

# Computer simulation studies of the aqueous solvation of ions and peptides



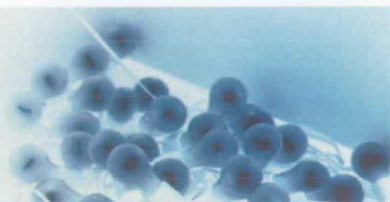
RAPHAEL ZACHARIAS TROITZSCH

A thesis submitted in fulfilment of the requirements  
for the degree of  
Doctor of Philosophy  
to  
The University of Edinburgh  
2007



---

## Abstract



---

Molecular Dynamics simulations have been carried out on systems of increasing complexity and biological relevance, in particular aqueous solutions of ions and peptides, in order to establish the local association patterns of solute and solvent molecules, as well as the structure formed by the solvent around the solvated ions/molecules.

Due to the high irregularity in the energy landscapes of the systems in question, techniques developed to mitigate against the challenges thus posed have been employed, including parallel tempering. To account for quantum effects not readily observed in classical MD, *ab initio* Car-Parrinello simulations have been carried out.

To establish a connection with experimental findings and thus underpin simulation results, techniques were developed to compute and refine structure factors from simulations to compare with experimental scattering data. Based on these findings, water models' qualities as a solvent have been evaluated, resulting in a firm disinclination to the TIP3P model.

Further tools to quantify local structure have been developed and implemented, including the development of an order parameter to gauge the four-fold coordination of water in pure and mixed form as a function of a parameter of interest, here temperature, and spatial density calculations to visualise the local neighbourhood of molecules.

A key objective was to elucidate previously disputed areas in the fundamental findings surrounding the systems in question. This has been achieved in the case of aqueous solution of L-proline with regards to the formation of macromolecular association and the formation of intermediate structure, as well as the question of close ion contact and local ion solvation structure and coordination of aqueous NaCl.

Beyond that, new insight into the mechanism of exchange of members of the first solvation shell of Na, as well as into the molecular mechanism based on the hydrogen bonding pattern of inter-proline and proline-water association as a facilitator to the cryoprotectant nature of proline was gained. Finally, the nature of the helix formed by

---

## Declaration

---



I do hereby declare that this thesis was composed by myself and that the work described within is my own, except where explicitly stated otherwise, and has not been submitted for any other degree or professional qualification.

---

---

## Acknowledgements

---



Clearly, over the course of a PhD, one picks up rather many people, institutions and things one feels good natured towards. Many know this, and need no separate mentioning (but might get it anyway). Others may be unsure or totally unaware, so they will find themselves here or in the small print, if only they care to look.

Among the chief players that do need mentioning are first and foremost the kat who is also a fish and Adelheid und Berta for all the egging on, and motivation-building over the last three years.

Naturally, a massive thanks goes out to Jason and Glenn, who have shown amazing patience with the sometimes hot-headed young upstart of a student that I am, and have given me a tremendous amount of support and splendidly sure-footed guidance. Furthermore, they have provided me with a source of good coffee and uncountable dinners respectively. . . thank you for it all.

Institutionally, both The University of Edinburgh and IBM Research at the T.J. Watson Research Lab in Yorktown, NY deserve mention. The latter also deserves special credit for the extended computer time courtesy of the Physical Sciences division, without which there would be nothing like the results herein. Similarly, all those involved in collaboration, be they at IBM, Rutherford Appleton Lab or NPL: without you I could not have done it, thank you!

Further: — Al and Diana, who really did *the* best job ever of making me feel welcome, at home, safe, and most importantly: fed. Thank you so much! — Troy, for many a PINY-question and the odd happy-hour — Yellmeister Dan — Max, for all that — Face, Lawrence, Caffy, Raj, and all them flatmates over the years, especially latterly those with the attic — Emily and associates for showing me her/their America — Simon, Jules and all them German boys — Cólín for it all — the lovely ladies guarding the tokens — Chris and Iain, and the tentofourometer — EPCC for taking such splendid care of me — La Grande Faffage — The Park in the Highlands, for keeping me in high spirits — Beyond that, in no particular order, I would like to thank: — blueriband, zeiss and zoom — Martin, for German nights in NYC — Carmen, for the welcoming chats at the till — Mags, for letting me get away with it — Patsy for all the coaching, as well as everyone at EFC and the old guard at EUFC — David Brent, for the opportunity and continued support in the work related arena — the Usual Suspect — the Grand Old

---

# Contents

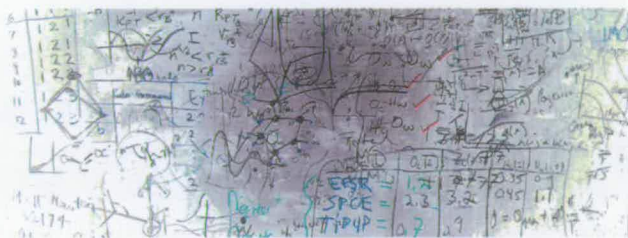


<b>Abstract</b>	i
<b>Declaration</b>	iii
<b>Acknowledgements</b>	v
<b>Contents</b>	vii
<b>1 Introduction</b>	<b>1</b>
1.1 Organisation Structure is the key to success . . . . .	1
1.1.1 Experiment alone does not tackle it – an example . . . . .	2
1.1.2 Simulation to the rescue – dents in shiny armour . . . . .	2
1.1.3 Joint ventures, all the rage . . . . .	3
1.2 Where does my work with Computer Simulation fit in the Grand Scheme?	3
1.3 Let's motivate the problems I've examined . . . . .	4
1.3.1 Water . . . . .	4
1.3.2 Ions . . . . .	5
1.3.3 Amino acids . . . . .	7
1.3.4 Poly-peptides . . . . .	7
1.3.5 Coming full circle? . . . . .	8
<b>2 Methods</b>	<b>9</b>
2.0.0 Very brief, Soft-Core, Rough-Guide to Simulation . . . . .	9
2.1 General Classical Molecular Dynamics . . . . .	12
2.1.1 Newtonian dynamics MD . . . . .	12
2.1.2 Connection to Thermodynamics and Statistical Mechanics . . . . .	13
2.1.3 Hamiltonian dynamics and MD . . . . .	14
2.1.4 Other ensembles . . . . .	21
2.2 Force Field . . . . .	26
2.2.1 Ingredients in the Force – brewing the right punch . . . . .	26
2.2.2 Where the nitty gets gritty: implementing a force field . . . . .	28
2.3 Parallel Tempering . . . . .	31
2.3.1 A look at phase space, or the Alps . . . . .	31
2.3.2 Replica exchange – a clever way . . . . .	33

---

<b>5</b>	<b>Natural antifreeze properties of Proline and Salt</b>	<b>105</b>
5.1	Motivation . . . . .	105
5.2	How this work explores antifreeze . . . . .	106
5.3	Mechanisms of Cryoprotection . . . . .	106
5.4	Conclusion . . . . .	110
5.5	Methods and Ingredients . . . . .	111
5.5.1	Proline solution . . . . .	111
5.5.2	Pure water . . . . .	112
5.5.3	NaCl solution . . . . .	112
5.6	Why the FFCR has value-add . . . . .	113
<b>6</b>	<b>Curbing HIV – using the virus’ structure against itself</b>	<b>115</b>
6.1	HIV infection on the molecular level . . . . .	115
6.2	The knowledge needed to make drugs . . . . .	116
6.2.1	Details of previous experimental findings on gp41 . . . . .	118
6.3	How this work studies gp41 <sub>659–671</sub> . . . . .	119
6.4	Structural revelations from simulations . . . . .	119
6.4.1	Comparison to experiment . . . . .	123
6.5	Conclusions . . . . .	124
6.6	Where to from here? . . . . .	125
6.7	Computational methods . . . . .	125
<b>7</b>	<b>Outroduction</b>	<b>127</b>
7.1	Conclusion . . . . .	127
7.1.1	Insights into local structure . . . . .	127
7.1.2	Structure with function . . . . .	128
7.1.3	Linking to experiment . . . . .	129
7.1.4	The significance of it all . . . . .	129
7.2	Vision . . . . .	129
	<b>Bibliography</b>	<b>133</b>

# Introduction



### 1.1 Organisation Structure is the key to success

As the title suggests, there is a bio-slant to this thesis<sup>1</sup>. Thus, any introduction should really start off by treating just that. At this time, phrases like ‘biological organisms’, ‘the genome project’, ‘DNA’, ‘genetic(-ally modified)’ etc. have very much become everyday conversation terminology. So if these have become common topic and knowledge, what is happening at the cutting edge of the research world? Where are we lacking in knowledge? Well, the obvious answer is that there is still a lot that can be derived from what we know, and similarly from what we know we don’t know. There is also a slightly more subtle point, which is that, despite having found out a lot about the workings of life, some intrinsic and fundamentally simple things remain in the realm of the unknown. The reasons for this are twofold. For one, it is often very hard to know where to look to find an answer to a well known problem, since the level of complexity is extremely high. Second, once we do find the place to look, we often lack the necessary tools to cast our probing scientific eye upon it. As a result, we are often left under-informed about particular structures of our problem-system.

The reason we should be bothered by this is that it is often based on our understanding of a specific structure that we learn to understand mechanisms that arise from, or are aided by, structural features. This becomes particularly important in cases where the systems become very complex, like when we seek to know just exactly how a drug enters a cell; exactly how a virus docks to a cell; exactly how certain molecules keep frogs from freezing. So that’s the sort of detail, and the sort of structure we seek to find.

---

<sup>1</sup>Clearly, the computer-slant is not negligible. It will be dealt with in due course.

### 1.1.3 Joint ventures, all the rage

Since it seems that simulation and experiment are hampered – broadly speaking – by equal and opposite problems (e.g. on timescale and detail reproduction), we would ideally simply join forces between them and live happily ever after. This is impeded by what one might call *language barriers*. The output from experiment is usually based on a bulk property, or an average, or represents a particular measurable variable, whereas simulation is *a priori* more flexible than that, but has a default output of different, small-scale detail based functions and distributions.

Finding an “Esperanto”<sup>2</sup> is not trivial, but the symbiosis of experiment and simulation allows for predictions and assessments to be made by scaling and extrapolating from small-scale simulation results via the reality check and verification by experiment to the bigger picture – and maybe one day even to a level where we can see the result with our bare eyes.

In order to achieve this joint venture, which I view as one of the most powerful scientific ones since the combination of cathode-ray tubes with logic to make the calculator, techniques have been developed, including the well established Empirical Potential Structure Refinement (EPSR)[94] and recent advances to meet with experiment from the simulation side via the static structure factor[104] and light scattering[106] (c.f. section 2.5).

## 1.2 Where does my work with Computer Simulation fit in the Grand Scheme?

Many previous studies involving computer simulations of biological systems have been conducted, so in order to contribute something worthwhile, I am operating at a number of interfaces in that realm. By choice of topic, I am already working in the tri-field area of biology, chemistry and physics. Physics, since it is the fundament of all science, and, in a less pompous manner, the experimental data with which I seek comparison is fundamentally obtained through techniques of physics. Chemistry, because in the systems I worked with, chemical bonds make the molecules of interest. Finally biology, because all of the molecules and molecular constellations I will describe have an impact on larger scale biological problems. I will extend the information from a small test sample to the bigger picture, so to speak, as the ultimate goal (see above).

Further, I am working on an interface, or maybe intermediary phase, between completely ordered things and completely disordered things (e.g. crystals and gases,

---

<sup>2</sup>an artificial language designed to connect people of different background without favouring one language over another[126]



key to this research and most all of my simulations have included water.

Water is still a puzzling substance in many ways[36], as alluded to before, and remains at the forefront of scientific interest. Since, for me, water was more the stage rather than an actor on the stage, I offer no additional insight into aspects of pure water, and instead recycle previous research results[27, 66, 79, 93]. Nonetheless, it is because of its ubiquity that water deserves first mention.

### 1.3.2 Ions

Ions, or more specifically salt ions, in the context of aqueous mixtures have been of particular interest to science, and indeed the general public as a result of that, for many decades, even centuries. This interest can be split in two: the interest in the effects of salt on water *per se*, such as the lowering of the freezing temperature in salty water, and the effect of the presence of salts in water on a tertiary component in a mixture, namely proteins. It is thus not surprising that the influence on aqueous systems, from sea water to boiling pasta to the stability of proteins in solution, has been studied particularly widely. (The term protein is used here, precariously, before a formal introduction of it. This will follow below.)

In terms of biological research, very early work by Hofmeister[45] organised various ions in a series according to the extent to which they influence protein solubility; the so called *Hofmeister series*. Referring to a diverse range of salt-induced phenomena in protein solutions, including denaturation, structure and enzyme kinetics[22, 81], *Hofmeister effects* have been the focus of active research, in both model and experimental techniques.

Setting the stage experimentally, one key hypothesis is that salt induces changes to the structure of water which then determine an effect on biomolecular properties. This is underpinned by neutron diffraction measurements of binary water-salt mixtures[57] and further so by recent work[64], which confirms that the perturbation of the water structure extends firmly beyond the first hydration shell.

However, recent calorimetric measurements[6], which have revealed no correlation between the influence that an ion has on water structure and its effect on protein stability, have challenged this widely held expectation. Additionally, recent dynamical measurements based on femtosecond pump-probe spectroscopies have shown that the rotational dynamics of water molecules outside the first hydration shell remain unperturbed by the presence of hydrated ions[82]. This seems to suggest that the extent of solvent restructuring is weaker than expected on the basis of the diffraction data[57].

Computer simulation is in a good position to arbitrate these disputes, and as a

### 1.3.3 Amino acids

Amino acids take the central rôle as both building blocks of proteins and intermediates in metabolism. The 17 proteinogenic amino acids, those of which proteins are made, display tremendous chemical versatility individually, ranging from acidic to basic, highly hydrophobic to hydrophilic (that is water-fearing and water-loving, respectively) and so forth.

It is the specific sequence of different amino acid types that defines a particular protein, which in turn is determined by the sequence of the bases in the gene that encodes that protein. Via the chemical properties of the amino acids, the biological activity of the protein is created.

Since, within their amino acid sequence, proteins also encode the required information to establish how they will fold into a three dimensional structure, research in protein folding – the holy grail of simulation and indeed the relevant scientific community – will invariably rely heavily on knowledge about individual amino acids.

The amino acids used in proteins fall into two classes. Those which the human body can produce on its own and those required in the diet. The latter are called the *essential* amino acids.

In addition to their rôle in proteins, amino acids have other features assigned to them. In my line of research, I focused particularly on proline (a non-essential amino acid), which will be introduced separately later. This specimen exhibits functions of bio-protective nature, and the molecular structure involved in the underlying process shall be explored in section 5 on page 105.

### 1.3.4 Poly-peptides

#### **what are they?**

A peptide is merely a short molecule consisting of at least two chemically bonded  $\alpha$ -amino acids. The link between one amino acid and the next is called an amide bond, or peptide bond.

Technically, proteins are merely polypeptides (i.e. more than two amino acids), or several groups of sub-polypeptides. Conventions for differentiating between polypeptides and proteins are muddy, but a fairly common sense one stipulates that peptides can be made synthetically from amino acid constituents; with the advancement of synthesising techniques, this seems to be muddied further.

## Chapter 2

---

# Methods



---

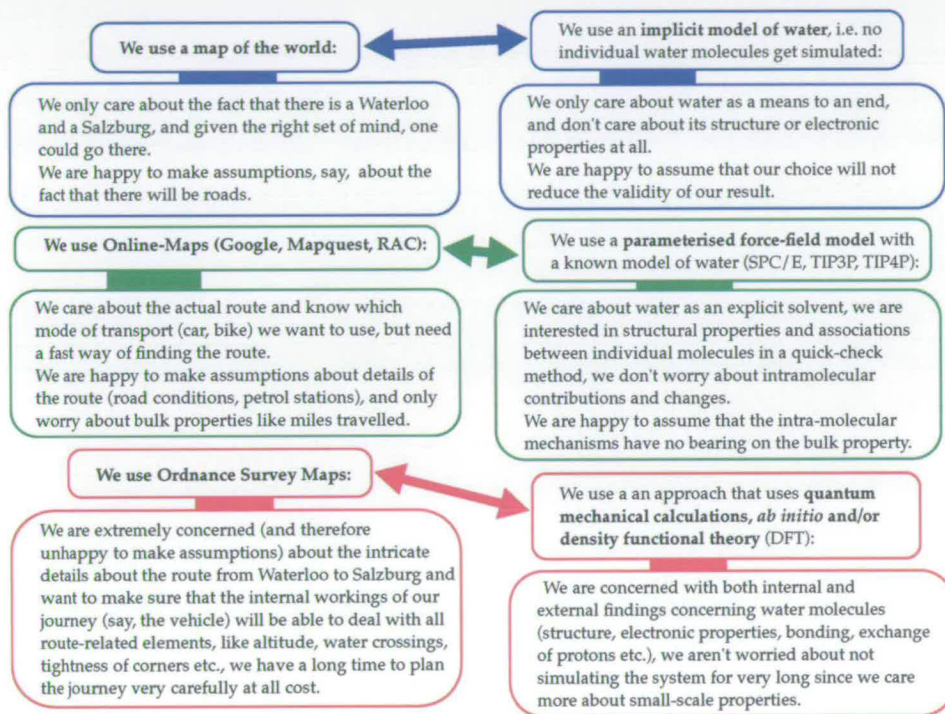
In this chapter we shall be exploring the basics of Molecular Dynamics methods, including some of the reasons why certain things are done the way they are. Only very few of these things are actually new, and there is no claim to completeness. Further reading should in particular include first Allen and Tildesley[2] and then Frenkel and Smit[38]. This presentation will hopefully prove useful for someone who has no prior exposure to the topic and provide depth for the more knowledgeable reader. My goal in this chapter is to author the type of introduction to Molecular Dynamics I would have liked for someone to have written before I first started in the field and yet provide a presentation that does not sacrifice insight.

### 2.0.0 Very brief, Soft-Core, Rough-Guide to Simulation

To simulate a “system” of atoms, molecules etc., one needs to define a set of interaction rules. Underlying the rules are, of course, physics and its laws, but depending on what level of detail is sought, one can choose different ways of defining a set of interaction rules.

Let me draw the comparison to finding “properties of car journey” from, say, Waterloo to Salzburg. We won’t actually take the journey, because that would be the experimental part, which we are not interested in for now.

We have a few options: We can look at a map of the world and get a good idea about where the two are with respect to one-another. We can log onto Google-Maps and have a route calculated for us. We can refer to loads of Ordnance Survey Maps and get a route from those, which will also have information about altitude at all points on the route. All methods will yield a route, but because we make varying degrees of assumptions about the general applicability of the information, there will be varying



**Figure 2.1:** Crib-sheet: A brief outline of the different techniques available, along with the parallels to an every-day life example.

$$x_i(2\Delta t) = x_i(\Delta t; x_1(\Delta t) \dots x_N(\Delta t), \dot{x}_1(\Delta t) \dots \dot{x}_N(\Delta t)) \quad (2.3)$$

As we anticipate  $\Delta t$  to be chosen fairly small, this implies many computations of the force. As  $N$  increases, so does the number of interactions; if we assume a pair potential, this scales as  $N^2$ . The combination of these two issues makes for a problem that is ultimately the core of MD.

### 2.1.2 Connection to Thermodynamics and Statistical Mechanics

This is going to be a fast crash-course, taking a few conditions implicitly as read. There are numerous publications on statistical mechanics and thermodynamics that go into deeper detail[35, 65, 73].

In order to arrive at a Hamiltonian formulation of MD (which we shall do in the next section), that connects back to thermodynamics and statistical mechanics and paves the way for important, but somewhat heavy mathing, let's start with the first law of thermodynamics ("Conservation of Energy, or Energy is a state function"), and the second law ("Entropy is a state function"), where a state function only depends on the conditions, not on history, to arrive at an expression for the infinitesimal change in energy

$$dE = TdS - pdV + \mu dN. \quad (2.4)$$

Here, as expected,  $p$  is the pressure,  $V$  the volume,  $\mu$  the chemical potential and  $N$  the particle number. Since energy is a state function,  $dE$  is an exact differential, and thus  $S, V$  and  $\mu$  must be variables of state. As entropy is the variable that's hard to set, it is often expressed as the independent variable, which yields

$$dS = \frac{1}{T}dE + \frac{p}{T}dV - \frac{\mu}{T}dN, \quad (2.5)$$

implying that  $S(NVE)$  can be used to obtain

$$\begin{aligned} \frac{1}{T} &= \left( \frac{\partial S}{\partial E} \right)_{V,N} \\ \frac{p}{T} &= \left( \frac{\partial S}{\partial V} \right)_{E,N} \\ \frac{\mu}{T} &= - \left( \frac{\partial S}{\partial N} \right)_{E,V}, \end{aligned} \quad (2.6)$$

from which we will develop a link to molecular, microscopic systems.

