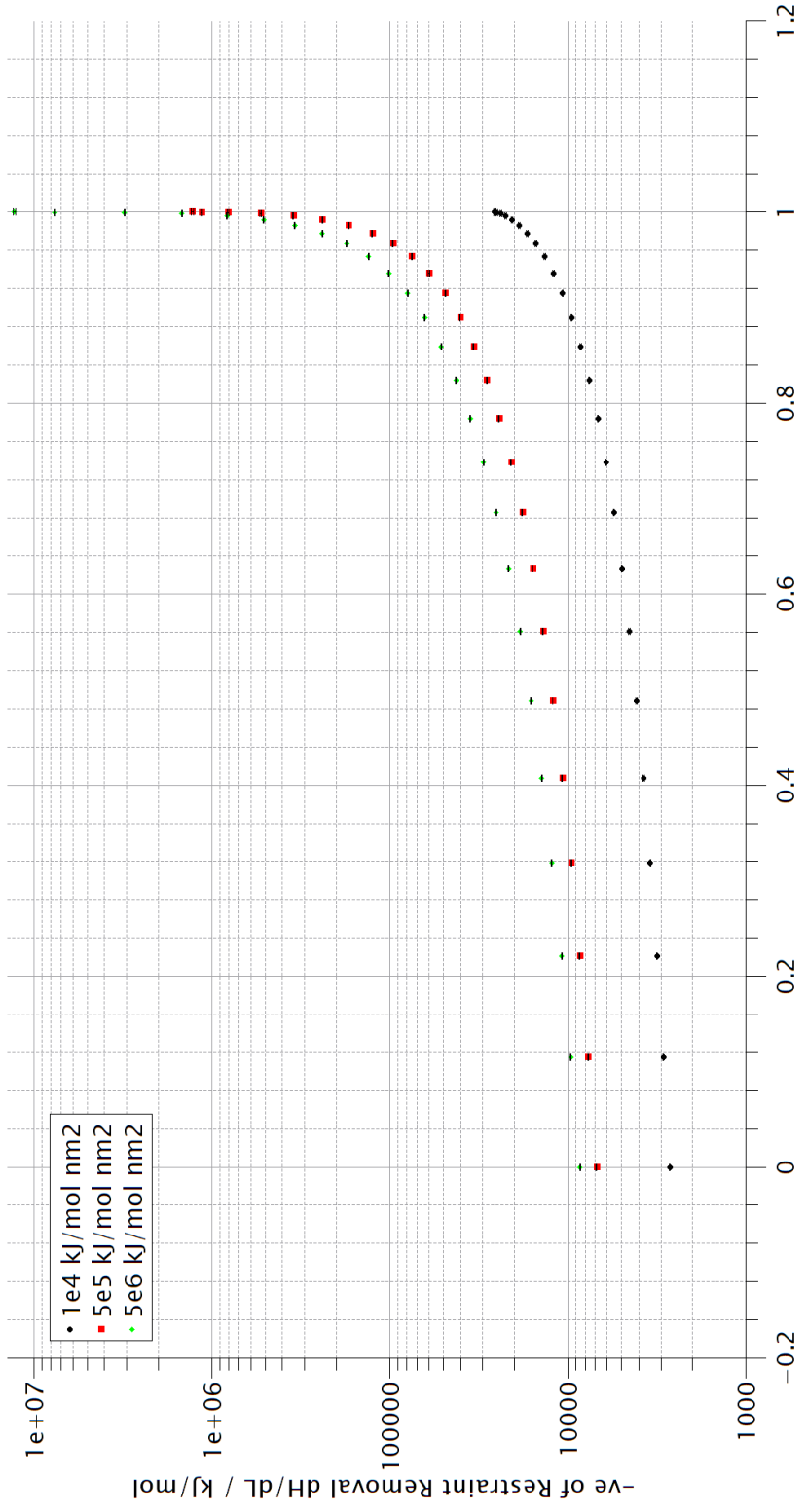


Figure 3.9: Plots of the negative of $dH/d\lambda$ of turning off the position restraints in the system according to the Özpınar model. Plotting the opposite allows use of a log-plot for clarity. As the restraint strength increases, the curve becomes more extreme at 1 and therefore the statistical errors increase. This increase in error counters the decrease in errors for the other contributions.

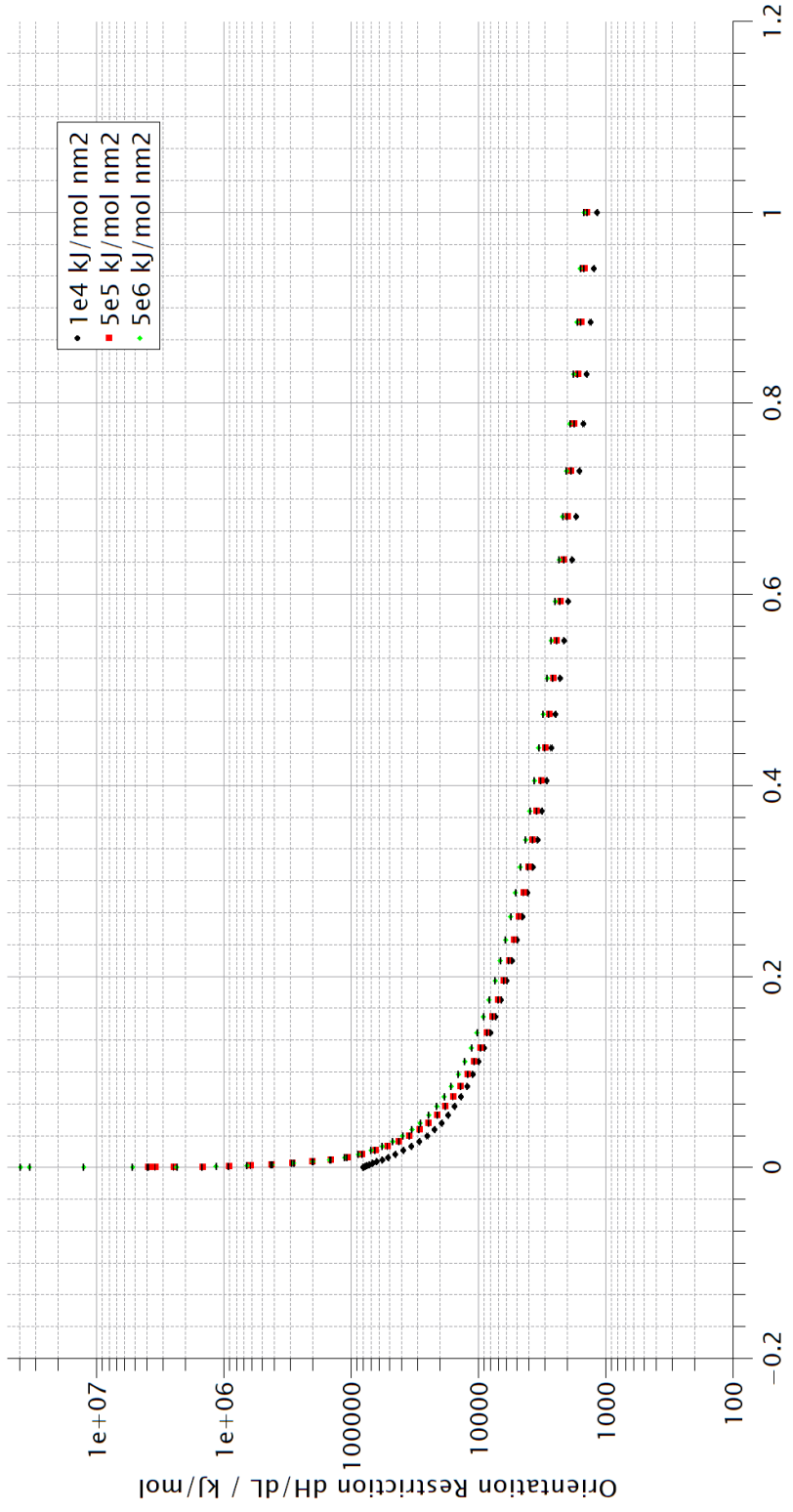


Figure 3.10: Plots of dH/dλ for restricting the orientation of the urea molecule with extra position restraints according to the Özpınar model. This shows the same behaviour as for removing the restraints in the last stage with the free energy and error increasing with restraint strength.

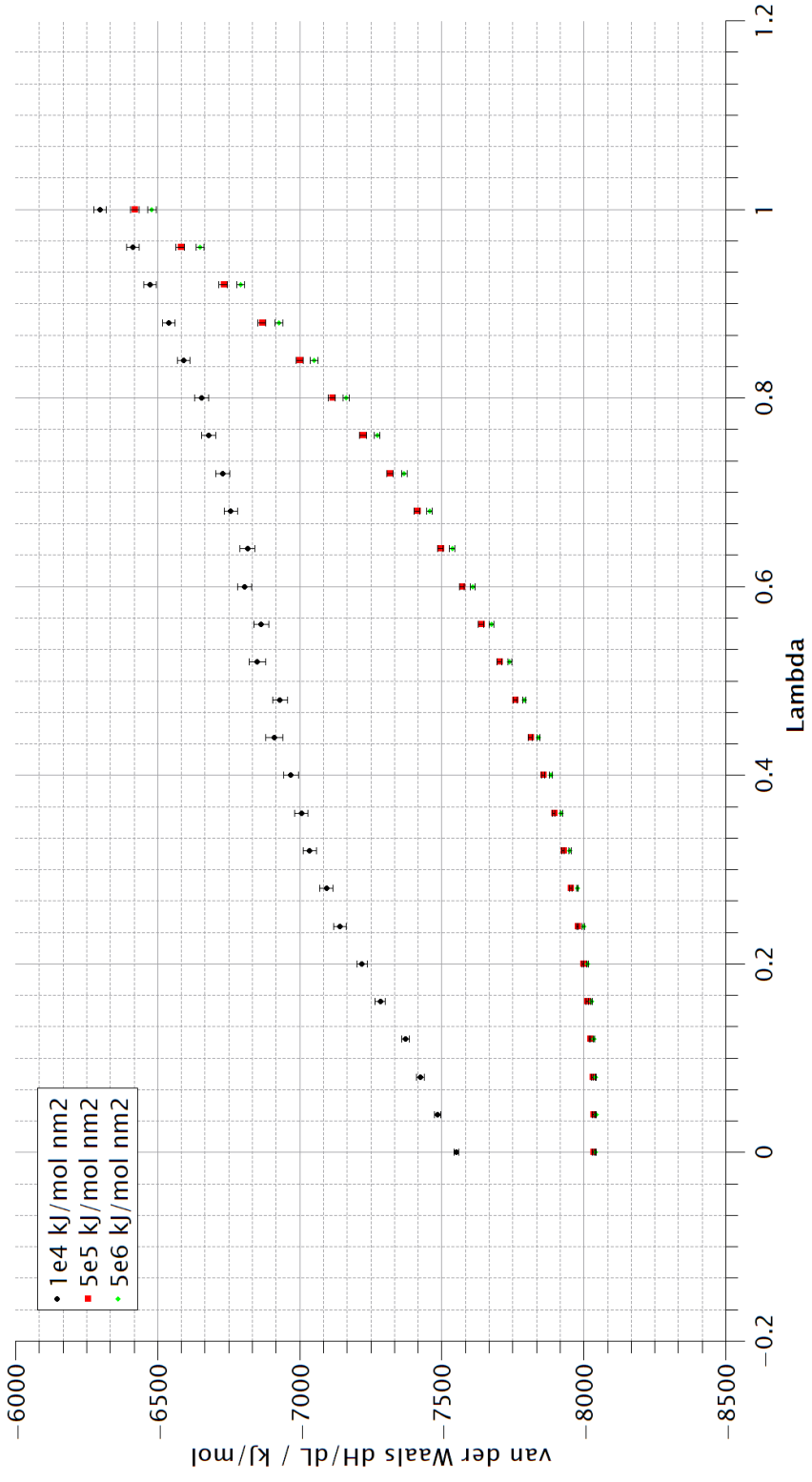


Figure 3.11: Plots of dH/dλ for turning on the intermolecular van der Waals interactions in the urea crystal according to the Özpınar model. The difference between low and medium strengths is significant but there are diminishing returns for a higher strength.

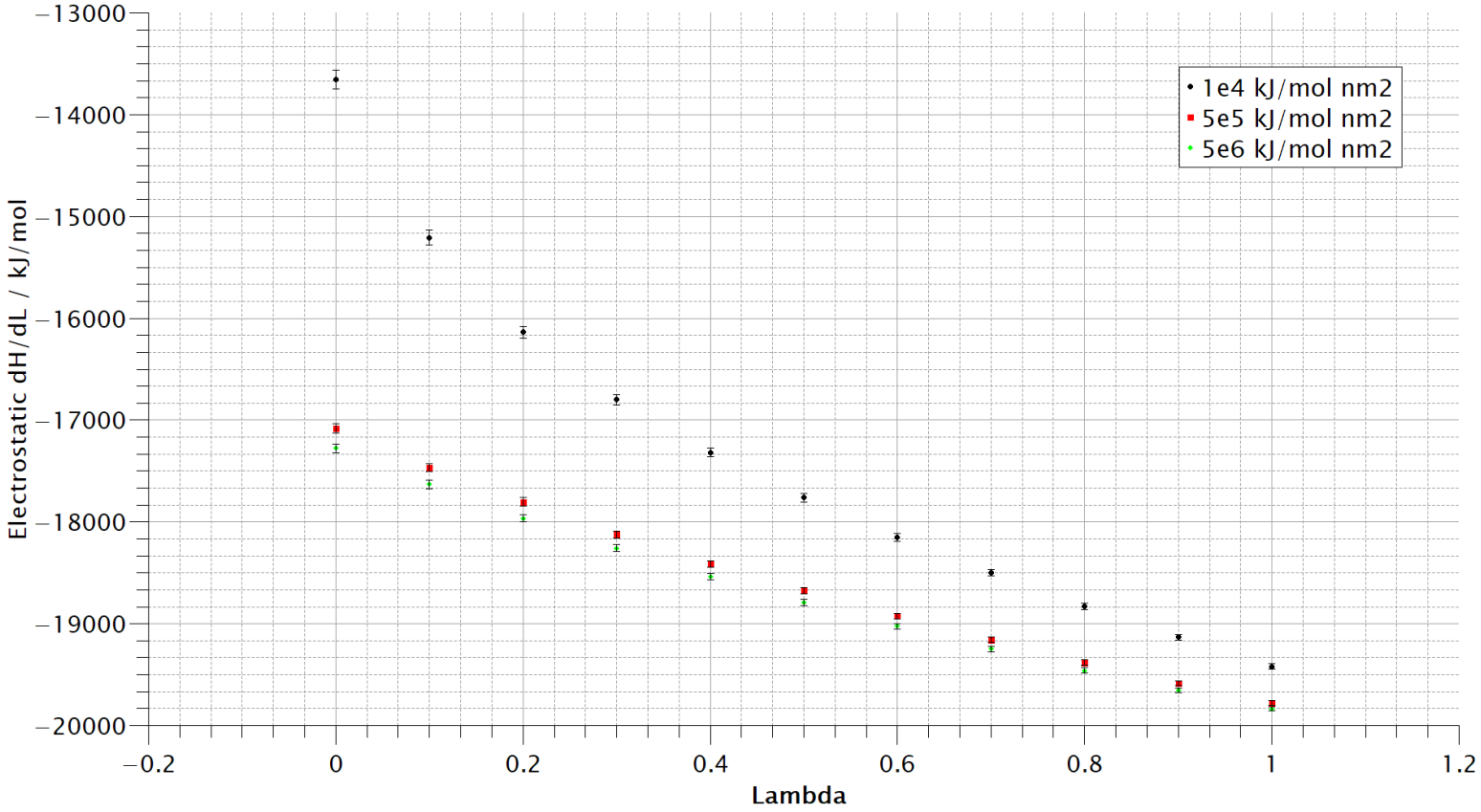


Figure 3.12: Plots of dH/dλ for turning on the intermolecular electrostatic interactions in the urea crystal according to the Özpınar model. The same convergence behaviour as for van der Waals interactions are shown with diminishing returns for high restraint strength.

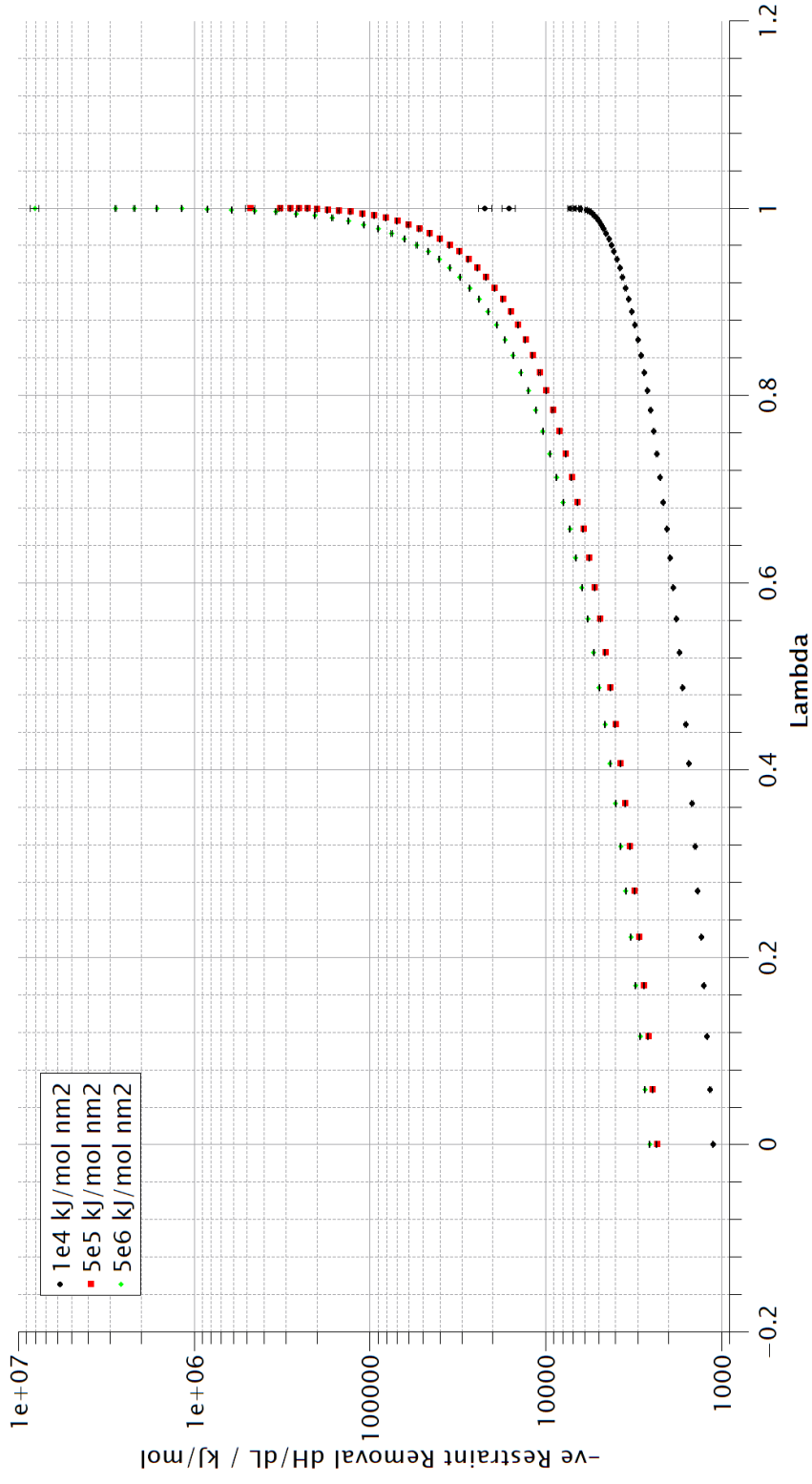


Figure 3.13: Plots of the negative of $dH/d\lambda$ of turning off the position restraints in the system according to the Özpınar model. Plotting the opposite allows use of a log-plot. Greater restraint strength leads to extreme behaviour at high λ with increased statistical errors.

The contributions to the chemical potential of the urea crystal according to the atomic and molecular routes are given in tables 3.3 and 3.4 respectively. The incomplete data for the atomic approach is due to the Gromacs build at the time having stability issues but there was not enough time to repeat the simulations with a later build.

After these simulations finished and the restraint strength was decided, new simulations were carried out for longer with a restraint strength of 500000 kJ/mol nm^2 to refine the statistical errors. The final estimates of the free energy contributions (in units of kJ/mol) in the atomic route were as follows:

- Einstein Crystal 133.49
- Bonds, angles and dihedrals 52.761 ± 0.001
- van der Waals -27.531 ± 0.00001
- Electrostatics -802.812 ± 0.0005
- Restraint release -80.732 ± 0.034

giving a total chemical potential of -723.091 ± 0.034 kJ/mol.

The contributions for the molecular route were:

- Einstein Crystal 14.127
- Orientation restriction 37.632 ± 0.023
- van der Waals -25.822 ± 0.0008
- Electrostatics -62.943 ± 0.005
- Restraint release -32.67 ± 0.005

giving a total chemical potential of -69.676 ± 0.024 kJ/mol.

Table 3.3: Chemical potential data for the urea crystal through the atomic approach according to the Özpınar model. Energies in kJ/mol.

Restraint strength / kJ/mol nm^2	Analytical EC	Ideal Gas	Restraint	R Error	vdW	vdW Error	Electrostatic	ES Error	Restraint Removal	RR Error	Total	Total Error
10000	147.52	−110.798	271.396	0.034	−26.434	0.0027	−797.78	0.0292	−18.3519	0.0197	−724.901	0.069
20000	168.27	−110.798	299.602	0.060								
50000	195.7	−110.798	347.181	0.128								
100000	216.44	−110.798	356.959	0.014	−27.283	0.00058	−799.955	0.0196	−48.2443	0.0425	−723.209	0.053
200000	237.19	−110.798	389.9	0.026	−27.412	0.0004	−801.244	0.0168	−61.2726	0.052	−723.326	0.058
500000	264.62	−110.798	440.99	0.059	−27.53	0.00027	−802.82	0.013	−81.4063	0.0669	−723.779	0.069
1000000	285.37	−110.798	454.008	0.006	−27.588	0.00019	−803.76	0.0102	−98.739	0.0805	−724.364	0.082
2000000	306.11	−110.798	495.69	0.011	−27.626	0.00014			−117.74	0.0966		
5000000	333.54	−110.798	548.849	0.025	−27.653	9e-05	−804.98	0.0053	−145.245	0.123	−727.276	0.123

Table 3.4: Chemical potential data for the urea crystal through the molecular approach according to the Özpınar model. Energies in kJ/mol.

Restraint strength / kJ/mol nm^2	Analytical EC	Ideal Gas	Restraint	R Error	Orientation	O Error	vdW	vdW Error	Electrostatic	ES Error	Restraint Removal	RR Error	Total	Total Error
10000	−138.52	−170.81	32.631	0.000587	17.441	0.014	−23.062	0.011	−58.124	0.0371	−6.7266	0.00104	27.179	0.041
20000	−135.93	−170.81	35.467	0.00105	20.067	0.0163	−24.138	0.007	−59.34	0.0321	−9.6334	0.0014	27.2	0.037
50000	−132.5	−170.81	39.439	0.00225	23.622	0.0198	−24.672	0.0056	−60.5	0.0282	−14.424	0.002	27.698	0.035
100000	−129.9	−170.81	41.72	0.00012	26.471	0.0226	−24.931	0.0049	−61.114	0.0263	−18.679	0.00252	28.013	0.035
200000	−127.31	−170.81	44.698	0.000223	29.525	0.0261	−25.09	0.0045	−61.573	0.025	−23.409	0.0033	28.311	0.036
500000	−123.88	−170.81	48.879	0.000487	33.952	0.027	−25.219	0.0042	−61.984	0.024	−30.277	0.00546	28.760	0.036
1000000	−121.29	−170.81	51.128	2.33e-05	37.506	0.0363	−25.277	0.0041	−62.18	0.0237	−35.828	0.0077	29.102	0.044
2000000	−118.7	−170.81	54.285	4.67e-05	41.302	0.0423	−25.313	0.004	−62.301	0.0233	−41.586	0.0146	29.576	0.048
5000000	−115.27	−170.81	58.766	0.000103	46.792	0.0521	−25.336	0.0039	−62.374	0.0232	−49.409	0.0273	30.575	0.057

Solution Chemical Potential

TI Curves

There is very little change in behaviour of the TI curve for electrostatics as a function of composition, as shown in figures 3.14 to 3.17. This makes sense as both water and urea engage in strong O$\cdots$H hydrogen bonding interactions. The behaviour of the van der Waals interactions are more interesting. The peak of the curve for the concentrated solution is higher than in pure water and the curves cross with the concentrated solution showing more favourable behaviour in the high λ region. This could be an artifact of the soft-core potential. These behaviours are shown in both atomic and molecular routes.

