Theory and simulation of molecular interactions in biological systems

by

Sadish Karunaweera

B.S., University of Peradeniya, Sri Lanka, 2006
M.S., University of Peradeniya, Sri Lanka, 2008

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Chemistry
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2017

Abstract

The impact of computer simulations has become quite significant especially with the development of supercomputers during the last couple of decades. They are used in a wide range of purposes such as exploring experimentally inaccessible phenomena and providing an alternative when experiments are expensive, dangerous, time consuming, difficult and controversial. In terms of applications in biological systems molecular modeling techniques can be used in rational drug design, predicting structures of proteins and circumstances where the atomic level descriptions provided by them are valuable for the understanding of the systems of interest. Hence, the potential of computer simulations of biomolecular systems is undeniable. Irrespective of the promising uses of computer simulations, it cannot be guaranteed that the results will be realistic. The precision of a molecular simulation depends on the degree of sampling achieved during the simulation while the accuracy of the results depends on the satisfactory description of intramolecular and intermolecular interactions in the system, i.e. the force field. Recently, we have been developing a force field for molecular dynamics simulations of biological systems based on the Kirkwood Buff (KB) theory of solutions, not only with an emphasis on the accurate description of intermolecular interactions, but also by reproducing several physical properties such as partial molar volume, compressibility and composition dependent chemical potential derivatives to match with respective experimental values. In this approach simulation results in terms of KB integrals can be directly compared with experimental data through a KB analysis of the solution properties and therefore it provides a simple and clear method to test the capability of the KB derived force field. Initially, we have provided a rigorous framework for the analysis of experimental and simulation data concerning open and closed multicomponent systems using the KB theory of solutions. The results are illustrated using computer simulations for various concentrations of the solutes Gly,

Gly_2 and Gly_3 in both open and closed systems, and in the absence or presence of NaCl as a cosolvent. Then, we have attempted to quantify the interactions between amino acids in aqueous solutions using the KB theory of solutions. The results are illustrated using computer simulations for various concentrations of the twenty zwitterionic amino acids at ambient temperature and pressure. Next, several amino acids were also studied at higher temperatures and pressures and the results are discussed in terms of the preferential (solute over solvent) interactions between the amino acids. Finally, we have described our most recent efforts towards a complete force field for peptides and proteins. The results are illustrated using molecular dynamics simulations of several tripeptides, selected peptides and selected globular proteins at ambient temperature and pressure followed by replica exchange molecular dynamics simulations of a few selected peptides.

Theory and simulation of molecular interactions in biological systems

by

Sadish Karunaweera

B.S., University of Peradeniya, Sri Lanka, 2006
M.S., University of Peradeniya, Sri Lanka, 2008

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Chemistry
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2017

Approved by:

Major Professor
Paul E. Smith

Copyright

Abstract

The impact of computer simulations has become quite significant especially with the development of supercomputers during the last couple of decades. They are used in a wide range of purposes such as exploring experimentally inaccessible phenomena and providing an alternative when experiments are expensive, dangerous, time consuming, difficult and controversial. In terms of applications in biological systems molecular modeling techniques can be used in rational drug design, predicting structures of proteins and circumstances where the atomic level descriptions provided by them are valuable for the understanding of the systems of interest. Hence, the potential of computer simulations of biomolecular systems is undeniable. Irrespective of the promising uses of computer simulations, it cannot be guaranteed that the results will be realistic. The precision of a molecular simulation depends on the degree of sampling achieved during the simulation while the accuracy of the results depends on the satisfactory description of intramolecular and intermolecular interactions in the system, i.e. the force field. Recently, we have been developing a force field for molecular dynamics simulations of biological systems based on the Kirkwood Buff (KB) theory of solutions, not only with an emphasis on the accurate description of intermolecular interactions, but also by reproducing several physical properties such as partial molar volume, compressibility and composition dependent chemical potential derivatives to match with respective experimental values. In this approach simulation results in terms of KB integrals can be directly compared with experimental data through a KB analysis of the solution properties and therefore it provides a simple and clear method to test the capability of the KB derived force field. Initially, we have provided a rigorous framework for the analysis of experimental and simulation data concerning open and closed multicomponent systems using the KB theory of solutions. The results are illustrated using computer simulations for various concentrations of the solutes Gly,