There are two more functions with different independent variables, the Helmholtz Free Energy  $A(NVT)$  and the Gibbs Free Energy  $G(NpT)$ , which can be defined via Legendre transformations to yield

$$A \equiv E - TS \text{ and } G \equiv A - TS + pV \quad (2.7)$$

the positions and momenta. Let us also define the vector function  $\xi(\Gamma, t)$  as

$$\xi(\Gamma, t) = \dot{\Gamma}. \quad (2.15)$$

Let us further consider a normalised distribution function,  $f(\Gamma, t)$ , describing an ensemble of states whose evolution obeys Hamilton's equations. An equation of motion for this distribution function can be derived by balancing the rate of change of the number of ensemble members inside a phase space volume  $V$  by the flux through the boundary surface

$$\frac{d}{dt} \int_V d\Gamma f(\Gamma, t) = - \int_S dS_\Gamma [\hat{n} \xi(\Gamma, t)] f(\Gamma, t) = - \int_V \sum_i \xi_i(\Gamma, t) \nabla_i f(\Gamma, t) \quad (2.16)$$

where phase space is assumed to be Euclidean[73] and Liouville's theorem has been used (i.e. phase space volume is preserved, so that the volume element is unity, which is embodied by the condition  $\sum_i \nabla_i \xi_i = 0$  valid for Hamiltonian dynamics. In 1D  $\sum_i \nabla_i = \sum_i \dot{q}_i d/dq_i + \dot{p}_i d/dp_i$ ;  $i = 1 \dots N$ ) If this result is to hold for all possible volumes, the local result holds

$$\frac{\partial f(\Gamma, t)}{\partial t} + \sum_i \xi_i(\Gamma, t) \nabla_i f(\Gamma, t) = \frac{df}{dt} = 0. \quad (2.17)$$

This is called the Liouville equation.

We are interested in equilibrium solutions,  $\frac{\partial f}{\partial t} = 0$ . In this case,  $f = f(H)$ , since

$$\frac{df}{dt} = \frac{df}{dH} \frac{dH}{dt} = 0. \quad (2.18)$$

Consistent with Gibbs' postulations of statistical mechanics, we need to construct a distribution function which allows us to visit *all* states, created equal in the eye of [insert favourite deity or equivalent here], in phase space, subject to the constraints imposed by the Hamiltonian,  $H$ , with equal *a priori* probability. Evidently

$$f(H(\Gamma)) = C \delta(H(\Gamma) - E), \quad (2.19)$$

with  $C$  a constant. We can thus define the phase space volume and ensemble partition function

$$\Omega(NVE) = \frac{C_N}{h^N} \int d\Gamma \delta(H(\Gamma) - E), \quad (2.20)$$

such that an average quantity

$$\langle O \rangle = \frac{C_N}{h^N \Omega(NVE)} \int d\Gamma \delta(H(\Gamma) - E) O(\Gamma) \quad (2.21)$$

and the definition of the Poisson Bracket

$$\frac{dA}{dt} = \dot{q} \frac{\partial A}{\partial q} + \dot{p} \frac{\partial A}{\partial p} = \frac{\partial H}{\partial p} \frac{\partial A}{\partial q} - \frac{\partial H}{\partial q} \frac{\partial A}{\partial p} \equiv \{A, H\} \quad (2.27)$$

leads to the definition of the Liouville operator

$$\begin{aligned} i\mathcal{L} &= \sum \left[ \dot{q}_i \frac{\partial}{\partial q_i} + \dot{p}_i \frac{\partial}{\partial p_i} \right] \\ &= \sum \left[ \frac{p_i}{m} \frac{\partial}{\partial q_i} - F(q_i) \frac{\partial}{\partial p_i} \right] = \sum \left[ \frac{\partial H}{\partial p_i} \frac{\partial}{\partial q_i} - \frac{\partial H}{\partial q_i} \frac{\partial}{\partial p_i} \right] \\ &\equiv \{ \dots, H \}. \end{aligned} \quad (2.28)$$

These results hold for arbitrary  $H(\mathbf{q}, \mathbf{p})$ , not just equation (2.26)[40]. We have as before, the state of the system,  $\Gamma$ , defined as the set of positions and momenta,  $\{\mathbf{q}, \mathbf{p}\}$ . The state of the system,  $\Gamma$ , is a function of time, and evolves in time according to

$$\begin{aligned} \dot{\Gamma} &= i\mathcal{L}\Gamma \\ \Gamma(\Delta t) &= \mathcal{T}(\Delta t)\Gamma(0), \end{aligned} \quad (2.29)$$

where  $\Gamma(0)$  is the state at time  $t = 0$  and  $\mathcal{T}(\Delta t)$  is the classical time evolution operator

$$\mathcal{T}(\Delta t) = e^{i\Delta t \mathcal{L}}. \quad (2.30)$$

Clearly,  $\mathcal{T}(\Delta t)$  is a unitary operator and equation (2.29) is time reversible, since

$$\Gamma(0) = \mathcal{T}(-\Delta t)\Gamma(\Delta t) = \mathcal{T}(-\Delta t)\mathcal{T}(\Delta t)\Gamma(0) = e^{-i\Delta t \mathcal{L}}e^{i\Delta t \mathcal{L}}\Gamma(0) = \Gamma(0). \quad (2.31)$$

Hamilton's equations can, in general, not be solved analytically, and numerical integration schemes must be employed. These are subject to approximation, and thus involve the introduction of error, as we move from the exact trajectory  $\Gamma(t)$  to the approximate trajectory  $\tilde{\Gamma}(n\Delta t)$ , where  $\Delta t$  is the finite time-step, such that  $t = n\Delta t$ .

In order to generate numerical integrators within controlled approximations, we exploit the freedom to decompose the Liouville operator into two parts, such that

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 = \{ \dots, h_1 \} + \{ \dots, h_2 \}, \quad (2.32)$$

where we have defined

$$H = h_1 + h_2 \quad (2.33)$$

at the same time. It follows that

$$\mathcal{T}(\Delta t) = e^{i\Delta t \mathcal{L}} = e^{i\Delta t(\mathcal{L}_1 + \mathcal{L}_2)} \neq e^{i\Delta t \mathcal{L}_1} e^{i\Delta t \mathcal{L}_2}, \quad (2.34)$$

Campbell-Hausdorff (BCH) expansion[121] of equation (2.35) on the preceding page, the integrator can be shown<sup>1</sup> to yield the solution to the continuous time equations of motion

$$\dot{\tilde{\Gamma}}(t) = \sum_{i=0}^{\infty} \Delta t^{2i} \tilde{\xi}^{(i)}(\tilde{\Gamma}(t)), \quad (2.40)$$

where  $\tilde{\xi}(\tilde{\Gamma}(t)) = \dot{\tilde{\Gamma}}(t)$  and  $\tilde{\Gamma}(t) \neq \Gamma(t)$ . Recall that the original equations have a conserved quantity

$$\frac{dH(\Gamma(t))}{dt} = 0, \text{ or } i\mathcal{L} \cdot H(\Gamma) = \xi(\Gamma) \cdot \nabla_{\mathbf{q}} H(\Gamma) = 0. \quad (2.41)$$

If we can determine conditions under which equation (2.40) will have a corresponding conservation law  $\tilde{H}(\Delta t, \tilde{\Gamma})$ , this will bound the error in equation (2.36) and allow us to consider MD as sampling  $\int d\Gamma \delta(E - \tilde{H}(\Delta t))$ , assuming ergodicity. In this way, we can compute any experimental observable as an ensemble average *à la* Gibbs from the approximate dynamics.

Since the integrator is reversible by construction, the expansion of  $\tilde{H}$  must occur in even powers of  $\Delta t$ , and we can expand formally as

$$\tilde{H}(\Gamma(0); \Delta t) = \sum_{i=0}^{\infty} \Delta t^{2i} \tilde{H}^{(i)}(\Gamma(0)). \quad (2.42)$$

So now we must find conditions for  $\tilde{H}$  to exist. First, the perturbation dynamics defined by decompositions of the classical Hamiltonian as in equations (2.32,2.33) can be shown to be of the form

$$\tilde{\xi}^{(i)}(\Gamma(0)) = \nabla_{\mathbf{p}} \tilde{H}^{(i)} \cdot \mathbf{g}^{-1}, \quad (2.43)$$

where we introduce the constant skew symmetric matrix  $\mathbf{g}^{-1}(\Gamma(0))$ .

$$\mathbf{g}^{-1} = \begin{pmatrix} 0 & \mathbf{I} \\ -\mathbf{I} & 0 \end{pmatrix}. \quad (2.44)$$

In this case, each perturbation dynamics clearly has the conservation law  $\tilde{\xi}_{(i)}(\Gamma) \cdot \tilde{H}^{(i)}(\Gamma) \equiv 0$ , and equation (2.42) holds.

Now let us connect the decomposition of  $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 = \{\dots, H_1\} + \{\dots, H_2\}$  to equations (2.42,2.43) to square the circle. Since the original equations are Hamiltonian,

<sup>1</sup>equation (2.24) in [76]



property of Hamiltonian systems, and integrators of Hamiltonian systems with this behaviours are referred to as *symplectic integrators*. Any integrator derived via a Liouville operator decomposition of a Hamiltonian dynamics, as in equations (2.32,2.33), will be a symplectic integrator, possess a  $\tilde{H}$  and have a guarantee of long stability and the existence of well defined phase space averages.

### 2.1.4 Other ensembles

Equations of motion that generate other ensembles, such as *NVT* and *NPT*, are more complex, and so in order to understand them we need to explore some more mathing first.

Let us express the traversal along the trajectory of a (for simplicity sake 1-dimensional) system

$$x(0) \rightarrow x(t) \tag{2.50}$$

as a variable transformation, which will involve a Jacobian,  $J(x_t; x_0)$ , such that

$$dx_t = J(x_t; x_0)dx_0. \tag{2.51}$$

In a system involving more than one particle/dimension, the Jacobian will be the determinant of a matrix. In a Hamiltonian system,  $J(x_t; x_0) = 1$  for all  $t$ , a property referred to as phase space volume preservation, which is a consequence of Liouville's theorem, as discussed above.

It can be shown (see Ref. [109] for more detail) that

$$\begin{aligned} \frac{d}{dt} J(x_t; x_0) &= J(x_t; x_0) \nabla_{x_t} \cdot \dot{x}_t \\ \frac{d \log J(x_t; x_0)}{dt} &= \nabla_{x_t} \cdot \dot{x}_t \end{aligned} \tag{2.52}$$

which we can integrate, formally,

$$\begin{aligned} J(x_t; x_0) &= \exp\left(\int_0^t \nabla_{x_s} \cdot \dot{x}_s ds\right) \\ &= \exp(w(x_t, t) - w(x_0, 0)), \end{aligned} \tag{2.53}$$

and substitute into equation (2.51) to yield, first,

$$\exp(-w(x_t, t))dx_t = \exp(-w(x_0, 0))dx_0 \tag{2.54}$$

and then

$$\sqrt{g(x_t, t)}dx_t = \sqrt{g(x_0, 0)}dx_0, \tag{2.55}$$

where we have introduced the metric determinant  $\sqrt{g(x_t, t)}$ . Therefore, for a general

## Canonical ensemble

Starting from the Nosé-Hoover-chain equations of motion postulated by Martyna, Tobias, Klein

$$\begin{aligned}
 \dot{\mathbf{q}}_i &= \frac{\mathbf{p}_i}{m_i} \\
 \dot{\mathbf{p}}_i &= \mathbf{F}_i - \frac{\mathbf{p}\eta_1}{Q} \\
 \dot{\eta}_k &= \frac{p\eta_k}{Q}, & \text{where } k = 1, \dots, M \\
 \dot{p}\eta_1 &= \left[ \sum_{i=1}^N \frac{\mathbf{p}_i^2}{m_i} - dNkT \right] - \frac{p\eta_2}{Q_2} p\eta_1 \\
 \dot{p}\eta_k &= \left[ \frac{p\eta_{k-1}^2}{Q_{k-1} - kT} \right] - \frac{p\eta_{k+1}}{Q_{k+1}} p\eta_k & \text{where } k = 2, \dots, M-1 \\
 \dot{p}\eta_M &= \left[ \frac{p\eta_{M-1}^2}{Q_{M-1}} - kT \right],
 \end{aligned} \tag{2.61}$$

where the system has  $N$  particles, and the forces  $\mathbf{F}_i = -\nabla_{\mathbf{q}_i} \phi$  are derived from the  $N$ -body potential function  $\phi$ , and  $dN$  is the number of degrees of freedom. The extra variables  $\{\eta, p\eta\}$  are considered to be a chain of *thermostats*, which control the fluctuations in the total kinetic energy of the system about  $kT$ . Finally,  $Q_k$  determines the timescale on which the thermostat evolves, and is found to be best described by  $Q_1 = dNkT\tau^2, Q_{k=2\dots M} = kT\tau^2$ . Here,  $\tau$  is the frequency upon which the chain removes/adds energy from/to the physical system.

Our conserved quantity for equation (2.61) case is

$$H' = H(\mathbf{p}, \mathbf{q}) + \sum_k \frac{p\eta_k^2}{2Q_k} + kT\eta_c, \tag{2.62}$$

where  $H(\mathbf{p}, \mathbf{q})$  is the physical Hamiltonian and  $\eta_c = dN\eta_1 + \sum_{k=2}^M \eta_k/Q_k$ . The dynamics has the phase space metric factor

$$\sqrt{g} = \exp(\eta_c) \tag{2.63}$$

Following the derivation of the previous section

$$\begin{aligned}
 \Omega_T(N, V, H') &= \int d^N \mathbf{p} \int_{D(V)} d^N \mathbf{q} \int dp\eta \int d\eta_c \exp(\eta_c) \\
 &\quad \times \delta \left( H(\mathbf{p}, \mathbf{q}) + \frac{p\eta^2}{2Q} + kT\eta_c - H' \right),
 \end{aligned} \tag{2.64}$$

As before, we apply the non-Hamiltonian procedure, whereby the microcanonical function is given as

$$\Omega_{T,P_{\text{ext}}}(N, H') = \int dV \int d^N \mathbf{p} d^N \mathbf{q} d^M p_\eta d\eta_1 d\eta_c dp_\epsilon \times \exp(\eta_c) \delta(H' - E). \quad (2.69)$$

Performing the  $\delta$ -function integration, we finally arrive at

$$\begin{aligned} \Omega(N, P_{\text{ext}}, E, T) &\propto \int d^M p_\eta \exp\left(\frac{\beta p_\eta^2}{2Q}\right) \int dp_\eta \exp\left(\frac{\beta p_\eta^2}{2W}\right) \int dV \exp(\beta P_{\text{ext}} V) \\ &\quad \times \int d^N \mathbf{p} \int_{D(V)} d^N \mathbf{q} \exp(-\beta H(\mathbf{p}, \mathbf{q})) \\ &\propto \Delta(N, P_{\text{ext}}, T), \end{aligned} \quad (2.70)$$

which, again assuming ergodicity, is the correct reproduction of the isotropic  $NPT$  ensemble!

of these as pseudo-springs, we can assign a potential energy to a deviation from these equilibrium distances, angles and so forth by means of a spring-constant, or better a force constant, which we then just put in with the rest of the parameters. We arrive at

$$\begin{aligned}
 V_{\text{bonds}} &= \sum_i^N k_{b_i} (b_i - b_{0_i})^2 \\
 V_{\text{bends}} &= \sum_i^N k_{\theta_i} (\theta_i - \theta_{0_i})^2 \\
 V_{\text{dihedrals}} &= \sum_i^N k_{\phi_i} [1 + \cos(n\phi_i - \delta_i)] \\
 V_{\text{improper}} &= \sum_i^N k_{\omega_i} (\omega_i - \omega_{0_i})^2 \\
 V_{\text{Urey-Bradley}} &= \sum_i^N k_{u_i} (u_i - u_{0_i})^2,
 \end{aligned}$$

where all the  $k$  represent the force constant,  $b, \theta, \phi, \omega$  and  $u$  are bondlength, bond angle between three bonded atoms, dihedral angle, the out of plane angle, and the distance between the 1,3 atoms in a harmonic bond, respectively.  $n$  is the multiplicity of the system (selected to describe the number of minima/maxima in the torsional surface, e.g. butane C-C-C-C torsion has 3 minima and 3 maxima, thus  $n = 3$ ), and  $\delta$  is a phase shift. The  $x_0$  represent the equilibrium value of the variables. Of course, these are all intramolecular energy terms, nevertheless, we'll need them to get the overall force.

Since some of these contributions are recognised to evolve in time far slower than others (e.g. bond vibrations need to be sampled at a higher frequency than, say, the Lennard-Jones interaction), we can make use of multiple timescale techniques already alluded to in section 2.1.3 by exploiting equation 2.36, which allow us to compute the computationally more expensive forces less frequently, e.g. by writing  $H = \sum_k h_k$  and introducing multiple time-steps at which to apply their associated Liouville operators. For further reading, please refer to Refs. [3, 69, 70, 87].

There is one more contribution to the energy, which is the electrostatic interaction, here modelled by a Coulombic potential,

$$V_{\text{electrostatic}} = \frac{q_a q_b}{r_{ab}} \quad (2.73)$$

where  $q$  is the charge of the atom in question,  $r_{ab}$  is the distance between atoms  $a$  and  $b$ . The electrostatic potential is a long range interaction, which in 3D and including

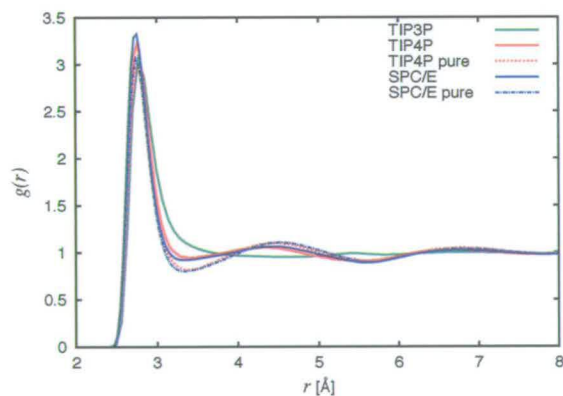
## Strengths

The power of a force field such as CHARMM is its extreme versatility dealing with many different types of systems quite efficaciously. Its ability to treat amino acids and other common building block molecules, as well as whole proteins, explains its wide use.

## Pitfalls

It may already have sounded a little bold to simply assume that we can decide on a parameter for our model for a category of atoms, and forget about it. Consider now the fact that in addition to this generalisation, we will often also, mainly to allow for a large  $\Delta t$  in the numerical integration, make certain bonds/bends rigid through the introduction of Lagrange multipliers[3, 40, 87]. This works fairly well in many cases, and has the bonus of speeding the simulation up significantly, but will also introduce artificial results which have to be eyed carefully.

As mentioned before, water is an extremely important part in life, and hence an important one in simulation. Sadly water is also rather unique in many bulk property respects, as well as in its intermolecular interactions, which makes it hard to find a good simple model. Particularly, the polarisation of water molecules, that is the redistribution of the atoms' charges across the water molecules, due to the presence of other charged particles (and water molecules), is a big factor. No force-field fully captures the polarisation of water to be included in the interactions, although polarisable force fields are being built.



**Figure 2.2:**  $O_w - O_w$  Radial distribution functions for three water models for aqueous proline are shown in the three solid colour lines. The corresponding functions for pure water are shown as dashed lines.

The water model that comes as the house-brand of CHARMM, TIP3P, describes the water molecule as a rigid three-charge site molecule with fixed bond length  $b_{OH} =$

## 2.3 Parallel Tempering

Let us spend a brief moment thinking about what we will actually be simulating, in terms of where the system in question can evolve. There is a large number of different states that the system could potentially occupy. We might call this “state space”, but it is commonly referred to as *phase space*, as a point or state is defined by  $\{\mathbf{p}, \mathbf{q}\}$ . Different states will cost different amounts of energy or entropy, this is as a result of the distances, angles, torsions etc. between atoms and within molecules. Loosely expressed, one of the underlying principles of physics is, that, given a choice, a system will try and be in the lowest energy state it can.

In the canonical ensemble, due to fluctuations in a complex system consisting of several hundreds or thousand of atoms, a system may sometimes go from a low energy state to one that is slightly higher and, as a result, might go to a different one which is lower than the one before.

In real life, the processes of evolution between states happen rapidly, so that one can assume that overall, the system is in an adequate state. When we set up a liquid simulation, we often start from a rather contrived structure, not uncommonly in fact from a lattice configuration of the molecules involved. Even though we then equilibrate, that doesn't give us confidence that the resulting simulation will in fact be running in a state that is the representation of the equilibrium distribution. It may well be the case, that we are stuck in a local minimum of energy, from which we cannot escape. “We” in this case means the system, and as a result the simulator, since he/she is stuck with a potentially incomplete or wrong result. Commonly, we refer to this as “not sampling enough of phase space”, or lack of ergodicity, which we assumed in previous subsections (c.f. equation (2.22) on page 16).

### 2.3.1 A look at phase space, or the Alps

In order to visualise the problem at hand, let us briefly hark back to the introduction with its concept of journeys. If we now make our journey an expedition in the Alps and consult figure 2.3 for an elevation map of the *Berner Oberland* in the Alps, then it is easy to see that we might have trouble getting to the top of the *Jungfrau* in our [insert model of really lame car here], and in fact we may not even get over the pass of a given mountain and to the next valley, say, from Brig to Interlaken. So we end up stuck driving around a valley, over and over again. We may see many interesting things there, but it may not be representative of Switzerland. In fact, we might have been stuck in a particularly special or even peculiar part of the country, and we'll end up knowing that part really well, but won't have the correct impression of the country as a whole.

to that, but let's dwell on that example.

If we were a little team in the Alps, we could do the following: We could have different cars (different replica of our system) starting in different valleys (states) with different engines (temperatures). Some would be strong (hot) enough to make it over the hills (energy boundaries). Now and then we could share information on the current scenery (configuration, motifs).

The analogy is starting to lack somewhat, since *per se* there is no drawback to having a LandRover in the Alps (other than maybe price, fuel efficiency), whereas a system whose properties we would like to explore at ambient conditions we cannot very well simulate at 100 K hotter. We need to find a clever way.

### 2.3.2 Replica exchange – a clever way

Freshly back from the Alps, we can now think about the intricacies of molecular dynamics and the trick we hinted at. Parallel tempering (PT) is first sourced in a paper by Swendsen and Wang[98] and is based on the idea of simulating  $M$  replicas in a series of temperatures (again, other variables like pressure could similarly be used). The highest temperature is picked such that easy traversal through phase space is given, while low temperatures will sample very accurately (since trapped in) a small area of phase space. This is visualised in figure 2.3.