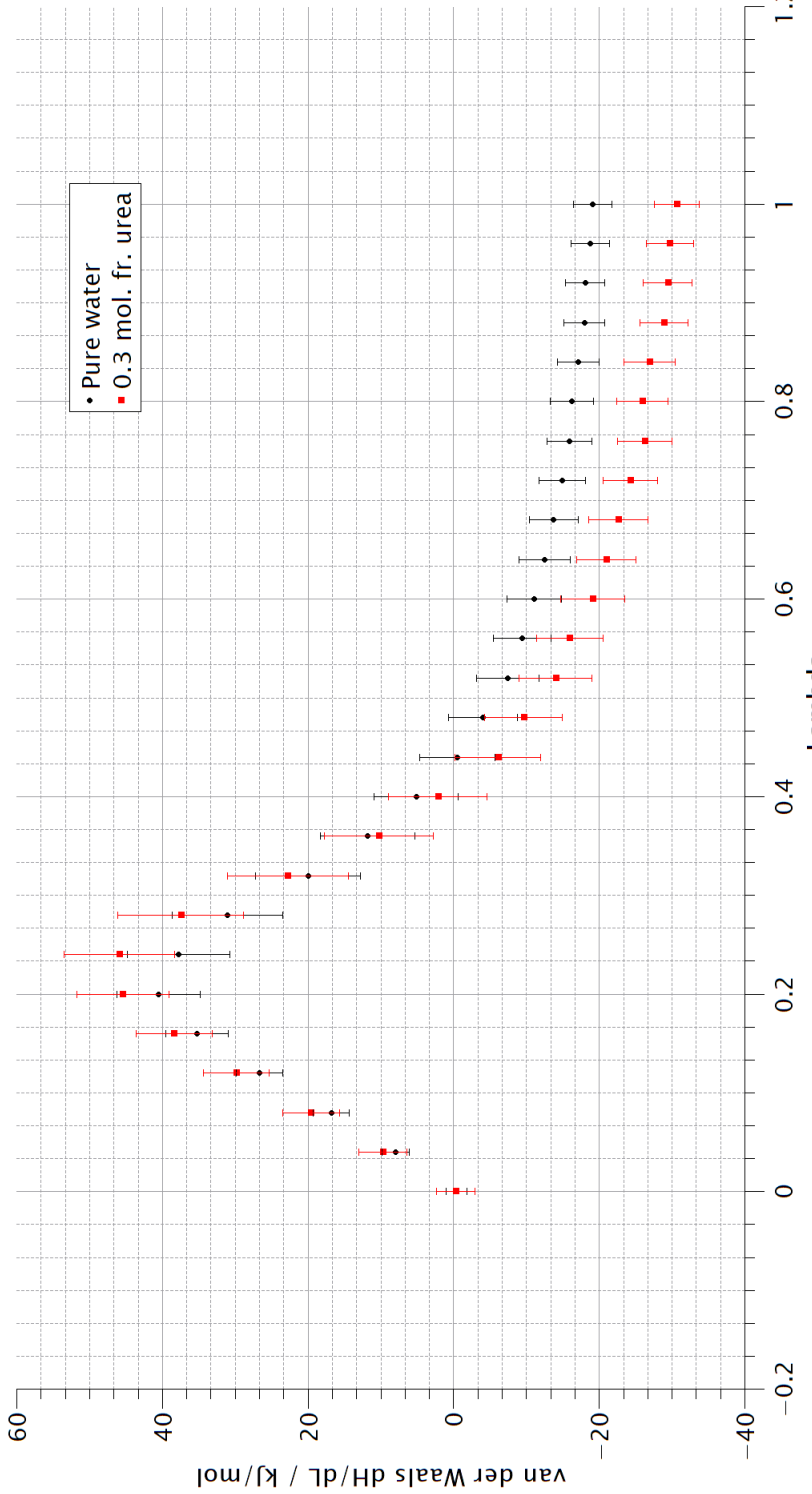


Figure 3.14: Plots of dH/dλ for turning on the van der Waals interactions in the urea molecule according to the Özpınar model. The range of dH/dλ is greater at higher concentration.

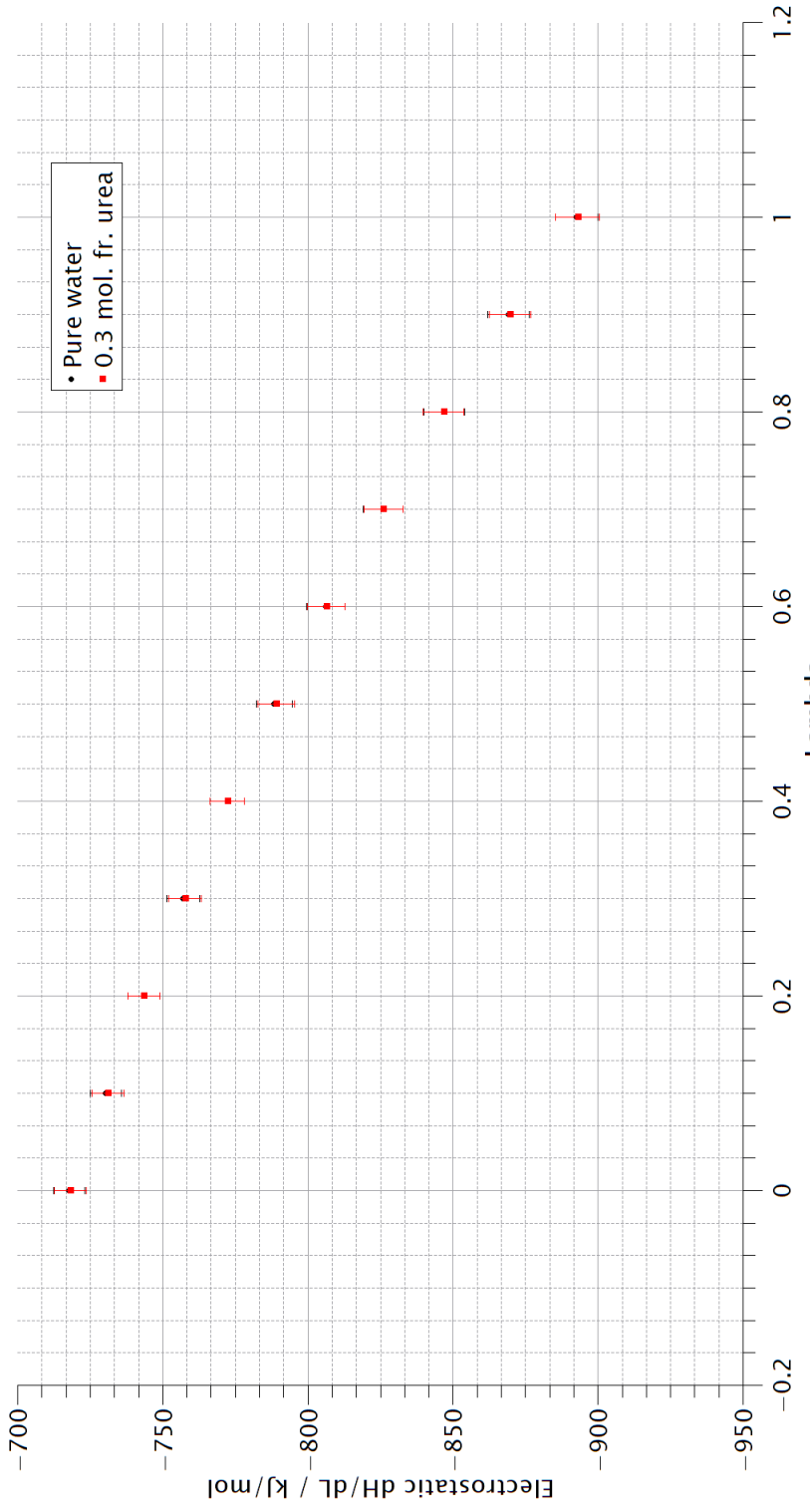


Figure 3.15: Plots of $dH/d\lambda$ for turning on the electrostatic interactions in the urea molecule according to the Özpınar model. There is negligible difference with respect to concentration.

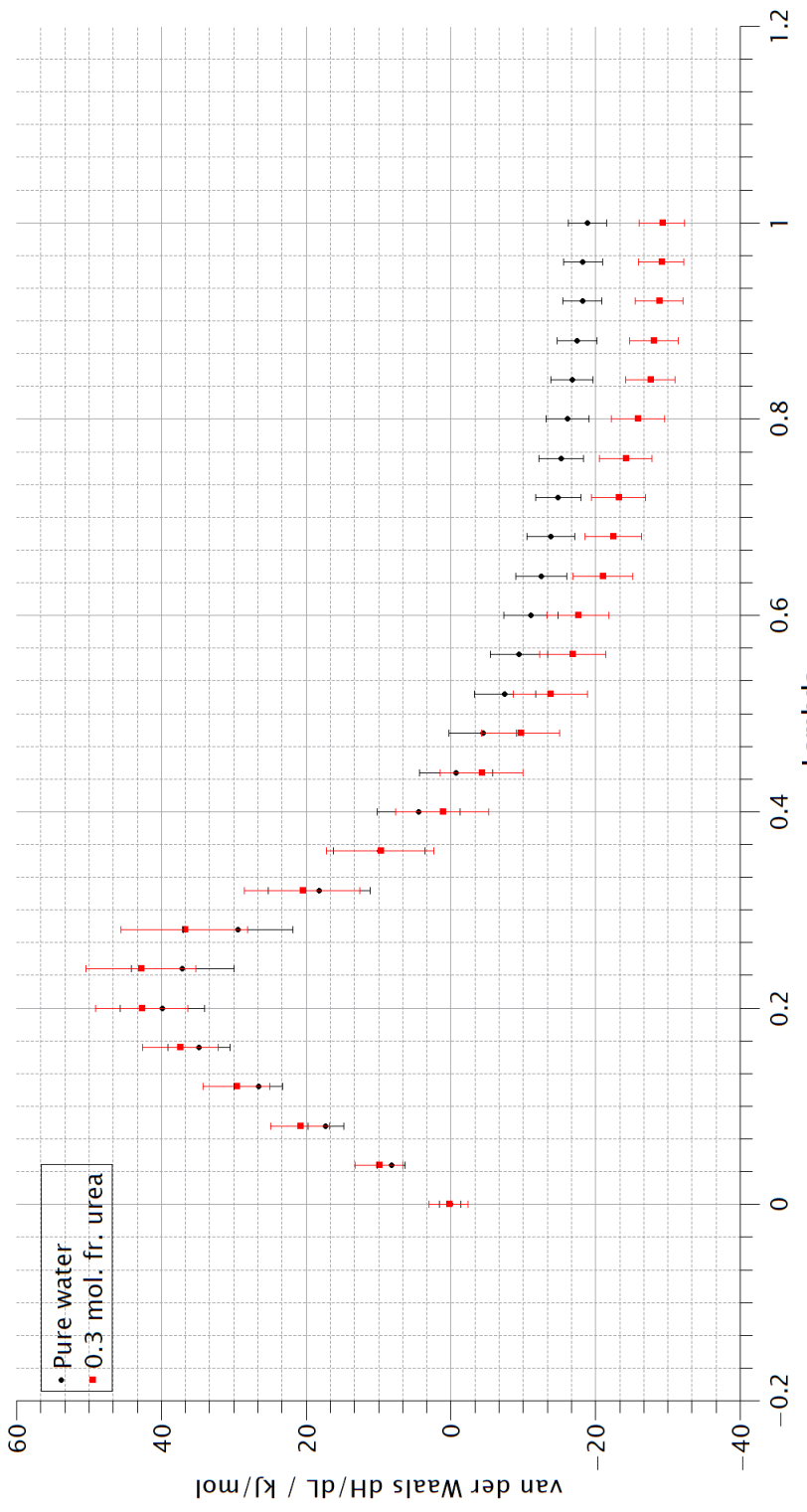


Figure 3.16: Plots of dH/dλ for turning on the intermolecular van der Waals interactions in the urea molecule according to the Özpınar model. The concentration behaviour is similar to turning on all interactions.

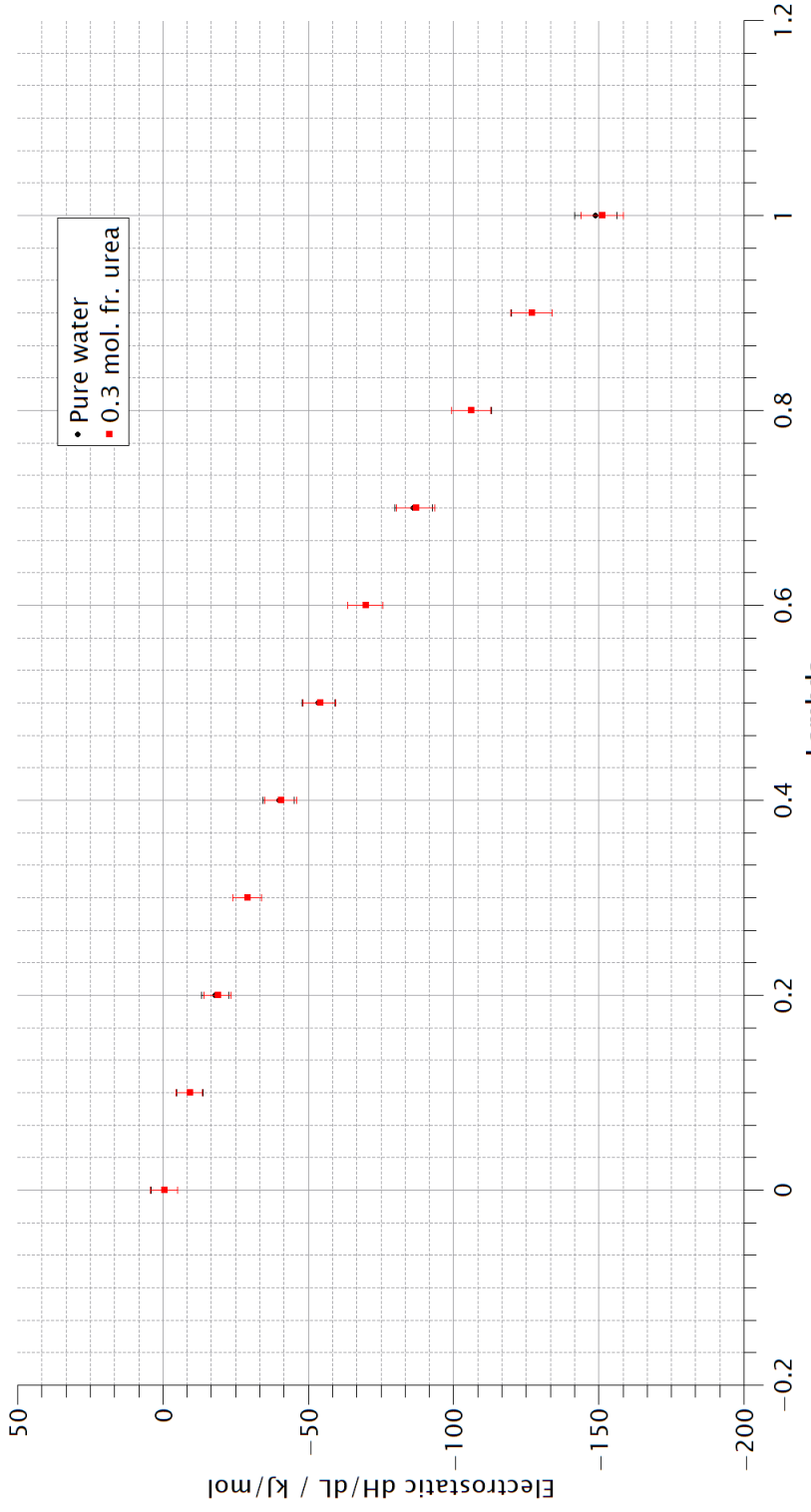


Figure 3.17: Plots of dH/dλ for turning on the intermolecular electrostatic interactions in the urea molecule according to the Özpınar model. The behaviour of the curve is qualitatively very similar to turning on all interactions.

Table 3.5: Chemical potential data for an aqueous urea solution totalling 1000 molecules through the atomic approach according to the Özpınar and TIP3P models. Energies in kJ/mol. The chemical potential is dominated by electrostatic interactions.

x(urea)	Volume / nm^3	Ideal Gas	Analytical Bond	Angle/Dihedral	AD Error	vdW	vdW Error	Electrostatic	ES Error	Total	Total Error
0	30.3771	−245.034	179.466	105.578	0.031	1.913	0.049	−794.369	0.038	−752.446	0.069
0.05	32.1959	−236.388	180.481	105.578	0.031	0.651	0.051	−794.528	0.040	−744.206	0.072
0.1	34.0484	−235.799	181.458	105.578	0.031	−0.194	0.058	−794.557	0.043	−743.514	0.079
0.15	35.9006	−235.853	182.383	105.578	0.031	−0.750	0.065	−794.728	0.045	−743.370	0.085
0.2	37.7777	−236.157	183.273	105.578	0.031	−1.480	0.066	−794.821	0.049	−743.607	0.088
0.25	39.6586	−236.572	184.121	105.578	0.031	−1.721	0.070	−794.621	0.052	−743.215	0.093
0.3	41.5683	−237.058	184.942	105.578	0.031	−2.058	0.074	−794.717	0.061	−743.313	0.1

Table 3.6: Chemical potential data for an aqueous urea solution totalling 1000 molecules through the molecular approach according to the Özpınar and TIP3P models. Energies in kJ/mol. Electrostatics dominate to a lesser extent compared to the atomic route.

x(urea)	Volume / nm^3	Ideal Gas	vdW	vdW Error	Electrostatic	ES Error	Total	Total Error
0	30.3695	−34.986	1.564	0.063	−61.173	0.064	−94.595	0.090
0.05	32.1775	−25.324	0.215	0.072	−61.314	0.066	−86.423	0.098
0.1	34.0099	−23.759	−0.309	0.076	−61.133	0.065	−85.201	0.100
0.15	35.8739	−22.888	−1.011	0.083	−61.345	0.067	−85.244	0.107
0.2	37.7335	−22.302	−1.291	0.090	−61.361	0.067	−84.954	0.112
0.25	39.6234	−21.869	−1.609	0.093	−61.557	0.072	−85.035	0.118
0.3	41.497	−21.527	−2.028	0.100	−61.417	0.072	−84.972	0.123

3.3.2 Hölzl Model

Direct Coexistence

The kinetics for the Hölzl and TIP4P/2005 system were much slower than for the Özpınar and TIP3P combination and does not display the same ”layer-by-layer” dissolution behaviour. Subsequently a much wider range of solubility for urea of 0.03–0.11 molar fraction was given by the pure and supersaturated simulations as shown in figures 3.18 and 3.19. These simulations were performed very late in the study so there was not enough time for fully ascertain whether equilibration had been reached.

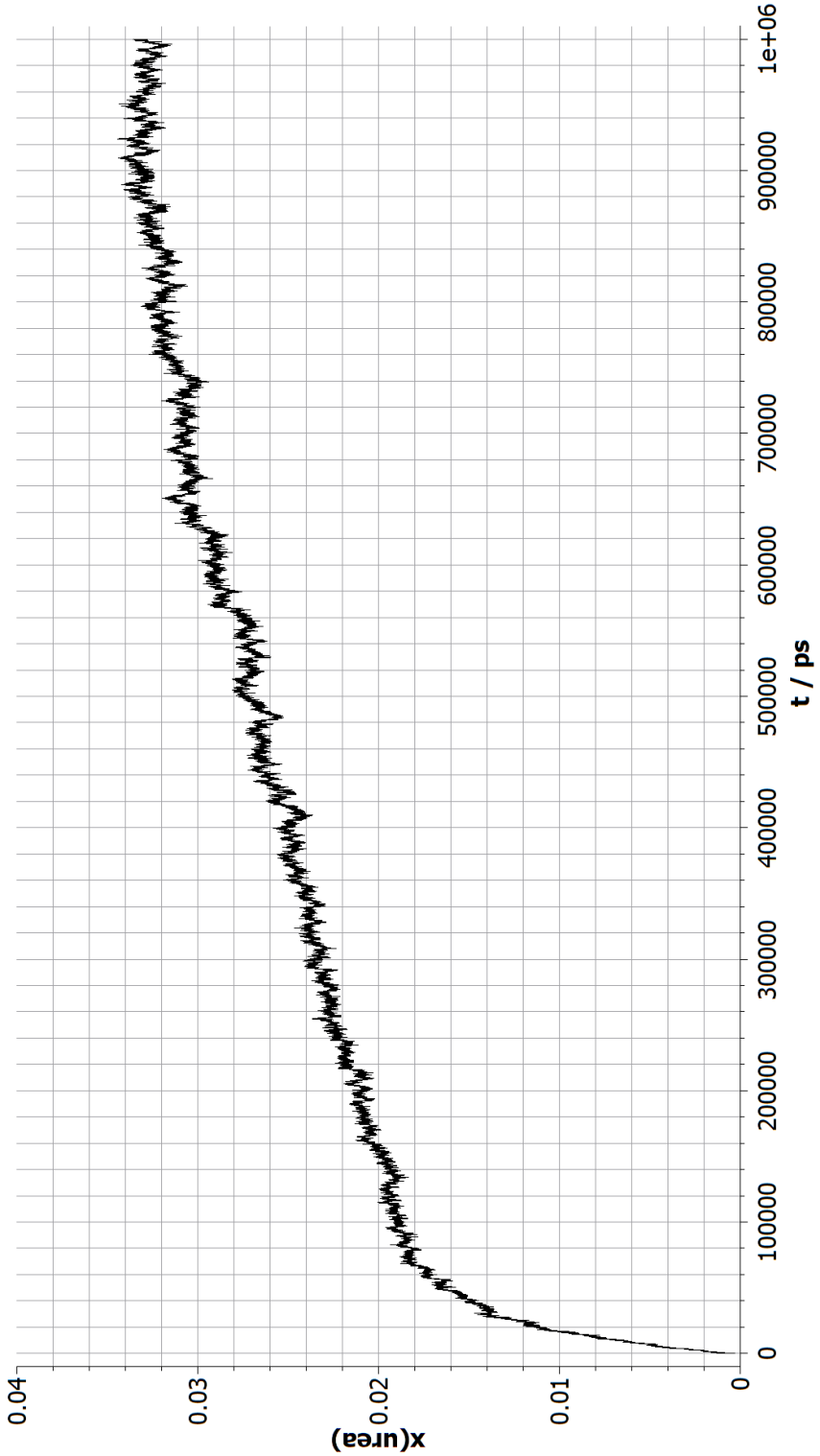


Figure 3.18: Evolution of molar fraction of urea in aqueous solution through the dissolution of a urea crystal into pure water according to the Hölzl and TIP4P/2005 models.

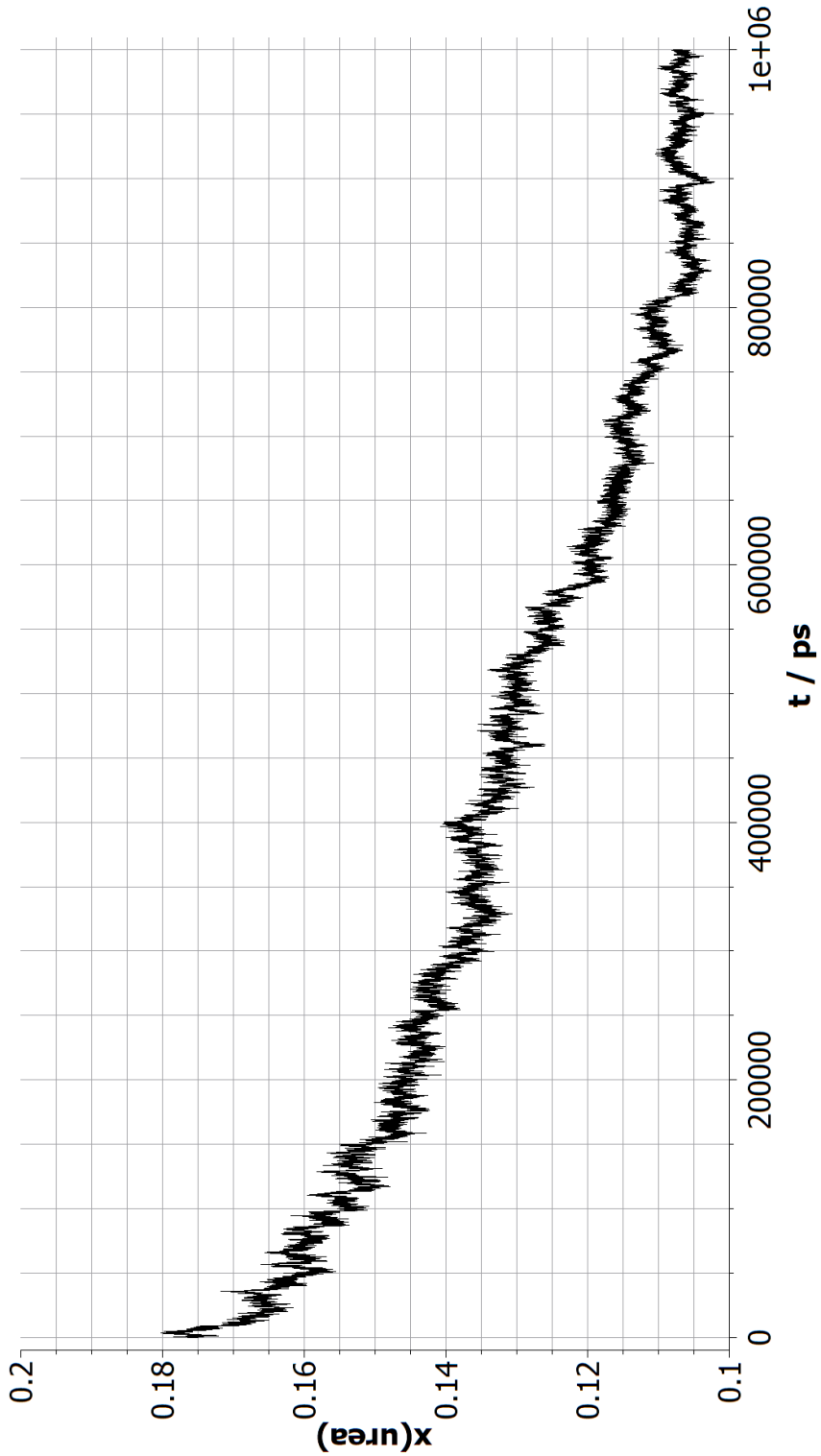


Figure 3.19: Evolution of molar fraction of urea in a supersaturated aqueous solution in contact with a crystal according to the Hölzl and TIP4P/2005 models.

Crystal Chemical Potential

The free energy contributions (in units of kJ/mol) to the chemical potential of the Hölzl model with a restraint strength of 500000 kJ/mol nm^2 through the atomic route are:

- Einstein Crystal 133.49
- Bonds, angles and dihedrals 48.901±0.002
- van der Waals −17.984±0.001
- Electrostatics −825.465±0.001
- Restraint release −76.459±0.043

giving a total chemical potential of −737.513±0.043 kJ/mol.

The contributions for the molecular route are:

- Einstein Crystal 14.127
- Orientation restriction 34.666±0.027
- van der Waals −15.199±0.001
- Electrostatics −103.344±0.005
- Restraint release −27.621±0.022

giving a total chemical potential of −97.372±0.035 kJ/mol.

3.3.3 Solution and Crystal Chemical Potential Comparison

The results of the four protocols are given in Figure 3.20. They all give very different results with the atomic route giving a lower chemical potential difference than the molecular route for both models. The Özpınar free energy simulations failed to give chemical potential differences that agreed with the direct coexistence simulations.