Gly_2 and Gly_3 in both open and closed systems, and in the absence or presence of NaCl as a cosolvent. Then, we have attempted to quantify the interactions between amino acids in aqueous solutions using the KB theory of solutions. The results are illustrated using computer simulations for various concentrations of the twenty zwitterionic amino acids at ambient temperature and pressure. Next, several amino acids were also studied at higher temperatures and pressures and the results are discussed in terms of the preferential (solute over solvent) interactions between the amino acids. Finally, we have described our most recent efforts towards a complete force field for peptides and proteins. The results are illustrated using molecular dynamics simulations of several tripeptides, selected peptides and selected globular proteins at ambient temperature and pressure followed by replica exchange molecular dynamics simulations of a few selected peptides.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to heartily express my gratitude to my major supervisor, Prof. Paul E. Smith, for giving his fullest support, guidance, patience, and encouragement throughout the course of my Ph. D. program. I cordially extend my thankfulness to the Ph. D. committee members, Prof. Christine M. Aikens, Prof. Stefan H. Bossmann and Prof. Jianhan Chen. Also, I am grateful to the outside chairperson, Prof. Roman R. Ganta. I want to acknowledge the faculty and staff members of the Department of Chemistry and also the funding sources; National Institute of General Medical Sciences, National Institute of Health and Kansas State University. I really thank my Computational Chemistry group members, including former members; Dr. Elizabeth A. Ploetz, Dr. Yuanfang Jiao, Dr. Shu Dai and Dr. Gayani N. Pallewela as well as present members; Mohomed Nawavi Mohomed Naleem and Nilusha Lakmali Kariyawasam Manachchige. I sincerely thank all my friends and relatives who supported me throughout the Ph. D. program. Finally, I am very grateful to my mother, father, brother, wife and in-laws for their love, support, and encouragement.

Dedication

To my loving parents and brother, who always pave the way for my success:

Rajendra Karunaweera

Shriyani Karunaweera

Chamaal Karunaweera

To my ever loving wife:

Madhubhashini B. Galkaduwa

To my parents-in-law for giving me strength and love:

Tikiri Banda Galkaduwa

Bandara Manike Galkaduwa

Chapter 1 - Introduction

1.1 Computational Chemistry

The foundations of computational chemistry are laid in quantum chemistry,[1] where the ultimate goal is to solve the time dependent Schrodinger equation,

$$-i\hbar \frac{\partial \Psi(r, r_e, t)}{\partial t} = \mathcal{H}\Psi(r, r_e, t) \tag{1.1}$$

Here the wave function Ψ is a function of the instantaneous positions of all the nuclei (r) and electrons (r_e) of the system, and $\mathcal{H}$ is the Hamiltonian operator. The postulates of quantum mechanics state that this wave function contains all possible information regarding the system. Unfortunately, Equation (1.1) is not solvable for all but the smallest (eg. single electron) systems and approximate solutions are computationally expensive, rising rapidly with the number of atoms/electrons in the system. Even with contemporary computers, this limits the system sizes that can be treated with quantum methods to 10^2 - 10^3 atoms.

In many cases we wish to study systems which comprise a much larger number of atoms. Consequently, the limitation of quantum chemistry has initiated the development of a host of computational techniques capable of simulating larger system sizes. In most of these techniques the behavior of nuclei and electrons are only considered in an averaged manner. The central concept involved is the ability to decompose the phenomena of interest into discrete size scales and model the system to a granularity commensurate with the phenomena of interest (Figure 1.1). For example, the lipid membranes comprising biological cell and organelle walls have often been treated as elastic sheets to study their undulations. In this case individual nuclei and electrons are unimportant, other than their contribution to the flexibility and compressibility of the membrane;

molecules which interact strongly will result in a stiffer membrane, which can be accounted for through the use of bending and stretching moduli as a part of the elastic model.