The ingenious trick is that at regular intervals during the simulation adjacent replica, or *temperers*, are allowed to swap configurations (as in the Alps people might swap cars), which was first formulated by Geyer in 1991[39]. This way, the states accessed at higher temperature have a chance of trickling down into a lower temperatures' replica. The exchange between temperers  $i$  and  $j$  is accepted with probability[23]

$$P_{ij} = \min \left\{ 1, \frac{\exp[\Delta\beta_{ij}\Delta E_{ij}]}{1 + \exp[\Delta\beta_{ij}\Delta E_{ij}]} \right\} \quad (2.76)$$

where  $\Delta\beta_{ij}$  and  $\Delta E_{ij}$  are the difference in inverse temperature and total energy respectively. Since the temperatures enter into the acceptance criterion explicitly, the probability of accepting a move decreases exponentially with a widening temperature gap between temperers. This is also the reason why commonly only neighbouring temperers are swapped, even though in principle a swap could be attempted between any two temperers.

#### Picking temperatures – or: eating the broth as hot as it is boiled?

Exactly how one should choose the temperatures is what one might call a dark art, but there have been advances (some quite rigorous) about how to pick them ideally[52].

At the time of analysis, we can refer back to these probabilities, and fold them into the calculation of, say, a radial distribution function. The distance distribution between two types of atoms, which for one temperer at temperature  $T$  is  $A(r, t)$ , and integrated over time becomes

$$A(r) = \frac{1}{t} \int_0^t dt' A(r, t'). \quad (2.77)$$

Note that, due to the switches,  $A(r, t)$  does not vary smoothly. In addition, since we use a finite integrator,

$$A(r) = \frac{1}{t} \int_0^t \dots \rightarrow \frac{1}{N} \sum_0^N \dots, \quad (2.78)$$

where  $t = N\Delta t$ . In our new scheme, where we include all  $M$  temperers, we find that the distribution at temperature  $T_j$  is  $A_j(r)$  as

$$A_j(r) = \int_0^t dt' \sum_{i=1}^M P_{ji}(t') A_i(r, t') \quad (2.79)$$

which is properly normalised as long as we ensure that

$$\sum_{i=1}^M P_{ji}(t) = 1 \quad (2.80)$$

for all  $t$ . Note, if the system is ergodic, we recover phase space averages.

So as a matter of fact, what we achieve this way is a double whammy of PT power. Not only do we improve sampling for our target temperature, we also include the results of the non-target temperatures legitimately into our analysis of the target temperature. An example of the resulting difference in a 1D radial distribution function can be seen in figure 2.4.



## 2.4 Car-Parrinello Molecular Dynamics

So far, we have treated atoms and molecules as simple ball and stick constructs, which is also the way we usually visualise an atom, as is in fact the case in all illustrations in this work. This is, of course, an extreme simplification. What really goes on under the bonnet of an atom and a covalent bond is rather more complicated, and arguably should enter the way we model the interactions.

Let us briefly sum up the drawbacks of relying on a force-field model of point particles. First of all, charges appear as static attributes in the force field, and thus effects due to electronic polarisation are not captured, except in a mean field sense. Attempts to overcome this problem have been made, and include *polarisable models*, in which charges and induced dipoles respond to changes in the (local) environment. Recent efforts to make a polarisable water model that will otherwise incorporate a regular force field are promising. Another huge limitation, certainly in a general force field applicable to all systems, is the presumed connectivity of molecules. Clearly, chemical reactions are thus *a priori* excluded from being sampled, unless a specific (thus not transferable) technique is included, which may in turn bias towards some particular reaction inadvertently.

Fortunately, the (at this point still fairly) recent technique of *ab initio* molecular dynamics (AIMD) method has been developed to overcome some of these problems. It includes the forces obtained from an approximate electronic structure, computing this contribution as it goes along. Because of this explicit treatment of the electronic structure, electronic polarisation, bond forming and breaking events, and many-body forces are expressed to within the accuracy of that treatment. Sadly, despite heroic efforts to scale small systems onto large buildings full of computers (namely IBM's BlueGene/L)[10, 114], this treatment only allows for small ( $\mathcal{O}(10^3)$  atoms) systems to be simulated long enough to generate ensemble averages.

For the real enthusiast, the formulation of AIMD is also compatible with Feynman's path integral approach, and can thus be extended to include quantum effects of the nucleus in the ground state surface. Since even with computing power available these days, that approach is rather the slow (if very exact) one, we won't look into it any further. Of course, path integrals can also be used in conjunction with classical force fields.

### 2.4.1 Complicating the story

Previously, we had stated that we are treating ions as point particles in the Born-Oppenheimer approximation on the ground state electronic structure, and we still are.

The treatment of the electronic structure has already been alluded to briefly, but it

Indeed, a series of approximations are used to employ DFT in practice. First, largely for computational efficiency, only the valence electrons are treated, and the core electrons are replaced by norm-conserving nonlocal *pseudopotentials*. Next, a plane-wave basis set is introduced to describe the states

$$\Psi(\mathbf{r}) = \sum_{\frac{1}{2}|\mathbf{g}|^{\frac{1}{2}} < E_{\text{cut}}} \Psi_i(\mathbf{g}) \exp(i\mathbf{g} \cdot \mathbf{r}), \quad (2.84)$$

where  $\{\Psi(\mathbf{g})\}$  is a set of expansion coefficients and  $\mathbf{g}$  is a reciprocal lattice vector. The expansion is truncated using an energy cutoff  $E_{\text{cut}} > \frac{1}{2}|\mathbf{g}|^2$ . Last, we take the local density approximation, whereby

$$E_{\text{ex-corr}}[\rho] \approx \int d\mathbf{r} \rho(\mathbf{r}) \epsilon_{\text{ex-corr}}(\rho(\mathbf{r}), \nabla\rho(\mathbf{r})), \quad (2.85)$$

where  $\epsilon_{\text{ex-corr}}(\rho(\mathbf{r}), \nabla\rho(\mathbf{r}))$  only depends on the electron density  $\rho(\mathbf{r})$  and its gradients[7, 58]. This type of approximation to the exchange-correlation functional is referred to as a generalised gradient approximation, or a ‘‘GGA’’. In this thesis, we will be working with KS-GGA-DFT in general, and the B-LYP[7, 58] functional in particular.

### Car-Parrinello, arriving at the equations of motion again

We now have all the tools to finally formulate the Car-Parrinello approach[15], which is quite inventive. In fact, it starts by proposing a fictitious dynamics for the plane-wave coefficients  $\{\Psi(\mathbf{g})\}$  of the states. This allows an electronic configuration that is initially minimised to co-move with the nuclei to produce nuclear dynamics on the ground state along the Born-Oppenheimer surface. This is accomplished by positing an adiabatic separation between the fictitious state dynamics and the real dynamics of the nuclei, which allows for the electrons and nuclei to be treated on equivalent grounds mathematically. The electrons can be kept at a far lower temperature than the nuclei, and can also be allowed to move far faster. This is not an immediately obvious thing to do, but it does mean that under these conditions the electrons will approximately minimise the density functional as the nuclei propagate slowly along their trajectory. We arrive at a new expression for the equations of motion

$$\begin{aligned} \mu(\mathbf{g})\ddot{\Psi}_i(\mathbf{g}) &= \frac{\partial E}{\partial \Psi_i^*(\mathbf{g})} + \sum_j \Lambda_{ij} \Psi_j(\mathbf{g}) \\ &= \sum_j F_{\Psi_i}(\mathbf{g}) + \sum_j \Lambda_{ij} \Psi_j(\mathbf{g}) \\ M_I \ddot{\mathbf{R}}_I &= -\frac{\partial E}{\partial \mathbf{R}_I} = \mathbf{F}_I. \end{aligned} \quad (2.86)$$

Here,  $\mu(\mathbf{g})$  is a mass-like parameter with units [energy-time<sup>2</sup>] for the electronic motion, and the orthogonality condition is enforced by a set of Lagrange multipliers  $\Lambda_{ij}$  [3, 40,

## 2.5 Analysis Tools

In this part, we will briefly dive into the techniques employed to, once we have obtained simulation data, pull out significant information from the trajectory. To this end, we will introduce briefly the idea of the radial distribution function – the mothership of all analysis in a way –, and continue on with an extension into multidimensional distribution function, radial as well as angular. I will then outline the way we compute the static scattering structure factor, which is a key quantity in connecting to experiment. Finally, a novel way of probing water structure is introduced.

### 2.5.1 1D- $g(r)$ s — the power-house of analysis

When we talk about radial distribution functions (RDF,  $g(r)$ ), we refer to the quintessential tool in the analysis tool box of any simulator. It is, mathematically, the probability density as a function of the distance,  $r$ , of two atoms, or, vernacularly, the probability density of finding an atom at distance  $r$ , given that you sit on an atom (at distance  $r = 0$ ). In a liquid, we expect that probability density to approach unity at large separation, which arises from the relative disorder at large separation. In contrast to a regular solid (the best example being a monatomic crystal), this is not the case, since even at long distances, the positions of atoms are well defined with reference to the “origin”, i.e. the atom we currently sit on.

Since we are dealing with a probability density function, we can find the expectation value of an operator  $A(r)$ , dependent only on  $r$ , as

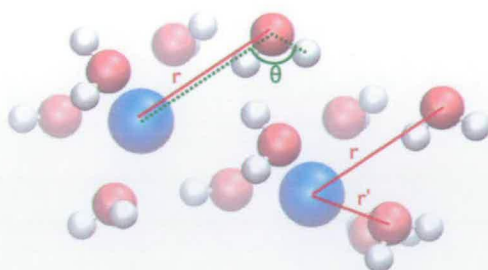
$$\langle A \rangle = 4\pi\rho \int_0^\infty dr A(r)g(r)r^2, \quad (2.87)$$

where  $\rho$  is the density of the system (Note that if we are looking at only a specific species in a multi component system, say oxygen, we must use the density of that species). This may prove useful in many instances, an obvious one being that we can compute the coordination number  $n(r_1, r_2)$ , i.e. the expected number of “guys” in a certain interval,  $[r_1, r_2]$ , as

$$n(r_1, r_2) = 4\pi\rho \int_{r_1}^{r_2} dr g(r)r^2. \quad (2.88)$$

Importantly, we can now find the number of atoms in a first coordination shell, i.e. the number of “guys” which are those on closest approach (and cause the first peak in figure 2.2 on page 29).

The concept of the  $g(r)$  is paramount to any operation involving simulation to elucidate structure, since it is an easy and cheap way to extract basic correlations



**Figure 2.5:** Schematics of potential 2D distributions one might consider of interest. Of course, any distribution is possible, and the only limitation beyond creativity is the computing time available to the analyst.

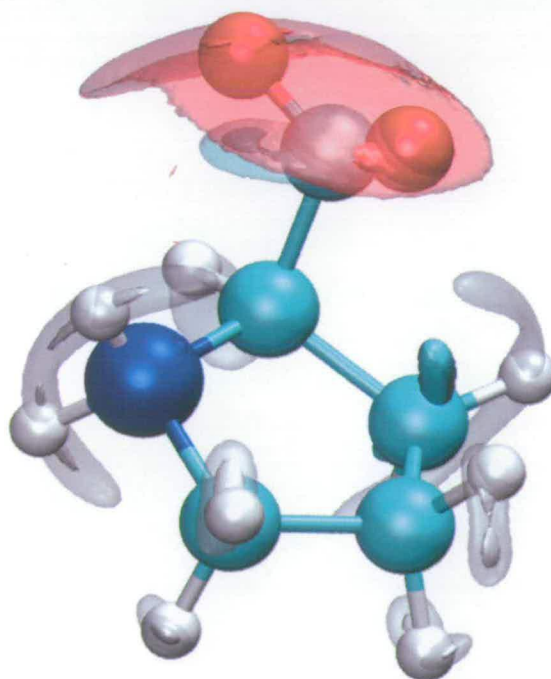
Since these methods are now well established, I believe it sufficient to refer to Whitfield *et al.* [120, 118], who introduced them formally in the papers cited.

It's time to look at an example of this, which is displayed in figure 2.6 on the next page. Shown there is the example of the distribution of radial and angular components as depicted in figure 2.5, along with the standard 1D- $g(r)$  overlaid to show the value-add. The ensuing results chapters are strewn with the types of plots seen in figure 2.6, as the 2D distribution has proven to be an invaluable tool.

### 2.5.3 Radial density visualisation in 3D – the swooshy

While the distribution functions discussed in the previous sections serve well to find particular contacts, pull out coordination numbers, and use them as an easy and cheap way to represent structural features, sometimes we just want to be able to look at a molecule and see where its constituents were likely to “swoosh” to over the course of a simulation. Or where a particular neighbour molecule has its likely location in a contact.

We can do this using a spatial function, by defining a set of axes from sites on a molecule, and then for each occurrence of it, we rotate the whole system to make the molecule sit on those same points. Then we look where the sought neighbours or, intramolecularly, where the other atoms of the molecule are. With a little careful rasterisation of space we can then calculate a statistical density function at every point in space in and around the molecule, much like in the 1D and 2D cases. The resulting files are huge and tedious, but they can be used to visualise a molecule. An example is shown in figure 2.7. Via these distributions it is possible to probe the spatial



**Figure 2.7:** Example of a spatial density function of the internal movements of a proline molecule. The solid and translucent areas correspond to 10 and 5 times uniform density respectively. Because the representation allows the visualisation of the swooshing movement of the atoms, the image has been dubbed “swooshy”.

### 2.5.4 Evaluation of the static structure factor from simulation

Neutron diffraction (particularly on a set of isotopically labelled samples) is perhaps the most powerful technique for the study of liquid structure in aqueous solutions. Experiment yields the *total* static structure factor,  $S(Q)$ , directly from the scattered neutron intensity, once various corrections for container scattering, multiple scattering and inelasticity are applied to the raw data. Structural information is obtained in the form of the total real space pair correlation function,  $G(r)$ , related to  $S(Q)$  via Fourier inversion in an infinite system[8, 31]. The total pair correlation function is a linear combination of partial atomic pair correlation functions defined by

$$G(r) = \sum_{i,j} b_i b_j c_i c_j g_{ij}(r) \quad (2.92)$$

where the  $b_i$  and  $c_i$  represent the coherent neutron scattering lengths and mole fractions of atomic species  $i$ , respectively. The site-site distributions  $g_{ij}(r)$  are therefore not directly accessible from a single measurement, and angular correlations, which cannot be reconstructed from molecular connectivity, are not included at all.

The total distribution is related to  $S(Q)$  via a continuous sine transform of

$$H(r) = \sum_{i,j} b_i b_j c_i c_j [g_{ij}(r) - 1] \quad (2.93)$$

only in thermodynamic limit,  $N \rightarrow \infty$  where  $N$  is the total number of particles in the system. Comparison between experiment and simulation at low  $Q$  is therefore most properly explored by direct evaluation of  $S(Q)$  from the molecular dynamics trajectories and comparison of the result to the diffraction data, as discussed in section 2.5.5. At high  $Q$ , the continuous transform relation represents a good approximation in the finite system.

#### Composite partial structure factor

As mentioned above, experimentally it is not possible to extract the real-space site-site correlations directly. However, by making use of partial physical substitutions in the sample (hydrogen/deuterium conventionally), in simple systems it is viable to extract some of them at least.

In the example of aqueous NaCl, there are only 4 components in the solution. Since the experimental scattering lengths,  $b_i$ , for materials are well known, partial substitution of H/D allows for the extraction of distributions concerning hydrogen, H, and everything else, X (here O, Na and Cl). These functions, referred to as composite partial structure factors, are then  $S_{HH}$ ,  $S_{XH}$  and  $S_{XX}$ .

system can be re-expressed as a configurational average over a partition function as

$$n_2(r) = \frac{1}{N} \left\langle \sum_{ij\mathbf{l}} \delta(|\mathbf{r}_i - \mathbf{r}_j - \mathbf{h}\mathbf{l}| - r) \right\rangle, \quad (2.96)$$

where  $N$  is the number of particles in the system,  $\mathbf{r}_i$  is the position of the  $i^{\text{th}}$  particle,  $\mathbf{h}$  is the matrix describing the simulation parallelepiped with volume  $V = \det \mathbf{h}$  and  $\mathbf{l}$  are the lattice indices such that  $\mathbf{h} = L\mathbf{I}$  and  $V = L^3$  in a cubic box of edge,  $L$ . In this way,  $n_2(r)$  is defined at all distances although 'spurious' correlations or finite size effects will be present when  $r > V^{1/3}$  in comparison to the result obtained in the thermodynamic limit.

The real space expression for the pair distribution function of a finite system can now be transformed into reciprocal space. Combining equation (2.96) and the Poisson summation formula,

$$\sum_{\mathbf{l}} F(\mathbf{h}\mathbf{l}) = \frac{1}{V} \sum_{\hat{\mathbf{Q}}} \tilde{F}(\hat{\mathbf{Q}}\mathbf{h}^{-1}) \quad (2.97)$$

$$\tilde{F}(\hat{\mathbf{Q}}\mathbf{h}^{-1}) = \int_{\text{allspace}} d\mathbf{r}'' \exp(-i2\pi\hat{\mathbf{Q}}\mathbf{h}^{-1}\mathbf{r}'') F(\mathbf{r}''),$$

where  $2\pi\hat{\mathbf{Q}}\mathbf{h}^{-1}$  is the reciprocal lattice vector with reciprocal lattice index,  $\hat{\mathbf{Q}}$ , yields

$$n_2(r) = \frac{4\pi r^2}{VN} \sum_{\hat{\mathbf{Q}}} \langle |n(\hat{\mathbf{Q}})|^2 \rangle \frac{\sin Qr}{Qr}. \quad (2.98)$$

Here, the atomic (single particle) density function in its reciprocal space form is given by

$$n(\mathbf{Q}) = \sum_j e^{i\mathbf{Q}\cdot\mathbf{r}_j}. \quad (2.99)$$

Taking  $\Omega_{\hat{\mathbf{Q}}}$  to be the number of reciprocal lattice vectors,  $\hat{\mathbf{Q}}$ , with magnitude,  $Q = |2\pi\hat{\mathbf{Q}}\mathbf{h}^{-1}|$ , in the finite system, it is possible to define the angle and ensemble averaged structure factor,

$$S(Q) \equiv \frac{1}{N\Omega_{\hat{\mathbf{Q}}}} \sum'_{\hat{\mathbf{Q}}} \langle |n(\hat{\mathbf{Q}})|^2 \rangle, \quad (2.100)$$

where the prime restricts the sum such that  $|\hat{\mathbf{Q}}'| = Q$ . Hence, in a finite system,

$$g(r) + \frac{\delta(r)}{2\pi r^2 \rho} = \sum_{\hat{\mathbf{Q}}} \bar{\Omega}_{\hat{\mathbf{Q}}} S(Q) \frac{\sin Qr}{Qr}, \quad (2.101)$$

where  $\bar{\Omega}_{\hat{\mathbf{Q}}} = \Omega_{\hat{\mathbf{Q}}}/N$ . Note, the sum over lattice vector magnitude is discrete with

systems to the thermodynamic limit. It is a well defined mathematical operation to compute the Fourier Sine transform of equation (2.101),

$$\int_0^R dr 4\pi r^2 \rho [g(r) - 1] \frac{\sin Qr}{Qr} = \left\{ \rho \sum_{\hat{Q}' \neq 0} \bar{\Omega}_{\hat{Q}'} S(Q') \left[ 4\pi \int_0^R dr \frac{\sin Qr}{Q} \frac{\sin Q'r}{Q'} \right] \right\} - 1, \quad (2.102)$$

where  $R$  is the prelimit factor that takes the system size to infinity. Clearly, the limit  $R \rightarrow \infty$  cannot be taken without changing the sum over the discrete  $Q'$  to an integral or equivalently taking the system to size infinity limit in reciprocal space simultaneously. In order to develop a correct limiting formalism, it is useful to define the prelimit  $\delta$ -function,

$$\Delta(Q, Q'; \sigma_Q) \equiv \frac{2}{\pi} \int_0^R dr \sin Qr \sin Q'r, \quad (2.103)$$

where  $\sigma_Q \sim R^{-1}$ . Inserting equation (2.103) into equation (2.102) yields the useful result

$$\int_0^R dr 4\pi r^2 \rho [g(r) - 1] \frac{\sin Qr}{Qr} = \frac{1}{4\pi Q^2} \left\{ \frac{8\pi^3}{V} \sum_{\hat{Q}' \neq 0} \Omega_{\hat{Q}'} S(Q') \left( \frac{Q}{Q'} \right) \Delta(Q, Q'; \sigma_Q) \right\} - 1. \quad (2.104)$$

In the thermodynamic limit, it is possible to simplify equation (2.104) using

$$\begin{aligned} \frac{8\pi^3}{V} \sum_{\hat{Q}' \neq 0} \Omega_{\hat{Q}'} &\rightarrow 4\pi \int dQ' Q'^2 \\ \Delta(Q, Q'; \sigma_Q) &\rightarrow \delta(Q - Q') \end{aligned} \quad (2.105)$$

to generate the standard result

$$\int_0^\infty dr 4\pi r^2 \rho [g(r) - 1] \frac{\sin Qr}{Qr} = S(Q) - 1 \quad Q > 0. \quad (2.106)$$

It is interesting to consider the approach of the prelimit form of the Fourier Sine Transformation of the radial distribution function, equation (2.104), to the thermodynamic limit. Using the Euler-Maclaurin summation techniques, it is clear that if the density of states is large around some  $Q$  and the prelimit  $\delta$ -function is strongly peaked at  $Q' = Q$ , then the continuum limit will be approximately valid. Thus, it can be expected that equation (2.104) will generate an accurate representation of  $S(Q)$  at large  $Q$  and fail to generate an accurate representation of  $S(Q)$  at small  $Q$ . In figure 2.9, Fourier Sine transformations of radial distribution functions are compared to a direct computation of  $S(Q)$  in finite systems of various size. Indeed, the agreement





where  $N_\alpha$  is the number of atoms of type  $\alpha$ . The partial radial distribution function is given by

$$g_{\alpha\beta}(r) = \left[ \frac{1}{4\pi\rho r^2 c_\alpha c_\beta} \right] \left[ \frac{1}{N} \left\langle \sum_{i_\alpha=1}^{N_\alpha} \sum_{j_\beta=1}^{N_\beta} \sum_{\mathbf{l} \neq \mathbf{0}; i_\alpha=j_\beta} \delta(|\mathbf{r}_{i_\alpha} - \mathbf{r}_{j_\beta} - \mathbf{hl}| - r) \right\rangle \right] \quad (2.108)$$

which approaches unity as  $r \rightarrow \infty$ . The notation  $\mathbf{r}_{i_\alpha}$  denotes the position of the  $i_\alpha^{\text{th}}$  atom of type  $\alpha$ . The associated partial structure factor is defined to be

$$c_\alpha c_\beta S_{\alpha\beta}(Q) \equiv \frac{1}{N\Omega_{\hat{Q}}} \left\langle \sum'_{\hat{Q}'} n_\alpha(\mathbf{Q}') n_\beta^*(\mathbf{Q}') \right\rangle, \quad (2.109)$$

where

$$n_\alpha(\mathbf{Q}) = \sum_{j_\alpha=1}^{N_\alpha} e^{i\mathbf{Q}\cdot\mathbf{r}_{j_\alpha}}, \quad (2.110)$$

so that

$$g_{\alpha\beta}(r) + \left[ \frac{\delta(r)}{2\pi r^2 \rho c_\alpha c_\beta} \right] \delta_{\alpha\beta} c_\alpha = \frac{1}{N} \sum_{\hat{Q}} \Omega_{\hat{Q}} S_{\alpha\beta}(Q) \frac{\sin Qr}{Qr}, \quad (2.111)$$

in a finite system. The quantity  $S_{\alpha\beta}(Q)$  is real due to the “time reversal symmetry” of the restricted sum. In the thermodynamic limit,

$$\int_0^\infty dr 4\pi r^2 \rho c_\alpha c_\beta [g_{\alpha\beta}(r) - 1] \frac{\sin Qr}{Qr} = c_\alpha c_\beta S_{\alpha\beta}(Q) - \delta_{\alpha\beta} c_\alpha \quad Q > 0, \quad (2.112)$$

as expected.