Table 3.7: Chemical potential data for an aqueous urea solution with the Hölzl model and TIP4P/2005 water totalling 1000 molecules from the atomic approach. Electrostatics dominate more than with the Özpınar model. Energies in kJ/mol.

x(urea)	Volume / nm^3	Ideal Gas	Analytical Bond	Angle/Dihedral	AD Error	vdW	vdW Error	Electrostatic	ES Error	Total	Total Error
0	29.6867	−247.070	179.501	107.166	0.043	4.353	0.070	−801.264	0.049	−757.314	0.096
0.05	31.7517	−238.605	180.675	107.166	0.043	4.428	0.074	−801.658	0.052	−747.994	0.100
0.1	33.922	−238.220	181.829	107.166	0.043	4.955	0.076	−801.930	0.054	−746.200	0.103
0.15	36.1197	−238.469	182.925	107.166	0.043	5.005	0.081	−802.059	0.058	−745.432	0.109
0.2	38.4025	−238.979	183.996	107.166	0.043	4.523	0.087	−802.121	0.062	−745.415	0.115
0.25	40.5578	−239.514	184.949	107.166	0.043	4.590	0.083	−802.149	0.064	−744.958	0.113
0.3	42.8175	−240.143	185.896	107.166	0.043	4.429	0.066	−801.827	0.062	−744.479	0.100

Table 3.8: Chemical potential data for an aqueous urea solution with the Hölzl model and TIP4P/2005 water totalling 1000 molecules with the molecular approach. Energies in kJ/mol.

x(urea)	Volume / nm^3	Ideal Gas	vdW	vdW Error	Electrostatic	ES Error	Total	Total Error
0	30.3695	−34.929	4.360	0.111	−61.301	1.301	−106.826	0.133
0.05	32.1775	−25.289	3.779	0.116	−61.425	1.304	−98.235	0.138
0.1	34.0099	−23.750	3.776	0.118	−61.353	1.303	−96.903	0.141
0.15	35.8739	−22.903	3.707	0.122	−61.441	1.307	−95.946	0.145
0.2	37.7335	−22.343	3.701	0.135	−61.477	1.311	−95.379	0.158
0.25	39.6234	−21.925	3.857	0.139	−61.632	1.298	−94.819	0.162
0.3	41.497	−21.607	3.549	0.145	−61.611	1.31	−95.108	0.17

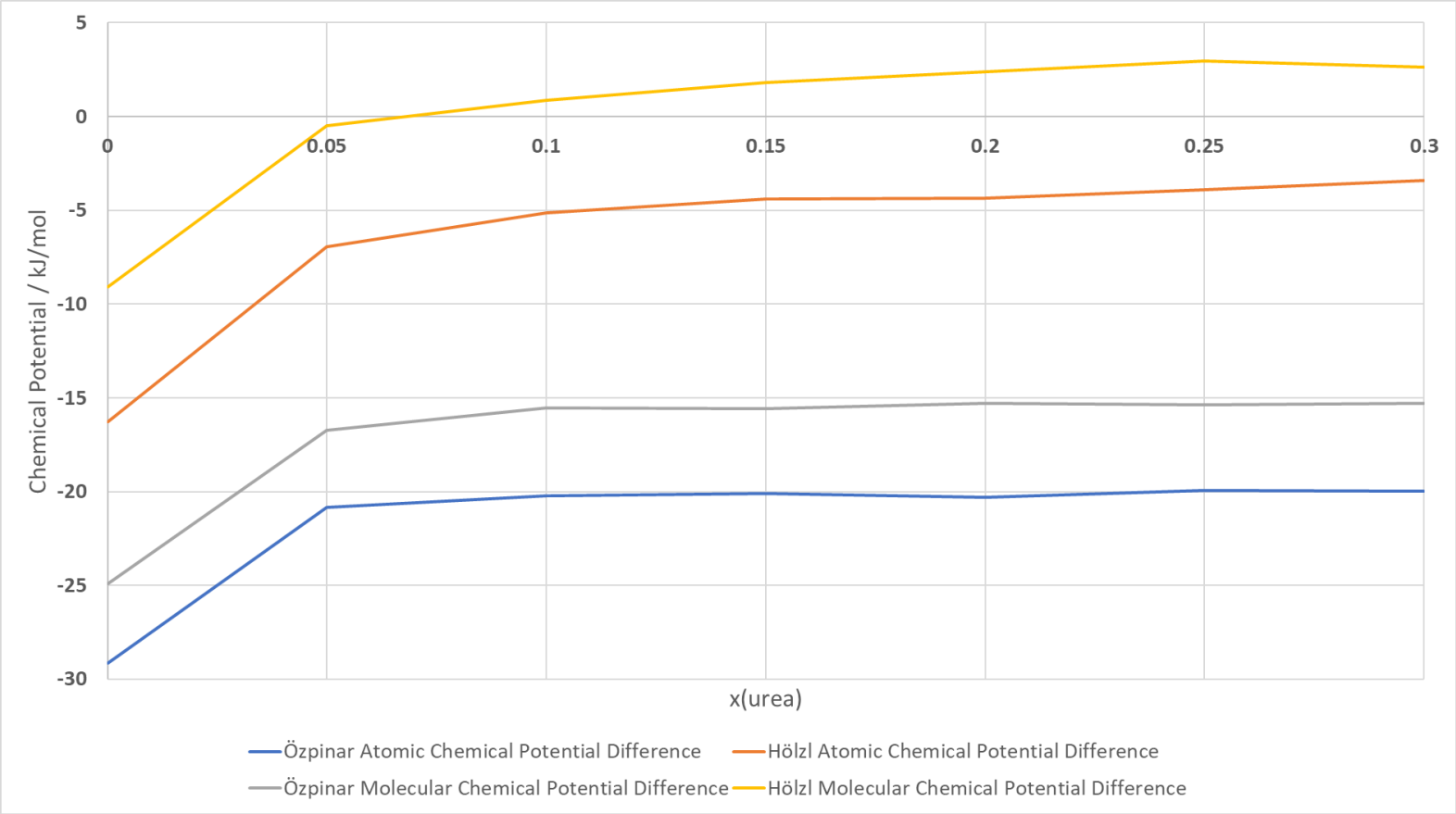


Figure 3.20: The difference between the solution and crystal chemical potentials according to the four routes in this study. For better visibility, error bars are not included. There are significantly different results from each method and only the Hölzl model and the molecular route give a sensible result with a limit on solubility.

3.4 Discussion

The molecular chemical potential pathway with the Hölzl model was the only combination of the four force field/pathway combinations that gave a sensible estimate of the chemical potential differences of the urea crystal and solution between 0.05 and 0.1 molar fraction. This agrees very well with the direct coexistence simulation range of 0.03-0.11 molar fraction.

Unfortunately, there have been major issues that have not been able to be resolved. Both urea models were subject to exactly the same simulation protocols yet the Özpınar model failed to produce chemical potential differences that made sense. The chemical potentials for the Özpınar model are significantly higher than the Hölzl model, up about 25 kJ/mol higher for the crystal in the molecular route. Human error cannot be excluded as a possibility. There are large differences in each contribution to the free energy from the two models so locating and identifying erroneous data proved very difficult.

The difference in chemical potentials are significantly lower in the atomic route for both force field sets, which indicates an inherent error in data production for free energies particular to this route. Until this issue can be resolved, the atomic route cannot be recommended for chemical potential calculations. It is suspected that the issue lies in the transformation of the harmonic functions in the force field as the calculations of their free energy contributions can be particularly difficult at high strengths.

The chemical potential difference for both models show a similar drop in chemical potential difference in the atomic thermodynamic pathway compared to the molecular pathway. This indicates that there could be an issue with the handling of intramolecular forces in free energy calculations.

Even if the unknown protocol errors are corrected, there are still valid observations to be made. Most importantly, the precision of the crystal chemical potential calculations are excellent regardless of force field or thermodynamic pathway. Solution chemical potential errors are an order of magnitude larger but still very good at below 0.2 kJ/mol — this is because of only being able to transform a single molecule so sampling efficiency

is reduced. For both urea models, the van der Waals contribution to the chemical potential in solution decreases significantly as a function of concentration. This leads to a very flat behaviour for the chemical potential at higher concentrations such that statistical error can easily have a large effect on the solubility estimate. The electrostatic free energy contribution to the solution chemical potential is surprisingly flat as a function of concentration, indicating that the interactions of urea molecules in the solution phase with water molecules and other urea molecules are of similar strength. While water and urea are both strongly polar with high hydrogen bonding potential, that consistency was not anticipated.

Chapter 4

Prediction of the Mutual Solubility of Butanol and Water as a Function of Temperature

4.1 Introduction

Mutual solubility is of fundamental interest to a large variety of applications including ionic liquids [117], geology of the Earth's mantle [118] and fuel engineering [119]. The mutual solubility of liquids is more complicated in comparison to solvation of solids as there are multiple competing solvation processes in thermodynamic balance. Computational methods are particularly important for investigating systems like the Earth's mantle as they are impossible to recreate in lab conditions.
The butanol-water system was chosen for its chemical simplicity and interesting phase behaviour where the solubility of butanol in water at low temperatures decreases as a function of temperature, reaches a minimum and then the two become miscible close to the boiling point of water [120, 121]. This study seeks to recreate this phase behaviour through chemical potential calculations and direct coexistence simulations. Direct coexistence simulations should be particularly effective as equilibration of fluid systems is rapid.

4.2 Methods

4.2.1 Direct Coexistence

The initial direct coexistence system comprised pure blocks of 2400 butanol molecules and 12000 water molecules in a tetragonal box with the z-axis around 3 times the x and y-axes. At 293 K the equilibrated box dimensions were $6.23 \times 6.23 \times 18.68$ nm. The system was simulated at a series of temperatures from 233 K to 413 K (-40°C to 140°C) in 20 K steps. Simulation was continued until the concentrations in the water and butanol-rich phases were stable for a significant amount of time. Molar fractions of water and butanol were tracked by slicing the system along the z-axis and over time using *gmx select* to count how many water and butanol molecules were present in each slice. The slicing is used to determine where the phase boundary is during the simulation as the two phases may significantly drift in the z-axis. Once equilibrium was achieved and a significant amount of time had passed, the average number density was then determined which was then converted to other quantities to compare with experimental data.

4.2.2 Free Energy Calculations

Because of the complication of mutual solubility, the chemical potentials of butanol and water have to be determined across the whole composition range from pure water to pure butanol. The same series of temperatures as for direct coexistence were used. All free energy calculation systems were 3000 atoms to maintain a consistent system size and the compositions used are given in Table 4.1.

The systems were equilibrated for 10 ns (5,000,000 $\times$ 2 fs) before the free energy simulations with equilibrium determined by consistent volume and potential energy. Each free energy simulation was 10 ns with the first 0.5 ns of thermodynamic data ignored in free energy calculations to account for any equilibration due to the perturbed molecule. The stochastic dynamics integrator was used for improved sampling with a friction constant of 1 ps. The Parrinello–Rahman barostat was used with a coupling constant of 2 ps and compressibility of 4.5×10^{-5} bar. A potential-switch function was used for van der Waals interactions and Particle-Mesh Ewald for electrostatics with a cut-off distance of 1.2/nm. Sampling of energies required for free energy analysis occurred every 1000 steps. Butanol bonds were constrained with the LINCS algorithm with water handled by the SHAKE algorithm. In total, this study used 1,776 simulations for free energy calculations. The calculated excess free energies were taken to create mixing free energies and fitted to Equation 2.16.

Table 4.1: Butanol/water system compositions used in free energy calculations. N is the number of molecules.

N(butanol)	N(water)	x(butanol)
0	1000	0
1	995	0.0010
2	990	0.0020
4	980	0.0041
6	970	0.0061
8	960	0.0083
10	950	0.0104
20	900	0.0217
40	800	0.0476
60	700	0.0790
80	600	0.1176
100	500	0.1667
120	400	0.2308
140	300	0.3182
160	200	0.4444
180	100	0.6429
184	80	0.6970
188	60	0.7581
192	40	0.8276
196	20	0.9074
197	15	0.9292
198	10	0.9519
199	5	0.9755
200	0	1

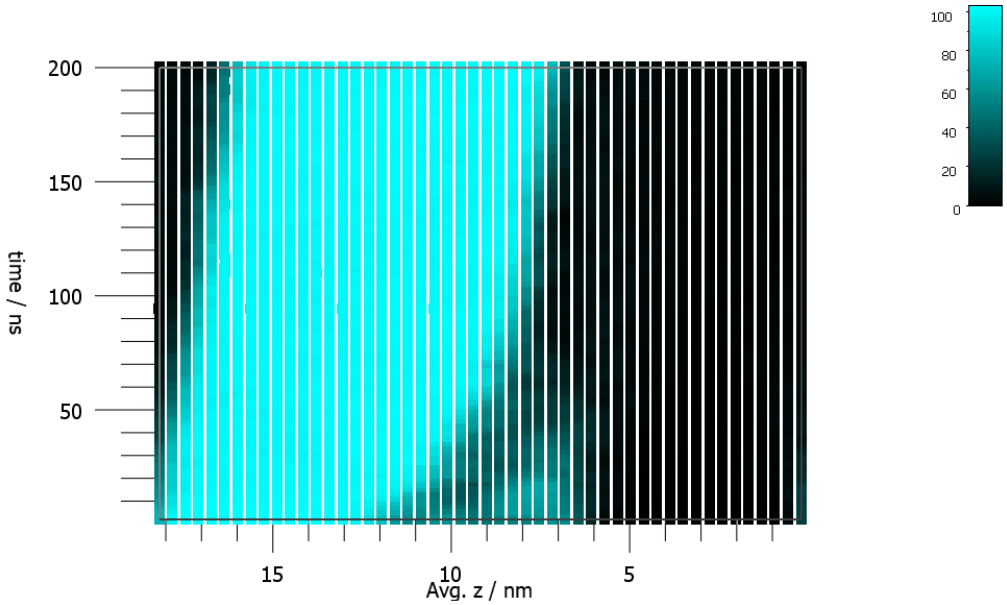


Figure 4.1: Count of water molecules in a butanol-water direct coexistence simulation at 233 K as a function of time and average z-coordinate with cyan representing the water rich phase and black the butanol rich phase. There is an initial separation of water phases which quickly merged and then on there is a sharp phase boundary which slightly drifted with time. The points away from the phase boundaries after 100 ns were used to determine the average number densities of water in each phase. The same process was used for butanol where the colours are essentially switched.

4.3 Results

4.3.1 Direct Coexistence

The butanol and water distributions as a function of time and z-coordinate are given visually for various temperatures (Figures 4.1 to 4.4). Numerically, these figure are graphs of counts of water molecules as a function of z-coordinate and time but they show why stratifying the number density in both space and time is important as the phase boundaries clearly drift and equilibrium is not a straightforward process. All the simulations were essentially equilibrated after 100 ns and the points after this time, determined to be significantly away from were used to determine the average number densities of butanol and water as a function of temperature.

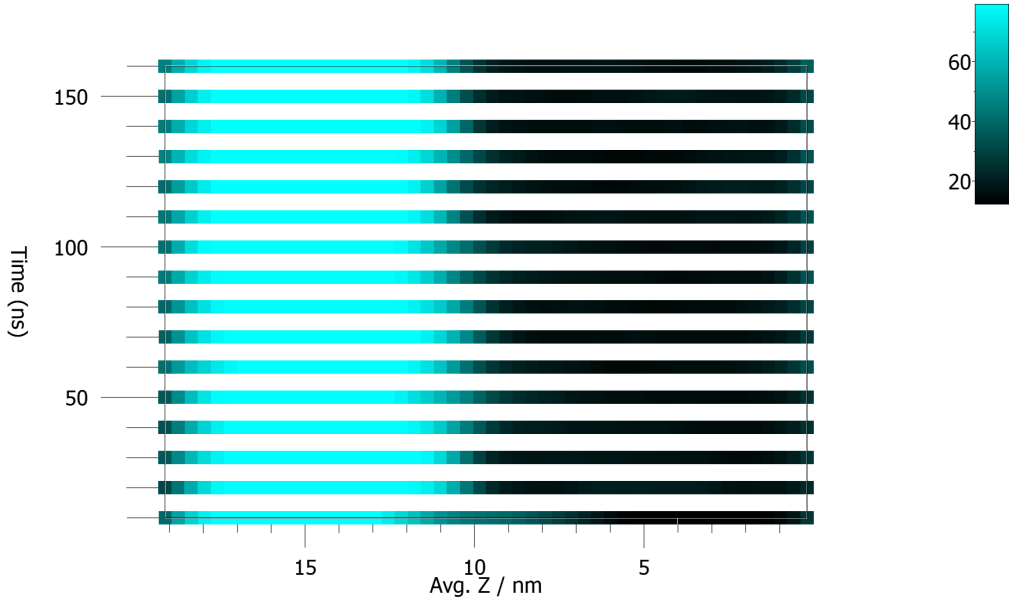


Figure 4.2: Count of water molecules in a butanol-water direct coexistence simulation at 373 K as a function of time and average z-coordinate with cyan representing the water rich phase and black the butanol rich phase. The phase behaviour is much simpler than at 233 K with no significant drift in the phase boundary. The difference may merely be coincidental.

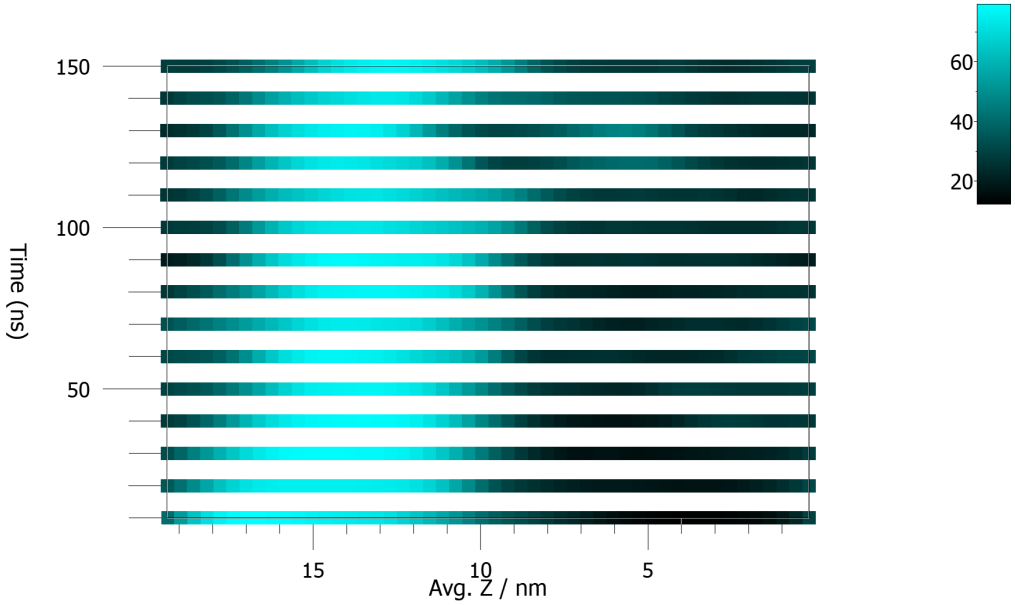


Figure 4.3: Count of water molecules in a butanol-water direct coexistence simulation at 393 K as a function of time and average z-coordinate with cyan representing the water rich phase and black the butanol rich phase. The phase boundary has become very diffuse with a lot of drift making statistical analysis difficult. This is very close to the critical temperature of miscibility.

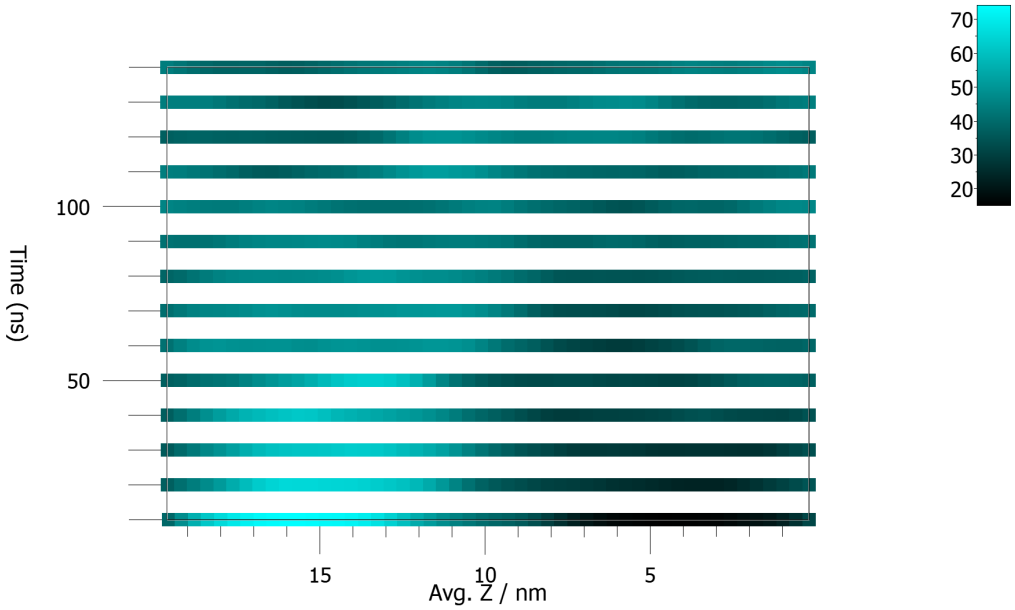


Figure 4.4: Count of water molecules in a butanol-water direct coexistence simulation at 413 K as a function of time and average z-coordinate with cyan representing the water rich phase and black the butanol rich phase. Phase separation has completely broken down and the two liquids are miscible.

It can be seen that as the temperature increases, the phase boundary becomes more diffuse. For lower temperatures the change in phase boundary behaviour is slight but as the system approaches the critical temperature, the boundary quickly becomes broader until the whole system becomes miscible. The simulation derived solubilities of butanol and water are presented as mass fractions in Table 4.2 and compared with experimental data. Unfortunately, there is little numerical agreement of the direct coexistence data with the experimental data. However, the qualitative behaviour is similar, particularly with the solubility of butanol in water showing a local minimum. Estimation of solubility close to the critical temperature of miscibility becomes very imprecise as the phase boundary broadens. The critical temperature of miscibility has not been precisely defined but it appears to roughly agree with experimental data showing it to be just above 100°C.

4.3.2 Free Energy Calculations

The free energy of mixing for a binary mixture was obtained by calculating the excess chemical potentials of both components. In this study the excess chemical potential is separated into van der Waals and electrostatic components. Representative $dH/d\lambda$ curves for each transformation are shown in Figure 4.6 and the excess chemical potential components are given in Figure 4.8. The van der Waals $dH/d\lambda$ curves are mostly unremarkable with the water curves showing a stronger rejection at low λ. The Coulombic $dH/d\lambda$ curves in the butanol rich phase show an interesting two-part behaviour where at high λ the interaction distinctly becomes more favourable. This is maybe the region where the perturbed water molecule or butanol hydroxyl group is preferentially associated with the hydroxyl chains. The water-rich systems are more favourable for electrostatic interactions while the butanol-rich systems are more favourable for van der Waals interactions. Remarkably, for both interactions, the majority of the change occurs in the water-rich region with a smaller gradient in the butanol-rich phase.

The systems used in the free energy simulations were stable against phase separation throughout the composition range. The density analysis as a function of composition (Figure 4.7) shows an uncomplicated relationship and statistical errors were consistently very small (smaller than the symbol sizes) which corroborated the stability.

Table 4.2: Simulated values of saturated mass fractions (W) of butanol in the water rich phase and water in the butanol rich phase as a function of temperature.

Temperature	Simulated w(butanol)	St. dev. w(butanol)	Simulated w(water)	St. dev. w(water)
-40	0.1075	0.0306	0.0232	0.0131
-20	0.08097	0.0281	0.0395	0.00958
0	0.0817	0.0254	0.051	0.0114
20	0.09951	0.027	0.0621	0.0115
40	0.1155	0.0296	0.0836	0.013
60	0.1516	0.0375	0.104	0.0164
80	0.2057	0.0378	0.149	0.0209
100	0.2896	0.077	0.218	0.0301

Table 4.3: Experimental values of saturated mass fractions (W) of butanol in the water rich phase and water in the butanol rich phase as a function of temperature according to Stephenson [121].

Temperature	Experimental w(butanol)	Experimental w(water)
0	0.1033	0.19
9.6	0.0898	0.197
20	0.0803	0.201
30.8	0.0707	0.206
40.1	0.0677	0.214
50	0.0654	0.222
60.1	0.0635	0.24
70.2	0.0673	0.248
80.1	0.0704	0.274
90.6	0.0726	0.306

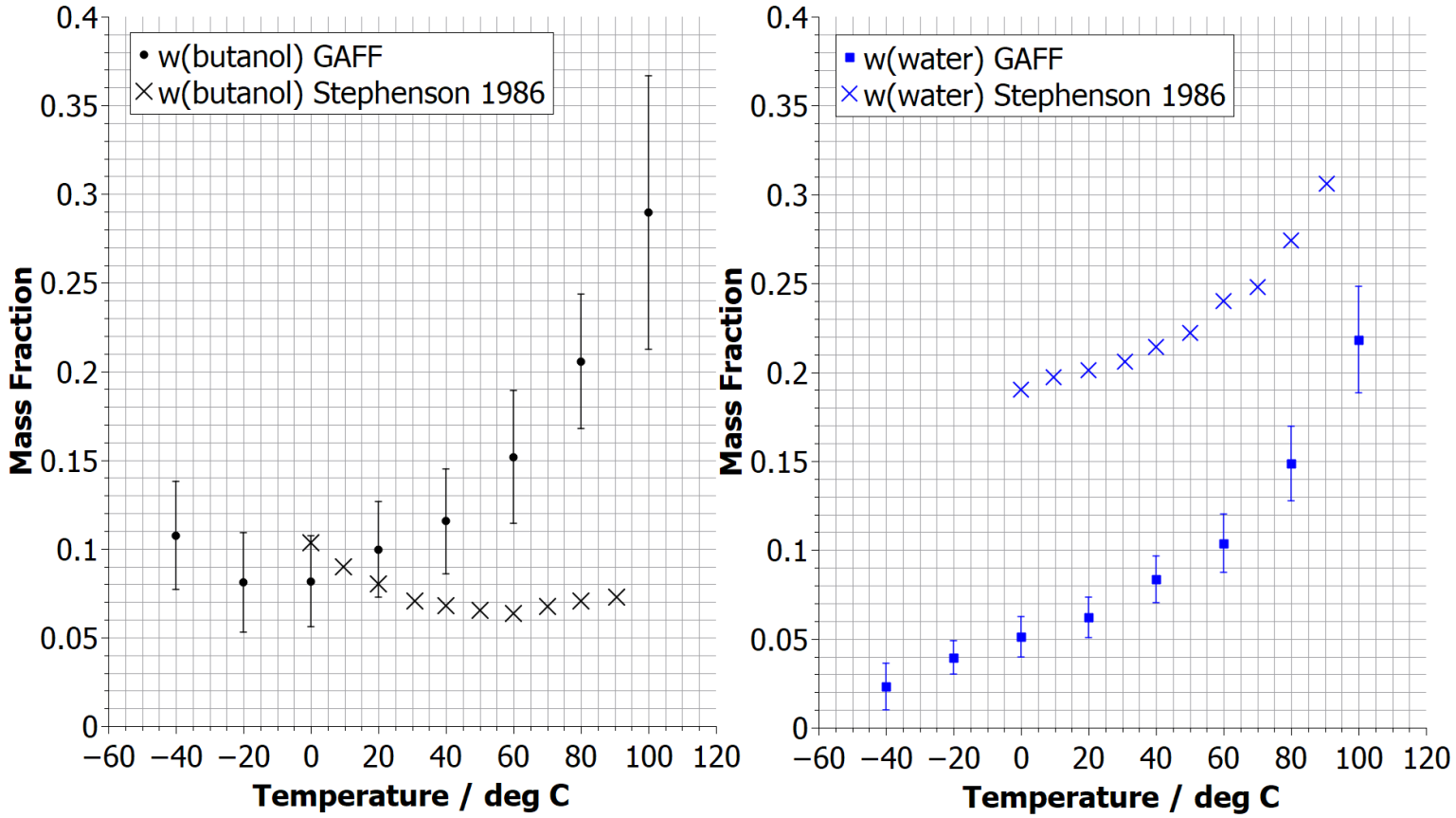


Figure 4.5: Mass fraction at solubility of water in the butanol rich phase and butanol in the water-rich phase as a function of temperature according to this study's direct coexistence simulations and experimental data [121]. As the direct coexistence simulation at 120°C produced a miscible system, the critical temperature for GAFF butanol and TIP3P water is somewhere between 100 and 120°C.