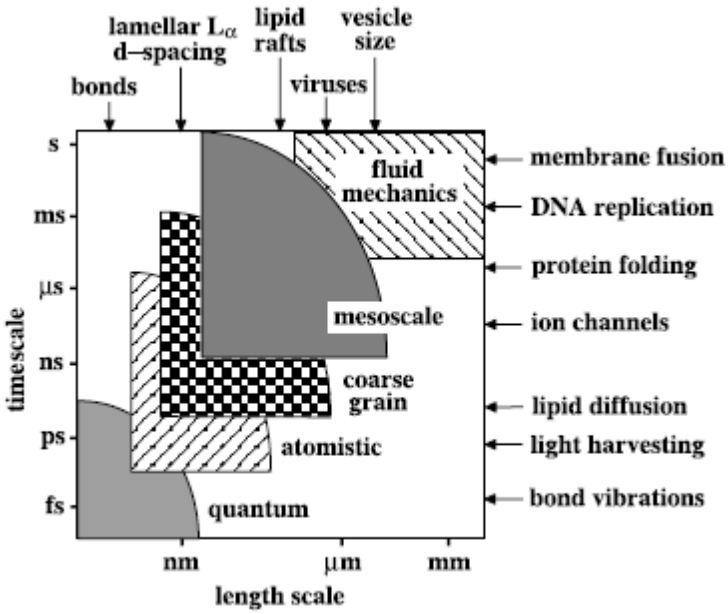


Figure 1.1 A schematic diagram of temporal and spatial scales accessible by simulation techniques. (Figure taken without modification from Nielsen's paper) [2]

The computational method suitable to study a system depends on the phenomena we are interested in studying. Figure 1.1 shows the length and time scales related to some chosen physical processes and common simulation techniques that may be used to study them. The nature of the interactions also helps in this decomposition because the length scale of interactions tends to follow the timescale of processes, allowing for example, the study of molecular vibrations using quantum calculations in the femto to pico-second length scale and vesicle fusion using continuum models in the milli-second scale.

1.2 Molecular Dynamics

Molecular dynamics (MD) is a computer simulation method which mainly uses classical mechanics (sometimes can also use quantum mechanics) to evolve a many-bodied system with time and, through statistical mechanics, enables the evaluation of equilibrium and transport properties.[3] Although molecules do not strictly follow classical mechanics, for many applications we are not interested in electron wave-functions generated by quantum mechanics, but rather the resultant interactions arising between atoms and the consequent molecular behavior.

Conceptually, molecular dynamics consists of repeating steps of evaluating the force on every atom, and moving them in space according to Newton's second law, F = ma. The application of classical mechanics to molecular systems relies on two main assumptions. The first assumption is the applicability of a force field, which maps atomic coordinates to a potential energy value. Common force fields are built on models which treat atoms as beads and bonds as springs and decompose the total potential energy into independent terms. The force field determines the physics of the systems and consequently the correlation between simulations and real systems. Due to their importance a comprehensive discussion of force fields is given in Section 1.3. The second assumption is the applicability of Newton's second law in simulating dynamics which is discussed in the following section.

1.2.1 Time Evolution

Integral to the method of molecular dynamics is the ability to approximate continuous movement through the use of small discrete time steps, wherein the system coordinates are updated. The time step, δt, must be adequately small so the change in force experienced by the atoms over a time step is negligible. While this assumption holds, at any given time the application

of a force field together with information of the molecular topology of a system yields the potential energy U(r), which can be used to calculate the force acting on each atom, $F = -\nabla U(r)$. These relations in conjunction with a numerical integrator allow the simulation of a many-bodied system. The most common integrators for molecular dynamics are based on a truncated Taylor expansion of spatial position,[4]

$$r(t + \delta t) = 2r(t) - r(t - \delta t) + \delta t^2 \ddot{r}(t) \tag{1.2}$$

where δt, $r(t), \dot{r}(t),$ and $\ddot{r}(t)$ represent integration time step, current position, velocity, and acceleration, respectively. This expression is usually cast as a 'velocity Verlet' algorithm, which comprises four simple steps,

i. Update positions: $r(t + \delta t) = r(t) + \delta t \dot{r}(t) + \frac{1}{2}\delta t^2 \ddot{r}(t)$

ii. Calculate velocities at half step: $\dot{r}\left(t + \frac{1}{2}\,\delta t\right) = \dot{r}(t) + \frac{1}{2}\,\delta t \ddot{r}(t)$

iii. Evaluate forces and acceleration at $(t + \delta t)$

iv. Calculate new velocities at full step: $\dot{r}(t + \delta t) = \dot{r}\left(t + \frac{1}{2}\,\delta t\right) + \frac{1}{2}\,\delta t \ddot{r}(t + \delta t)$