The total cross section in a neutron scattering experiment is

$$S_{\text{neutron}}(Q) = \frac{1}{N\Omega_{\hat{Q}}} \left\langle \sum'_{\hat{Q}'} |n_{\text{neutron}}(\mathbf{Q}')|^2 \right\rangle, \quad (2.113)$$

where

$$n_{\text{neutron}}(\mathbf{Q}) = \sum_j b_j e^{i\mathbf{Q}\cdot\mathbf{r}_j}. \quad (2.114)$$

is the (single particle) atomic density function weighted by the coherent neutron scattering of each atom  $b_j$  following the first Born approximation. The total cross-section can be decomposed into the partials

$$S_{\text{neutron}}(Q) = \sum_{\alpha\beta} c_\alpha c_\beta b_\alpha b_\beta S_{\alpha\beta}(Q). \quad (2.115)$$

Note,  $S_{\text{neutron}}(Q)$  approaches  $\sum_\alpha b_\alpha^2 c_\alpha$  as  $Q \rightarrow \infty$ . Therefore, the unit-less cross-

### 2.5.6 Probing water and its surrogates – Ratio for 4

Water has been subject to much scrutiny by both experiment and simulation. While molecular dynamics simulation models still struggle to reproduce all the micro-, and macroscopic features satisfactorily at the same time[67], advances have been made in the study of local structure based on well established models[25]. The order parameters proposed in[25] give a clearer picture of the quality of the coordination shell around a water molecule, albeit on the premise that the molecule indeed has four neighbours. My efforts are primarily concerned with supramolecular, structural motifs; I am particularly interested to see how the presence of a second type of molecule might influence the water structure. Since it is known that water forms extended networks of tetrahedral type four-fold coordination even in the liquid, finding the ratio of molecules coordinated in such a manner gives direct insight into the structure and its change.

Since often the tools that have a lot of clout afford us an *intuitive* insight, the quantity I propose to extract from simulation data is in effect and by design a simple one: it is the ratio of molecules which have four distinct (tetrahedral-style) neighbours within a cutoff,  $r_c$ , to the number of molecules in the system, thus

$$R_{\text{ffc}} = \frac{N_c}{N}. \quad (2.119)$$

Because life is not quite that simple, we need to discuss some details.

First, neighbours need to be distinct, which is achieved trivially. The method shown here, much like Debenedetti's order parameter [25], is based only on the position of the oxygen atoms in the system. A cutoff distance  $r_c$ , corresponding to the first minimum in the radial distribution function  $g(r_{\text{O-O}})$ , is imposed. Any angular information was deliberately not taken into consideration.

For pure water, this quantity is indeed very intuitive, since it relies solely on the H-bonding distance between water molecules and the resulting O-O separation.

In a system where water is one of  $n$  types of molecules, an interesting opportunity arises, since water will solvate the other ion or molecule. Potential H-bonds between water and molecule type  $x$  can be identified, and, resulting from these, atoms which can act as *surrogate* neighbours in a four-fold coordination (FFC) can be found and included in the calculation. For instance, in a system containing water and (let's randomly pick) the proline amino acid, the potential for H-bond formation between its polar groups and water molecules arises. Hence, the amide group's  $N_P$  can act as a *surrogate* FFC site.

Of course, unlike the 2D- $g(r)$  discussed in section 2.5.2 on page 42, this parameter is rather reducing the whole system's information content to one number. As a result, we need to vary a parameter of interest and do several simulations (or indeed use the

---

## Findings



---

In the following you shall find the fruit of my work and research. The ensuing chapters are based on the papers listed below.

R.Z. Troitzsch, G.J. Martyna, and J. Crain. Structure and hydration shell mobility in concentrated aqueous NaCl solutions. *J. Am. Chem. Soc.*, submitted, 2007.

R.Z. Troitzsch, G.J. Martyna, S.E. McLain, A.K. Soper, and J. Crain. Structure of aqueous proline via parallel tempering molecular dynamics and neutron diffraction. *J. Phys. Chem. B*, 111(28):8210–8222, 2007.

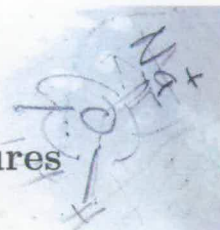
R.Z. Troitzsch, J. Crain, and G.J. Martyna. A simplified model of local structure in aqueous proline amino acid revealed by first principles molecular dynamics simulations. *J. Am. Chem. Soc.*, submitted, 2007.

R.Z. Troitzsch, H. Vass, W.J. Hossack, G.J. Martyna, and J. Crain. Molecular mechanisms of cryoprotection in aqueous proline: light scattering and molecular dynamics simulations. *J. Phys. Chem. B*, accepted, 2007.

R.Z. Troitzsch, G.J. Martyna, S. Thobhani, E. Cerasoli, G. Tranter, and J. Crain. Solution structure of an hiv-1 antibody epitope: Parallel tempering molecular dynamics and circular dichroism spectroscopy of peptide gp41<sub>659–671</sub>. *Biophys. J.*, submitted, 2007.

# Salt in solution

## Novel structural and dynamical features

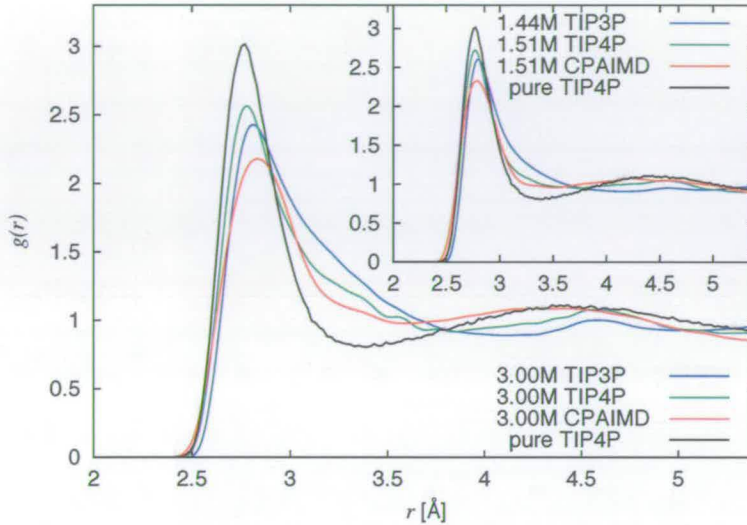


---

### 3.1 Motivation for studying salts

As alluded to in section 1.3, there remain open questions concerning aqueous saline solutions and the resulting effects and impact on biotic systems.

In principle, computer simulation is well placed to resolve these issues and, consequently, the structure of aqueous salts has received considerable attention. In recapitulation, atomistic simulation can reveal (1) the degree of water restructuring, (2) the hydration shells (including coordination number) around individual ions, (3) the nature of ion-ion pair correlations and (4) mechanisms for exchange of waters of ion solvation with the bulk. With the exception of the third point, the majority of prior investigations have focused on a single ion or ion pair in solution. However, ion-ion correlations are directly related to the potential of mean force between them and they are therefore an important structural property[59]. Previous reports have led to contradictory conclusions concerning the nature of this contact under ambient conditions. The majority of simulations show two minima in the anion-cation effective interaction potential, leading to the formation of close contact ion pairs and to more distant solvent-separated configurations[59, 84, 127] (see figure 3.1). However, there is very wide variation in the height and position of the first maximum in anion-cation radial distribution function. The origin of this variation may lie in sensitivity to interaction potentials, although a systematic exploration has not been performed. Moreover, slow ion diffusion and the presence of energy barriers between the close-contact and various solvent separated anion-cation structures leads to inefficient configurational sampling and the equilibrium between solvated and de-solvated ion pairs may be difficult to achieve in conventional MD simulations where the force fields used are empirical, typically involve non-polarisable treatments and thereby neglect



**Figure 3.2:** Radial distribution function for O-O contact. The hydrogen bond network of the pure water system is significantly perturbed in the solutions, as expected.

As expected, compared to the pure water distribution, the second maximum in the distribution is lowered and the density is shifted into the minimum. It has been suggested that the addition of salt to water introduces an effective pressure[64].

**Hydration** Let us next consider the local hydration structure around the  $\text{Na}^+$  ion represented in figure 3.3 by the joint distribution  $g(r_{\text{Na-O}}; \cos(\theta_{\text{Na-O-H}}))$ . This gives the distance between Na and O atoms versus the cosine of the angle formed by the triplet Na-O-H (O at vertex). As discussed in section 2.5.2, integration over the angular degrees of freedom recovers the properly normalised 1D distribution,  $g(r_{\text{Na-O}})$ , which is also shown on the 2D plot. This description allows the orientational contributions to the radial structure to be separated. It is evident that the second hydration shell of sodium (giving a peak in the radial function at around  $4.5 \text{ \AA}$ ) arises from two distinct orientational contributions. The primary one occurs as a diffuse feature in the range  $\cos(\theta) = -1 \dots 0.3$  and corresponds to the expected motif in which the oxygen atoms point toward the  $\text{Na}^+$  ion. The weaker feature near  $\cos(\theta) = 1$  is clearly separated from the primary one and corresponds to a concentration-dependent minority population of water molecules (also referred to as the “special guy”) in the second shell having a

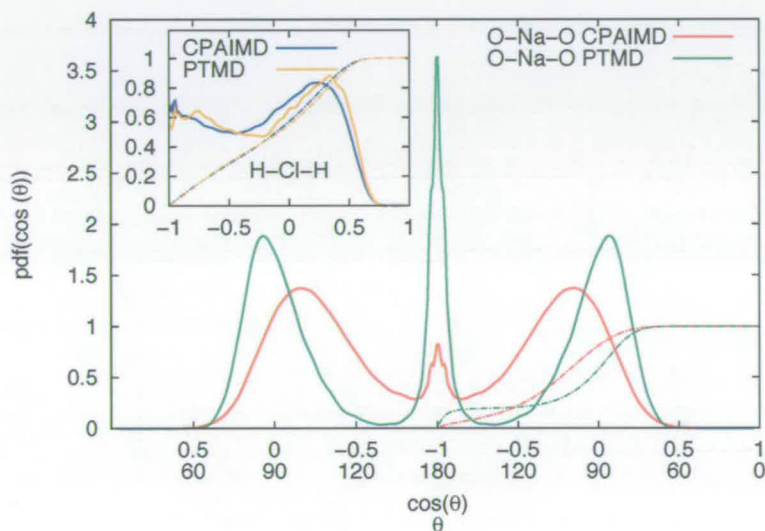
distinct orientational relationship relative to the ion they hydrate: according to the sign convention, the peak at  $\cos(\theta) \approx 1$  corresponds to a configuration in which the H atom of a water molecule points toward the Na atom. The *ab initio* calculations reveal the same feature in approximately the same position (see figure 3.3). Visual inspection of the MD trajectories clearly identifies the inverted orientation as a persistent motif in the hydration structure. Table 3.1 offers a full list of coordination numbers as obtained from the simulations and from the EPSR data, including, where applicable, a population number for the inverted feature, labelled Na-O<sub>W</sub><sup>if</sup>, which we extract by integration from the relevant area  $r = 3 \dots 5, \cos(\theta) = 0.7 \dots 1$ . The coordination number corresponding to this motif is lowest at the highest concentration, but it seems that each Na<sup>+</sup> has approximately one “special guy” on average.

distribution	Cl-H <sub>W</sub>		Na-O <sub>W</sub>		O <sub>W</sub> -O <sub>W</sub>		O <sub>W</sub> -H <sub>W</sub>		Na-Cl		Na-O <sub>W</sub> <sup>if</sup>	
conc.[M]	1.5	3.0	1.5	3.0	1.5	3.0	1.5	3.0	1.5	3.0	1.5	3.0
	coordination numbers											
$r_{\min}$ [Å]	2.86		3.13		3.37		2.38		3.56			
TIP3P	6.59	6.44	5.93	5.86	4.92	4.45	1.59	1.30	.002	.003	1.32	1.16
TIP4P	6.63	6.53	5.98	5.93	4.74	4.33	1.65	1.38	.003	.002	1.37	1.13
CP	5.54	5.52	4.61	4.51	4.54	4.01	1.66	1.42	.000	.013	1.31	1.14
EPSR	5.60	5.29	5.00	4.70	4.44	4.47	1.62	1.51	.483	.401		
CP <sub>∞</sub>	N/A		4.29		4.31		1.73		N/A		N/A	
	first peak position $r_{\max}$ [Å]											
TIP3P	2.42	2.42	2.29	2.28	2.83	2.81	1.87	1.86	5.16	5.19		
TIP4P	2.37	2.39	2.32	2.30	2.78	2.28	1.84	1.84	4.98	4.83		
CP	2.23	2.24	2.10	2.07	2.81	2.84	1.82	1.85	4.95	4.82		
EPSR	2.17	2.17	2.34	2.33	2.74	2.75	1.77	1.77	2.69	2.70		
CP <sub>∞</sub>	N/A		2.06		2.80		1.77		N/A		N/A	

**Table 3.1:** Coordination numbers of contacts in Na-Cl solution for different models used in MD, CPAIMD, and in comparison to EPSR results. Further listed is the re-computed result for the , infinitely diluted Na<sup>+</sup> system described in Ref. [117].

Evolution of the trajectories shows that water molecules in this orientation participate in exchange events between water molecules in the first and second hydration shells (see figure 3.4). The inverted orientation allows for hydrogen bonding between the members of first and second shells and subsequent exchange of positions. Solvent exchange between the first and second hydration shells around the Na<sup>+</sup> ion has also been suggested in first principles simulations of a single ion[117] and may be the origin of dynamical processes recently reported in the power spectral analysis of NaCl solutions[78]. Finally, the inverted water molecule forms an O-O contact at about 3.5 Å, which is not a normal water-water distance and the structures involved in the solvent exchange process possibly contribute to the well known blurring of the  $g(r_{\text{O}_W-\text{O}_W})$  on

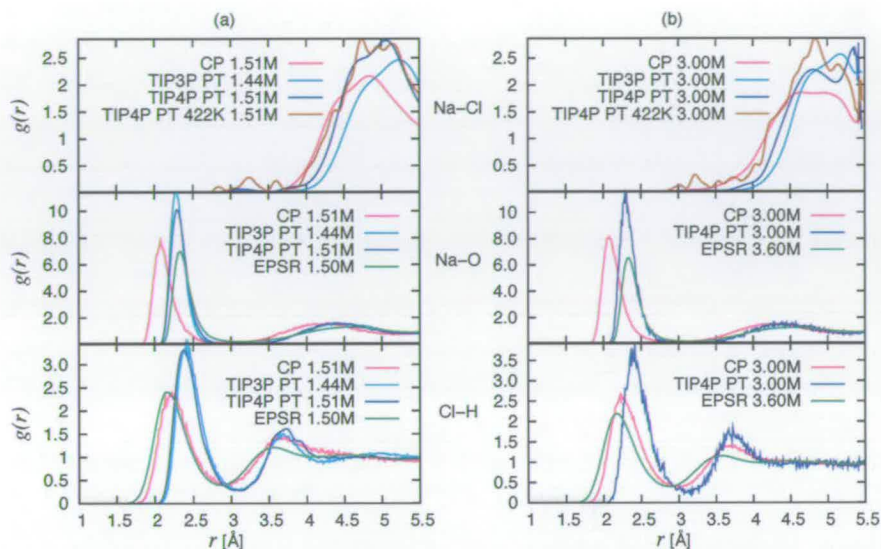
Fluctuations in the geometry of the coordination shells of the ions are examined through the probability distribution of the angles  $\angle O - Na^+ - O$  and  $\angle H - Cl^- - H$ , which are shown in figure 3.5.



**Figure 3.5:** Probability density of  $\cos(\theta_{O-Na^+-O})$  and  $\cos(\theta_{H-Cl^- - H})$ , with running coordination numbers (dashed lines) at 3.0 M for PTMD and CPAIMD.

In the case of  $Na^+$  two distinct shell geometries can be identified, (1) tetragonal bipyramidal and (2) regular tetrahedral, described by the angle pairs  $(180^\circ, 90^\circ)$  and  $(\sim 109^\circ, \sim 109^\circ)$  and with coordination numbers 6 and 4 respectively. The PTMD result has two well-defined peaks at  $180^\circ$  and  $90^\circ$ , indicating the bipyramidal geometry. Since the  $90^\circ$ -peak is asymmetric and much broader relative to the  $180^\circ$ -peak, this suggests a second, not clearly separated feature around  $109^\circ$ , given rise to by the tetrahedral geometry. Careful integration under the PDF yields a probability of being in the bipyramidal configuration of 90%, which cross-checks with the observed average coordination number of 5.8-5.9 very well. Under CPAIMD, peaks are less pronounced, and the population at  $180^\circ$  is clearly less compared with PTMD. Performing the integration as before yields a probability of being in the bipyramid of 30%, which suggests an average coordination number of 4.66, which is very close to the observed result. Interestingly, the *ab initio* result favours the tetrahedral hydration geometry, while the classical result is dominated by the tetragonal bipyramid. This dependence

panel (b) in figure 3.1 on page 60.

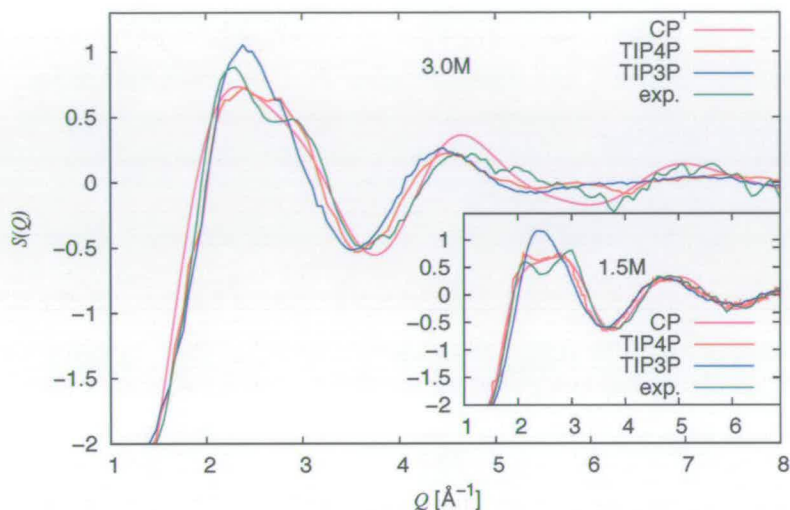


**Figure 3.6:** Comparison of  $g(r)$  between EPSR[64] results and simulation for (a) 1.5 M and (b) 3.0 M. In the top panels a comparison to the highest temperature in the PT simulation is also shown. All distributions labelled “PT” are based on the virtual move PT technique proposed in Ref. [23]

In previous work Lyubartsev and Laaksonen Ref. [59] observed that the formation of the close-contact ion pair was a sensitive function of concentration. At 0.9 M approximately 10% of the ions are in close contact. This rises to about 50% in 4 M solutions and decreases to about 4% at 0.5 M. These authors also comment on the instability of the NaCl radial distribution function in their simulations which they attribute to high energy barrier (several  $k_B T$ ) between solvated and de-solvated pairs. This situation is significantly improved in the PTMD simulations at hand, where sampling efficiency is greater. The results also differ from the recent simulations by Uchida and Matsuoka[113] who report coexistence of both close contact and solvent separated species.