Unfortunately, the free energy data for butanol-rich systems had very high statistical uncertainty (as seen in figures 4.9 to 4.14) and this made constructing a chemical potential curve as a function of composition very difficult with poor R^2 for all fitted curves. The statistical errors follow the opposite trend to direct coexistence as they decrease with increasing temperature. As a result of the large errors, in this study it was not possible to obtain sensible estimates of solubility from free energy calculations. Despite this poor data, there is still some indication that water is more soluble in butanol-rich systems than butanol in water-rich systems, which is consistent with the direct coexistence data.

4.4 Discussion

The butanol-water direct coexistence simulations were effective for deriving statistically robust solubility data (Table 4.2). Despite the numerical inaccuracy of the simulation data versus experimental data (Figure 4.5), there was a decent reproduction of the qualitative behaviour of the solubility of butanol and water in each other as shown in the closeness of the critical temperature of miscibility at just above 100°C and the local minimum solubility of butanol in water which is only about 0.02 mass fraction off but disagreeing in temperature. This is a good step forward for the use of direct coexistence simulations for mutual solubility of liquids as it can be seen from this study that the inaccuracy of the data comes from the molecular force fields rather than the simulation protocols. The high variance of the molecular counts for systems near miscibility is potentially an issue but one could look at the standard error and extend simulation until the standard error is reduced below an acceptable threshold.

The free energy calculations of butanol and water across the composition range unfortunately failed to produce reliable data from which to derive solubilities. The variance of the free energy data are simply far too large for what turned out to be a small range of mixing free energy to cover (see figures 4.9 to 4.14). There were persistent difficulties with obtaining precise and reliable free energy estimates for butanol-rich systems in particular. The hydroxyl groups of butanol form chains dividing the system into hydrophilic, electrostatic-dominated and hydrophobic, van der Waals-dominated regions. This transient chaining behaviour may be a factor of frustration in the pre-

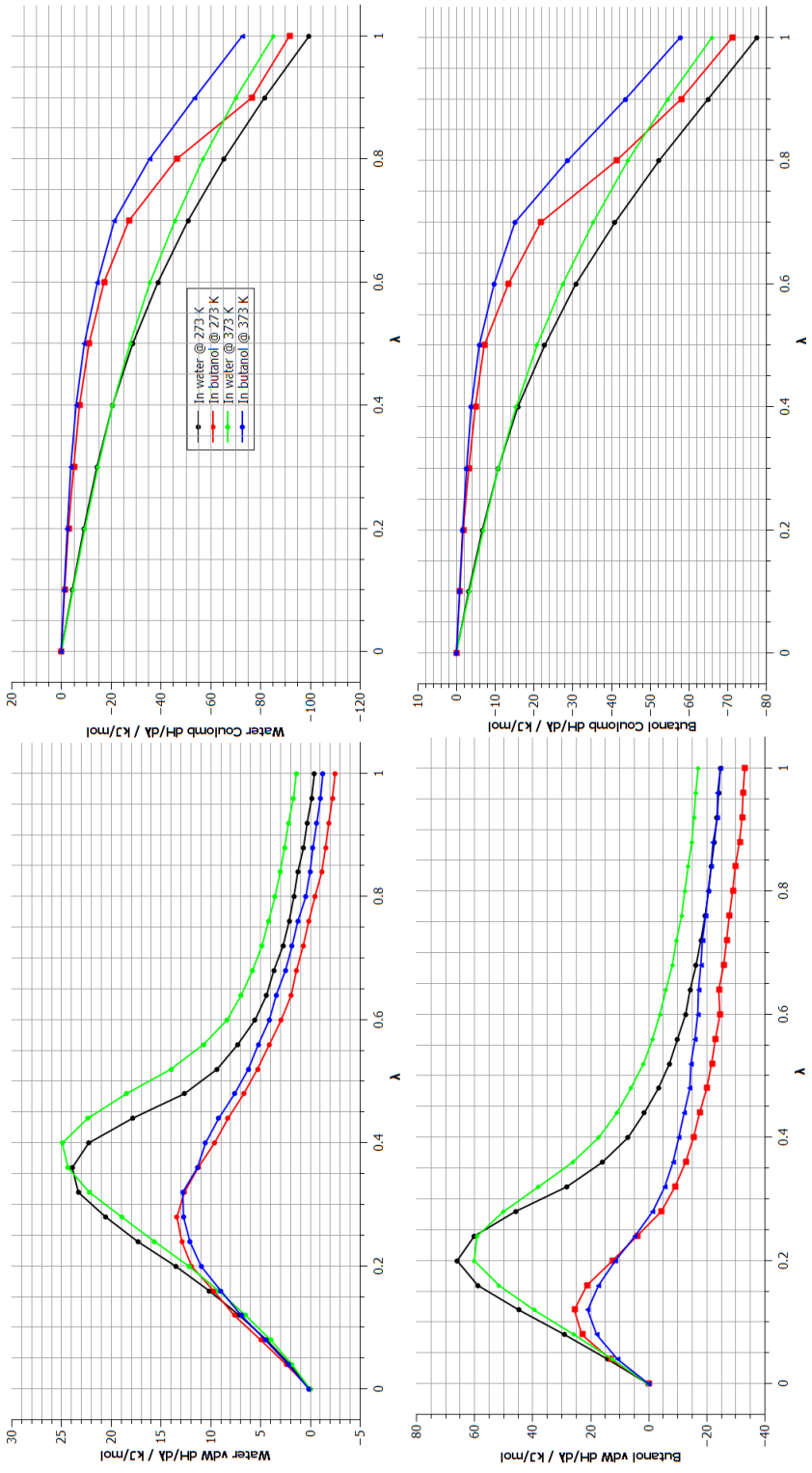


Figure 4.6: The dH/dλ curves at the extreme temperatures for water and butanol in pure phases. Water induces a stronger initial rejection response from the environment than butanol. In pure butanol, a peculiar shifting behaviour is shown in the electrostatic curves. This may be where the probe molecule becomes preferentially attracted to the hydroxyl chains in butanol.

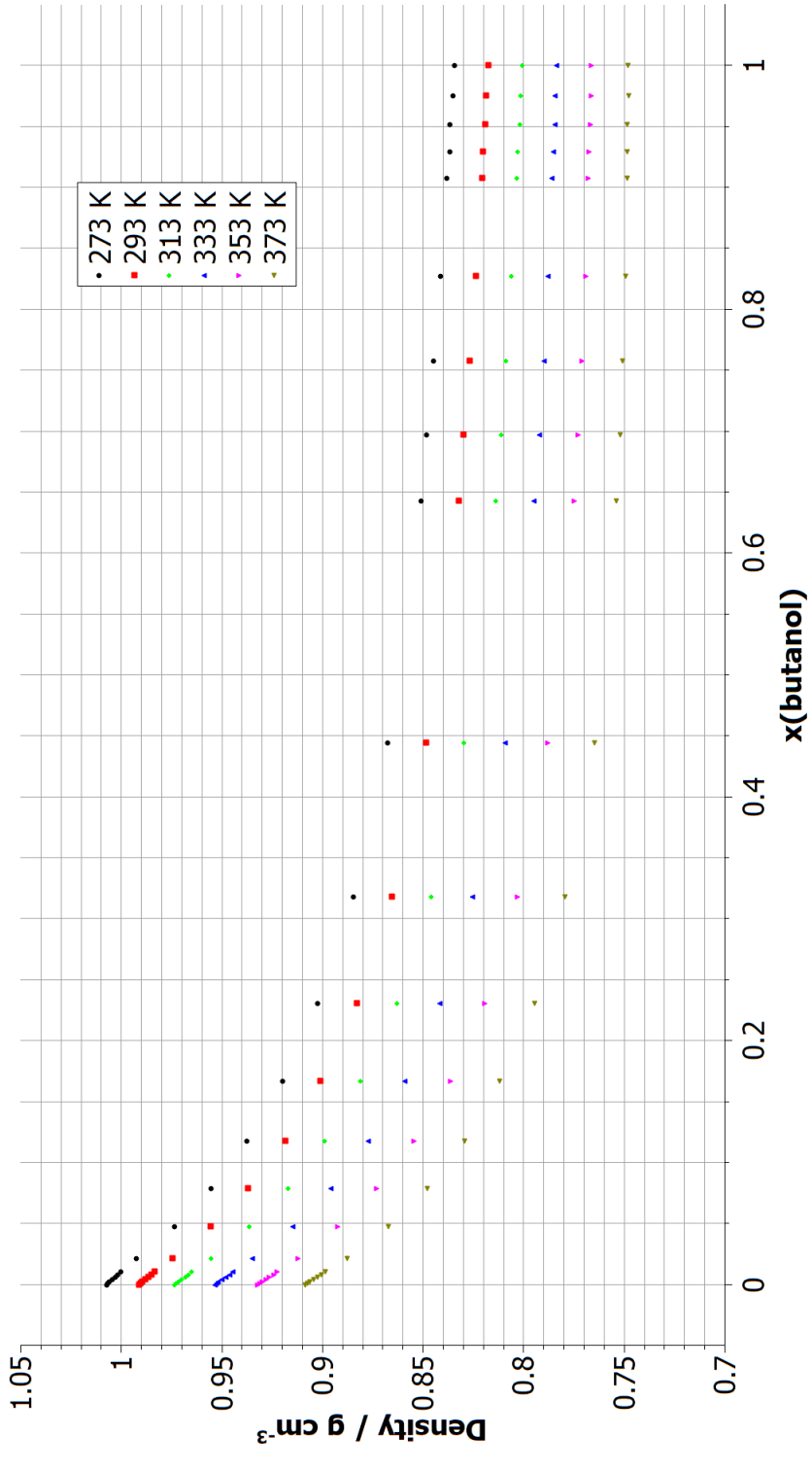


Figure 4.7: The density of the butanol-water system as a function of composition and temperature. The change of density with temperature is uncomplicated. As a function of composition, density is lower than a simple linear combination of the densities of the two components.

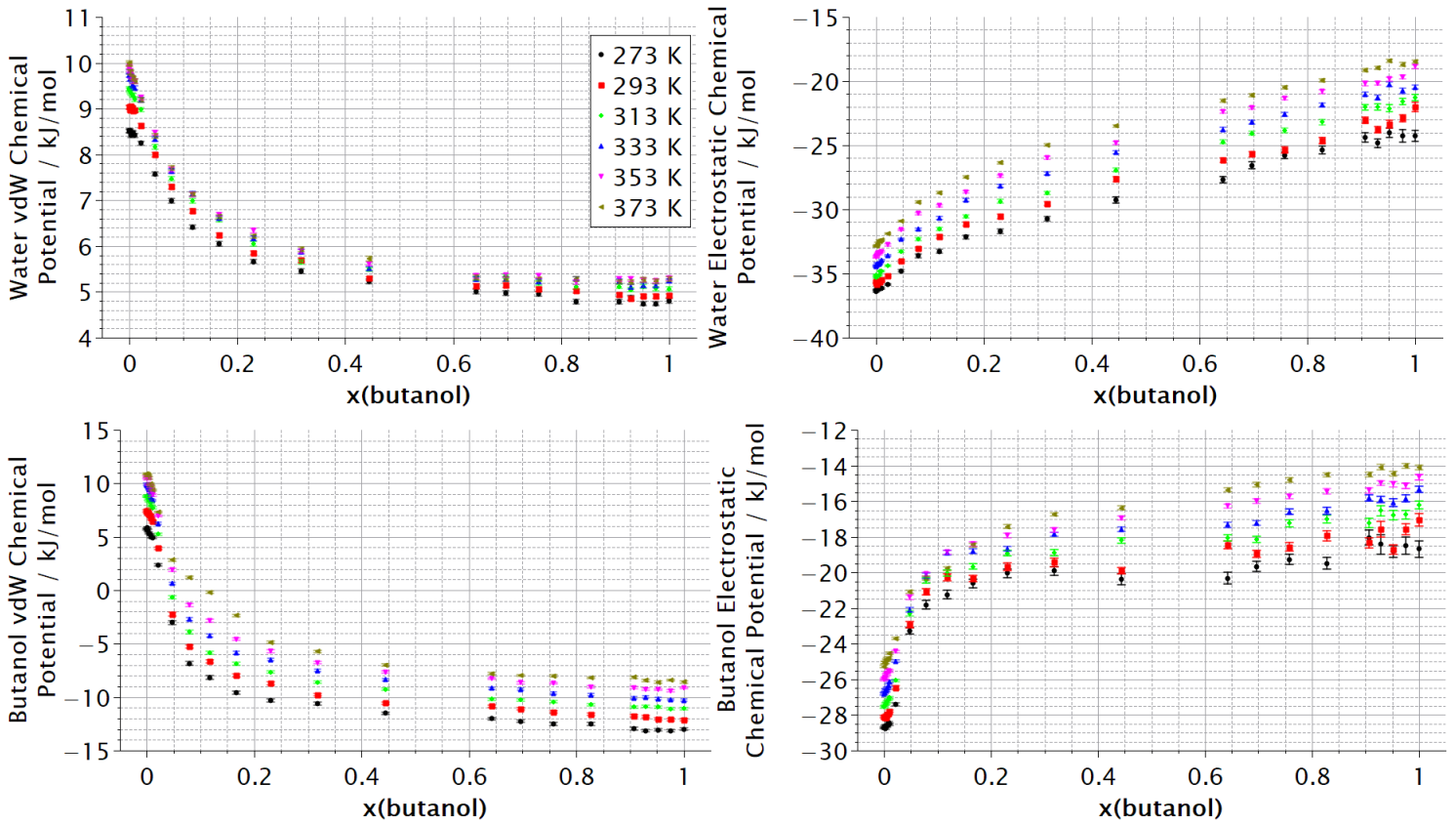


Figure 4.8: The van der Waals and electrostatic contributions to the excess chemical potential of butanol and water in the butanol/water mixture as a function of composition and temperature. It can be seen that chemical potential increases with temperature. The vdW contribution is more favourable in butanol rich systems while electrostatic interactions are more favourable in water rich systems, behaviours which are consistent with chemical intuition.

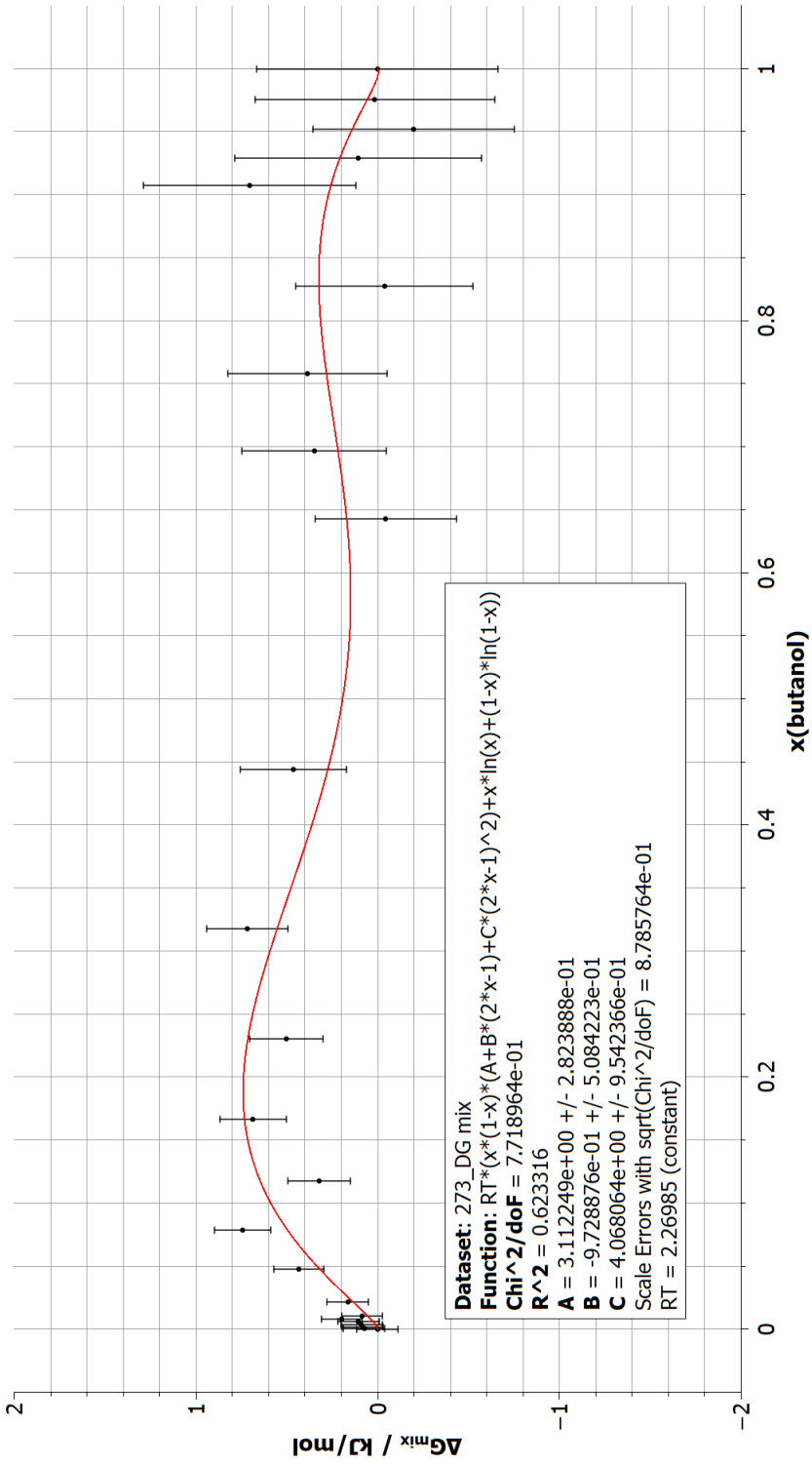


Figure 4.9: The mixing free energy change of butanol and water as a function of composition at 273 K. No minima close to the pure states are immediately visible.

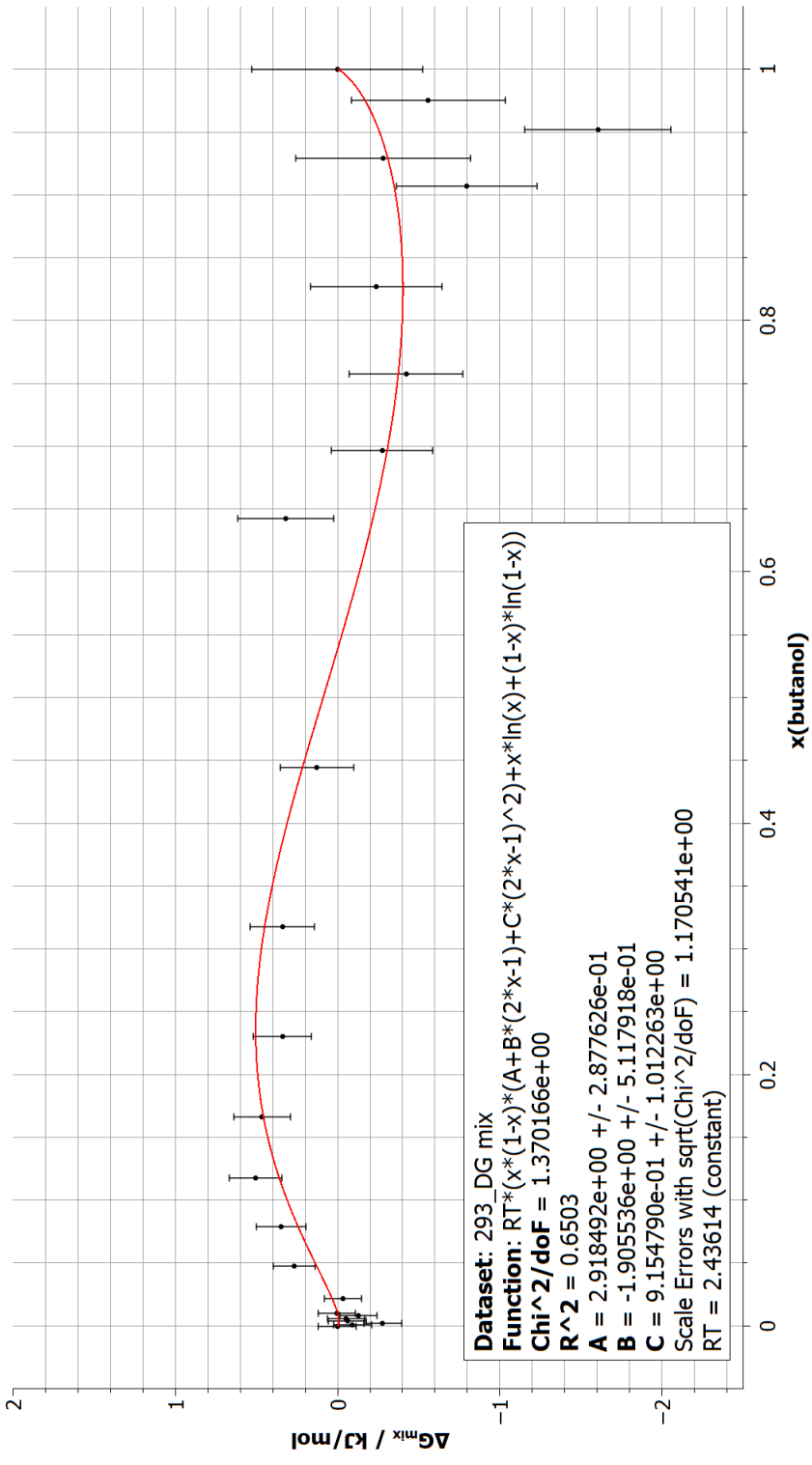


Figure 4.10: The mixing free energy change of butanol and water as a function of composition at 293 K. The spread of data at the butanol end prevents a sensible fit.

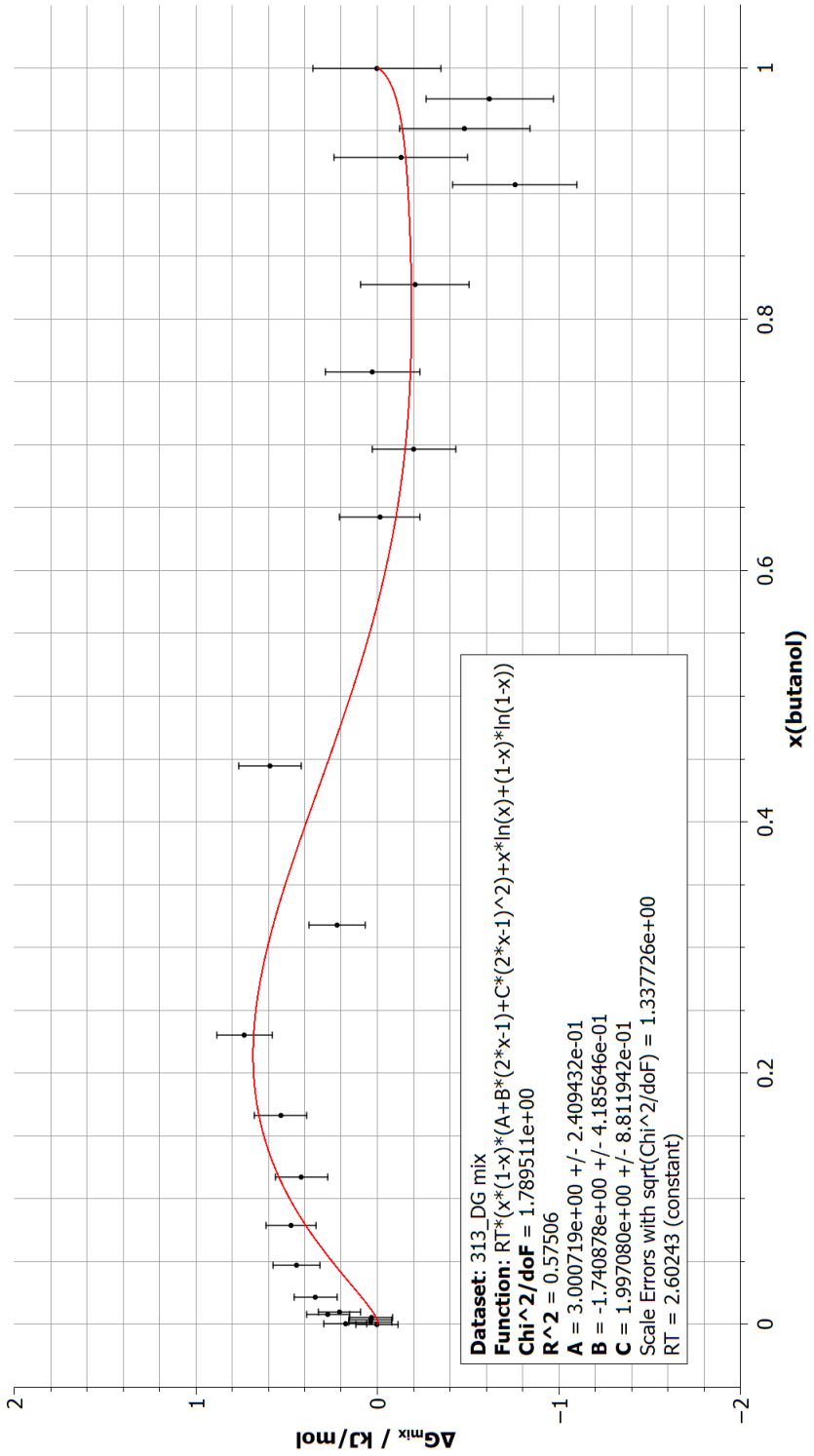


Figure 4.11: The mixing free energy change of butanol and water as a function of composition at 313 K. The shape of the fitted curve is not consistent with the others.