Essentially, molecular dynamics is carried out by iterating these steps repeatedly. Most codes give the option of writing out the trajectory of the system, i.e. the coordinates of atoms in the system, intermittently. This allows for visualization and analysis after the completion of the MD run.

The time represented by a simulation is simply the number of time steps multiplied by the time step interval. The choice of the time step used is dictated by the gradient of the potential energy surface of the system. For atomic systems with relatively steep changes in potential energy with atomic positions, a time step of 1 - 2 fs is common. However, for systems with softer potentials such as coarse grained systems a larger time step (in the range of ~10 - 40 fs) can safely be used while conserving energy. Furthermore, multi-time step integrators are also common, where

the interactions which change rapidly with position are integrated over small steps, while the more slowly changing interactions are integrated over longer time steps, and therefore less often.[2]

1.2.2 Periodic Boundary Conditions

In computer simulations the molecules are localized to a spatial region designated as a simulation 'box' or 'cell'. Due to computer resources being finite, we are limited to modeling a finite box size. In most cases these system sizes do not approach experimental size scales. Consequently, simulation cells have a high ratio of surface atoms to bulk atoms, leading to unnatural behavior.

This problem is avoided by enforcing periodic boundary conditions, where the simulation cell is replicated throughout space to form an infinite lattice.[4] The positions of the atoms in these replicas are not stored, but mirror the movement of atoms in the central box. As an atom or molecule moves through the edge of the central simulation box, the corresponding atom or molecule from a neighboring cell moves in from the opposite side. The mass and density of the simulation cell is conserved, and furthermore because atoms at the edge of the central simulation cell feel a homogeneous environment due to replica cells, no edges are present in the simulation and the system models an infinitely extended 'bulk' system.

The disadvantage of using periodic boundaries arises from the artificial constraints they can impose upon the system. Particularly in the presence of long ranged interactions a periodic system may induce artificial forces or suppress natural ones. For example, in the simulation of lipid membranes the size of the simulation cell limits the longest wavelength of the undulation modes. In small simulation cells the presence of periodic boundaries may even affect the structuring of liquid[4] due to indirect interactions which span the length of the box.

1.2.3 Control of Temperature and Pressure

Since most of the fundamental formulations of molecular dynamics obey Newton's laws, they conserve energy. In this sense they can be thought of as sampling from the microcanonical (NVE) ensemble. However, for better correlation with experimental conditions MD techniques have been developed to sample from the isothermal-isobaric (NPT), canonical (NVT) and even the grandcanonical (μVT) ensembles. The use of thermostats and barostats is particularly important for chemistry applications where experiments are carried out under constant temperature and pressure conditions.

The control of temperature in simulations is performed through the coupling of the simulation cell with an imaginary external heat bath. Temperature is a measure of the kinetic energy of the molecules within a system, which can be manipulated in several ways to maintain a constant temperature. The simplest scheme of 'velocity rescaling' uniformly scales the velocities of all the molecules in the system to obtain the required temperature. Stochastic thermostatting methods aim to replicate random collisions of the system atoms with those of the heat bath, of which the Andersen scheme is a common example.[5] Presently the most widespread thermostatting method is the Nose-Hoover chain, which is the least perturbative to the natural dynamics (that is, NVE dynamics) of MD simulations. This method couples the atomic degrees of freedom with external variables which propagate with the system.[6–8]