While the detection that the contact ion pair is not a significant structural feature of the solution is in conflict with several previous simulations, it is consistent with experimental measurements of the dielectric spectra in the microwave region[13, 127]. Here, the expectation is that, because the dipole moment of the contact ion pair is approximately three times that of a water molecule, there should be an appreciable signature at low frequencies. This is not observed and to this author’s knowledge no direct experimental evidence exists for the formation of such pairs under ambient conditions.





**Figure 3.7:** Comparison of partial structure factors as obtained experimentally[64] and those drawn from simulations. Shown here is the scattering off all non-hydrogenous species,  $S_{XX}$ . Agreement between experiment and simulation is particularly notable for the TIP4P and CPAIMD calculations.

According to the prescription given in section 2.5.4,  $S_{XX}(Q)$  is computed directly from MD trajectories. At the concentrations of interest here,  $S_{XX}(Q)$  is dominated by water-water correlations ( $\approx 90\%$ ) near 1.5 M and is therefore an effective measure of solvent restructuring that does not involve invoking inverse procedures such as the empirical potential structure refinement scheme (EPSR). In figure 3.7, the experimental data and the results of CPAIMD and classical PTMD simulations using the TIP3P and TIP4P empirical potentials are given. It is clear that the TIP3P system does not capture the main features of the experimental data. The width and structure of the first peak is not reproduced and the position of the second main peak ( $Q \approx 4.5 \text{ \AA}^{-1}$ ) and first minimum ( $Q \approx 3.7 \text{ \AA}^{-1}$ ) is underestimated. In contrast, the CPAIMD and TIP4P-PTMD systems perform significantly better. However, only the CPAIMD system shows a clearly resolved shoulder on the first peak. This shoulder becomes more prominent at lower salinity in reasonable agreement with the experimental  $S(Q)$  (see inset). The CPAIMD system also gives the first minimum and second peak maximum position in best agreement with the measured  $S_{XX}(Q)$  but the height is overestimated

SHAKE[87], RATTLE[3] and ROLL[69] were used in combination with multiple time-step integrators, constraint algorithms and a 15.999 a.u. hydrogen mass, to maintain an average 5.0 fs “outer” time-step[68, 69, 70]. Periodic boundary conditions were assumed and Ewald summation was employed to compute long range interactions, enabled by the Smooth Particle Mesh[32] approximation to the reciprocal space part.

### 3.5.2 Parallel tempering

Systems were replicated  $n_T$ -fold for parallel tempering in the canonical ensemble and distributed equally across the temperature range. For  $n_T$  and the relevant ranges, see table 3.2. The probability of acceptance was defined as

$$P = \exp(\Delta\beta\Delta E)/(1 + \exp(\Delta\beta\Delta E)) \quad (3.1)$$

where  $\Delta\beta$  and  $\Delta E$  are the differences of the inverse temperatures and energies of the replicas, respectively. Data were collected over a period of 3.5 ns, with an attempt to switch configurations every 1 ps. Statistics were collected on the traversal of the temperers through temperature space and used to ensure adequate sampling, as well as the rate of switch-acceptance, which was found to average at  $\sim 50\%$  across all temperatures. Finally, the statistics required for the virtual move technique proposed by Colluzza and Frenkel[23] were collected and used in the ensuing analysis.

### 3.5.3 *Ab initio* model

CPAIMD simulations of aqueous NaCl using norm conserving pseudopotentials[4] and a plane wave basis set with the energy cutoff set to 70 Ry were carried out. The gradient-corrected B-LYP functional for exchange and correlation is used. Car-Parrinello *ab initio* molecular dynamics was performed in the canonical ensemble with a 0.125 fs

molarity [M]	water model	$n_{\text{Na}^+, \text{Cl}^-}$	$n_{\text{H}_2\text{O}}$	$n_T$	$T_{\text{low}}$ [K]	$T_{\text{high}}$ [K]	$\rho$ [kg m <sup>-3</sup> ]	$t_{\text{tot}}$ [ns]
1.44	TIP3P	48	1840	32	275	423	1077	0.6
1.51	TIP4P	3	110	44	281	398	1083	1.0
1.51	CP	3	110	1	300	300	1083	0.08
3.00	TIP3P	4	66	44	281	398	1066	20
3.00	TIP4P	4	66	44	281	398	1066	0.4
3.00	CP	4	66	1	300	300	1066	0.185
$\infty$	CP	1 Na <sup>+</sup>	53	1	300	300	1000	0.015

**Table 3.2:** Outline of the systems’ parameters.  $n_{\text{Na}^+, \text{Cl}^-}$ : number of salt ions each,  $n_{\text{H}_2\text{O}}$ : number of water molecules,  $n_T$ : number of parallel temperers,  $t_{\text{tot}}$ : total run-time of simulation.

#### 3.5.4 Empirical Potential Structure Refinement (EPSR)

Empirical Potential Structure Refinement (EPSR)[94] of neutron diffraction data has been established as a powerful data analysis tool for extracting 3D structural information on disordered matter. The basic technique relies on construction of a plausible 3D model of a liquid using a reference potential that removes unphysical contacts and provides a plausible starting configuration. The experimental diffraction data (composite partial structure factors) are then introduced as a constraint and the 3D structure is iteratively updated until an ensemble of structures are produced which are consistent with the available diffraction data.

# Proline

## Structural motifs in aqueous solution

---



### 4.1 Motivation to study the amino acid proline

In several respects, proline is unique among the twenty amino acids. In high concentration aqueous solutions (over 3.5 M) it exhibits hydrotropism – the property which allows it to increase the solubility of hydrophobic compounds in water[86]. Linked to this is the assumption that proline may act as a non-complex chemical pharmacological chaperone, assisting the process of protein folding[17, 21]. Both *in vivo* and *in vitro* experiments have shown that proline disfavours protein aggregation and misfolding, which cause conditions such as Alzheimer’s disease[85, 89]. Perhaps most intriguing, and linked to the above chaperone characteristics, is the expression of proline in plants and other organisms (bacteria, protozoa, frogs) under inauspicious conditions, like desiccation or extreme temperatures, thus acting as a bio-protectant[42, 55, 77, 80, 83, 125]. (This particular feature is explored further in chapter 5).

Structurally, proline is the only cyclic amino acid in which the side chain is covalently linked to the backbone amine forming a pyrrolidine ring<sup>1</sup>. This has a highly restrictive impact on the conformational freedom, resulting in the prevalence of proline in the bend regions of polypeptides and proteins.

Further, among the amino acids proline exhibits the highest solubility in water (up to 6.5 M). These solutions are signified by unusual colligative properties[91, 92]; transport properties are also affected, and it has been shown that solutions of > 3.5 M acquire viscosities that are unexpectedly large for such low molecular weight species[18].

Proline’s unusual properties have led to the suggestion that the mixtures may form

---

<sup>1</sup>This difference further means that formally proline is considered an *imino* acid, since its amino group is of secondary rather than primary nature, but that shall not keep us from calling it an amino acid throughout.

supported by evidence from protein structural studies which suggest that proline residues frequently occur on the *exterior* of proteins, in contact with solvent[74].

## 4.2 How this work explores proline

To address some of the above contradictions, in this section is presented work which has been partly published in Ref. [104] and partly submitted for publication in Ref. [102].

The solution state structure and properties of aqueous L-proline are examined using classical parallel tempering molecular dynamics simulations (PTMD) of large systems (6860/343 water/proline molecules respectively) to ensure good sampling, while comparisons to solution phase neutron diffraction experiments and x-ray crystal data are employed to probe model accuracy. The x-ray crystal (of two molecular crystals, L-proline and L-proline monohydrate) data is compared to Parrinello-Rahman type constant pressure MD simulations. The comparison between the neutron diffraction data and simulation is made directly through the static structure factor,  $S(Q)$ . The three empirical water models TIP3P, TIP4P and SPC/E are used and the effects of nuclear delocalisation on  $S(Q)$  are included approximately through path integral molecular dynamics simulations on gas phase molecules. The objective is to draw firm conclusions about the local and mesoscale structure of aqueous proline mixtures by combining simulation and diffraction data, to explore the sensitivity of those conclusions to the empirical water potential and the force field, and to assess the effect of quantum corrections on the liquid structure.

Further, preempting the shortfalls of the empirical model calculations of the liquid phase, Car-Parrinello *ab initio* simulations (CPAIMD) are performed on two different system sizes. This makes it possible to assess the influence of flexibility and polarisability on the structural properties of this important system.

In order to present clearly the results of the studies, it is useful to adopt a naming convention for the atoms in the proline molecule; the one used in this thesis is given in figure 4.1.

## 4.3 Simulation Results and Comparison to Experiment

In order to develop a complete and convincing picture of proline-water solutions via experiment-simulation comparisons, a variety of results is presented. First, the force field models are validated by comparison to structural data on molecular crystals. Second, in selected cases, parallel tempering MD (PTMD) results are employed to validate the sampling of the force field at room temperature in the more challenging liquid state. Next, the validity of the force field is stringently probed by detailed

prominent. The amine H appear to make no water O contacts. Further, in proline-proline contact  $H_C \cdots O_C$  bonding is more prominent than the ring-ring contact, or hydrophobic association; the prominence of each of these in the context of aqueous proline shall be discussed later.

In contrast, the CHARMM22 model does not yield L-proline crystallographic parameters in satisfactory agreement with experiment (see table 4.2). The force field model predicts a structure that is much too expanded, particularly along the a and b crystal axes. This indicates that the non-hydrogen bonded structural motifs involved in crystal packing given in figure 4.3, are not properly described. It is therefore likely that the CHARMM22 model will at least slightly underestimate proline-proline ring association.

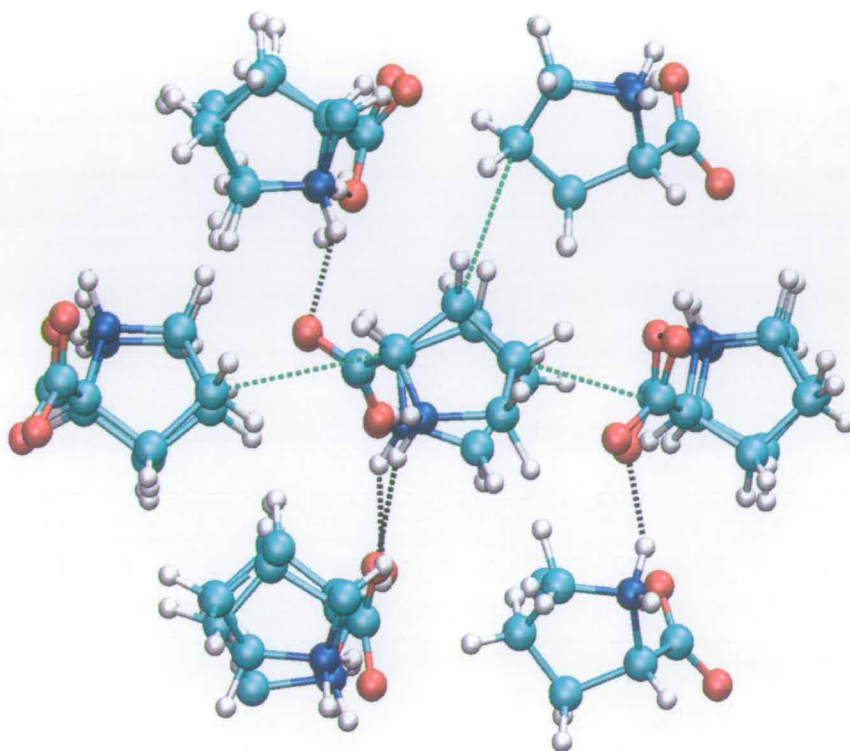
### 4.3.2 Effect of quantum corrections on $S(Q)$

The total  $S(Q)$  for a fully hydrogenous system is shown in figure 4.4, using the TIP4P model as an example and with quantum corrections applied as described in section 4.5.4. The effects of the corrections are visible across the entire  $Q$  range but are most prominent in the high  $Q$  region, where the intramolecular contributions dominate. In general, the peaks in  $S(Q)$  are lowered. As will be shown later, this behaviour tends to bring the quantum-corrected structure factor into closer agreement with the experimental diffraction data. While only the TIP4P data is shown here, the results are representative of all the models and the effects are largest in the fully hydrogenous samples, as expected. The effect of the quantum correction on the intramolecular structure factors is shown in figure 4.4, insets (a) and (b).

The assumption intrinsic to the quantum correction is that intramolecular liquid and gas phase structures are similar. This is tested by extracting an average intramolecular liquid state structure factor from the MD simulations of the aqueous solution and comparing it to the corresponding gas-phase structure factor. The results of the comparison are shown for proline in figure 4.4, inset (c), where it is evident that there is negligible perturbation of the gas phase structure in solution. Therefore, the subtraction and substitution procedure outlined in section 4.5.4 on page 103 is indeed justified.

	a	b	c	$\alpha$	$\beta$	$\gamma$
experiment	20.4	6.2	5.1	90.0	95.8	90.0
SPC/E	20.8	6.4	5.2	90.0	91.0	89.7
TIP3P	20.8	6.4	5.2	89.7	89.6	90.3
TIP4P	20.8	6.4	5.2	89.8	89.6	90.3

**Table 4.1:** Unit cell lengths in Å and angles in degrees, for the monohydrate crystal

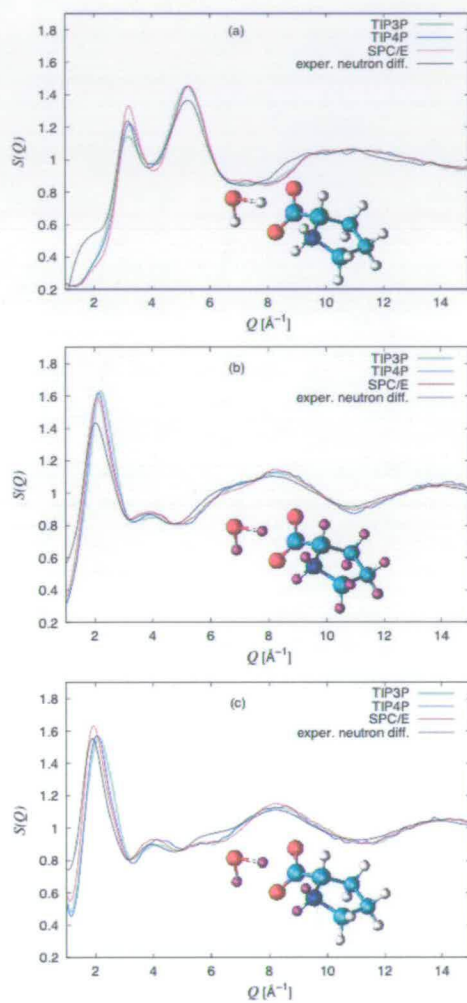


**Figure 4.3:** The backbone contacts in the L-proline crystal, visible as stacking of the pyrrolidine rings. The indicated distances (green) are  $\approx 3.8 \text{ \AA}$ .

dominant contributions to the total structure factor. The neat liquid water data is given in figure 4.6 for comparison.

First, based on the discussion in the preceding section, in all cases, the agreement between simulated and experimental  $S(Q)$  is improved upon application of the quantum correction procedure. There are, however, a number of notable differences that remain: First, in the fully hydrogenated sample (figure 4.5(a)) the position of the first peak is well described by all the models but the peak height is best accounted for by the TIP4P solvated system. Hydration by TIP3P and SPC/E water leads to slight over- and under-structuring, respectively. The height of the second peak at  $Q \approx 5.5 \text{ \AA}^{-1}$  is slightly overestimated, albeit by the same amount for all the models, relative to the diffraction data, but the peak position is in uniformly good agreement.

Second, in the fully deuterated sample (figure 4.5(b)) the first peak in  $S(Q)$  is observed to be distinctly higher in all the simulated systems than is observed in the experimental data. The other features in  $S(Q)$  appear to be reasonably well described



**Figure 4.5:** Three isotopic substitutions of the proline–water mixture. (a) fully hydrogenated samples, (b) fully deuterated samples, (c) deuteration only on proline amine hydrogen and water sites. The adjacent diagrams indicate these substitutions graphically, where purple and white spheres correspond to deuterium and hydrogen respectively.

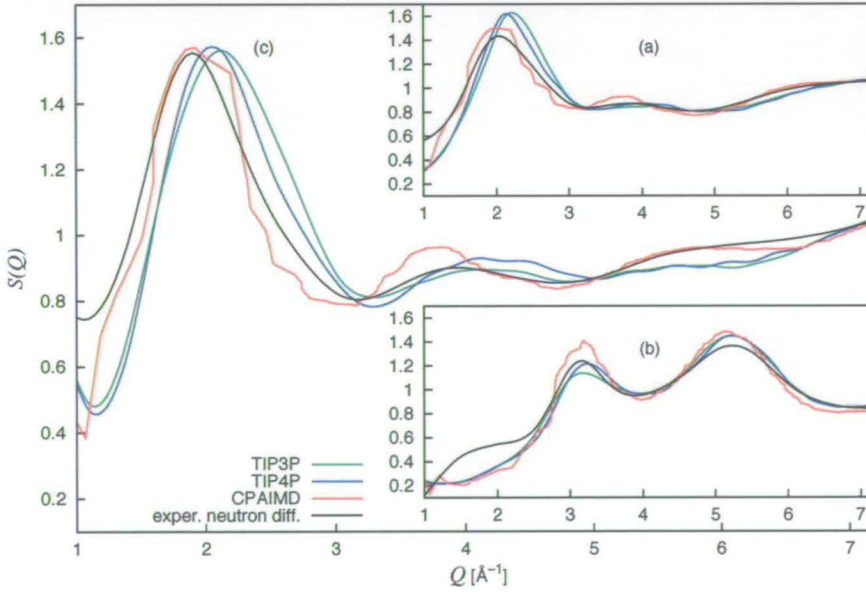


Isolation of  $S(Q)$  contributions

Subtraction of total  $S(Q)$ s with different isotopic substitutions from each other can reveal contributions to the scattering data, arising from particular atom types (e.g. particular partials). Let us define

$$\Delta S(Q)_{\text{red}} = \frac{2N}{\bar{b}_A^2 + \bar{b}_B^2} (\bar{b}_A^2 S_{\text{red}}^{(A)}(Q) - \bar{b}_B^2 S_{\text{red}}^{(B)}(Q)), \quad (4.1)$$

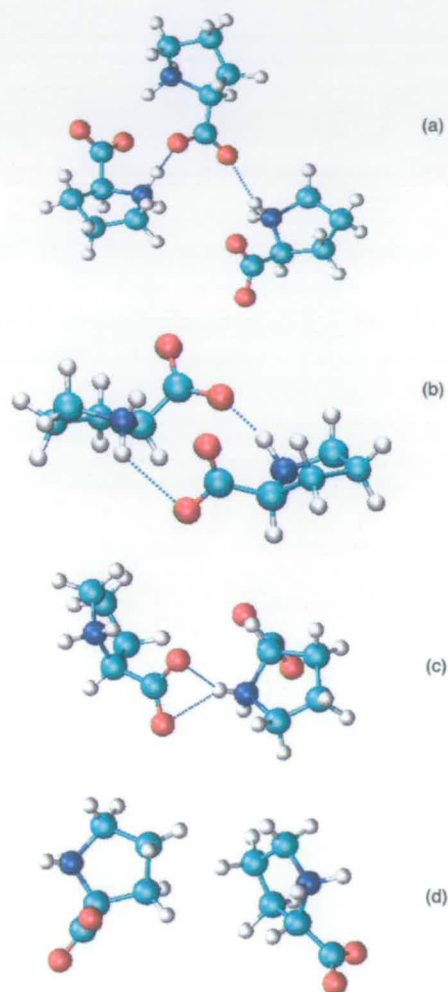
where the weights,  $\bar{b}_{A/B}^2 = \sum_i^N b_{A/B,i}^2 / N$  are required because unitless  $S(Q)$  are plotted as discussed in section 2.5.4. Figure 4.6(c) emphasises the polar contacts, obtained by subtracting figure 4.5(c) from (a) and correcting for water contributions by subtracting the difference of figure 4.6(a) and (b) with appropriate weight. Figure 4.6(d) shows the backbone interactions, obtained by subtraction of figure 4.5(c) from (b). The discrepancies between experiment and simulation are evident in the region  $Q < 2.5 \text{ \AA}^{-1}$ , and absent outside this region. This indicates that any misrepresentations originate from intermediary scale structures.



**Figure 4.7:** Static structure factor,  $S(Q)$ , as obtained from the trajectories for classical and *ab initio* calculations, as well as from neutron experiments[72]. Panels show (a) fully hydrogenous, (b) fully deuterated system and (c) the system fully deuterated, apart from the proline backbone.

water[116].

Finally, close approaches between the pyrrolidine ring segments of the proline molecule occur without facilitating hydrogen bonding interactions. These are considered to be primarily hydrophobically driven contacts (figure 4.8(d)).