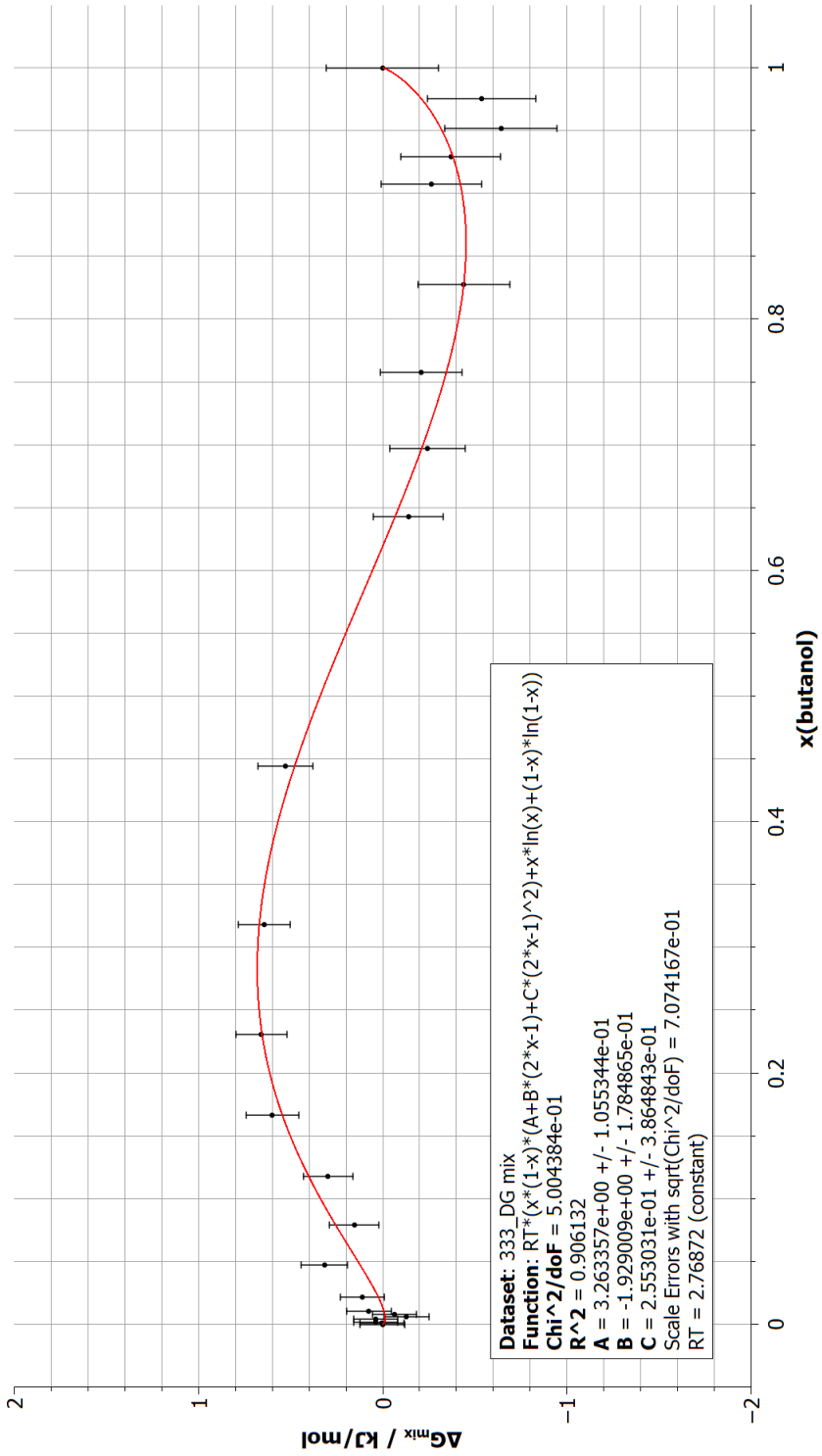


Figure 4.12: The mixing free energy change of butanol and water as a function of composition at 333 K. This seems to have given the best behaved fit but it is still very poor.

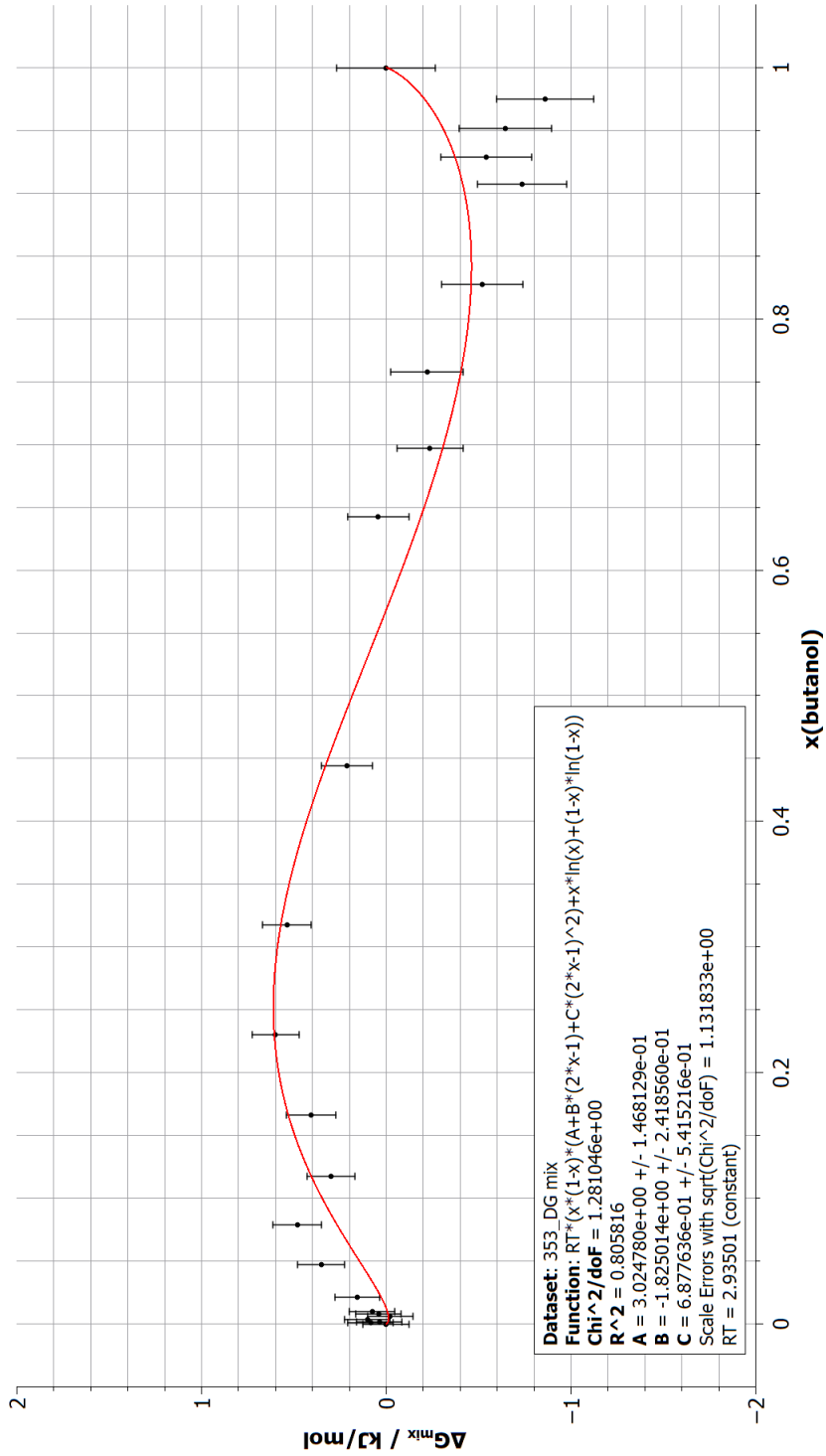


Figure 4.13: The mixing free energy change of butanol and water as a function of composition at 353 K. It has similar behaviour to 333 K.

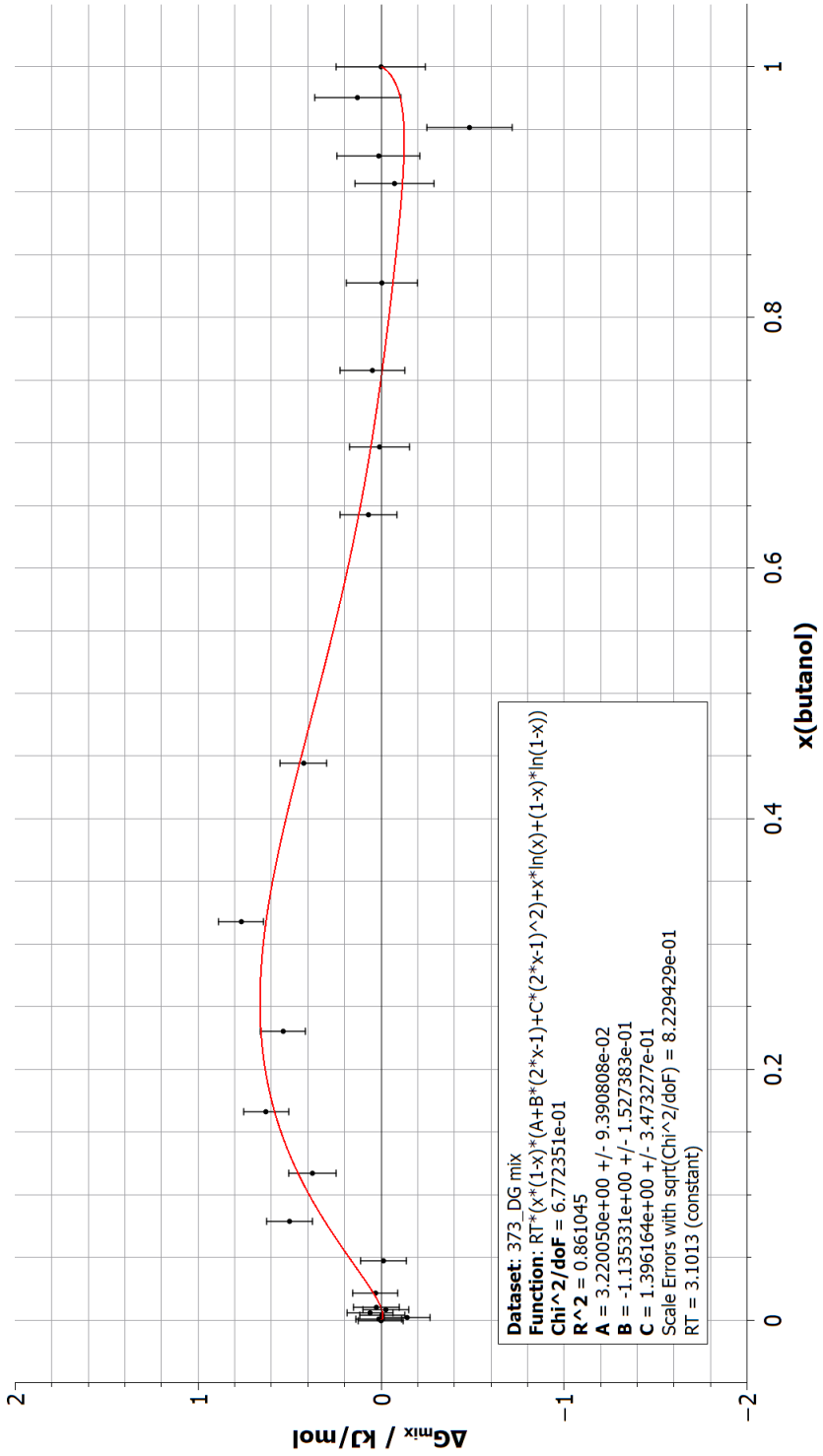


Figure 4.14: The mixing free energy change of butanol and water as a function of composition at 373 K. The minimum at the butanol end is much shallower than the fits at lower temperatures.

cision of these free energy calculations that normally require relatively homogeneous systems. Evidence of this is in the distinctive two-part behaviour of the electrostatic $dH/d\lambda$ curve and the significant decrease in statistical error at higher temperatures, where these chains have lower barriers to being created and destroyed. The butanol-rich systems were significantly perturbed by the probe molecule. As butanol is a much larger molecule than water, a system which is butanol-rich will not have as much freedom for the molecules to accommodate the probe molecule than a water-rich system of the same size. Therefore, systems composed of larger molecules require much larger systems to compensate for this lack of freedom.

The construction of a curve for the free energy of mixing two components requires a large number of simulations that take a correspondingly large amount of CPU time. If one is only interested in solubility of fluids without needing the free energy data, it will be more efficient to run direct coexistence simulations that only take one simulation for a particular temperature and pressure. Looking back at figures 4.9 to 4.14, it can be seen that a holistic approach to generating the free energy of mixing is not necessary for obtaining the solubility compositions as they will usually be close to the pure compositions. A more efficient approach would be to start with the pure system and add more of the solute until the minimum is seen in the free energy of mixing. The common tangent can then be generated from these two parts of the curve.

In conclusion, this study shows that direct coexistence simulations are preferable to free energy calculations for determining the mutual solubility of liquids from the perspective of ease-of-use and time efficiency. Calculating the free energy of mixing of two components requires a lot of simulations providing data with precision that may be very difficult to achieve, especially without dedicated software to expedite the process and avoid human errors.

Chapter 5

Comparison of Aqueous Solubilities of Polymorphs of Carbamazepine

5.1 Introduction

Polymorphism is a phenomenon in the solid state in which the molecules of a compound can have different stable arrangements of molecules depending on temperature and pressure. A polymorph is more stable than another when it has a lower chemical potential. There are two modes of polymorphism, monotropism and enantiotropism. The chemical potentials of monotropic polymorphs do not cross before the melting point and the transformation from one polymorph to the other is irreversible. The chemical potentials of enentiotropic polymorphs do cross before the melting point and can be reversibly transformed by adjusting the temperature.

Polymorphism is relevant to many areas of science and engineering. Understanding the polymorphs of silicate and oxide minerals is key to understanding the structure of the Earth's mantle — there are discontinuities in the structure of the mantle determined by the phase boundaries of mineral polymorphs [122]. The pigment 5-methyl-2-[(2-nitrophenyl)-amino]thiophene-3-carbonitrile gained the nickname ROY (Red Orange Yellow) because it can form a large range of polymorphs with various colours. It holds the record for the most polymorphs isolated with 11 as of 2020 [123]. Polymorphism in explosives has a significant effect on their energetic properties [124].

In pharmaceuticals, polymorphism is a critical priority in formulation as regulatory approval can only be given to one polymorph at a time and the sudden appearance of a new polymorph can have disastrous effects on efficacy [125]. A notable example is ritonavir [126]. The second polymorph of ritonavir was not known until after formulation and regulatory approval proceeded without issue. The critical issue was that ritonavir is not bioavailable in the solid state so had to be delivered as a solution or gel. In 1998, liquid capsules started failing solubility tests and eventually form II started appearing throughout the manufacturing process as the formulation concentration was 400% supersaturated with respect to Form II, forcing a new formulation to be developed. Therefore, the ability to determine, isolate, and control polymorphism is critical in determining and ensuring the efficacy of liquid and gel drug formulations.

The precision of computational free energy calculations for crystals is such that the stability of polymorphs should be able to be differentiated. The majority of polymorph pairs have a chemical potential difference below 2 kJ/mol [127].

Here we attempt to use molecular dynamics with a common force field to differentiate polymorph stabilities for a molecule and establish the stability hierarchy of the polymorphs. This study looks at carbamazepine, a drug molecule used to treat various mental and neurological conditions that exists in several polymorphs at ambient conditions [128, 129, 130, 131, 132]. Carbamazepine was chosen as a test case for these well characterised polymorphs and relative rigidity of the molecule which will help reduce noise in free energy sampling. The unit cell parameters for the four well-characterised polymorphs are given in Table 5.1.
These polymorphs have different stabilities (III>I>IV>II at standard conditions) with a corresponding inverse ranking in solubility. The aim of this study was to recreate this stability hierarchy and estimate how different the solubilities are with the aim of establishing a robust protocol that can be used for novel compounds.

5.2 Theory

The chemical potential of an extra molecule in a system of N solute molecules in a fluid of volume V_{IG} using the ideal gas reference is:

$$\mu_{sol} = -RT \ln \left(\frac{V_{IG}}{N+1} \right) + \mu_{sol}^{ex}. \tag{5.1}$$

For solutes with low solubility, the route to estimating solubility is simplified as the chemical potential is dominated by the ideal contribution and μ_{sol}^{ex} can be considered constant in the narrow concentration window of interest. With this assumption, only a single chemical potential determination at infinite dilution (single molecule in pure solvent) is required to compare with the estimated crystal chemical potential. The chemical potential for a solute molecule in a pure solvent simplifies to:

$$\lim_{N \to 0} \mu_{sol} = -RT \ln(V_{IG}) + \mu_{sol}^{ex}. \tag{5.2}$$

The chemical potential of a crystal using the Einstein crystal reference is:

Table 5.1: Experimental unit cell parameters for the four polymorphs of carbamazepine.

Polymorph	Space Group	Z	a / Å	b / Å	c / Å	α / deg	β / deg	γ / deg
Form I/triclinic	2	8	5.1705(6)	20.574(2)	22.245(2)	84.124(4)	88.008(4)	85.187(4)
Form II/trigonal	148	18	35.454(3)	35.454(3)	5.253(1)	90	90	120
Form III/P-monoclinic	14	4	7.534(1)	11.150(2)	13.917(3)	90	92.94(4)	90
Form IV/C-monoclinic	15	8	26.609(4)	6.9269(10)	13.957(2)	90	109.702(2)	90

$$\mu_{crys} = -RT \ln(V_{EC}) + \mu_{crys}^{ex} \tag{5.3}$$

The Einstein crystal reference is where the molecules are held in their average positions by harmonic restraints on a central atom. The available volume V_{EC} of an Einstein crystal with restraint strength K_E is given by:

$$V_{EC} = \left(\frac{2\pi kT}{K_E} \right)^{\frac{3}{2}}. \tag{5.4}$$

The contribution μ_{crys}^{ex} is from the transformation of the reference state to the normal crystal, proceeding through the application of two extra restraints to restrain rotation, turning on the intermolecular interactions and then removing the harmonic restraints. At solubility, $\mu_{sol}^{inf} = \mu_{crys}$ and therefore:

$$-RT \ln \left(\frac{V_{IG}}{V_{EC}} \right) + \Delta G_{sol}^{ex} - \Delta G_{crys}^{ex} = 0 \tag{5.5}$$

which can be rearranged to give an estimated volume per molecule:

$$V_{IG} = V_{EC}\, e^{\frac{\Delta G_{sol}^{ex} - \Delta G_{crys}^{ex}}{RT}} \tag{5.6}$$

Excess chemical potentials are calculated using the Bennett Acceptance Ratio method [106].

5.3 Methods

The thermodynamic conditions were 300 K and 100 kPa. The force field used for carbamazepine is GAFF [133]. There are no novel moieties in the molecule so the van der Waals parameters were taken from the standard GAFF set. Partial charges were generated using RESP with the HF/6-31G* basis set — the recommended process for GAFF [133]. The partial charges on the azepine and terminal amine moieties were adjusted by averaging the partial charges on opposite atoms to account for their symmetry. The differences between the partial charges on opposite sides were small enough that no significant effects on phase equilibria were anticipated. Partial charges are shown in Figure 5.1.

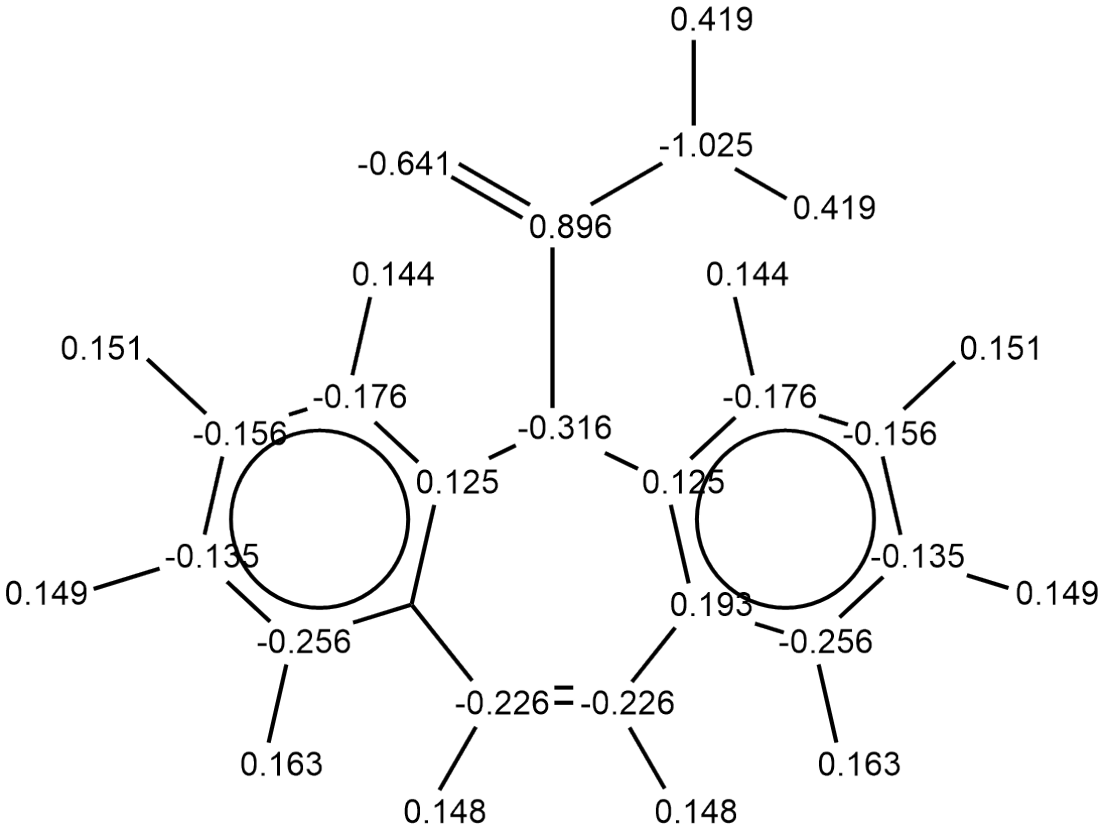


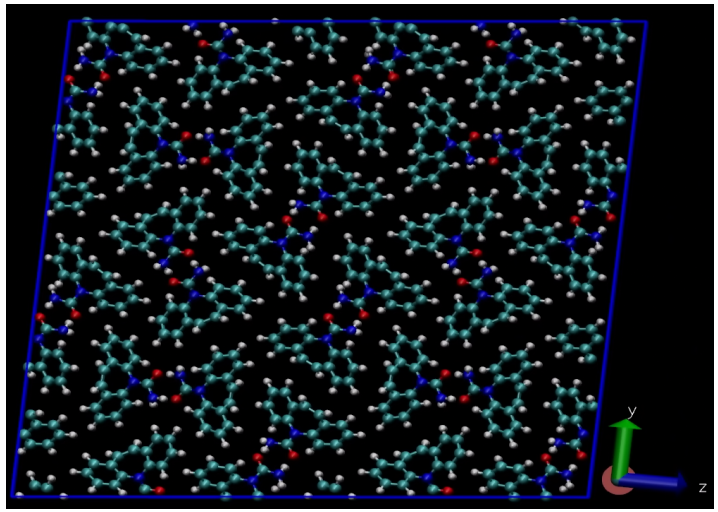
Figure 5.1: Partial charges generated for carbamazepine using RESP with the HF/6-31G* basis set. The charges on the amine hydrogens and opposite azepine atoms have been averaged to suit the symmetry of these moieties.