1.3 Replica Exchange Molecular Dynamics

The replica exchange (RE) method[9-12], also known as parallel tempering, has emerged as a relatively straightforward and powerful approach that can enhance conformational sampling. The basic idea is to simulate multiple replicas of the system at different temperatures independently

using either MC or MD. Periodically, replicas attempt to exchange simulation temperatures according to a Metropolis criterion that preserves the detailed balance and ensures canonical distributions at all temperatures. The resulting random walk in the temperature space helps the replicas to escape the energy local minima and sample a wider range of conformational space. Replica exchange molecular dynamics (REMD) in particular has been successfully applied to protein simulations [13-17]. On the other hand, the true efficiency of REMD in sampling large-scale protein conformational transitions needs to be rigorously benchmarked, and the dependence of REMD simulations on the protein system and key parameters needs to be explored.

1.4 Force Fields

A force field aims to separate the contributions to the total energy of the system into physically motivated independent terms in relation to the relative position of atoms to one another. These terms usually include bond stretching, bending, and dihedrals for the bonded interactions and Coulombic and van der Waals interactions for the non-bonded interaction.

The importance of a force field is that it maps atom positions ($r \equiv x_1, y_1, z_1, ..., x_N, y_N, z_N$), or more generally interaction sites to potential energy (U(r)). Importantly, this allows one to determine the negative gradient of the potential energy as a function of particle position, i.e. the force acting on the particle, a central quantity in molecular dynamics.

The functional form of a force field can be expressed by the following equation,

$$U_{total} = U_{bond} + U_{angle} + U_{\substack{proper \\ dihedral}} + U_{\substack{improper \\ dihedral}} + U_{electrostatics} + U_{van\ der\ Waals} \quad (1.3)$$

where the total energy, U_{total} of the system is the sum of the bonded and non-bonded terms, which are discussed below.

1.4.1 Bonded Interactions

These interactions operate on atoms which are specified to be bonded in the system topology and are within three bonds of one another. The bond stretching is modeled through a harmonic potential and operates only on directly bonded atoms,

$$U_{bond} = \Sigma \frac{1}{2} k_b (r - r_0)^2 \tag{1.4}$$

where r is the distance between two bonded atoms, k_b is the bond stretching force constant, and r_0 is the equilibrium bond distance. The harmonic potential assumes that the bond is always close to its equilibrium length, and hence does not account for anharmonicity or bond breaking. Consequently, this functional form is not suitable for high energy applications or chemical reactions.

The bond bending is also modeled through a harmonic function, although this terms acts on atoms separated by two bonds,

$$U_{angle} = \Sigma \frac{1}{2} k_a (\theta - \theta_0)^2 \tag{1.5}$$

where θ is the bond angle, k_a is the angle bending force constant, and θ_0 is the equilibrium bond angle.

The proper dihedral potential is evaluated through a Fourier series,

$$U_{\substack{proper \\ dihedral}} = \Sigma k_\phi [1 + \cos(n\phi - \phi_s] \tag{1.6}$$

where ϕ is the dihedral angle, k_ϕ is the dihedral force constant, n is the multiplicity of the torsion, and ϕ_s is the phase shift.

The last bonded term is the improper dihedral function,

$$U_{\substack{improper \\ dihedral}} = \Sigma \frac{1}{2} k_\xi (\xi - \xi_0)^2 \tag{1.7}$$

which is an additional term used to enforce planarity conjugation between the four specified atoms. Here k_ξ is the improper dihedral force constant, and $(\xi - \xi_0)$ is the out-of-plane angle.

1.4.2 Non-bonded Interactions

Non-bonded interactions are treated in two parts: electrostatics modeled through Coulombic interactions, and van der Waals interactions modeled through the Lennard-Jones (LJ) function.

Electrostatic interactions are modeled through Coulombic interactions,

$$U_{electrostatics} = \Sigma \frac{q_i q_j}{r_{ij}} \tag{1.8}$$

where q_i and q_j are the effective partial atomic charges on the ith and jth atoms, and r_{ij} is the interatomic distance.