**Figure 4.8:** Short range associations of proline: (a) (doubly bonded) dimer, (b) oligomeric chain segment, (c) bifurcated H-bond, (d) (hydrophobic) backbone contact

In order to quantify the importance of the H-bonded and hydrophobic contact motifs above, radial and joint distribution functions, designed such that the major features specific to these motifs can be assigned easily, have been computed. The construction

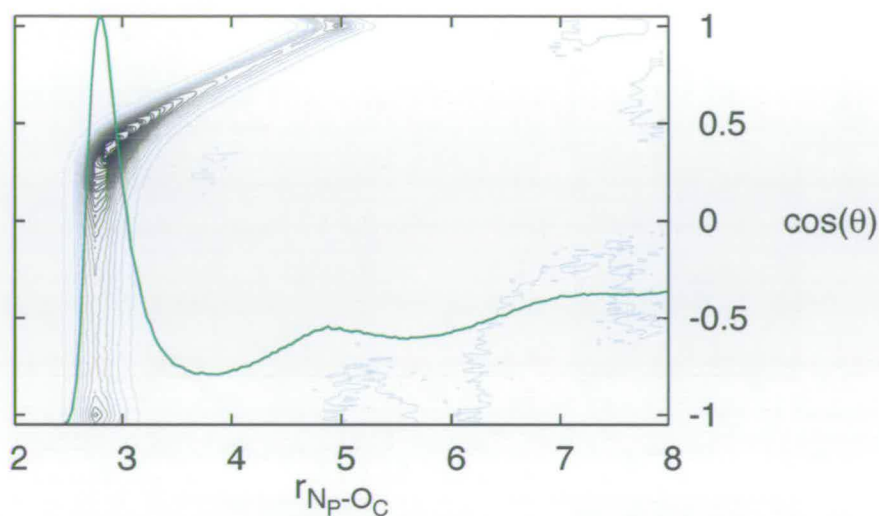


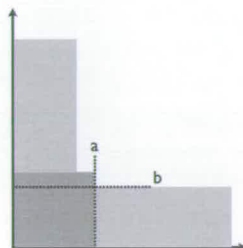
Figure 4.9: 1D and 2D distributions for  $r_{N_P-O_C}$  and  $\cos(\theta_{N_P-O_C-O_C})$  indicating the features in question.

### H-bond bifurcation

Let us further consider the H-bond patterns. A high level of bifurcated H-bonds formed with the amine hydrogens,  $H_C$ , is observed in the classical trajectory snapshots. The relative energetics of conventional two-centre and the rarer three-centre hydrogen bonds remain a matter of debate[124, 123]. In principle, two bifurcated bond types are possible: type I involves a single donor and two acceptors ( $O \cdots H \cdots O$ ) and type II involves a single acceptor shared by two donors ( $H \cdots O \cdots H$ ). In order to explore bifurcation in the solution, the corresponding areas in the 2D distributions (figure 4.10) are identified. In panels (a,d), the large classical system is explored and shows clear peakage in the bifurcation sensitive areas. The results for a smaller system size and shorter time classical simulation are shown in figure 4.10(b,e), where similar evidence for bifurcation is observed. This is in stark contrast to the *ab initio* calculations in panels (c,f), where the corresponding area is devoid of signal, indicating complete absence of bifurcated bonding. Coordination numbers drawn from the relevant distributions are given in table 4.5 for panels (a,c,d,f). They are calculated as given in figure 4.11, by integration under the shaded areas bounded as indicated. From this analysis, classically  $H_{C3}$  appears to predominantly donate bifurcated H-bonds, while  $H_{C2}$  displays no preference. A similar study, not depicted, shows that type II bifurcation does not form between prolines, neither classically nor from first principles simulation.

	TIP4P	SPC/E	CPAIMD	EPSR[72]	$a$	$b$
H <sub>C2</sub> bifurcated	0.01	0.01	0.00		2.4	2.1
H <sub>C2</sub>	0.03	0.03	0.06		2.4	2.1
H <sub>C3</sub> bifurcated	0.06	0.01	0.01		2.4	2.1
H <sub>C3</sub>	0.05	0.04	0.11		2.4	2.1
H <sub>W</sub> bifurcated	0.42	0.50	0.03		2.4	2.1
H <sub>W</sub>	2.29	2.28	1.53		2.4	2.1
total H <sub>W</sub>	2.71	2.78	1.56	1.67	2.4	
total bifurcated	0.49	0.52	0.04			
total non-bifur.	2.37	2.35	1.80			
total	2.86	2.86	1.84	1.96		

**Table 4.5:** Coordination numbers for bifurcated and conventional H-bonds in proline-proline and proline-water contacts at the carboxyl group. Figures 4.10, 4.16 show the corresponding distributions.  $a$  and  $b$  (in Å) delimit the symmetric shaded areas in figure 4.11 used for integration.

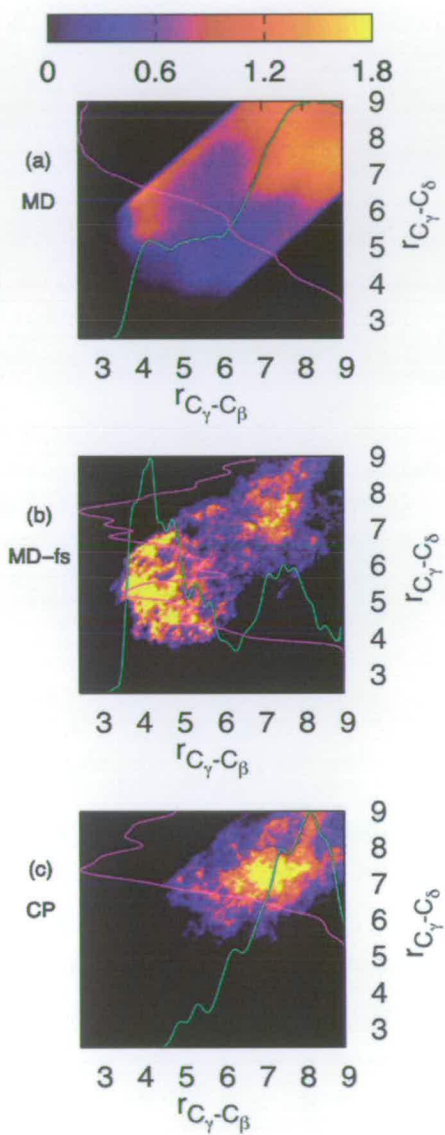


**Figure 4.11:** Areas of bifurcated and conventional H-bonds; symmetric around  $x = y$  and marked by dark and light shades respectively.  $a$  and  $b$  represent the edges for integration used in table 4.5.

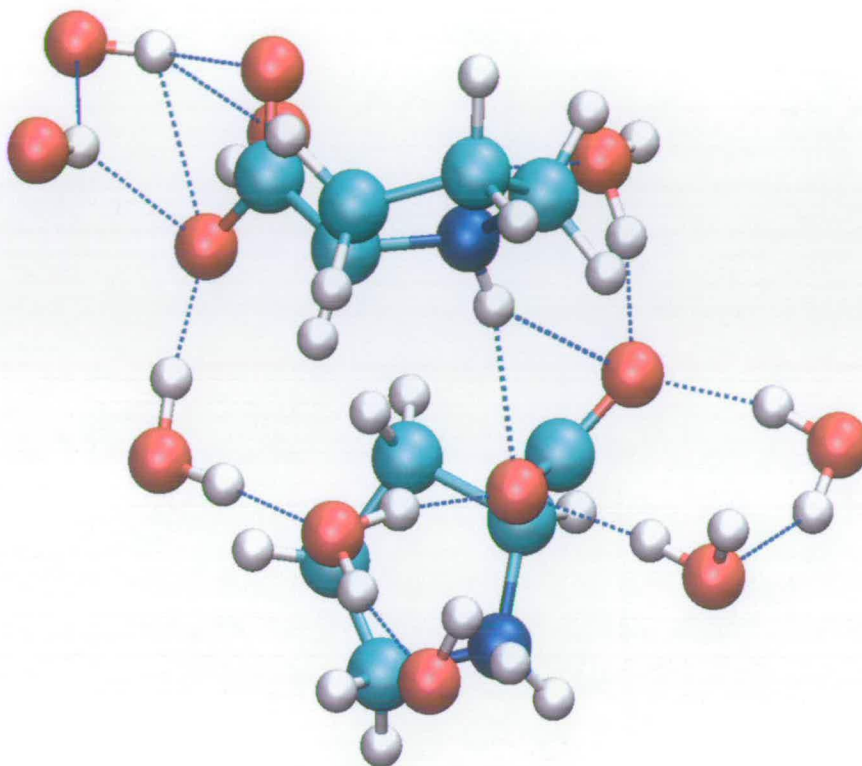
### Hydrophobic contacts

Defining a coordination number to describe the inter-pyrrolidine ring contact is not as simple as defining a coordination number to describe H-bonding. However, from examination of the molecular crystals, the  $C_\gamma$ - $C_\gamma$  distance gives a good description of the backbone inter-proline contact. That is, the  $C_\gamma$  carbon is furthest from the polar groups and therefore taken as the centre of the hydrophobic end (see figure 4.1). Figure 4.13(top) shows the  $g(r_{C_\gamma-C_\gamma})$  radial distribution function. Clearly, the extension of the plateau down to  $r = 3.8$  Å indicates the presence of a hydrophobic contact, since at this distance no water molecules can act as bridges between the rings.

In an effort to pin down orientation of the rings to one another, the joint distribution of the distances of  $C_\gamma$  to the carbons either side ( $C_\beta$ ,  $C_\delta$ ), is explored in figure 4.13(a-c). The corresponding 1D- $g(r)$  are overlaid. From visual inspection, the peakage corresponding to the hydrophobic contact is found to occur chiefly in the region between  $3.8 - 4.8$  Å on the  $x$ -axis and  $5.9 - 6.8$  Å on the  $y$ -axis in the classical cases. For the



**Figure 4.13:** 1D and 2D distributions for the distances between  $C_\gamma$  and the atoms located either side in the backbone ring of proline,  $C_\delta$  and  $C_\beta$ ; corresponding to a close range, dry contact of the backbone rings (figure 4.8 (d)). The corresponding 1D distributions are overlaid.

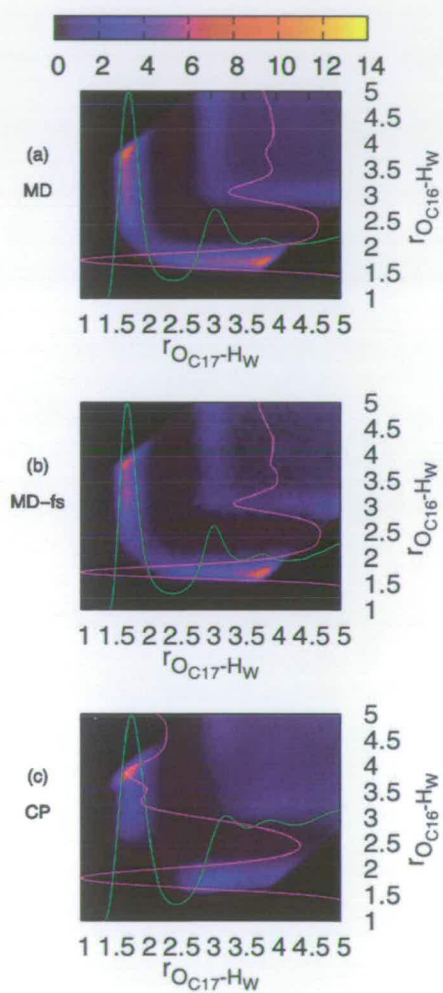


**Figure 4.14:** Network of proline-proline association, using water molecules as bridges. Of note is the bifurcated H-bond.

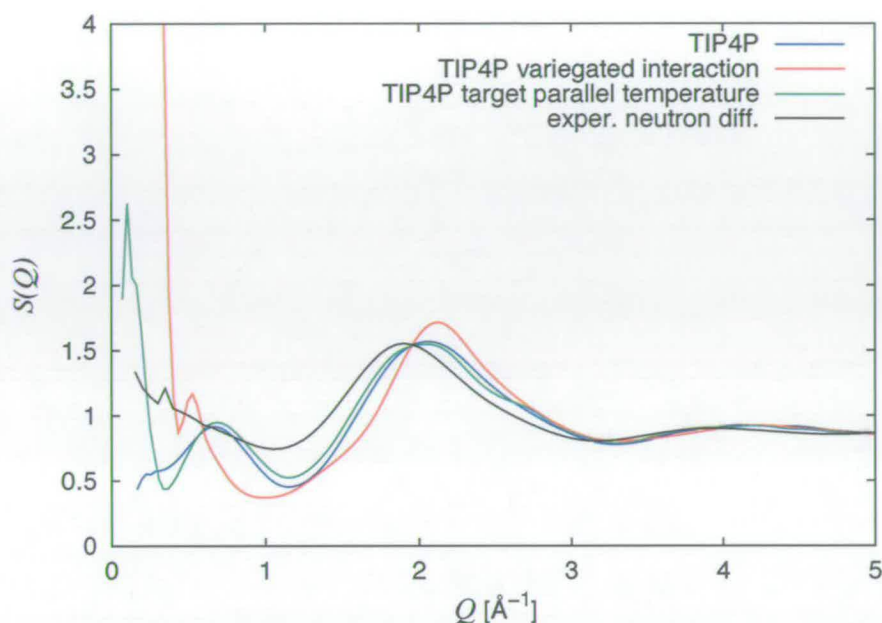
water network. This network is evidently most loosely associated in the TIP3P solvent, relative to either TIP4P or SPC/E. The energy penalty for hydrating the proline H-bonded dimer is, therefore, expected to be lowest in TIP3P water and the population of such dimers somewhat higher, as is indeed observed.

In the SPC/E and TIP4P water model mixture, the presence of the proline even at a concentration of 1:20, has a clear effect on  $g(r_{O_W-O_W})$ . Specifically, the second shell is contracted to shorter distance and the separation between the first and second shell peaks becomes less well defined.

The coordination numbers in table 4.7 indicate that both the simulations and the experimental inversion (EPSR) are close to one another. In all cases the coordination numbers from EPSR fits to the neutron data are higher than those from any of the models considered here. While TIP3P is closest to EPSR, the TIP3P  $g(r_{O_W-O_W})$  is significantly understructured, and overall provides a poor description of the water structure. A further point of note is that the SPC/E MD system is under-coordinated



**Figure 4.16:** 1D and 2D distributions indicating the distances of  $H_W$  amine hydrogen to carboxyl oxygens, labelled  $O_{C16}$  and  $O_{C17}$ . The main peak indicates a strong level of bifurcation.



**Figure 4.17:**  $S(Q)$  data corresponding to the fully deuterated isotopic substitution (c) in figure 4.5. Shown are the single temperature TIP4P simulation result, as well as the graph obtained from the target (ambient) temperature of the parallel tempering.

Indeed, the low  $Q$  neutron diffraction data provides a strong argument against such structures being present. However, the SANS measurements are sensitive to assemblies in which the characteristic lengthscales are at least  $30 \text{ \AA}$  in size. Smaller or very diffuse aggregates may still be present. Very recent optical measurements of a high Landau-Placzek ratio in proline solutions support a picture of strong concentration fluctuations[106]. At present, a more complete model of proline, perhaps one which takes into account polarisation effects, may be necessary to fully resolve the nature of any long-range associations.

## 4.4 Conclusion

An extensive series of classical and *ab initio* MD simulations has been performed to investigate the local and mesoscale structure of aqueous L-proline amino acid which has been the subject of considerable attention as well as conflicting reports. The simulated structures (obtained using several different empirical potentials for water) capture the main features of the experimentally-determined static structure factor,  $S(Q)$ , over most of the  $Q$  range of the measurement and for all three isotopic



(H/D) compositions investigated. Quantum corrections improve slightly the agreement between classical simulation and experiment, and CPAIMD results yield the best fit of  $S(Q)$  to experimental EPSR data. The resulting local structures comprise populations of well-defined motifs assignable as hydrogen-bonded dimers and chains, a proportion of which show H-bond bifurcation. Also present are hydrophobically associated assemblies in contact via the non-polar backbone. The relative populations of these show some sensitivity to both the empirical model for the water solvent and *ab initio* calculations. Longer-range correlations comprising stacked pyrrolidine rings that have long been implicated as the origin of the anomalous solution state properties of aqueous proline do not form spontaneously in the simulations starting from well-mixed configurations. Moreover, supramolecular assemblies of the proposed type can be created artificially by tuning the potential parameters to increase the hydrophobicity of the pyrrolidine rings. These motifs are annealed away when equilibrated under the correct CHARMM22 parameterisation and by PTMD simulation. Our MD simulations cannot rule out the presence of longer range correlations in the form of diffuse or transient aggregates. However existing SANS data shows no evidence of such structures.

## 4.5 Computational Models and Methods

### 4.5.1 Classical empirical force field model

Classical molecular dynamics (MD) simulations were performed to examine the properties of 2.75 M aqueous proline solution, a 20:1 mixture, at  $T = 300$  K using the CHARMM22 force field[60] and three well-known empirical potentials for water (TIP3P, SPC/E and TIP4P). These water potentials are all rigid (fixed geometry) and non-polarisable. TIP3P and SPC/E are both three-site potentials with charges and Lennard-Jones parameters assigned to each of the atom positions. TIP4P is a four-site model in which an additional site associated with the oxygen charge is displaced along the bisector of the HOH angle. It should be noted that the CHARMM22 force field is designed for use with TIP3P water model but it is nonetheless informative to explore the effect of changing the water model.

A simulation cell containing 6860 and 343 water and proline molecules, respectively, was prepared by placing the molecules on a cubic lattice with randomised orientations. The system was then equilibrated in the canonical ensemble, for 300 ps at a temperature of  $T = 300$  K in order to anneal out any unphysical high-energy contacts. At this stage the configuration was copied three-fold, to create initial conditions for independent simulations involving each of the three water models, TIP3P, TIP4P and SPC/E. After another 60 ps relaxation of each of the systems, production runs of 2 ns were

#### 4.5.4 Nuclear quantum effects on the static structure factor: gas phase path integral simulations


Compounds containing light atoms are known to exhibit quantum effects that impact their structural properties[56], and hence the structure factor. Therefore, a simple, approximate method of incorporating intramolecular quantum effects into  $S(Q)$  obtained from classical MD has been developed. First, classical gas phase simulations of proline and water are performed and the classical molecular structure factors  $S^{(cl,proline)}(Q)$  and  $S^{(cl,water)}(Q)$  generated. Assuming the gas phase structure is not significantly perturbed in the liquid, the classical [intra]molecular structure factors can be subtracted from the total  $S(Q)$  and molecular structure factors, obtained from gas phase simulations in which nuclear delocalisation is included using imaginary-time path integration[34],  $S^{(qm,proline)}(Q)$  and  $S^{(qm,water)}(Q)$ , can be added. Specifically, path integral molecular dynamics (PIMD)[108] computations with 64 pseudoparticles have been performed. The 1 molecule systems were equilibrated for 20 ps using PIMD. The time-step was set to 0.25 fs and data were collected during 200 ps of production for each isotopic composition. The quantum-corrected structure factor is defined by

$$\begin{aligned}
 S(Q)_{(qm,total)} = & S^{cl,total}(Q) \\
 & + c_{proline}[\tilde{S}^{(qm,proline)}(Q) - \tilde{S}^{(cl,proline)}(Q)] \\
 & + c_{water}[\tilde{S}^{(qm,water)}(Q) - \tilde{S}^{(cl,water)}(Q)]
 \end{aligned} \quad (4.2)$$

where  $c$  is the concentration. Further discussion is provided in section 2.5.4, along with a definition of  $\tilde{S}(Q)$ .

# Natural antifreeze!

## Ought Proline to be in your windscreen washer?



---

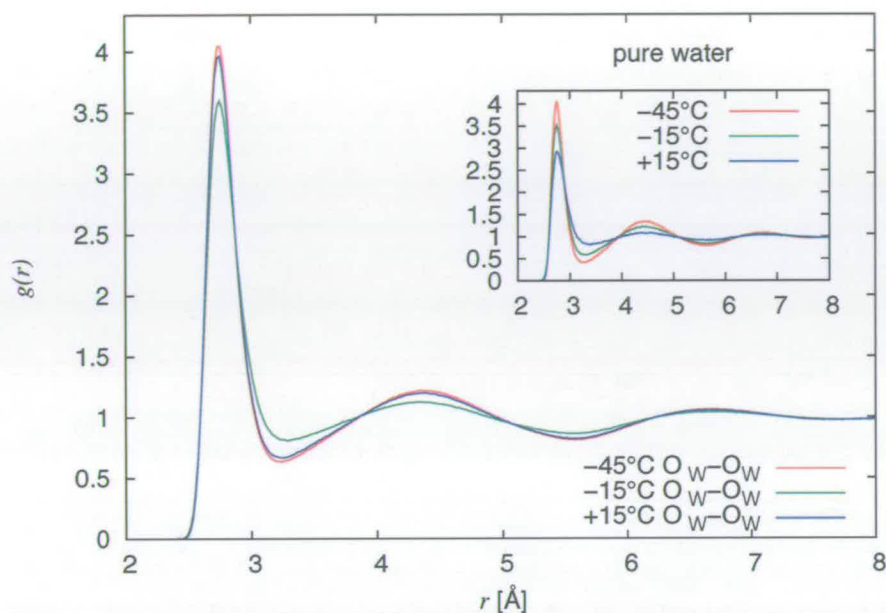
### 5.1 Motivation

As mentioned in section 4.1, proline is known to exhibit natural cryoprotectant properties. It is further commonly known that salts are used to de-ice roads in winter (a cause of sweat to many an environmentalist) and it is also relatively well known that saline water freezes at lower temperature, so I feel salt needs no extra justification or motivation in this context.