5.3.1 Simulation Details

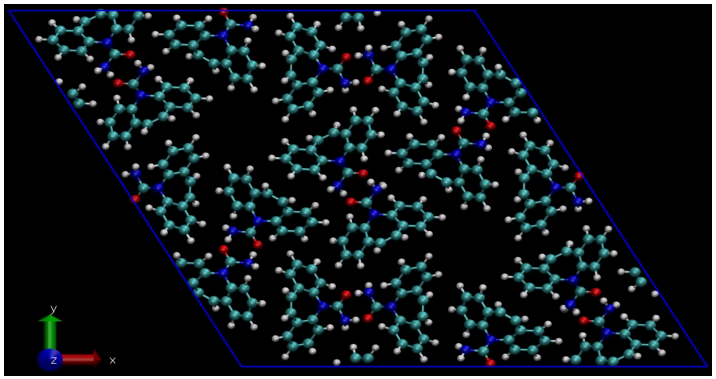
Simulations were carried out using the stochastic dynamics integrator for improved phase space sampling. To avoid statistical issues, the centre-of-mass was constrained at every step in crystal calculations. With bond constraints, a timestep of 2 fs was permitted. Data for free energy calculations were collected every 5000 steps. The energy potentials used were a potential-switch function for van der Waals interactions and Particle-Mesh Ewald for electrostatics with long range dispersion corrections for energy and pressure. The Berendsen barostat was used for equilibration simulations and the Parrinello-Rahman barostat for free energy simulations with a compressibility of 4.5×10^{-5} bar (matching that of water at standard conditions). Free energy calculations for van der Waals and electrostatics use 26 and 11 equidistant lambda states respectively. Calculations for bonds and position restraints unfortunately have to be tuned on an ad hoc basis with test runs. Soft-core parameters are left as Gromacs defaults — sc-alpha = 0.5, sc-r-power = 6, sc-power = 1 and sc-sigma = 0.3.
Each polymorph unit cell was extended to a system large enough to allow a 1.2 nm cut-off for non-bonded forces:

- Form I — $5a \times 2b \times 2c$ — 2.58525 nm $\times$ 4.1148 nm $\times$ 4.449 nm — 160 molecules
- Form II — $1a \times 1b \times 5c$ — 3.5454 nm $\times$ 3.5454 nm $\times$ 2.6265 nm — 90 molecules
- Form III — $4a \times 3b \times 2c$ — 3.036 nm $\times$ 3.345 nm $\times$ 2.7834 nm — 96 molecules
- Form IV — $1a \times 4b \times 2c$ — 2.6609 nm $\times$ 2.77076 nm $\times$ 2.7954 nm — 64 molecules

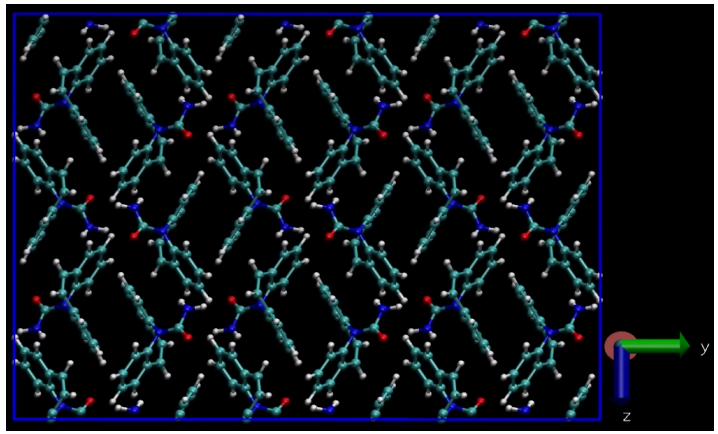
The force field was tested for Forms I, II and IV through NST simulations of 2 ns duration and testing whether the unit box dimensions were maintained within 5% of experimental values. Gromacs had stability issues with simulating Form II under NST conditions due to the extreme box angles so an NPT simulation was used instead. After equilibration, the average box parameters were taken forward and the crystals were then simulated in the NVT regime for 2 ns. The reference atom positions for the Einstein crystals were taken from the average atom positions in the NVT simulations. The reference crystals are shown in Figure 5.2.
For this study the Einstein molecule approach to free energy calculations was employed with a restraint strength of 500000 kJ/mol nm^2. As the de Broglie wavelength does not influence the chemical potential difference between crystal and solution at constant



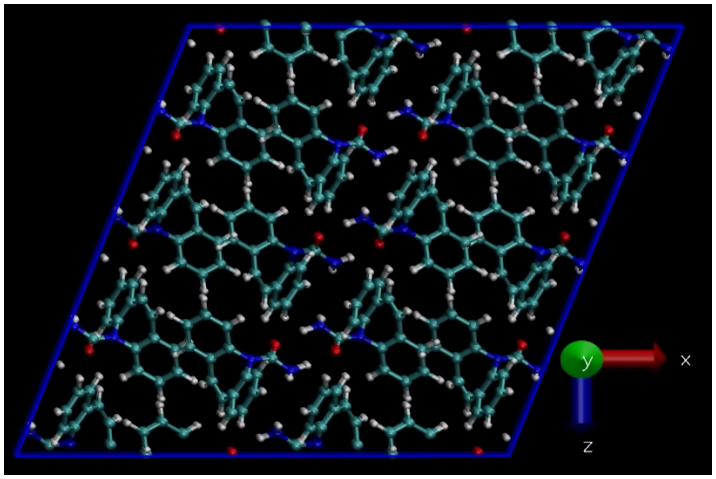
(a) GAFF reference crystal of triclinic Form I of carbamazepine viewed along the a-axis.



(b) GAFF reference crystal of trigonal Form II of carbamazepine viewed along the b-axis. The large voids that cause the instability of this polymorph can be easily seen.



(c) GAFF reference crystal of p-monoclinic Form III of carbamazepine viewed along the a-axis. The c-axis is not consistent with the real crystal



(d) GAFF reference crystal of c-monoclinic Form IV of carbamazepine viewed along the b-axis.

Figure 5.2: GAFF reference crystals for carbamazepine polymorphs.

Table 5.2: GAFF unit cell parameters for the four polymorphs of carbamazepine. The angles in Form II were not tested due to software issues.

Polymorph	a / Å	b / Å	c / Å	α / deg	β / deg	γ / deg
Form I/triclinic	5.203	21.229	22.474	83.81	88.81	84.47
Form II/trigonal	35.324	35.322	5.175	*90*	*90*	*120*
Form III/P-monoclinic	7.660	11.389	14.717	90	114.28	90
Form IV/C-monoclinic	26.926	7.010	14.123	90	109.7	90

Table 5.3: Percentage errors of GAFF unit cell parameters for the four polymorphs of carbamazepine in relation to the experimental values in Table 5.1

Polymorph	a	b	c	α	β	γ
Form I/triclinic	0.6	3.2	1.0	-0.4	0.9	-0.8
Form II/trigonal	-0.4	-0.4	-1.5	N/A	N/A	N/A
Form III/P-monoclinic	1.7	2.1	5.7	0.0	23.0	0.0
Form IV/C-monoclinic	1.2	1.2	1.2	0.0	0.0	0.0

temperature, the wavelength of the molecule was set at 1 nm for convenience. The central restraint was applied to the azepine nitrogen and orientation restraints were applied in the x and y dimensions to the outermost carbons on the azepine. The excess free energy of carbamazepine was determined for a single molecule in 1000 TIP3P water molecules. The volume given by an Einstein restraint of 500000 kJ/mol nm^2 is 1.75489×10^{-7} nm^3.

5.4 Results

5.4.1 Polymorph Tests

The GAFF force field successfully reproduced most of the unit cell parameters within 5% of the experimental values for all the polymorphs at standard conditions — their structures are shown in Figure 5.2. The notable exception is Form III which suffered significant distortion in the c-axis but was fine in the other axes. The parameters are given in Table 5.2 and errors in Table 5.3. Despite the distortion of Form III, free energy calculations were still carried out for the sake of completeness.

5.4.2 Chemical Potentials

The Einstein crystal free energy is 38.801 kJ/mol and the free energy change of restricting the orientation is 30.624 kJ/mol. The rest of the free energy contributions and total chemical potentials for each polymorph are given in Table 5.4. The free energy calculations for the four polymorphs have successfully reproduced the stability hierarchy of III>I>IV>II, though the chemical potential of Form III is suspect with its distortion. The dH/dλ curves for the transformations are given in Figure 5.3. The statistical errors are very small compared to the chemical potential differences — the largest error is 0.071 kJ/mol compared to the smallest chemical potential difference of about 1.3 kJ/mol between forms II and IV.

For the solvation free energy, the contributions of the van der Waals and electrostatic components respectively are 2.925$\pm$0.186 kJ/mol and $-$56.031$\pm$0.084 kJ/mol — a total of $-$53.106$\pm$0.204 kJ/mol. The dH/dλ curves are shown in Figure 5.4
Using Equation 5.6, the estimated solubilities of each polymorph are as follows:

- Form III/P-monoclinic — 1.5x10^{-5} M
- Form I/triclinic — 2.8x10^{-5} M
- Form IV/C-monoclinic — 5.7x10^{-4} M
- Form II/trigonal — 9.7x10^{-4} M

There is a simple exponential relationship between crystal chemical potential and solubility as shown in figure 5.6. The real solubility is 4.91x10^{-4} M [134]. The estimated solubility of the most stable polymorph is therefore more than an order of magnitude out.

5.5 Discussion

Not only did GAFF, a general purpose force field, faithfully reproduce three different polymorph crystal structures for carbamazepine, it also successfully produced the correct stability hierarchy. This is a great accomplishment for relatively little effort. The cause of the distortion in Form III is yet unknown but it may be based on interactions

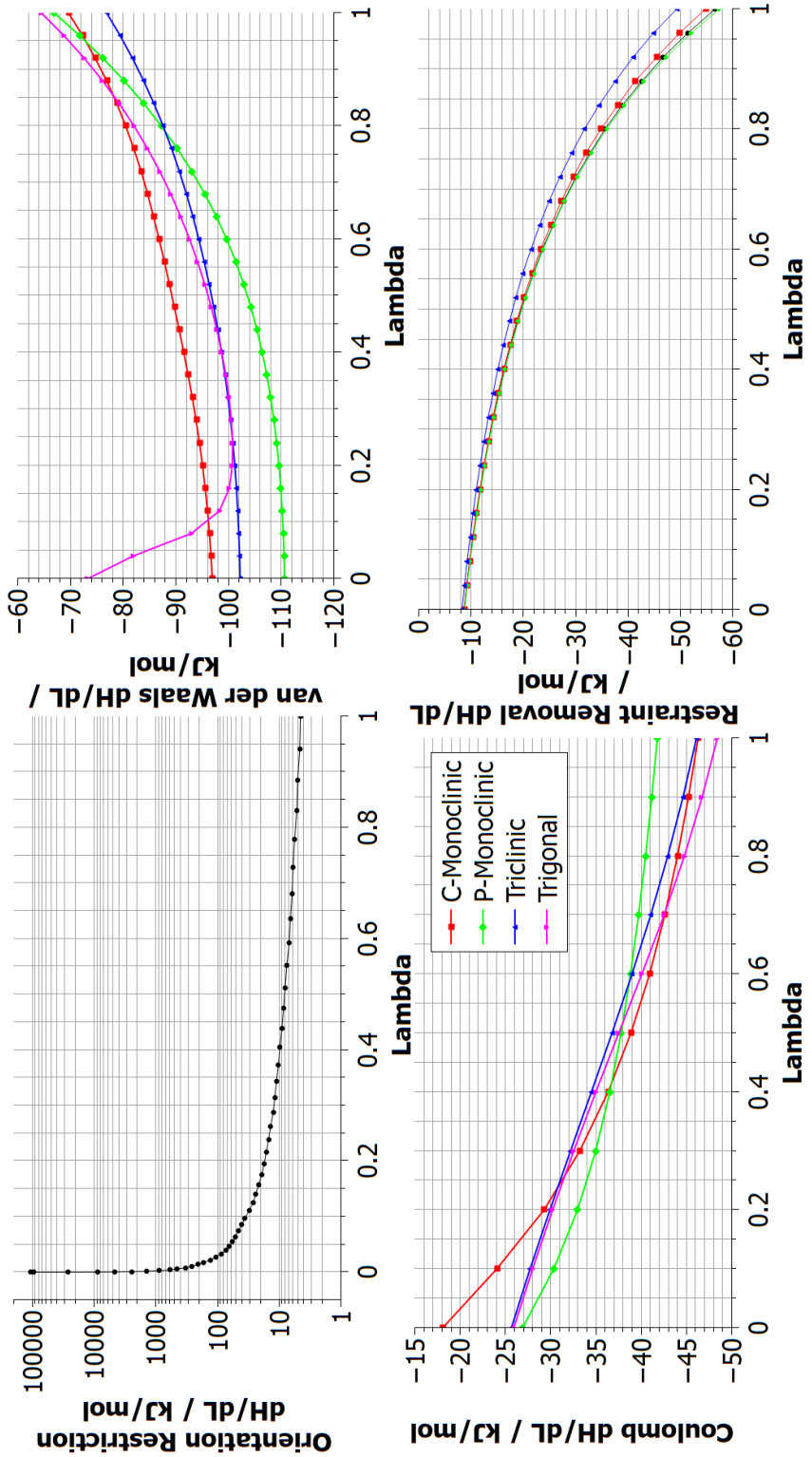


Figure 5.3: The dH/dλ curves for the crystal free energy transformations in the four polymorphs of carbamazepine. One can see the largest difference is made in the van der Waals forces, particularly for the trigonal form. Electrostatic forces have some differences too while the restraint removal only shows differences in magnitude.

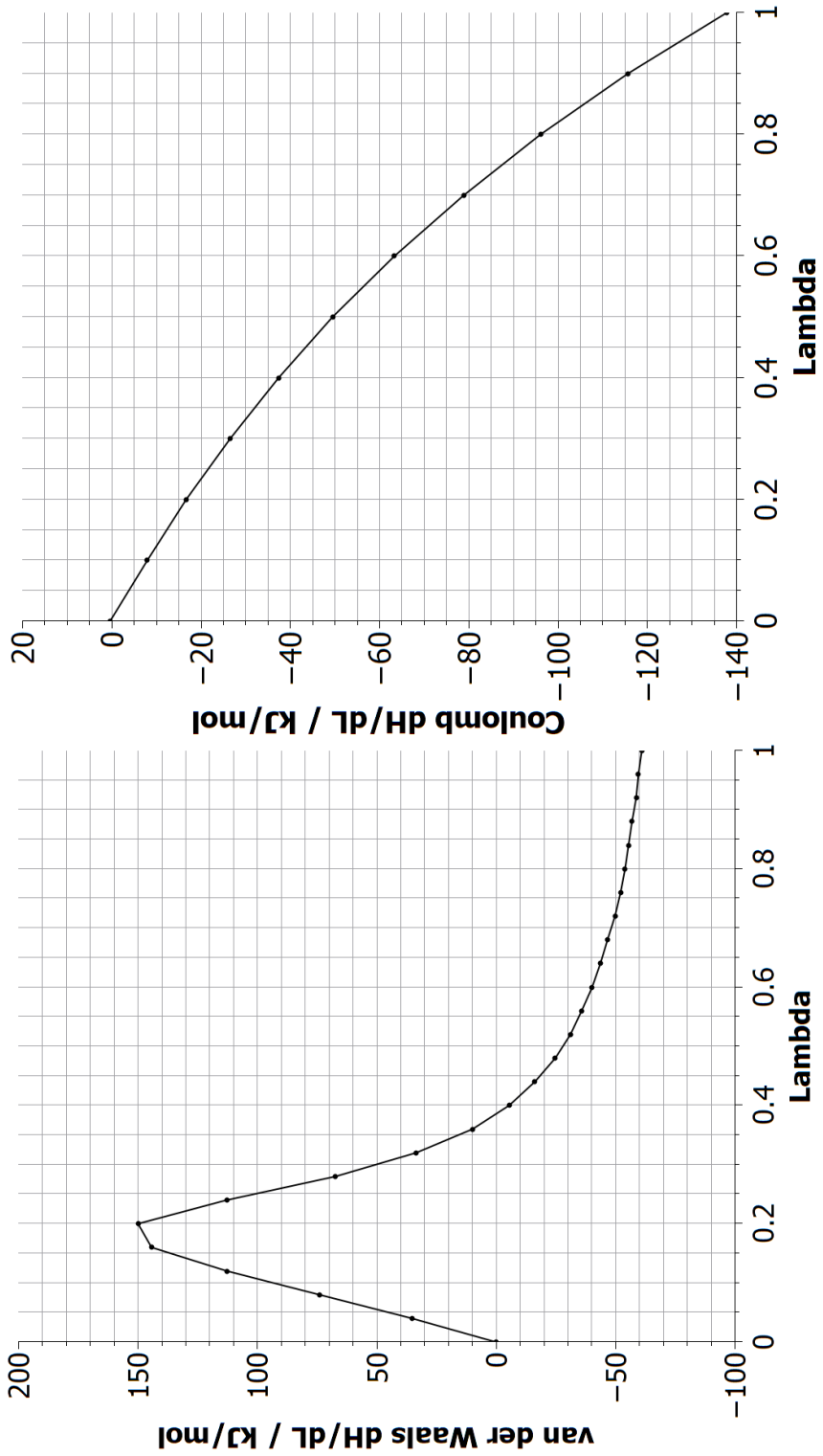


Figure 5.4: The dH/dλ curves for the crystal free energy transformations in the four polymorphs of carbamazepine. One can see the largest difference is made in the van der Waals forces, particularly for the trigonal form. Electrostatic forces have some differences too while the restraint removal only shows differences in magnitude.

Table 5.4: Excess free energy contributions per molecule for CBZ polymorphs. Energies in kJ/mol.

Polymorph	#mol	Orientation FE	O error	vdW	V error	Electrostatics	E error	Restraint Removal	RR error	Total Excess	X error
Form I/Triclinic	160	47.23	0.068	−94.562	0.003	−36.486	0.014	−35.462	0.008	−119.28	0.07
Form II/Trigonal	90	47.23	0.068	−90.114	0.014	−37.514	0.009	−30.083	0.009	−110.481	0.071
Form III/P-monoclinic	96	47.23	0.068	−98.823	0.004	−36.794	0.007	−32.479	0.012	−120.866	0.07
Form IV/C-monoclinic	64	47.23	0.068	−87.728	0.002	−36.748	0.005	−34.552	0.012	−111.798	0.069

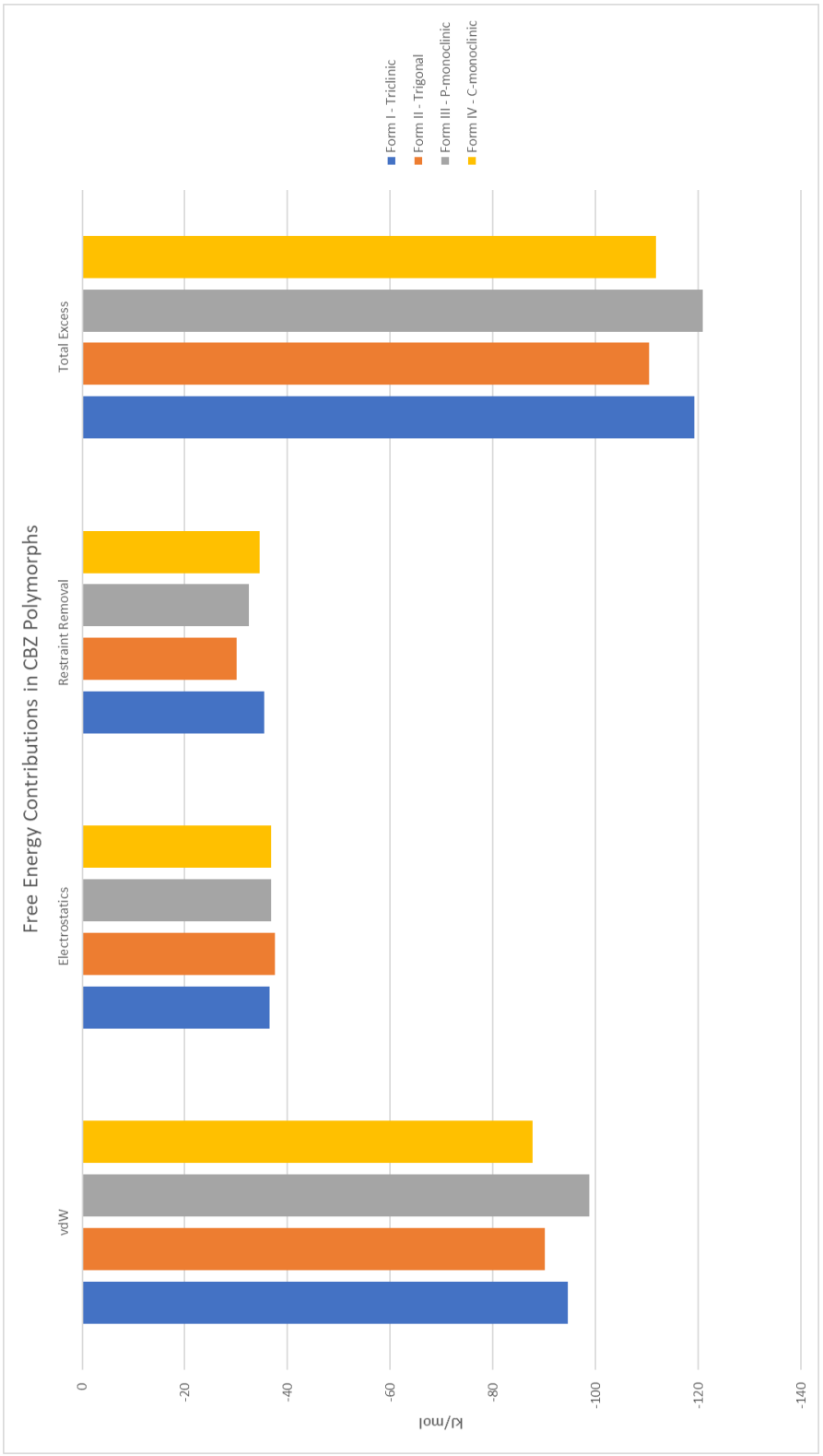


Figure 5.5: Free energy contributions and total excess free energy per molecule for each CBZ polymorph. It can be easily seen that the largest differences come from the van der Waals contribution, followed by the restraint release and then minor differences exist for electrostatics.

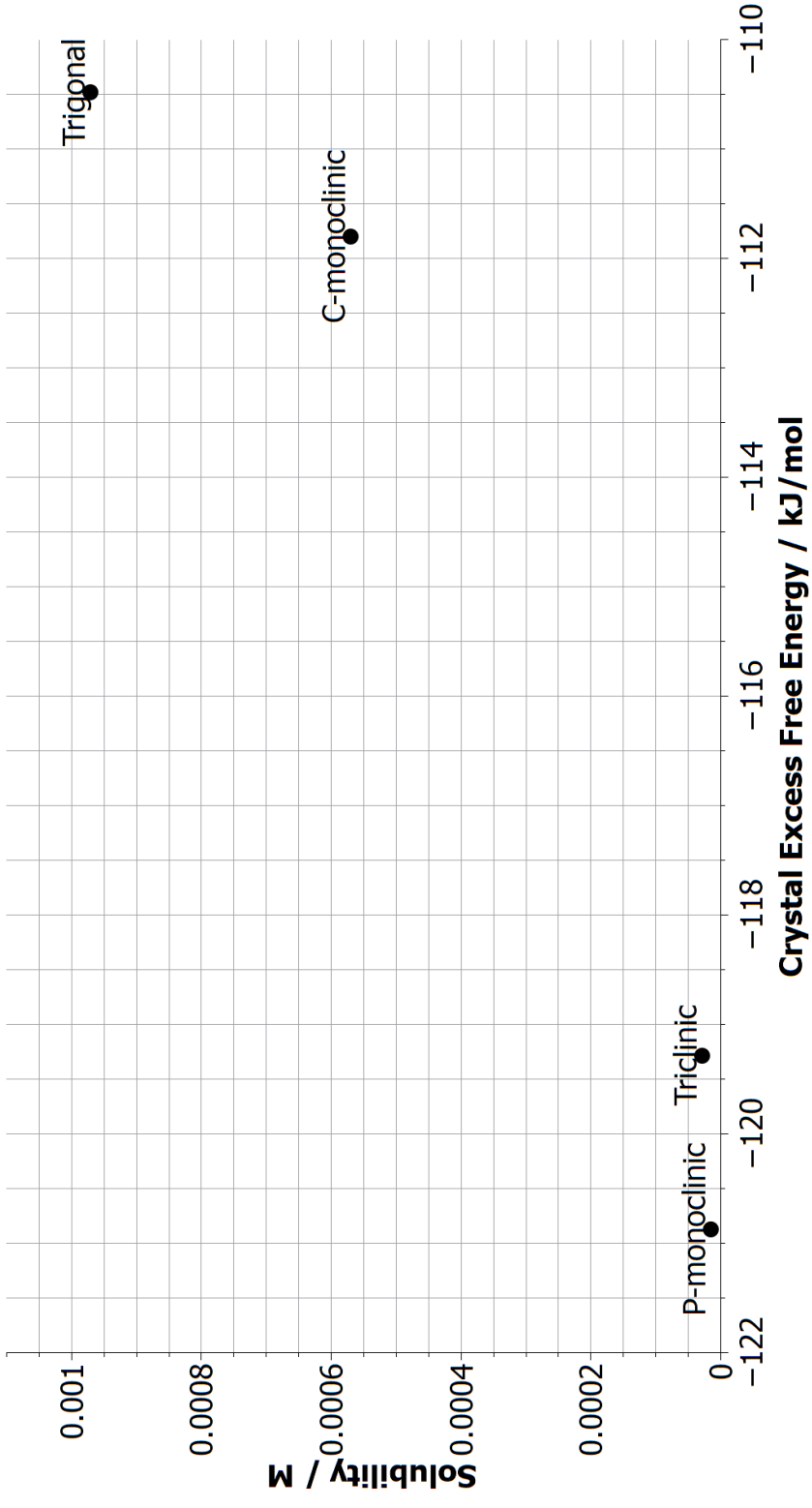


Figure 5.6: Estimated solubility of carbamazepine polymorphs compared with their excess chemical potentials (sum of simulation-derived free energy changes).

neglected in the force field — the structure of carbamazepine is controlled by both hydrogen bonding between the amide groups and π-stacking of the azepine rings which are not explicitly handled. It could also be an issue in the barostat parameters in the simulation settings. The statistical errors in the crystal free energy calculations are very small with most of the error coming from the imposition of orientation restraints. An investigation into whether weaker restraints could work to give an even lower error would be prudent.

The polymorphs have a range of solubilities spanning almost two orders of magnitude. While they have theoretically much higher solubility, the much lower stability for the C-monoclinic and trigonal polymorphs combined with solvation would probably lead to their transformation to the more stable P-monoclinic or triclinic polymorphs with a corresponding drop in solubility. A method to prevent formation of the more stable polymorphs would greatly increase the bioavailability of carbamazepine. The easy creation of the force field, only requiring quantum chemistry software to generate partial charges with everything else already taken from standardised parameters, means that it would not be difficult to extend this testing to other drug molecules with well characterised polymorphs and then possibly on to predictive studies.

Even with the successful thermodynamic hierarchy, there is a lot that isn't predicted by free energy calculations, particularly the kinetics of phase transitions. Transition from the trigonal and C-monoclinic forms to the P-monoclinic form are probably quicker than from the triclinic form but would require a dedicated study to verify this.

The greatest difference to the crystal polymorph free energy comes from the van der Waals contribution while there is very little difference due to partial charges. It is a surprise that there is little difference in electrostatic contributions to the free energy considering the role of hydrogen bonding in the crystal structures. This implies that the van der Waals parameters (dispersion forces in reality) have the largest influence on polymorph stability. While decades of optimisation have gone into these parameters, there are still moieties that are not covered. When addressing a novel moiety, the greatest care should therefore be taken in choosing the right van der Waals parameters.