For proline, it has been known for several decades that accumulation of free proline occurs in plants upon exposure to low temperature stress[42]. In numerous cases, a direct relationship has been established between the accumulation of proline during cold acclimation and freezing tolerance[55, 83]. It is also known that certain transgenic plants having enhanced proline levels show higher freezing tolerance than do the corresponding wild types[80]. In this sense, proline is considered a common natural cryoprotectant (osmolyte) and its expression under adverse conditions has also been observed in bacteria, invertebrates, protozoa and algae[125].

So far, considerable attention has been focused on the biochemical and signal transduction pathways governing proline level regulation under low-temperature conditions. However, the molecular properties of simple aqueous proline amino acid solutions at low temperatures remain largely unexplored. As a result, the mechanisms underlying its protective properties are not understood from this perspective. This is of course something computer simulation is potentially well suited for, since it can explore the structural mechanisms of the phenomena.

Without much further ado, let us dive into how the present work will explore the realm of natural antifreeze.

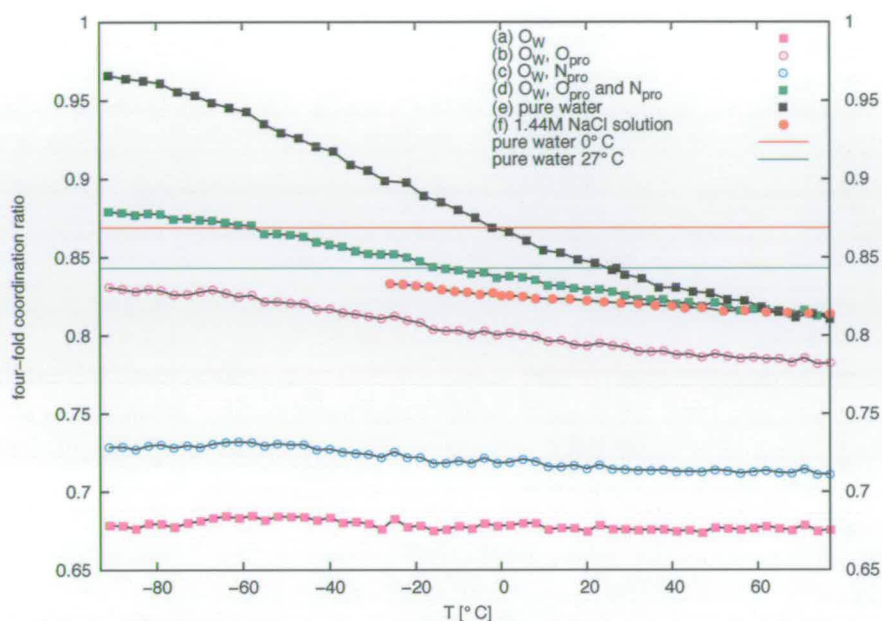


**Figure 5.1:** Radial distribution functions for water oxygen contacts in solution (main panel) and pure water (inset) at  $-45$ ,  $-15$  and  $15$  °C. In the pure liquid a distinct sharpening of the peaks is observed with decreasing temperature. This behaviour is absent in the solution where we find very modest structural changes for the lowest temperature, and, perhaps more surprisingly, a decrease in structure for the intermediate temperature. The results suggest that, proline suppresses the normal low-temperature evolution of water structure. The low temperature structure therefore corresponds to a higher effective temperature.

function,  $g(r)$ , we will work from the pure water case to the proline solution; table 5.1 lists the values of the ratio at the three key temperatures. As expected, the ratio of four-fold coordinated water molecules as a fraction of all molecules rises upon cooling in the pure water case. However, a similar study of the water constituent of the proline solution yields that the ratio remains at the same level, irrespective of temperature.

As indicated in section 2.5.6, the FFCR lends itself to exploring the assumption that the polar groups in non-water molecules may de facto act as “pseudo”-waters. The shape of the amide group in proline is such that we can approximate it to a water molecule protruding from a larger molecule, which itself is of no interest in the FFCR since it is beyond the cutoff,  $r_c$ . We can thus include the nitrogen site as a “substitute” oxygen.

Similarly, we can argue that the carboxyl oxygens are of similar quality in that they can serve as a “substitute” water oxygen by position and hydrogen bond proxy. Accordingly, the table 5.1 also lists the case where the carboxyl and amide group can act as coordinating partners (we will refer to this as the *proline*<sup>+</sup> configuration).



**Figure 5.2:** FFCR for various sets of "allowed" neighbours of four-fold in proline-water solution, as well as in the pure liquid. Refer to table 5.2 for details on sets.

different ways. In the case of the proline solution, this is achieved by including proline polar groups into the network, satisfying each water molecule's four-fold coordination need, but preventing the lattice structure from forming. In the case of salt, the presence of the ions, and thus strong local charge density, prevents that a sufficient number of water molecules are in the four-fold coordinated environment needed to form ice, by re-orientating them.

### Comparison to experimental evidence

In order to reconcile these conclusions with recently published Rayleigh-Brillouin light scattering and Raman spectra[106], let us briefly consider the conclusions from these experiments: upon cooling (1) the Brillouin spectra indicate glassy dynamics and an increase in viscoelasticity, and (2) the Raman spectra undergo the water-typical redshift, indicating a redistribution of the charge density away from the covalent bonds, into the intermolecular H-bonds, which weakens the covalent bonds as a result of stronger H-bonds; the ice-typical lattice phonons, however, are absent.

The Brillouin data is straightforwardly compatible with the simulation data FFCR conclusions, even though it can give no further indication of structural features beyond

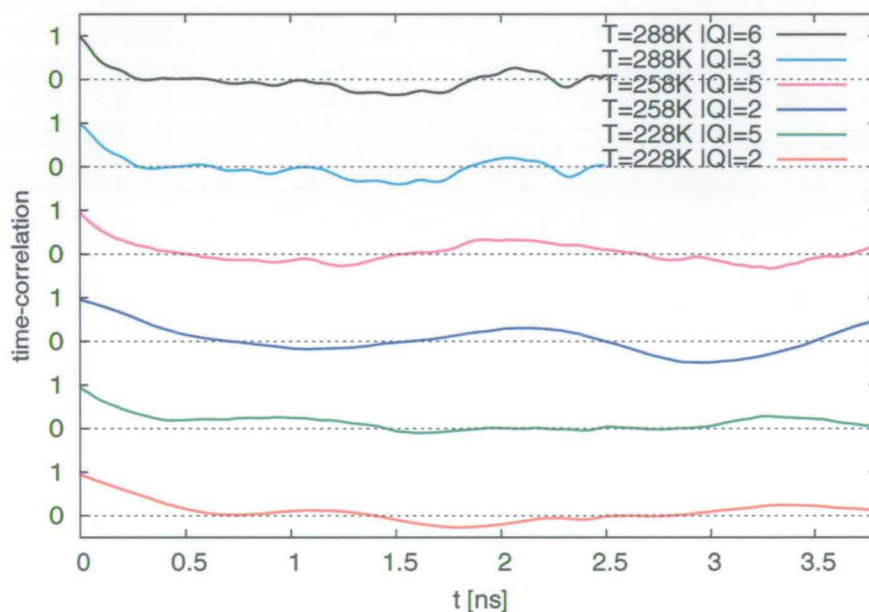
pure water. Even at very low temperatures ( $< -30^\circ\text{C}$ ) a vitreous liquid forms instead of an ice-like crystal structure.

Saline solution appears to achieve a similar effect by preventing the formation of a tetrahedral structure due to the strong local charge of the ions.

## 5.5 Methods and Ingredients

### 5.5.1 Proline solution

Classical molecular dynamics (MD) simulations were performed on the 2.75 M aqueous proline solution at  $T = -45, -15, +15^\circ\text{C}$  using the CHARMM 22[33, 60] force field and the TIP4P[44] empirical potentials for water. This well-known four-site model is rigid and non-polarisable.



**Figure 5.4:** Time-correlation function of neutron scattering in the proline solution at select  $Q$ -vectors and for the three target temperatures. Since all the curves decay satisfactorily on timescales small to the total run-time, the systems are taken to be well relaxed and in an equilibrium state.

A primitive cell containing 6860 and 343 water and proline molecules respectively, was initially placed on a cubic lattice with randomised molecular orientations in a simulation box of edge length  $62.8 \text{ \AA}$ . The system was then equilibrated to anneal out any unphysical high-energy contacts, using the canonical ensemble, for 300 ps at a temperature of  $T = 27^\circ\text{C}$ . To improve equilibration and minimise configurational

equilibration of 80 ps, the data were collected over simulation time of 400 ps.

## 5.6 Why the FFCR has value-add

Both the four-fold coordination ratio (FFCR) and the Debenedetti order number were calculated from the configurations, resulting in figure 5.2 on page 109 and figure 5.5, respectively.

For the pure water case, only one curve exists. For the solution, four distinct curves were computed; they correspond to different sets of “allowed” neighbours in the search for coordination (see above and table 5.2).

In examining figure 5.5, it becomes apparent that the temperature dependence of the Debenedetti order number is marginal, a change of less than 3% is observed over the entire temperature range. It is also noteworthy that there appears to be no difference in the gradient of what is in essence a straight line between the various ways of picking “allowed” neighbours, and indeed the pure water itself.

The scenario is different for the FFCR, where significant change is observed over the temperature range in both the pure liquid (~18%) and the various solution curves (1-7%) (see section 5.3 for more extensive analysis and physical interpretation).

curve in figure 5.2	“allowed” neighbours
(a)	O <sub>W</sub>
(b)	O <sub>W</sub> , O <sub>proline</sub>
(c)	O <sub>W</sub> , N <sub>proline</sub>
(d)	O <sub>W</sub> , O <sub>proline</sub> , N <sub>proline</sub> (“proline <sup>+</sup> ”)
(e)	O <sub>W</sub> (pure water system)

**Table 5.2:** The four sets of allowed sets of neighbours for the computation of FFCR and Debenedetti order numbers from the solution system.

# Curbing HIV

## Steps toward a vaccine



It is clear that HIV remains an acute and incurable super-virus affliction, and in some parts of the world dominates in presence and fatality (actually really assistance thereof) over many other horrid diseases. It is also fairly common knowledge that it is transferred between humans by the transfer of blood and other bodily fluids. In that respect, knowledge of both the virus, its transfer<sup>1</sup> and its symptoms, namely that it levers out the human immune system will be assumed.

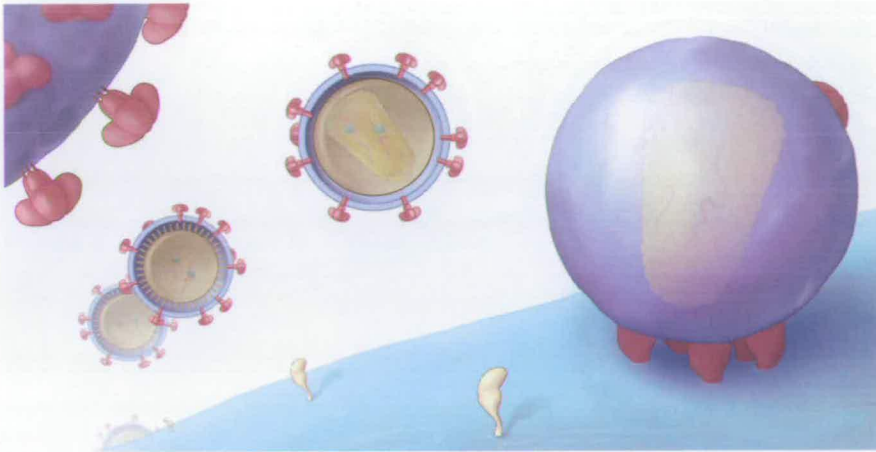
### 6.1 HIV infection on the molecular level

Once it has entered the bloodstream, the human immunodeficiency virus (HIV-1) infects the CD4+ T cells, which provide the immune system with “memory” of past infections. The mechanical work-flow of the infection of the cell is conceptually very simple. It occurs via the combined action of two envelope glycoproteins, gp120 and gp41, both of which sit on the outside of the virus (see figure 6.1). gp120 is responsible for initial host cell recognition and interaction with target CD4 and chemokine receptor sites on the target cell’s surface. The docking to the target happens via this protein. An artists impression of the docking process is displayed in figure 6.2. Once docked via the gp120, the gp41 peptide “shoots out” in a harpoon-like fashion and enters the target’s membrane. Structurally, the main features of the extracellular domain of gp41 consists of a glycine rich fusion domain at the N-terminus, followed by two highly conserved helical regions (N- and C- heptad repeats) running into the membrane proximal domain. Upon insertion of the fusion peptide into the target cell, the two helical domains of gp41 fold to form a 6-helix bundle with an open central channel, referred to as a hairpin

---

<sup>1</sup>an excellent and witty coverage in four languages, dealing with the topic can be found at [www.lovelife.ch](http://www.lovelife.ch)





**Figure 6.2:** Schematics of the HI virus docking to the target cell via the gp120 tentacles.

transmembrane) segment can prevent fusion[9, 88]. This region also spans the synthetic peptide T-20 (gp41<sub>638–673</sub>) derived from the C-terminal heptad repeat: referred to in the literature by various names, including *Fuzeon* and *Enfuviritide*, it is the first FDA-approved fusion inhibitor targeting the gp41 transmembrane subunit. Its mechanism is still debated but current thinking suggests it binds competitively to the N-helix and forms a steric blockade of the conformational changes required for infection[49]. A second strategy in vaccine development is to elicit broadly neutralising antibodies that target conserved epitopes of the gp120 or gp41 surface glycoproteins. The membrane proximal domain gp41<sub>659–671</sub> (containing a segment of T20) also spans the complete epitope ELDKWA for the 2F5 monoclonal antibody[101] which shows high binding affinity for this peptide[128]. The affinity is reduced after binding between gp120 and CD4, suggesting that the gp41<sub>659–671</sub> epitope is solvent exposed in the prefusogenic form but becomes less accessible or restructured after fusion[9, 24]. Also, Barbato suggests that the prohibition of fusion of the HI virus with the cell occurs in a transition phase of the gp41 peptide between turn, extended and helical structure[5]. The solution structure of this peptide (in prefusogenic form) is therefore recognised to be important for the development of peptide antigens, which links nicely back to the claim in the introduction chapter that structure is in fact the key to success. Very loosely speaking, if one knows not the shape of the shaft, making a sheath that fits is a little tricky.

However, attempts to elicit antibody responses using recombinant gp41 and synthetic sequences overlapping the 2F5 core epitope have not been effective[12]. One possible reason for their failure may be that current synthetic peptides do not adopt a structure sufficiently similar to that of real, prefusogenic viral gp41. Molecular

scattering[1] than is evident in the NMR data[9]. Using the spectra of the backbone Amide III bands as a reporter of secondary structure, these authors find evidence for significant populations of  $\beta$ -turn motifs as well as poly-proline II (pP<sub>II</sub>),  $3_{10}$ - and  $\pi$ -helices. Only small populations of  $\alpha$ -helices are inferred from the spectra. Moreover, the spectroscopic evidence implies that the relative populations of folded vs. unfolded conformations show an unexpectedly weak dependence on temperature[1]. It has to be noted that available circular dichroism data are not in agreement over the  $3_{10}$  propensities: The data of Biron *et al.* imply a higher degree of  $3_{10}$ -helical content than does that of Ahmed and Asher[1] and Barbato *et al.*

Also, in related systems, a longer peptide spanning residues 665 to 683 shows  $\alpha$ -helical folds in a membrane mimetic solvent of dodecyl phosphocholine micelles[90]. Finally, the epitope region spanning residues 662-667 is reported to form type-I  $\beta$ -turn crystalline complexes with the monoclonal antibody 2F5<sup>2</sup>.

### 6.3 How this work studies gp41<sub>659–671</sub>

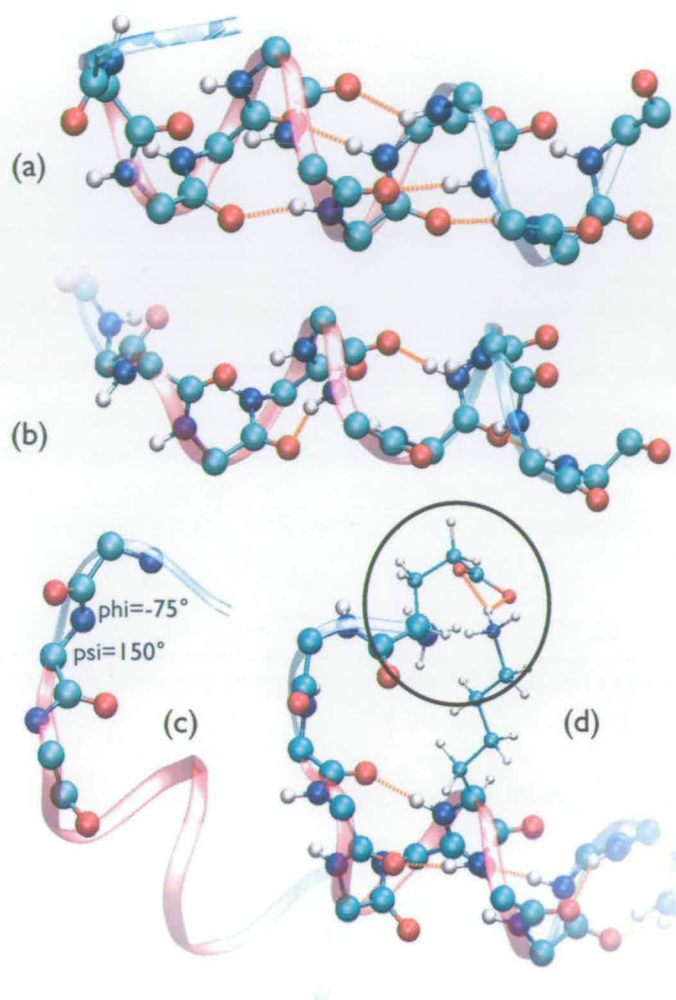
Here, the structure in aqueous solution of the gp41<sub>659–671</sub> peptide will be established using classical atomistic molecular dynamics simulations. As in previous chapters, the efficiency of configurational sampling is improved through the implementation of parallel tempering (or *replica exchange*) on the CHARMM22 empirical potential surface. Further, to account for sensitivity to water models, both TIP3P and TIP4P are employed, and the CMAP correction established by MacKerell *et al.* [61, 62, 63] is employed to account for backbone torsional corrections.

Based on data published in Ref. [105], comparison to circular dichroism spectroscopy (CD) of the peptide conformation will be drawn to elucidate various findings and put them in experimental context.

### 6.4 Structural revelations from simulations

In the structural analysis, the hydrogen bonding propensities between Amide proton based on a geometric definition of a hydrogen bond as well as the Ramachandran ( $\phi, \psi$ )-map of backbone angles is explored. First, the classification of hydrogen bonding propensities is addressed, based on the connectivity of residue  $i$  to residue  $i+n$ , with  $n = 3, 4, 5$  for  $3_{10}$ ,  $\alpha$  and  $\pi$  helices respectively, directly from the trajectories. A summary of results of this procedure for each residue is shown in figure 6.3. Here, the proportion of  $3_{10}$  and  $\alpha$  is given for each residue, averaged over the entire parallel tempering simulation (3.5 ns) at  $T = 300$  K. The occurrence (18% on average over all residues and

<sup>2</sup>World Intellectual Property Organisation patent WO-00/61618



**Figure 6.4:** Snapshots of the molecular trajectory showing examples of (a) alpha helical (b) 3-10 helical, (c) pP<sub>II</sub> motifs. Hydrogen bonding the side chain groups in the flexible N-terminal is shown in (d). In all panels the region of the epitope is signified by a red ribbon in the backbone.

prevalence of  $\alpha$ -helical contact is reduced with CMAP but, where present, is more tightly confined to the prescriptive angles of  $(\phi, \psi) = (-60, -45)$ .

The region of the peptide immediately preceding the ELDKWA emerges as being very flexible under ambient conditions. Visual inspection of the MD trajectories reveals occasional hydrogen bonding between side chain atoms. An example is shown in figure 6.4(d). Also in this region, a small population of conformers with angles  $(\phi, \psi) \approx (-75, 150)$  is found, which is similar to that expected for poly-proline II (pP<sub>II</sub>)

persistent motif. Outwith the epitope region, the peptide is vastly more flexible and contains elements of pP<sub>II</sub>, as well as hydrogen bonding of side chains.

#### 6.4.1 Comparison to experiment

The simulation results are in qualitative agreement with the results of both NMR and UV resonance Raman measurements[5, 9]. However, the NMR data suggest that the  $3_{10}$ -helix is the dominant secondary structure, with the  $\alpha$ -helix appearing as a minority population. In the simulation result this is found to be less stable than the  $\alpha$ -helical motif for all empirical potentials. As discussed above, however, there is considerable solvent and temperature sensitivity for certain residues, suggesting that buffer conditions in the experiment (which are not exactly matched in the simulations) may account for some of the difference in populations.

With regard to UV resonance Raman data[1], various Amide bands can be taken as indirect reporters of secondary structure. Much of the analysis is based on the features of the Amide III band (combination of C-N stretch and N-H bending modes in the range spanning  $1200\text{ cm}^{-1}$  to  $1350\text{ cm}^{-1}$ ) and the analysis of Mikhonin *et al.* [75], which gives a prescription by which the  $\phi$  dihedral angle can be inferred from the spectra. The physical origins of the connection lies in the apparent sinusoidal variation[75] in the C $_{\alpha}$ -H and N-H bending modes with  $\psi$  dihedral angle. By contrast, the Amide III bands vary only weakly with the  $\phi$  backbone angle. The analysis is complicated, however, because it is necessary to separate the hydrogen bonding dependence of the Amide III bands from that on backbone conformation[75]. In the spectroscopic data[1], a broad Amide III band is observed near  $1257\text{ cm}^{-1}$ . Whereas it is expected that a pure  $\alpha$ -helical backbone conformation gives rise to a narrow Amide III band, here the spectroscopic observation is taken as evidence for multiple conformations. The  $1254\text{ cm}^{-1}$  band represents a  $9\text{ cm}^{-1}$  redshift relative to  $\alpha$ -helical peptides and is assigned as the spectroscopic signature of the  $3_{10}$ -helix. The authors point out that while no strong spectroscopic evidence exists for high  $\alpha$ -helical populations, some proportion may be consistent with the breadth of the Amide III bands. Based on the temperature dependence of the Amide III band and that of C $_{\alpha}$ -H bending bands, it was noted that a similar behaviour was observed for water-exposed pP<sub>II</sub>-type peptide and therefore that pP<sub>II</sub> may also be present. These authors also observe no decrease in helicity over the temperature range  $1 - 30\text{ }^{\circ}\text{C}$ , in agreement with Biron[9]. Other spectral features (e.g., the  $1224\text{ cm}^{-1}$ ) indicate the presence of  $\beta$ -turn motifs. Also significant is the sub-band of the Amide III complex centred around  $1293\text{ cm}^{-1}$ . Again, following the data of Mikhonin *et al.* [75], this band is assigned to a dihedral angle  $\phi \approx 75^{\circ}$  which corresponds to  $\pi$ -helices ( $i \rightarrow i + 5$ ) hydrogen bonding patterns.

However the  $\alpha$ -helix is over-expressed relative to  $3_{10}$  and the number of other competing secondary structures (turns,  $\pi$ -helix) is fewer than that inferred from the optical spectroscopic data.

## 6.6 Where to from here?

All this is fine and well, but of course we are still a long way from routinely vaccinating against HIV. However, understanding the structure, and as a result the mechanisms of the “evil ways” of the virus from the molecular level up is an immense advancement in postulating ways to stop the disease. So, in a way the first step has now been taken: what does the solution structure of this vital part of the virus look like. Following on from this should surely come a simulation of 1) the structure of the “real” gp41 (see section 6.2), 2) the docking process of either the gp120 peptide, but just as importantly 3) insight into the *harpooning* of the gp41 into a cell membrane should be sought. Additionally, the docking process of the virus to one of the synthetically produced (and thus structurally known) drugs could and should be explored.

Of course, these are long shots, involving huge scale simulations, all hampered by the same problems as even the smaller systems, and ideally approached with *ab initio* calculations to account for the sensitivity of non-covalent, yet strong bonds involved in the docking.

There are many grants to be written, many collaborations to be struck, many hours spent developing more tools to accelerate simulations of this type, many inspired ways of linking simulation to experiment to be found, before we can truly formulate a script of how to effectively oust HIV. However, this approach is a promising one, and it has as a reward something no money can buy.

## 6.7 Computational methods

A system of one gp41 peptide and 1853 water molecules, as well as 4  $\text{Cl}^-$  ions to balance overall charge, with random molecular orientations was set up on a cubic lattice in a box of length 38 Å. The system was equilibrated at 300 K in the canonical ensemble for 300 ps with a time-step of 0.1 fs to anneal out unphysical contacts. The system was then run for 500 ps with a time-step of 1.5 fs in the isothermal-isobaric ensemble to allow for spatial relaxation. At this stage the system was triplicated and set aside for each of the three simulations (using water models TIP3P, TIP4P and TIP4P-CMAP (see below)). These rigid, non-polarisable models are used in combination with the CHARMM22 force field[60]. Periodic boundary conditions were applied and

# Outroduction



If I still have your attention at this point, it is either because I really managed to grip you, or because you skipped most of it and wanted to just get a quick impression of whether my work came to some fruition. Either way, this is indeed the time to sum it all up, to see if the grand plan outlined in the introduction was mostly realised and where one might opt to invest time beyond the contents of this thesis.

## 7.1 Conclusion

In the introduction, a few key objectives were laid out for examination. It is probably instructive to recapitulate them briefly. Foremost, we were interested to look into the specific structure formed of particular systems with biological impact. The scope of these systems was to grow from elementary problems to rather involved ones, all along driven by the idea that we may learn something on the way. Better yet, instead of just gathering knowledge, we strove to recycle results, be they new insight into applicability of certain approximations in modelling systems, or simply intermediary results which will aid us in future, understanding systems at the next level of complexity.

Further, comparison with experiment was heralded as a key player, since, as discussed in the introduction, we can only draw confidence about our results from verification with experiment.

### 7.1.1 Insights into local structure

Extensive insight was achieved into the structures of all the systems studied. In particular, the solvation of the  $\text{Na}^+$  ion in aqueous solution was of interest. Here, it was found that there are molecular mechanisms which, through a non-intuitive spatial conformation, facilitate the exchange of water molecules in the first solvation

### 7.1.3 Linking to experiment

In order to address the connection of simulation to experiment, the static structure factor was employed to compare simulation data to experimental neutron diffraction data. This was found to be a powerful means to confirm the quality of the simulation data, allowing for the assumption that subsequent statements about local structure were vindicated. Particularly in the case of aqueous proline, this method allowed for a high level of confidence. Quite generally, it was found that simulations based on lower levels of detail (or alternatively on a higher level of assumption at the outset; see section 2.0.0 on page 9) in the model yielded results that matched the experimental data less. While not surprising, this does indicate the need for the establishment of models that include more detail (like polarisability) while remaining computationally viable for large systems ( $> \mathcal{O}10^2$  atoms).

Where possible, reconciliation was sought of measurements carried out on bulk systems, like light scattering and Raman spectroscopy, and mechanisms found in the systems based on simulation. While of a qualitative nature, these links were key in elucidating the aforementioned bio-protective properties of proline, as well as the structural properties of the HIV's gp41 peptide.

### 7.1.4 The significance of it all

Has it, overall, made sense to do this work? What's the take-home message? What have we learned that won't go out of date?

Yes, it has made sense. Not just because scientific findings have surfaced, or because some decades down the line someone might say "ah, there was this guy, he did this, it applies like so, hence this can be explained like yae". It made sense because actual ways to link techniques were found; actual tools were developed and expanded. That should also be the take-home message: neither experiment nor simulation alone are the one and only. They need to feed off one another, they need each other to make a statement possible, or at the very least supportable.

If there is one thing I have learned that shall not go out of date, is that collaboration between people is as important as their input into a common project. Pooling expertise, across boundaries of scientific discipline, across language and continents, is vital and makes for a better picture and for a more harmonious environment. Heed it.

## 7.2 Vision

It may seem pompous to announce a vision in a thesis, since by the time it sits on a shelf for others to peruse, the vehicle of research will have rolled on and much of what

## **Epilogue**

As the hands on the clock of this PhD creak past the zenith position, and the shadows of the studies shorten and those of the long night of real life are looming ahead, the distant tolls of an Old Bell can be heard. All that remains to say is...



---

## Bibliography

---



- [1] Z. Ahmed and S.A. Asher. UV Resonance Raman Investigation of a  $3_{10}$ -Helical Peptide Reveals a Rough Energy Landscape. *Biochemistry*, 45(45):9068 – 9073, 2006.
- [2] M.P. Allen and D.J. Tildesley. *Computer Simulation of Liquids*. Oxford University Press, Great Clarendon Street, Oxford, OX2 6DP, UK, 2006.
- [3] H.C. Andersen. Rattle: A “velocity” version of the shake algorithm for molecular dynamics calculations. *J. Comp. Phys.*, 52:24–34, 1983.
- [4] G.B. Bachelet, D.R. Hamann, and M. Schlüter. Pseudopotentials that work: from H to Pu. *Phys. Rev. B*, 26:4199, 1982.
- [5] G. Barbato, E. Bianchi, P. Ingallinella, W.H. Hurni, Michael D. Miller, Gennaro Ciliberto, R. Cortese, R. Bazzo, J. W. Shiver, and A. Pessi. Structural Analysis of the Epitope of the Anti-HIV Antibody 2F5 Sheds Light into Its Mechanism of Neutralization and HIV Fusion. *J. Mol. Biol.*, 330(5):1101, 2003.
- [6] J.D. Batchelor, A. Olteanu, A. Tripathy, and G.J. Pielak. Impact of Protein Denaturants and Stabilizers on Water Structure. *J. Am. Chem. Soc.*, 126(7):1958, 2004.
- [7] A.D. Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A*, 38:3098–3100, 1988.
- [8] S.J.L. Billinge and M.F. Thorpe, editors. *Local Structure from Diffraction*. Plenum press, 2000.
- [9] Z. Biron, S. Khare, A. Samson, Y. Heyak, F. Naider, and J. Anglister. A Monomeric  $3_{10}$ -Helix Is Formed in Water by a 13-Residue Peptide Representing the Neutralizing Determinant of HIV-1 on gp41. *Biochemistry*, 41:12687, 2002.
- [10] E. Bohm, A. Bhatele, L.V. Kalé, M.E. Tuckerman, S. Kumar, J.A. Gunnels, and G.J. Martyna. Fine-grained parallelization of the Car-Parrinello ab initio molecular dynamics method on the Blue Gene/L supercomputer. *IBM J. Res. Dev.*, 52(1/2), 2007.
- [11] S. Bouazizi, S. Nasr, N. Jaidane, and M-C Bellissent-Funel. Local Order in Aqueous NaCl Solutions and Pure Water: X-ray Scattering and Molecular Dynamics Simulations Study. *J. Phys. Chem. B*, 110:23515–23523, 2006.
- [12] F.M. Brunel, M.B. Zwick, R.M.F. Cardoso, J.D. Nelson, I.A. Wilson, D.R. Burton, and P.E. Dawson. Structure-Function Analysis of the Epitope for 4E10, a Broadly Neutralizing Human Immunodeficiency Virus Type 1 Antibody. *J. Virol.*, 80:1680–1687, 2006.
- [13] R. Buchner, G.T. Hefter, and P.M. May. Dielectric Relaxation of Aqueous NaCl Solutions. *J. Phys. Chem. A*, 103(1):1–9, 1999.

- [33] M. Feig, A.D. Mackerell Jr., and III C.L. Brooks. Force Field Influence on the Observation of  $\pi$ -Helical Protein Structures in Molecular Dynamics Simulations. *J. Phys. Chem. B*, 117(12):2831 – 2836, 2003.
- [34] R.P. Feynman and A.R. Hibbs. *Quantum Physics and Path Integrals*. McGraw-Hill, New York, 1965.
- [35] C.B.P. Finn. *Thermal Physics*. Stanley Thornes, 2 edition, 1998.
- [36] J.L. Finney. Water? What's so special about it? *Phil. Trans. R. Soc. Lond. B*, 359:1145, 2004.
- [37] M.T. Fisher. Proline to the rescue. *Proc Natl Acad Sci*, 103(36):13265–13266, 2006.
- [38] D. Frenkel and B. Smit. *Understanding Molecular Dynamics*, volume 1 of *Computational Science Series*. Academic Press, 84 Theobald's Road, London WC1X 8RR, UK, 2 edition, 2002.
- [39] C.J. Geyer. Computing Science and Statistics Proceedings of the 23rd Symposium on the Interface. In *Computing Science and Statistics Proceedings of the 23rd Symposium on the Interface*, page 156, New York, 1991. American Statistical Association.
- [40] H. Goldstein. *Classical Mechanics*, volume 2. Addison-Wesley, July 1980.
- [41] G. Han, Y. Deng, J. Glimm, and G.J. Martyna. Error and timing analysis of multiple time-step integration methods for molecular dynamics. *Comput. Phys. Comm.*, 176:271–291, 2007.
- [42] P.D. Hare, W.A. Cress, and J. van Staden. Review article. Proline synthesis and degradation: a model system for elucidating stress-related signal transduction. *J. Exp. Bot.*, 50(333):413–434, 1999.
- [43] Y. Hayashi, M. Matsuzawa, J. Yamaguchi, S. Yonehara, Y. Matumoto, M. Shoji, Hashizume D., and H. Koshino. Large Nonlinear Effect Observed in the Enantiomeric Excess of Proline in Solution and That in the Solid State. *Angew. Chem Int. Ed.*, 45:4593, 2006.
- [44] M.Z. Hernandez, J.B.P. da Silva, and R.L. Longo. Chemometric study of liquid water simulations. Part I: the parameters of the TIP4P model potential. *J. Comp. Chem.*, 24:973–981, 2003.
- [45] F. Hofmeister. Zur Lehre von der Wirkung der Salze. II. *Arch. Exp. Pathol. Pharmacol.*, 24:247, 1888.
- [46] P. Hohenberg and W. Kohn. Inhomogeneous Electron Gas. *Phys. Rev.*, 136(3B):B864 – B871, 1964.
- [47] J. Hutter, M.E. Tuckerman, and M. Parrinello. Integrating the Car–Parrinello equations. III. Techniques for ultrasoft pseudopotentials. *J. Chem. Phys.*, 102:859, 1995.
- [48] G.P. Jones, L.G. Paleg, and D.Z. Winzor. Elimination of self-association as the source of the thermodynamic nonideality in aqueous proline solutions. *Biochim. Biophys. Acta.*, 1201:37, 1994.
- [49] J.G. Joyce, W.M. Hurni, M.J. Bogusky, V.M. Garsky, X. Liang, M.P. Citron, R.C. Danzeisen, M.D. Miller, J.W. Shiver, and P.M. Keller. Enhancement of  $\alpha$ -Helicity in the HIV-1 Inhibitory Peptide DP178 Leads to an Increased Affinity for Human Monoclonal Antibody 2F5 but Does Not Elicit Neutralizing Responses in Vitro. *J. Biol. Chem.*, 277:45811–45820, 2002.

- [67] Y.A. Mantz, B. Chen, and G.J. Martyna. Structural Correlations and Motifs in Liquid Water at Selected Temperatures: Ab Initio and Empirical Model Predictions. *J. Phys. Chem. B*, 110:3540, 2006.
- [68] G.J. Martyna, M.L. Klein, and M.E. Tuckerman. Nosé–hoover chains: The canonical ensemble via continuous dynamics. *J. Chem. Phys.*, 97:2635, 1992.
- [69] G.J. Martyna, D.J. Tobias, and M.L. Klein. Constant pressure molecular dynamics algorithms. *J. Chem. Phys.*, 101:4177, 1994.
- [70] G.J. Martyna, M.E. Tuckerman, D.J. Tobias, and M.L. Klein. Explicit reversible integrators for extended systems dynamics. *Mol. Phys.*, 87:1117, 1996.
- [71] M. Matsumoto, H. Tanaka, and K. Nakanishi. Acetonitrile pair formation in aqueous solution. *J. Chem. Phys.*, 99:6935, 1993.
- [72] S.E. McLain, A.K. Soper, A.E. Terry, and A. Watts. The structure and hydration of L-proline in aqueous solutions. *J. Phys. Chem. B*, 2007. in press.
- [73] D.A. McQuarrie. *Statistical Mechanics*. University Science Books, 2000.
- [74] H. Meirovitch, S. Rackovsky, and H.A. Scheraga. Empirical Studies of Hydrophobicity. 1. Effect of Protein Size on the Hydrophobic Behavior of Amino Acids. *Macromolecules*, 13:1398, 1980.
- [75] A.V. Mikhonin, S.V. Bykov, N.S. Myshakina, and S. Asher. Peptide Secondary Structure Folding Reaction Coordinate: Correlation between UV Raman Amide III Frequency,  $\Psi$  Ramachandran Angle, and Hydrogen Bonding. *J. Phys. Chem. B*, 110(2):1928, 2006.
- [76] T.F. Miller, M. Eleftheriou, P. Pattnaik, A. Ndirango, D. Newns, and G.J. Martyna. Symplectic quaternion scheme for biophysical molecular dynamics. *J. Chem. Phys.*, 116(20):8649–8659, 2002.
- [77] Y. Morito, S. Nakamori, and H. Takagi. L-Proline Accumulation and Freeze Tolerance in *Saccharomyces cerevisiae* Are Caused by a Mutation in the PRO1 Gene Encoding  $\gamma$ -Glutamyl Kinase. *Appl. Environ. Microbiol.*, 69(1):212–219, 2003.
- [78] A. Mudi and C. Chakravarty. Effect of Ionic Solutes on the Hydrogen Bond Network Dynamics of Water: Power Spectral Analysis of Aqueous NaCl Solutions. *J. Phys. Chem. B*, 110:8422–8431, 2006.
- [79] Yu.I. Naberukhin, V.P. Voloshin, and N.N. Medvedev. Geometrical analysis of the structure of simple liquids: percolation approach. *Molecular Physics*, 73:917, 1991.
- [80] T. M. Nanjo, Y. Kobayashi, Y. Yoshida, K. Kakubari, K. Yamaguchi-Shinozaki, and K. Shinozaki. Antisense suppression of proline degradation improves tolerance to freezing and salinity in *Arabidopsis thaliana*. *FEBS Lett.*, 461:205, 1999.
- [81] A. Neagu, M. Neagu, and A. Dér. Fluctuations and the Hofmeister Effect. *Biophys. J.*, 81(1285), 2001.
- [82] A.W. Omta, M.F. Kropman, S. Woutersen, and H.J. Bakker. Influence of ions on the hydrogen-bond structure in liquid water. *J. Chem. Phys.*, 119:12457, 2003.
- [83] R. Paquin and G. Pelletier. Acclimatation naturelle de la luzerne (*Medicago media* Pers.) au froid I. Variations de la teneur en proline libre des feuilles et des collets. *Physiolog. Veg.*, 19:103–117, 1981.
- [84] M. Patra and M. Karttunen. Systematic comparison of force fields for microscopic simulations of NaCl in aqueous solutions: Diffusion, free energy of hydration, and structural properties. *J. Comp. Chem.*, 25(5):678–689, 2004.

- [102] R.Z. Troitzsch, J. Crain, and G.J. Martyna. A simplified model of local structure in aqueous proline amino acid revealed by first principles molecular dynamics simulations. *J. Am. Chem. Soc.*, submitted, 2007.
- [103] R.Z. Troitzsch, G.J. Martyna, and J. Crain. Novel structural and dynamical features of concentrated aqueous NaCl. *Proc. Natl. Acad. Sci.*, submitted, 2007.
- [104] R.Z. Troitzsch, G.J. Martyna, S.E. McLain, A.K. Soper, and J. Crain. Structure of Aqueous Proline via Parallel Tempering Molecular Dynamics and Neutron Diffraction. *J. Phys. Chem. B*, 111(28):8210–8222, 2007.
- [105] R.Z. Troitzsch, G.J. Martyna, S. Thobhani, E. Cerasoli, G. Tranter, and J. Crain. Solution structure of an HIV-1 antibody epitope: Parallel tempering molecular dynamics and circular dichroism spectroscopy of peptide gp41<sub>659–671</sub>. *Biophys. J.*, submitted, 2007.
- [106] R.Z. Troitzsch, H. Vass, W.J. Hossack, G.J. Martyna, and J. Crain. Molecular mechanisms of cryoprotection in aqueous proline: light scattering and molecular dynamics simulations. *J. Phys. Chem. B*, accepted, 2007.
- [107] N. Troullier and J.L. Martins. Efficient pseudopotentials for plane-wave calculations. *Phys. Rev. B*, 43:1993 – 2006, 1991.
- [108] M.E. Tuckerman, B.J. Berne, G.J. Martyna, and M.L. Klein. Efficient molecular dynamics and hybrid Monte Carlo algorithms for path integrals. *J. Chem. Phys.*, 99:2796, 1993.
- [109] M.E. Tuckerman, Y. Liu, G. Ciccotti, and G.J. Martyna. Non-Hamiltonian molecular dynamics: Generalizing Hamiltonian phase space principles to non-Hamiltonian systems. *J. Chem. Phys.*, 115(4):1678, 2001.
- [110] M.E. Tuckerman and M. Parrinello. Integrating the Car–Parrinello equations. I. Basic integration techniques. *J. Chem. Phys.*, 101:1302, 1994.
- [111] M.E. Tuckerman and M. Parrinello. Integrating the Car–Parrinello equations. II. Multiple time scale techniques. *J. Chem. Phys.*, 101:1316, 1994.
- [112] M. Tuckermann, B.J. Berne, and G.J. Martyna. Reversible multiple time scale molecular dynamics. *J. Chem. Phys.*, 97(3):1990–2001, 1992.
- [113] H. Uchida and M. Matsuoka. Molecular dynamics simulation of solution structure and dynamics of aqueous sodium chloride solutions from dilute to supersaturated concentration. *Fluid Phase Equilibria*, 219(1):49–54, 2004.
- [114] R.V. Vadali, Y. Shi, S. Kumar, L.V. Kale, M.E. Tuckerman, and G.J. Martyna. Scalable Fine-Grained Parallelization of Plane-Wave-Based Ab Initio Molecular Dynamics for Large Supercomputers. *J. Comp. Chem.*, 25(16):2006–2022, 2007.
- [115] L. Verlet. Computer “Experiments” on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.*, 159:98–103, 1967.
- [116] G.E. Walrafen, M.S. Hokmabadi, W.H. Yang, Y.C. Chu, and B. Monosmith. Collision-induced Raman scattering from water and aqueous solutions. *J. Phys. Chem.*, 93:2909, 1989.
- [117] J.A. White, E. Schwengler, G. Galli, and F. Gygi. The solvation of  $\text{Na}^+$  in water: First-principles simulations. *J. Chem. Phys.*, 113(11):4668, 2000.
- [118] T.W. Whitfield, J. Crain, and G.J. Martyna. Structural properties of liquid N-methylacetamide via *ab initio*, path integral, and classical molecular dynamics. *J. Chem. Phys.*, 124, 2006.