The estimated solubility, being more than an order of magnitude out, is less promis-

ing than the precision of the crystal free energy calculations but there are clear ways to seek an improvement. This study used the very old TIP3P water model as it is the recommended model for GAFF. Better results may come from using more modern models such as TIP4P/2005, then investigating alternative force fields for carbamazepine if necessary to improve accuracy further. In addition, the partial charges are calculated without taking dielectric constants into account — in reality, the polarity of a molecule changes according to the permittivity of the environment. There are also many atomic interactions that are neglected by common force fields, such as polarisability and explicit handling of hydrogen bonding with its partial covalent character, that may be significant contributions to the free energy.

Corroborating the free energy derived estimates with a direct coexistence simulation would be useful but, despite being technically possible, it would likely be a major challenge in itself considering a very large crystal would be required to avoid finite size effects for a large molecule.

Chapter 6

Conclusion

This research was an exploration of various methods used to predict solubility, identifying limitations and ways to improve methodology for future studies. The research was split into three studies — estimating the solubility of urea in water using different force fields and thermodynamic pathways, estimating the mutual solubility of butanol and water using direct coexistence and free energy of mixing calculations, and finally predicting the stability heirarchy of the four carbamazepine polymorphs and estimating their solubilities.

The first study sought to recreate the solubility of urea in water using four different combinations of two theoretically valid thermodynamic pathways, one preserving molecular structure and the other breaking the perturbed molecules down to atoms, and two sets of force fields of different ages — Özpınar's urea force field from 2010 with TIP3P and Hölzl's urea force field from 2019 with TIP4P/2005. The free energy calculations had excellent precision but only the Hölzl set with the molecular pathway produced a sensible chemical potential difference between the crystal and solution, giving a solubility of urea in water between 0.05 and 0.1 molar fraction. This agreed with direct coexistence simulations giving a range of solubility for the Hölzl urea model of 0.03-0.11 molar fraction. The direct coexistence simulations for the Özpınar model gave a range of 0.035-0.06 molar ratio but the free energy calculations failed to agree with this. Unfortunately, this discrepancy could not be successfully troubleshooted.

The atomic thermodynamic pathway gave chemical potential differences several kJ/mol lower than the molecular route for both force fields — this indicates a possible issue

with how Gromacs calculates the data for free energy analysis and therefore the atomic pathway should not be used for research until it can be troubleshooted and validated. Other studies in the literature have encountered persistent issues with estimating solubility for a variety of molecules which indicates severe issues with force fields and free energy calculations. A separate study by Boothroyd on the solubility of urea in water gave an estimate of 0.46-50.0 mol/kg, a margin of error that is practically useless for further analysis [73]. Matos and Mobley could not get a reasonable estimate of the solubility of aspirin in water as the chemical potential difference was much too high [135]. These are recent studies with modern software and technology yet reconciliation of estimates and experimental data was not possible. These issues underpin the importance of direct coexistence simulations as they do not rely on complex free energy calculations and solubility is given explicitly — only finite-size effects have to be accounted for.

The second study sought to recreate the interesting solubility/temperature phase diagram of butanol and water using the GAFF and TIP3P force fields. Direct coexistence simulations and free energy calculations to construct curves of free energy of mixing were performed at a wide range of temperatures. The direct coexistence simulations had excellent performance due to both phases being liquid state and produced statistically robust estimates of the mutual solubilities of butanol and water, though the variance of the solubility became very large near the critical temperature of miscibility. The critical temperature was correctly estimated at just above 100°C. The estimated solubilities were far off experimental values but their qualitative behaviour was similar, showing a local minimum in the solubility of butanol in water. The free energy calculations failed to produce chemical potentials with the required precision needed to construct the curves of free energy of mixing. Qualitatively, the chemical potential calculations showed the correct behaviour as a function of composition but they were hampered by finite-size effects not being sufficiently suppressed, especially in butanol-rich systems. In future, significantly larger system sizes are needed.

The third study was a test to see if a general-use force field (GAFF in this study) could correctly represent polymorphism and to see how strongly it affects solubility. The test molecule was carbamazepine with four characterised polymorphs. GAFF successfully reproduced the lattice parameters at standard conditions for three polymorphs in NST

but had issues with one of the axes for the form III. Chemical potential calculations successfully reproduced the stability hierarchy of the polymorphs (III>I>IV>II) but the chemical potential of form III is suspect due to the lattice parameter issues. As in the urea study, the precision of the free energy calculations were excellent. The calculated solubilities span almost two orders of magnitude ($9.7x10^{-4}$ — $1.5x10^{-5}$ M) but the estimated solubility of Form III is an order of magnitude lower than the experimental solubility of $4.91x10^{-4}$ M.

It has been shown that direct coexistence simulations are now an effective method of determining solubilities. They are particularly effective for mutual solubilities of fluids where it can be quicker than the total time for the large number of simulations required for a comprehensive coverage of mixing free energy with respect to composition. Direct coexistence is less effective for solid state solvation as the kinetics are much slower and outcomes of dissolution and precipitation simulations are different but is still a useful diagnostic tool if the accuracy of free energy calculations are suspect.

Free energy calculations are very effective for crystals as applying the alchemical transformation to every molecule in the system helps the collected data quickly converge to a robust average due to the Central Limit Theorem, as shown with the small standard errors in the study data. Conversely, solution free energy calculations are less precise as the alchemical transformation is limited to one molecular unit to minimise perturbation of the rest of the system, leading to the free energy data having much greater variance. This is particularly problematic for mutual solubility of fluids as very high precision is needed to accurately represent the mixing free energy curve as a function of composition, particularly when the local minimum has small magnitude or high curvature. Another way to ameliorate statistical errors is to perform multiple simulations starting from different configurations on each lambda point to gain a set of statistically independent samples to average over. However, this may be onerous on time for a study on a subject like mutual solubility which already requires thousands of simulations at a minimum.

This research project demonstrated how difficult it is to represent many properties of molecules with one force field. Being able to reproduce the correct densities and crystal structures is important but is not enough for the purpose of solubility prediction.

Future work needs to be dedicated to refining force fields to reproduce chemical potentials. However, there may be only so much that can be achieved with the traditional combination of the Lennard-Jones potential and point charges.

It has been shown that large fluid molecules require larger systems to reduce the perturbation from the probe molecule. A study into the relationship between the size of a molecule and the amount of solvent required to minimise finite size effects to a consistent threshold would be welcome. We are not aware of any existing literature dedicated to this issue.

The time required for individual simulations for a free energy calculation are very short now. The free energy simulations for turning on charges in the urea crystal had a simulation rate of 1.2 hr/ns on a 16-core 2.6 GHz Intel Xeon CPU and the direct coexistence simulation for urea and water had a simulation rate of around 1.4 hr/ns on a 40-core 2.4 GHz Intel Xeon CPU. In the modern computing age, the concern is no longer how fast CPUs are but how many you can access and the cloud computing industry could be a huge boon for physical chemists. Molecular dynamics simulations are very easy to parallelise these days and they should be significantly faster with the newest versions of molecular dynamics software embracing the dedicated vector-computing power of GPUs.

Another major issue in this project was the lack of dedicated software for creating free energy calculations. All the simulation setting files were made by hand or expedited by small programs from old tutorials or written *ad hoc*. When a study requires thousands of simulations there are many risks for human error. If free energy calculations are to truly go mainstream in industry, there need to be software that automates and expedites the process.

In essence, there is now a robust theoretical and technological foundation for the use of free energy calculations and direct coexistence simulations to determine solubilities. Force fields need to be further developed to reliably recreate experimental solubility data and software need to be developed to make free energy calculations easier to set up and analyse. This is of particular importance for industrial applications such as pharmaceuticals and agrochemicals, which require seamless processes to analyse the

plethora of candidate molecules coming through the pipeline.

Appendices

A.1 System Topology Files

A.1.1 Özpınar Urea

```
[ defaults ]
; nbfunc comb-rule gen-pairs fudgeLJ fudgeQQ
1 2 yes 0.5 0.83333

[ atomtypes ]
; name mass charge ptype sigma eps
C 12.010 0.0 A 3.3997e-1 3.59800e-1
N 14.010 0.0 A 3.7500e-1 7.11300e-1
H 1.008 0.0 A 1.0691e-1 6.57000e-2
O 16.000 0.0 A 2.9599e-1 8.78600e-1

[ bondtypes ]
; i j func r k
C N 1 0.1383 354803.2
C O 1 0.1250 548940.8
H N 1 0.1010 363171.2

[ angletypes ]
; i j k func th k
N C O 1 120.9 669.4
N C N 1 118.6 585.8
C N H 1 120.0 251.0
H N H 1 120.0 292.9

[ dihedraltypes ]
; i j k l func th k n
H N C O 9 0.0000 8.368 1.0000
H N C O 9 180.0000 10.46 2.0000
X C N X 1 180.0000 10.46 2.0000
```

```
[ moleculetype ]
; Name nrexcl
URE 3

[ atoms ] ; nr type r# res atom cgnr charge mass
1 C 0 RES C 1 0.884 12.010
2 N 0 RES N 2 -0.888 14.010
3 H 0 RES H 3 0.388 1.008
4 H 0 RES H 4 0.388 1.008
5 O 0 RES O 5 -0.660 16.000
6 N 0 RES N 6 -0.888 14.010
7 H 0 RES H 7 0.388 1.008
8 H 0 RES H 8 0.388 1.008

[ bonds ]
; i j funct
1 2 1
1 5 1
1 6 1
2 3 1
2 4 1
6 7 1
6 8 1

[ pairs ]
; i j funct
2 7 1
2 8 1
3 5 1
3 6 1
4 5 1
4 6 1
5 7 1
5 8 1
```

```
[ angles ]
; i j k funct
2 1 5 1
2 1 6 1
5 1 6 1
1 2 3 1
1 2 4 1
3 2 4 1
1 6 7 1
1 6 8 1
7 6 8 1


[ dihedrals ]
; ai aj ak al funct
5 1 2 3 9
5 1 2 4 9
6 1 2 3 9
6 1 2 4 9
5 1 6 7 9
5 1 6 8 9
2 1 6 7 9
2 1 6 8 9
3 1 4 5 9 180.0 4.6024 2
6 1 7 8 9 180.0 4.6024 2
1 2 3 6 9 180.0 43.932 2


[ position_restraints ]
; ai f Akx Aky Akz
1 1 500000.0 500000.0 500000.0
2 1 500000.0 500000.0 0.0
6 1 500000.0 500000.0 0.0
```

A.1.2 Hölzl Urea

```
[ defaults ]
; nbfunc comb-rule gen-pairs fudgeLJ fudgeQQ
1 2 yes 0.5 0.83333

[ atomtypes ]
; name mass charge ptype sigma eps
C 12.010 0.0 A 0.36039 0.35982
N 14.010 0.0 A 0.34452 0.51114
H 1.008 0.0 A 0.11333 0.06569
O 16.000 0.0 A 0.31377 0.59432

[ bondtypes ]
; i j func r k
C N 1 0.13350 410032.0
C O 1 0.12290 476976.0
H N 1 0.10100 363171.2

[ angletypes ]
; i j k func th k
N C O 1 121.4 670
N C N 1 117.2 670
C N H 1 120.0 390
H N H 1 120.0 445

[ dihedraltypes ]
; i j k l func th k n
H N C O 9 0.0000 8.368 1.0000
H N C O 9 180.0000 10.46 2.0000
X C N X 1 180.0000 10.46 2.0000

[ moleculetype ]
; Name nrexcl
```

```
URE 3

[ atoms ] ; nr type r# res atom cgnr charge mass
1 C 0 RES C 1 0.6068 12.010
2 N 0 RES N 2 -0.8400 14.010
3 H 0 RES H 3 0.4026 1.008
4 H 0 RES H 4 0.4421 1.008
5 O 0 RES O 5 -0.6162 16.000
6 N 0 RES N 6 -0.8400 14.010
7 H 0 RES H 7 0.4026 1.008
8 H 0 RES H 8 0.4421 1.008

[ bonds ]
; i j funct
1 2 1
1 5 1
1 6 1
2 3 1
2 4 1
6 7 1
6 8 1

[ pairs ]
; i j funct
2 7 1
2 8 1
3 5 1
3 6 1
4 5 1
4 6 1
5 7 1
5 8 1

[ angles ]
```

```
; i  j  k  funct
2  1  5  1
2  1  6  1
5  1  6  1
1  2  3  1
1  2  4  1
3  2  4  1
1  6  7  1
1  6  8  1
7  6  8  1


[ dihedrals ]
; ai  aj  ak  al  funct
5  1  2  3  9
5  1  2  4  9
6  1  2  3  9
6  1  2  4  9
5  1  6  7  9
5  1  6  8  9
2  1  6  7  9
2  1  6  8  9
5  1  2  6  4  180.00  43.93200  2
2  1  3  4  4  180.00  4.18400  2
6  1  7  8  4  180.00  4.18400  2


[ position_restraints ]
; ai  f  Akx  Aky  Akz
1  1  500000.0  500000.0  500000.0
2  1  500000.0  500000.0  0.0
6  1  500000.0  500000.0  0.0
```

A.1.3 Butanol

```
; 1-Butanol

[ defaults ]
; nbfunc comb-rule gen-pairs fudgeLJ fudgeQQ
1 2 yes 0.5 0.8333

[ atomtypes ]
;name bond_type mass charge ptype sigma epsilon
c3 c3 0.0000 0.0000 A 3.39967e-01 4.57730e-01
hc hc 0.0000 0.0000 A 2.64953e-01 6.56888e-02
h1 h1 0.0000 0.0000 A 2.47135e-01 6.56888e-02
oh oh 0.0000 0.0000 A 3.06647e-01 8.80314e-01
ho ho 0.0000 0.0000 A 0.00000e+00 0.00000e+00

[ bondtypes ]
; a b f r k
c3 hc 1 1.0920e-01 2.8225e+05
c3 c3 1 1.5350e-01 2.5363e+05
c3 h1 1 1.0930e-01 2.8108e+05
c3 oh 1 1.4260e-01 2.6284e+05
oh ho 1 9.7400e-02 3.0928e+05

[ angletypes ]
; a b c f theta k
hc c3 hc 1 1.0835e+02 3.2970e+02
hc c3 c3 1 1.1005e+02 3.8828e+02
c3 c3 c3 1 1.1063e+02 5.2886e+02
c3 c3 h1 1 1.1007e+02 3.8828e+02
c3 c3 oh 1 1.0943e+02 5.6651e+02
c3 oh ho 1 1.0816e+02 3.9413e+02
h1 c3 h1 1 1.0955e+02 3.2803e+02
h1 c3 oh 1 1.0988e+02 4.2677e+02
```

```
[ dihedraltypes ]
; a b c d f
c3 c3 c3 c3 3 3.68192 3.09616 -2.09200 -3.01248 0.00000 0.00000
c3 c3 c3 hc 3 0.66944 2.00832 0.00000 -2.67776 0.00000 0.00000
c3 c3 c3 h1 3 0.65270 1.95811 0.00000 -2.61082 0.00000 0.00000
c3 c3 c3 oh 3 0.65270 1.95811 0.00000 -2.61082 0.00000 0.00000
c3 c3 oh ho 3 1.71544 0.96232 0.00000 -2.67776 0.00000 0.00000
hc c3 c3 hc 3 0.62760 1.88280 0.00000 -2.51040 0.00000 0.00000
hc c3 c3 h1 3 0.65270 1.95811 0.00000 -2.61082 0.00000 0.00000
hc c3 c3 oh 3 1.04600 -1.04600 0.00000 0.00000 0.00000 0.00000
h1 c3 oh ho 3 0.69873 2.09618 0.00000 -2.79491 0.00000 0.00000

[ moleculetype ]
; Name nrexcl
1-butanol 3

[ atoms ]
; nr type resnr residue atom cgnr charge mass
1 c3 1 MOL C1 1 -0.16680 12.000000
2 hc 1 MOL H1 2 0.04020 1.000000
3 hc 1 MOL H2 3 0.04020 1.000000
4 hc 1 MOL H3 4 0.04020 1.000000
5 c3 1 MOL C2 5 0.03140 12.000000
6 hc 1 MOL H4 6 -0.00080 1.000000
7 hc 1 MOL H5 7 -0.00080 1.000000
8 c3 1 MOL C3 8 -0.00650 12.000000
9 hc 1 MOL H6 9 0.03340 1.000000
10 hc 1 MOL H7 10 0.03340 1.000000
11 c3 1 MOL C4 11 0.29770 12.000000
12 h1 1 MOL H8 12 -0.03040 1.000000
13 h1 1 MOL H9 13 -0.03040 1.000000
14 oh 1 MOL O1 14 -0.71430 16.000000
15 ho 1 MOL H10 15 0.43350 1.000000
```

```
[ bonds ]
; ai aj funct r k
1 2 1
1 3 1
1 4 1
5 6 1
5 7 1
8 9 1
8 10 1
11 12 1
11 13 1
14 15 1
1 5 1
5 8 1
8 11 1
11 14 1

[ pairs ]
; ai aj funct
1 9 1
1 10 1
2 6 1
2 7 1
2 8 1
3 6 1
3 7 1
3 8 1
4 6 1
4 7 1
4 8 1
5 12 1
5 13 1
6 9 1
```

```
6 10 1
6 11 1
7 9 1
7 10 1
7 11 1
8 15 1
9 12 1
9 13 1
9 14 1
10 12 1
10 13 1
10 14 1
12 15 1
13 15 1
1 11 1
5 14 1

[ angles ]
; ai aj ak funct theta cth
1 5 6 1
1 5 7 1
2 1 3 1
2 1 4 1
2 1 5 1
3 1 4 1
3 1 5 1
4 1 5 1
5 8 9 1
5 8 10 1
6 5 7 1
6 5 8 1
7 5 8 1
8 11 12 1
8 11 13 1
```

```
9 8 10 1
9 8 11 1
10 8 11 1
11 14 15 1
12 11 13 1
12 11 14 1
13 11 14 1
1 5 8 1
5 8 11 1
8 11 14 1

[ dihedrals ]
;i j k l func C0 ...  C5
1 5 8 9 3
1 5 8 10 3
2 1 5 6 3
2 1 5 7 3
2 1 5 8 3
3 1 5 6 3
3 1 5 7 3
3 1 5 8 3
4 1 5 6 3
4 1 5 7 3
4 1 5 8 3
5 8 11 12 3
5 8 11 13 3
6 5 8 9 3
6 5 8 10 3
6 5 8 11 3
7 5 8 9 3
7 5 8 10 3
7 5 8 11 3
8 11 14 15 3
9 8 11 12 3
```

```
9 8 11 13 3
9 8 11 14 3
10 8 11 12 3
10 8 11 13 3
10 8 11 14 3
12 11 14 15 3
13 11 14 15 3
1 5 8 11 3
5 8 11 14 3
```

A.1.4 Carbamazepine

```
[ defaults ]
; nbfunc comb-rule gen-pairs fudgeLJ fudgeQQ
1 2 yes 0.5 0.833

[ atomtypes ]
; name bond_type mass charge ptype sigma eps
c 12.01000 0.000 A 3.39967e-1 3.59824e-1
c2 12.01000 0.000 A 3.39967e-1 3.59824e-1
ca 12.01000 0.000 A 3.39967e-1 3.59824e-1
ha 1.00800 0.000 A 2.59964e-1 6.27600e-2
hc 1.00800 0.000 A 2.64953e-1 6.56888e-2
hn 1.00800 0.000 A 1.06908e-1 6.56888e-2
n 14.01000 0.000 A 3.25000e-1 7.11280e-1
o 16.00000 0.000 A 2.95992e-1 8.78640e-1

[ bondtypes ]
; i j func
c n 1 0.1379 357815.68
c o 1 0.1218 533627.36
c2 c2 1 0.1373 419153.12
c2 ca 1 0.1456 322251.68
```

```
c2 hc 1 0.1084 292126.88
ca ca 1 0.1398 385848.48
ca n 1 0.1412 321498.56
ca ha 1 0.1086 289365.44
hn n 1 0.1013 337397.76

[ angletypes ]
; i j k func
n c n 1 113.56 610.3188
n c o 1 123.05 621.36984
c n ca 1 123.71 534.30104
c n hn 1 117.55 404.61876
c2 c2 ca 1 113.51 566.11464
c2 c2 hc 1 121.76 406.37688
ca c2 hc 1 124.04 383.68876
c2 ca ca 1 120.79 544.34744
ca ca ca 1 120.02 557.74264
ca ca n 1 120.19 568.20764
ca ca ha 1 119.88 403.36296
ca n ca 1 117.37 542.25444
hn n hn 1 117.95 331.19632

[ dihedraltypes ]
; i j k l func
X ca ca X 1 180.0000 15.1670 2.0000
X c n X 1 180.0000 10.4600 2.0000
X ca n X 1 180.0000 1.8828 2.0000
hn n c o 9 0.0000 8.3680 1.0000
hn n c o 9 180.0000 10.4600 2.0000
X c2 c2 X 1 180.0000 16.736 2.0000
X c2 ca X 1 180.0000 2.9288 2.0000

[ moleculetype ]
; Name nrexcl
```

```
CBZ 3

[ atoms ]
; nr type resnr res atom cgnr charge mass
1 o 0 RES o 1 -0.641 16
2 n 0 RES n 2 -1.025 14.01
3 hn 0 RES hn 3 0.419 1.008
4 hn 0 RES hn 4 0.419 1.008
5 n 0 RES n 5 -0.316 14.01
6 c 0 RES c 6 0.896 12.01
7 ca 0 RES ca 7 0.125 12.01
8 ca 0 RES ca 8 -0.176 12.01
9 ha 0 RES ha 9 0.144 1.008
10 ca 0 RES ca 10 -0.156 12.01
11 ha 0 RES ha 11 0.151 1.008
12 ca 0 RES ca 12 -0.135 12.01
13 ha 0 RES ha 13 0.149 1.008
14 ca 0 RES ca 14 -0.256 12.01
15 ha 0 RES ha 15 0.163 1.008
16 ca 0 RES ca 16 0.193 12.01
17 c2 0 RES c2 17 -0.226 12.01
18 hc 0 RES hc 18 0.148 1.008
19 c2 0 RES c2 19 -0.226 12.01
20 hc 0 RES hc 20 0.148 1.008
21 ca 0 RES ca 21 0.193 12.01
22 ca 0 RES ca 22 -0.256 12.01
23 ha 0 RES ha 23 0.163 1.008
24 ca 0 RES ca 24 -0.135 12.01
25 ha 0 RES ha 25 0.149 1.008
26 ca 0 RES ca 26 -0.156 12.01
27 ha 0 RES ha 27 0.151 1.008
28 ca 0 RES ca 28 -0.176 12.01
29 ha 0 RES ha 29 0.144 1.008
30 ca 0 RES ca 30 0.125 12.01
```

```
[ bonds ]
; ai aj f
1 6 1
2 3 1
2 4 1
2 6 1
5 6 1
5 7 1
5 30 1
7 8 1
7 16 1
8 9 1
8 10 1
10 11 1
10 12 1
12 13 1
12 14 1
14 15 1
14 16 1
16 17 1
17 18 1
17 19 1
19 20 1
19 21 1
21 22 1
21 30 1
22 23 1
22 24 1
24 25 1
24 26 1
26 27 1
26 28 1
28 29 1
```

```
28 30 1

[ pairs ]
; ai aj f
1 3 1
1 4 1
1 7 1
1 30 1
2 7 1
2 30 1
3 5 1
4 5 1
5 9 1
5 10 1
5 14 1
5 17 1
5 19 1
5 22 1
5 26 1
5 29 1
6 8 1
6 16 1
6 21 1
6 28 1
7 11 1
7 12 1
7 15 1
7 18 1
7 19 1
7 21 1
7 28 1
8 13 1
8 14 1
8 17 1
```

```
8 30 1
9 11 1
9 12 1
9 16 1
10 15 1
10 16 1
11 13 1
11 14 1
12 17 1
13 15 1
13 16 1
14 18 1
14 19 1
15 17 1
16 30 1
16 20 1
16 21 1
17 22 1
17 30 1
18 20 1
18 21 1
19 23 1
19 24 1
19 28 1
20 22 1
20 30 1
21 25 1
21 26 1
21 29 1
22 27 1
22 28 1
23 25 1
23 26 1
23 30 1
```

```
24 29 1
24 30 1
25 27 1
25 28 1
27 29 1
27 30 1

[ angles ]
; ai aj ak funct
1 6 2 1
1 6 5 1
3 2 4 1
3 2 6 1
4 2 6 1
2 6 5 1
6 5 7 1
6 5 30 1
7 5 30 1
5 7 8 1
5 7 16 1
5 30 21 1
5 30 28 1
8 7 16 1
7 8 9 1
7 8 10 1
7 16 14 1
7 16 17 1
9 8 10 1
8 10 11 1
8 10 12 1
11 10 12 1
10 12 13 1
10 12 14 1
13 12 14 1
```

```
12 14 15 1
12 14 16 1
15 14 16 1
14 16 17 1
16 17 18 1
16 17 19 1
18 17 19 1
17 19 20 1
17 19 21 1
20 19 21 1
19 21 22 1
19 21 30 1
22 21 30 1
21 22 23 1
21 22 24 1
21 30 28 1
23 22 24 1
22 24 25 1
22 24 26 1
25 24 26 1
24 26 27 1
24 26 28 1
27 26 28 1
26 28 29 1
26 28 30 1
29 28 30 1

[ dihedrals ]
; ai aj ak al f phi k n
1 6 2 3 9 0.0 0.0 1
1 6 2 3 9 180.0 0.0 2
1 6 2 4 9 0.0 0.0 1
1 6 2 4 9 180.0 0.0 2
1 6 5 7 1
```

1 6 5 30 1
3 2 6 5 1
4 2 6 5 1
2 6 5 7 1
2 6 5 30 1
6 5 7 8 1
6 5 7 16 1
6 5 30 21 1
6 5 30 28 1
8 7 5 30 1
16 7 5 30 1
7 5 30 21 1
7 5 30 28 1
5 7 8 9 1
5 7 8 10 1
5 7 16 14 1
5 7 16 17 1
5 30 21 19 1
5 30 21 22 1
5 30 28 26 1
5 30 28 29 1
9 8 7 16 1
10 8 7 16 1
8 7 16 14 1
8 7 16 17 1
7 8 10 11 1
7 8 10 12 1
7 16 14 12 1
7 16 14 15 1
7 16 17 18 1
7 16 17 19 1
9 8 10 11 1
9 8 10 12 1
8 10 12 13 1

```
8 10 12 14 1
11 10 12 13 1
11 10 12 14 1
10 12 14 15 1
10 12 14 16 1
13 12 14 15 1
13 12 14 16 1
12 14 16 17 1
15 14 16 17 1
14 16 17 18 1
14 16 17 19 1
16 17 19 20 1
16 17 19 21 1
18 17 19 20 1
18 17 19 21 1
17 19 21 22 1
17 19 21 30 1
20 19 21 22 1
20 19 21 30 1
19 21 22 23 1
19 21 22 24 1
19 21 30 28 1
23 22 21 30 1
24 22 21 30 1
22 21 30 28 1
21 22 24 25 1
21 22 24 26 1
21 30 28 26 1
21 30 28 29 1
23 22 24 25 1
23 22 24 26 1
22 24 26 27 1
22 24 26 28 1
25 24 26 27 1
```

```
25 24 26 28 1
24 26 28 29 1
24 26 28 30 1
27 26 28 29 1
27 26 28 30 1
6 2 3 4 9 180.0 0.0 2
6 5 7 30 9 180.0 0.0 2
1 6 2 5 9 180.0 0.0 2
```

Bibliography

[1] S. H. Yalkowsky, Y. He, and P. Jain. *Handbook Of Aqueous Solubility Data*. 2nd edition, 2010.

[2] H. H. Wasserman, R. W. DeSimone, K. R. X. Chia, and M. G. Banwell. *Singlet Oxygen*. American Cancer Society, 2013.

[3] A.I. Vogel and G. Svehla. *Vogel's Qualitative Inorganic Analysis*. Longman, 1996.

[4] I. F. McConvey and P. Nancarrow. Chapter 10 liquid–liquid extraction for process development in the pharmaceutical industry. In *Pharmaceutical Process Development: Current Chemical and Engineering Challenges*, pages 209–237. The Royal Society of Chemistry, 2011.

[5] B. Swain, C. Mishra, H. S. Hong, and S.-S. Cho. Treatment of indium-tin-oxide etching wastewater and recovery of In, Mo, Sn and Cu by liquid–liquid extraction and wet chemical reduction: a laboratory scale sustainable commercial green process. *Green Chem.*, 17:4418–4431, 2015.

[6] M. Filiz, N. A. Sayar, and A. A. Sayar. Extraction of cobalt(II) from aqueous hydrochloric acid solutions into alamine–m-xylene mixtures. *Hydrometallurgy*, 81(3):167–173, 2006.

[7] K. M. Elkins. Chapter 4 – DNA extraction. In K. M. Elkins, editor, *Forensic DNA Biology*, pages 39–52. Academic Press, San Diego, 2013.

[8] P. Burger, H. Plainfossé, X. Brochet, F. Chemat, and X. Fernandez. Extraction of natural fragrance ingredients: History overview and future trends. *Chemistry & Biodiversity*, 16(10):e1900424, 2019.

[9] D. Musha. Studies on Body Water in Man. *The Tohoku Journal of Experimental Medicine*, 63(4):309–317, 1956.

[10] A. F. A. Cros. *Action de l'alcool amylique sur l'organisme*. Ph.d., University of Strasbourg, 1863.

[11] B. Alberts, A. Johnson, J. Lewis, D. Morgan, M. Raff, K. Roberts, and P. Walter. *The Shape and Structure of Proteins*, chapter 3, pages 109–134. W. W. Norton & Co., 2014.

[12] G. P. Michel and A. A. Bühlmann. *Decompression — Decompression Sickness*. Springer Berlin Heidelberg, 2013.

[13] J. M. Long and D. M. Holtzman. Alzheimer disease: An update on pathobiology and treatment strategies. *Cell*, 179(2):312–339, 2019.

[14] A. Fick. Ueber diffusion. *Annalen der Physik*, 170(1):59–86, 1855.

[15] L. Di and E. H. Kerns. Biological assay challenges from compound solubility: strategies for bioassay optimization. *Drug Discovery Today*, 11(9-10):446–451, 2006.

[16] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 64(SUPPL.):4–17, 2012.

[17] D. Schuster, C. Laggner, and T. Langer. Why Drugs Fail — A Study on Side Effects in New Chemical Entities. *Current Pharmaceutical Design*, 11(27):3545–3559, 2005.

[18] A. M. Palmer. New horizons in drug metabolism, pharmacokinetics and drug discovery. *Drug News Perspect*, 16(1):57–62, 2003.

[19] E. H. Kerns and L. Di. *Drug-like Properties: Concepts, Structure Design and Methods from ADME to Toxicity Optimization*. Academic Press, 2nd edition, 2016.

[20] S. N. Bhattachar, L. A. Deschenes, and J. A. Wesley. Solubility: it's not just for physical chemists. *Drug Discovery Today*, 11(21-22):1012–1018, 2006.

[21] W. Dai, C. Pollock-Dove, L. C. Dong, and S. Li. Advanced screening assays to rapidly identify solubility-enhancing formulations: High-throughput, miniaturization and automation. *Advanced Drug Delivery Reviews*, 60(6):657–672, 2008.

[22] S. L. McGovern, E. Caselli, N. Grigorieff, and B. K. Shoichet. A Common Mechanism Underlying Promiscuous Inhibitors from Virtual and High-Throughput Screening. *Journal of Medicinal Chemistry*, 45(8):1712–1722, 2002.

[23] H. van de Waterbeemd, D. A. Smith, K. Beaumont, and D. K. Walker. Property-Based Design: Optimization of Drug Absorption and Pharmacokinetics. *Journal of Medicinal Chemistry*, 44(9):1313–1333, 2001.

[24] H. Keppler and N. Bolfan-Casanova. *Thermodynamics of water solubility and partitioning*. 2018.

[25] B. Robinson, N. Bolan, S. Mahimairaja, and B. Clothier. *Solubility, mobility, and bioaccumulation of trace elements: Abiotic processes in the rhizosphere*. 2005.

[26] E. D. Clarke and J. S. Delaney. Physical and Molecular Properties of Agrochemicals: An Analysis of Screen Inputs, Hits, Leads, and Products. *CHIMIA International Journal for Chemistry*, 57(11):731–734, 2003.

[27] Y. Lu, S. Song, R. Wang, Z. Liu, J. Meng, A. J. Sweetman, A. Jenkins, R. C. Ferrier, H. Li, W. Luo, and T. Wang. Impacts of soil and water pollution on food safety and health risks in China. *Environment International*, 77:5–15, 2015.

[28] L. Jiao. Water Shortages Loom as Northern China's Aquifers Are Sucked Dry. *Science*, 328(5985):1462–1463, 2010.

[29] World Health Organization and United Nations Environment Programme. World Health Organization (1990) – public health impact of pesticides used in agriculture, 1990.

[30] R. Kühne, R.-U. Ebert, F. Kleint, G. Schmidt, and G. Schüürmann. Group contribution methods to estimate water solubility of organic chemicals. *Chemosphere*, 30(11):2061–2077, 1995.

[31] J. C. Dearden and N. M. Shinnawei. Improved prediction of fish bioconcentration factor of Hydrophobic Chemicals. *SAR and QSAR in Environmental Research*, 15(5-6):449–455, 2004.

[32] R. R. dos Reis, S. C. Sampaio, and E. B. de Melo. An alternative approach for the use of water solubility of nonionic pesticides in the modeling of the soil sorption coefficients. *Water Research*, 53:191–199, 2014.

[33] R.F. Tylecote and Institute of Materials (Great Britain). *A History of Metallurgy*. Book (Institute of Materials (Great Britain))). Institute of Materials, 1992.

[34] J. D. Verhoeven. *Steel Metallurgy for the Non-Metallurgist*. EngineeringPro collection. ASM International, 2007.

[35] W. E. Bachmann. The mechanism of reduction by sodium amalgam and alcohol. i. the reduction of aromatic ketones to hydrols. *Journal of the American Chemical Society*, 55(2):770–774, 1933.

[36] P. Ham. *Zinc Amalgam*. American Cancer Society, 2001.

[37] R. Bharti, K. Wadhwani, A. Tikku, and A. Chandra. Dental amalgam: An update. *Journal of Conservative Dentistry*, 13(4):204–208, 2010.

[38] C. Vargel. *Corrosion of Aluminium*, page 158. Elsevier Science, 2004.

[39] Á. Könczöl and G. Dargó. Brief overview of solubility methods: Recent trends in equilibrium solubility measurement and predictive models. *Drug Discovery Today: Technologies*, 27:3 – 10, 2018. Physicochemical characterisation in drug discovery.

[40] H. Birch, A. D. Redman, D. J. Letinski, D. Y. Lyon, and P. Mayer. Determining the water solubility of difficult-to-test substances: A tutorial review. *Analytica Chimica Acta*, 1086:16–28, 2019.

[41] A. Veseli, S. Žakelj, and A. Kristl. A review of methods for solubility determination in biopharmaceutical drug characterization. *Drug Development and Industrial Pharmacy*, 45(11):1717–1724, 2019. PMID: 31512934.

[42] C. A. Lipinski. Drug-like properties and the causes of poor solubility and poor permeability. *Journal of Pharmacological and Toxicological Methods*, 44(1):235 – 249, 2000. Current Directions in Drug Discovery:A Review of Modern Techniques.

[43] C. Nantasenamat, C. Isarankura–na–ayudhya, and T. Naenna. A Practical Overview Of Quantitative Structure-Activity Relationship. *Excli J*, 8(7):74–88, 2009.

[44] A. Tropsha. Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics*, 29(6-7):476–488, 2010.

[45] A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. I. Baskin, M. Cronin, J. C. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, V. Consonni, V. E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard, and A. Tropsha. QSAR Modeling: Where Have You Been? Where Are You Going To? *Journal of Medicinal Chemistry*, 57(12):4977–5010, 2014.

[46] J. C. Dearden. In silico prediction of aqueous solubility. *Expert Opinion on Drug Discovery*, 1(1):31–52, 2006.

[47] E. J. Mills. XXIII. On melting-point and boiling-point as related to chemical composition. *Philosophical Magazine Series 5*, 17(105):173–187, 1884.

[48] H. Fühner. The water solubility in homologous series. *Reports of the German Chemical Society (A and B Series)*, 57(3):510–515, 1924.

[49] V. A. Filov, A. A. Golubev, E. I. Liublina, and N. A. Tolokontsev. *Quantitative Toxicology*. John Wiley & Sons, Ltd, New York, 1979.

[50] D. Mackay, R. Mascarenhas, W. Y. Shiu, S. C. Valvani, and S. H. Yalkowsky. Aqueous solubility of polychlorinated biphenyls. *Chemosphere*, 9(5–6):257–264, 1980.

[51] C. Hansch, P. P. Maloney, T. Fujita, and R. M. Muir. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature*, 194(4824):178–180, 1962.

[52] N. Bodor, A. Harget, and M. J. Huang. Neural network studies. 1. Estimation of the aqueous solubility of organic compounds. *Journal of the American Chemical Society*, 113(25):9480–9483, 1991.

[53] M. H. Abraham and J. Le. The correlation and prediction of the solubility of compounds in water using an amended solvation energy relationship. *Journal of Pharmaceutical Sciences*, 88(9):868–880, 1999.

[54] D. Cao, Q. Xu, Y. Liang, X. Chen, and H. Li. Prediction of aqueous solubility of druglike organic compounds using partial least squares, back-propagation network and support vector machine. *Journal of Chemometrics*, 24(9):584–595, 2010.

[55] J. G. Topliss. Utilization of operational schemes for analog synthesis in drug design. *Journal of Medicinal Chemistry*, 15(10):1006–1011, 1972.

[56] J. M. Sutter and P. C. Jurs. Prediction of Aqueous Solubility for a Diverse Set of Heteroatom-Containing Organic Compounds Using a Quantitative Structure-Property Relationship. *Journal of Chemical Information and Computer Sciences*, 36(1):100–107, 1996.

[57] H. Gao, V. Shanmugasundaram, and P. Lee. Estimation of Aqueous Solubility of Organic Compounds with QSPR Approach. *Pharmaceutical Research*, 19(4):497–503, 2002.

[58] A. Mauri, V. Consonni, M. Pavan, and R. Todeschini. DRAGON software: An easy approach to molecular descriptor calculations. *MATCH Communications in Mathematical and in Computer Chemistry*, 56(2):237–248, 2006.

[59] W. L. Jorgensen. QSAR/QSPR and Proprietary Data. *Journal of Chemical Information and Modeling*, 46(3):937–937, 2006.

[60] J. L. McDonagh, N. Nath, L. De Ferrari, T. van Mourik, and J. B. O. Mitchell. Uniting Cheminformatics and Chemical Theory To Predict the Intrinsic Aqueous Solubility of Crystalline Druglike Molecules. *Journal of Chemical Information and Modeling*, 54(3):844–856, 2014.

[61] J. C. Dearden, M. T. D. Cronin, and K. L. E. Kaiser. How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). *SAR and QSAR in Environmental Research*, 20(3-4):241–266, 2009.

[62] S. B. Bunally, C. N. Luscombe, and R. J. Young. Using physicochemical measurements to influence better compound design. *SLAS DISCOVERY: Advancing the Science of Drug Discovery*, 24(8):791–801, 2019. PMID: 31429385.

[63] R. E. Skyner, J. L. McDonagh, C. R. Groom, T. van Mourik, and J. B. O. Mitchell. A review of methods for the calculation of solution free energies and the modelling of systems in solution. *Phys. Chem. Chem. Phys.*, 17(9):6174–6191, 2015.

[64] J. Alsenz and M. Kuentz. From quantum chemistry to prediction of drug solubility in glycerides. *Molecular Pharmaceutics*, 16(11):4661–4669, 2019. PMID: 31518142.

[65] J. L. Aragones, E. Sanz, and C. Vega. Solubility of NaCl in water by molecular simulation revisited. *The Journal of Chemical Physics*, 136(24):244508, 2012.

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

[67] W. C. Swope and H. C. Andersen. A molecular dynamics method for calculating the solubility of gases in liquids and the hydrophobic hydration of inert–gas atoms in aqueous solution. *The Journal of Physical Chemistry*, 88(26):6548–6556, 1984.

[68] D. Frenkel and A. J. C. Ladd. 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*, 81(7):3188–3193, 1984.

[69] E. J. Meijer, D. Frenkel, R. A. LeSar, and A. J. C. Ladd. Location of melting point at 300 k of nitrogen by monte carlo simulation. *The Journal of Chemical Physics*, 92(12):7570–7575, 1990.

[70] M. Ferrario, G. Ciccotti, E. Spohr, T. Cartailler, and P. Turq. Solubility of KF in water by molecular dynamics using the Kirkwood integration method. *The Journal of Chemical Physics*, 117(10):4947, 2002.

[71] C. Vega, E. Sanz, J. L. F. Abascal, and E. G. Noya. Determination of phase diagrams via computer simulation: methodology and applications to water, electrolytes and proteins. *Journal of Physics: Condensed Matter*, 20(15):153101, 2008.

[72] M. A. Bellucci, G. Gobbo, T. K. Wijethunga, G. Ciccotti, and B. L. Trout. Solubility of paracetamol in ethanol by molecular dynamics using the extended einstein crystal method and experiments. *The Journal of Chemical Physics*, 150(9):094107, 2019.

[73] S. Boothroyd, A. Kerridge, A. Broo, D. Buttar, and J. Anwar. Solubility prediction from first principles: a density of states approach. *Phys. Chem. Chem. Phys.*, 20:20981–20987, 2018.

[74] M. S. Sellers, M. Lísal, and J. K. Brennan. Free-energy calculations using classical molecular simulation: application to the determination of the melting point and chemical potential of a flexible rdx model. *Phys. Chem. Chem. Phys.*, 18:7841–7850, 2016.

[75] G. A. Özpınar, F. R. Beierlein, W. Peukert, D. Zahn, and T. Clark. A test of improved force field parameters for urea: molecular-dynamics simulations of urea crystals. *Journal of Molecular Modeling*, 18(8):3455–3466, 2012.

[76] C. Hölzl, P. Kibies, S. Imoto, J. Noetzel, M. Knierbein, P. Salmen, M. Paulus, J. Nase, C. Held, G. Sadowski, D. Marx, S. M. Kast, and D. Horinek. Structure and thermodynamics of aqueous urea solutions from ambient to kilobar pressures: From thermodynamic modeling, experiments, and first principles simulations to an accurate force field description. *Biophysical Chemistry*, 254:106260, 2019.

[77] D. D. Nolte. The tangled tale of phase space. *Physics Today*, 63(4):33–38, 2010.

[78] R. Becker and G. Leibfried. *Theory of Heat*. Springer Berlin Heidelberg, 2012.

[79] T. A. Kaplan. The chemical potential. *Journal of Statistical Physics*, 122(6):1237–1260, 2006.

[80] T. Cheng, F. Li, J. Dai, and H. Sun. Prediction of the mutual solubility of water and dipropylene glycol dimethyl ether using molecular dynamics simulation. *Fluid Phase Equilibria*, 314:1–6, 2012.

[81] R. Iftimie, P. Minary, and M. E. Tuckerman. Ab initio molecular dynamics: Concepts, recent developments, and future trends. *Proceedings of the National Academy of Sciences*, 102(19):6654–6659, 2005.

[82] D. Paschek. Heat capacity effects associated with the hydrophobic hydration and interaction of simple solutes: A detailed structural and energetical analysis based on molecular dynamics simulations. *The Journal of Chemical Physics*, 120(22):10605–10617, 2004.

[83] J. W. Ponder and D. A. Case. Force fields for protein simulations. In *Protein Simulations*, volume 66 of *Advances in Protein Chemistry*, pages 27 – 85. Academic Press, 2003.

[84] N. J. Boyd and M. R. Wilson. Optimization of the GAFF force field to describe liquid crystal molecules: the path to a dramatic improvement in transition temperature predictions. *Phys. Chem. Chem. Phys.*, 17:24851–24865, 2015.

[85] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. D. Mackerell Jr. Charmm general force field: A force field for drug-like molecules compatible with the charmm all-atom additive biological force fields. *Journal of Computational Chemistry*, 31(4):671–690, 2010.

[86] C. Oostenbrink, A. Villa, A.E. Mark, and W.F. Van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: The gromos force-field parameter sets 53a5 and 53a6. *Journal of Computational Chemistry*, 25(13):1656–1676, 2004.

[87] M. J. Robertson, J. Tirado-Rives, and W. L. Jorgensen. Improved peptide and protein torsional energetics with the opls-aa force field. *Journal of Chemical Theory and Computation*, 11(7):3499–3509, 2015. PMID: 26190950.

[88] M. Chaplin. Water Models. `http://www1.lsbu.ac.uk/water/water_models.html`, 2020.

[89] M. Chaplin. Anomalous Properties of Water. `http://www1.lsbu.ac.uk/water/water_anomalies.html`, 2020.

[90] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[91] C. Störmer. Sur les trajectoires des corpuscules électrisés dans l'espace sous l'action du magnétisme terrestre, avec application aux aurores boréales. *Radium (Paris)*, 9(11):395–399, 1912.

[92] L. Verlet. Computer "experiments" on classical fluids. i. thermodynamical properties of lennard–jones molecules. *Phys. Rev.*, 159:98–103, 1967.

[93] W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *The Journal of Chemical Physics*, 76(1):637–649, 1982.

[94] N. Goga, A. J. Rzepiela, A. H. De Vries, S. J. Marrink, and H. J. C. Berendsen. Efficient algorithms for langevin and dpd dynamics. *Journal of chemical theory and computation*, 8(10):3637–3649, 2012.

[95] H. C. Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *The Journal of chemical physics*, 72(4):2384–2393, 1980.

[96] H. J. C. Berendsen, J. P. M. van Postma, W. F. van Gunsteren, A. R. H. J. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *The Journal of chemical physics*, 81(8):3684–3690, 1984.

[97] W. G. Hoover. Canonical dynamics: Equilibrium phase–space distributions. *Physical review A*, 31(3):1695, 1985.

[98] G. Bussi, D. Donadio, and M. Parrinello. Canonical sampling through velocity rescaling. *The Journal of chemical physics*, 126(1):014101, 2007.

[99] S. Nosé and M. L. Klein. Constant pressure molecular dynamics for molecular systems. *Molecular Physics*, 50(5):1055–1076, 1983.

[100] J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3):327–341, 1977.

[101] S. Miyamoto and P. A. Kollman. Settle: An analytical version of the shake and rattle algorithm for rigid water models. *Journal of computational chemistry*, 13(8):952–962, 1992.

[102] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije. LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, 18(12):1463–1472, 1997.

[103] P. P. Ewald. The calculation of optical and electrostatic grid potentials. *annals of physics*, 369(3):253–287, 1921.

[104] T. Darden, D. York, and L. Pedersen. Particle mesh Ewald: An N log (N) method for ewald sums in large systems. *The Journal of chemical physics*, 98(12):10089–10092, 1993.

[105] R. W. Zwanzig. High–Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *The Journal of Chemical Physics*, 22(8):1420, 1954.

[106] C. H. Bennett. Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics*, 22(2):245–268, 1976.

[107] M. R. Shirts, E. Bair, G. Hooker, and V. S. Pande. Equilibrium free energies from nonequilibrium measurements using maximum–likelihood methods. *Phys. Rev. Lett.*, 91:140601, 2003.

[108] E. Lindahl, M. J. Abraham, B. Hess, and D. van der Spoel. Gromacs 2021.2 source code, 2021.

[109] J. A. Schellman. A simple model for salvation in mixed solvents. *Biophysical Chemistry*, 37(1–3):121–140, 1990.

[110] M. C. Stumpe and H. Grubmüller. Aqueous Urea Solutions: Structure, Energetics, and Urea Aggregation. *The Journal of Physical Chemistry B*, 111(22):6220–6228, 2007.

[111] T. Yamazaki, T. Tanabe, and T. Sugahara. Enclathration of ethane, propane, and propylene into urea clathrates and roles of methanol on urea clathrate formation. *ACS Omega*, 3:13154–13159, 2018.

[112] F. Leng, K. Robeyns, and T. Leyssens. Urea as a cocrystal formerâ€"study of 3 urea based pharmaceutical cocrystals. *Pharmaceutics*, 13(5), 2021.

[113] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1-2:19–25, 2015.

[114] P. V. Klimovich, M. R. Shirts, and D. L. Mobley. Guidelines for the analysis of free energy calculations. *Journal of Computer-Aided Molecular Design*, 29(5):397–411, 2015.

[115] J. E. Worsham, H. A. Levy, and S. W. Peterson. The positions of hydrogen atoms in urea by neutron diffraction. *Acta Crystallographica*, 10(4):319–323, 1957.

[116] S. Piana and J. D. Gale. Understanding the Barriers to Crystal Growth: Dynamical Simulation of the Dissolution and Growth of Urea from Aqueous Solution. *Journal of the American Chemical Society*, 127(6):1975–1982, 2005.

[117] V. Mazan, M. Y. Boltoeva, E. E. Tereshatov, and C. M. Folden III. Mutual solubility of water and hydrophobic ionic liquids in the presence of hydrochloric acid. *RSC Adv.*, 6:56260–56270, 2016.

[118] P. Bertrand and J.-C. C. Mercier. The mutual solubility of coexisting ortho- and clinopyroxene: toward an absolute geothermometer for the natural system? *Earth and Planetary Science Letters*, 76(1):109–122, 1985.

[119] C. Jin, X. Zhang, Han W., Z. Geng, M. T. M. Thomas, A. D. Jeffrey, G. Wang, J. Ji, and H. Liu. Macro and micro solubility between low-carbon alcohols and rapeseed oil using different co-solvents. *Fuel*, 270:117511, 2020.

[120] B. Marongiu, I. Ferino, R. Monaci, V. Solinas, and S. Torrazza. Thermodynamic properties of aqueous non-electrolyte mixtures. alkanols + + water systems. *Journal of Molecular Liquids*, 28(4):229–247, 1984.

[121] R. Stephenson and J. Stuart. Mutual binary solubilities: water–alcohols and water–esters. *Journal of Chemical and Engineering Data*, 31(1):56–70, 1986.

[122] C. M. R. Fowler. *The Solid Earth: An Introduction to Global Geophysics*. Cambridge University Press, 2005.

[123] A. Lévesque, T. Maris, and J. D. Wuest. Roy reclaims its crown: New ways to increase polymorphic diversity. *Journal of the American Chemical Society*, 142(27):11873–11883, 2020. PMID: 32510946.

[124] P. Zhang, X. Zhao, Y. Du, M. Gozin, S. Li, and S. Pang. Polymorphism, phase transformation and energetic properties of 3-nitro-1,2,4-triazole. *RSC Adv.*, 8:24627–24632, 2018.

[125] M. von Raumer and R. Hilfiker. *Polymorphism in the Pharmaceutical Industry*. John Wiley & Sons, Ltd, 2018.

[126] J. Bauer, S. Spanton, R. Henry, J. Quick, W. Dziki, W. Porter, and J. Morris. Ritonavir: An Extraordinary Example of Conformational Polymorphism. *Pharmaceutical Research*, 18(6):859–866, 2001.

[127] J. Nyman and G. M. Day. Static and lattice vibrational energy differences between polymorphs. *CrystEngComm*, 17:5154–5165, 2015.

[128] W. Czernicki and M. Baranska. Carbamazepine polymorphs: Theoretical and experimental vibrational spectroscopy studies. *Vibrational Spectroscopy*, 65:12 – 23, 2013.

[129] A. L. Grzesiak, M. Lang, K. Kim, and A. J. Matzger. Comparison of the four anhydrous polymorphs of carbamazepine and the crystal structure of form i. *Journal of Pharmaceutical Sciences*, 92(11):2260–2271, 2003.

[130] M. M. J. Lowes, M. R. Caira, A. P. Lötter, and J. G. van der Watt. Physicochemical properties and x-ray structural studies of the trigonal polymorph of carbamazepine. *Journal of Pharmaceutical Sciences*, 76(9):744–752, 1987.

[131] J.N. Lisgarten, R.A. Palmer, and J.W. Saldanha. Crystal and molecular structure of 5-carbamyl-5h-dibenzo[b,f] azepine. 19(4):641–649.

[132] M. Lang, J. W. Kampf, and A. J. Matzger. Form IV of carbamazepine. *Journal of Pharmaceutical Sciences*, 91(4):1186–1190, 2002.

[133] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case. Development and testing of a general amber force field. *Journal of Computational Chemistry*, 25(9):1157–1174, 2004.

[134] U.S. Environmental Protection Agency. Carbamazepine - comptox chemicals dashboard. `https://comptox.epa.gov/dashboard/DTXSID4022731`, 2021. [Online; accessed 28-May-2021].

[135] G. Duarte Ramos Matos and D. L. Mobley. Challenges in the use of atomistic simulations to predict solubilities of drug-like molecules. *F1000Research*, 7:686, 2018.