The evaluation of electrostatics in computer simulations is not straightforward due to their extended decay length combined with the use of periodic boundary conditions. Accurate evaluation of the electrostatic energy must include interactions spanning over several periodic images. Conventional approaches to efficiently solve this problem rely on splitting the Coulombic interaction into a short ranged component with a cutoff similar to the LJ cutoff and long ranged

component which is solved in inverse (Fourier) space. The first method which was widely adopted was the Ewald summation method,[3] which has since been supplanted through the use of particle-mesh methods.[3]

Lennard-Jones potentials model van der Waals interactions between two non-bonded atoms by using two opposing terms which account for the Pauli repulsion and the attraction due to dispersion.

$$U_{van\ der\ Waals} = \Sigma 4\varepsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right) \tag{1.9}$$

where r_{ij} is the interatomic distance, ε_{ij} is the interaction strength parameter and σ_{ij} is the size parameter defining the distance below which the total potential becomes repulsive. Both ε_{ij} and σ_{ij} depend on the chemical nature of the two atoms involved.

Due to the rapid decay of the attractive $1/r^6$ term the potential becomes negligible with increasing r_{ij} (within a few σ lengths), though the potential never actually reaches zero. A common practice to avoid evaluating non-bonded interactions which do not significantly contribute to the potential energy is to use a distance cutoff beyond which you do not calculate interactions.

1.5 Biomolecular Force Fields

There are numerous force fields which have been used to simulate proteins over the last couple of decades. The historical discussion of protein force fields dates back to early 1980s where molecular dynamics and Monte Carlo simulations of proteins were starting to develop. The advancement of protein force fields was not unique and they were gradually developed based on force fields which have been used in organic chemistry. Of particular importance were the

Empirical Conformational Energy Program for Peptides (ECEPP) potentials from Scheraga and co-workers[18,19] and the Consistent Force Field (CFF) developments from the Lifson group.[20-23] They initiated the development of potential energy functions in the general area of organic chemistry.[24,25]

1.5.1 The Amber Force Fields

In early 1980s, enough experience had accumulated with earlier parameterizations to begin fairly systematic projects to develop a new generation of force fields. The earliest of these efforts were still done at a time when the limited power of computers made it attractive to not include all hydrogen atoms as explicit force centers. The importance of hydrogen bonding, however, led many investigators to adopt a compromise whereby polar hydrogens were explicitly represented but hydrogens bonded to carbon were combined into united atoms. A widely used force field at this level was developed in 1984 in the Kollman group[26] and incorporated into the Amber molecular mechanics package, which was at an early stage of development as well.[27] The key ideas in this initial work were to be used repeatedly in later efforts by this group. Charges were derived from quantum chemistry calculations at the Hartree-Fock STO-3G level, via fitting of partial atomic charges to the quantum electrostatic potential; these are generally called electrostatic potential (ESP) charges. The van der Waals terms were adapted from fits to amide crystal data by Lifson's group[28,29] and from liquid-state simulations pioneered by Jorgensen.[30] Force constants and idealized bond lengths and angles were taken from crystal structures and adapted to match normal mode frequencies for a number of peptide fragments. Finally, torsion force constants were adjusted to match torsional barriers extracted from experiment or from quantum chemistry calculations. Since it is only the total potential energy, as a function of torsion angle, that needs to agree with

the target values, and since these barriers have significant electrostatic and van der Waals interactions between the end atoms, the k values are closely coupled to the non-bonded potentials used and are hardly transferable from one force field to another.

Three problems with this polar hydrogen only approach, along with improvements in the speed of available computers, led many researchers to move to an all-atom approach. First, aromatic rings such as benzene have a significant quadrupolar charge distribution, with an effective positive charge near the hydrogens and an effective negative charge nearer to the middle of the ring. This effect can be crucial in determining the ways in which aromatic side chains in proteins interact with other groups. For example, ''T-shaped'' geometries between rings are stabilized relative to ''stacked'' geometries that optimize van der Waals interactions.[31] Also important are π-cation interactions, where positive groups are found directly above the centers of aromatic rings.[32,33] Second, the forces that affect the pseudo rotation between conformations, or ''pucker'' of five-membered aliphatic rings[34] are difficult to describe when only the heavy atoms are available as force centers. This affects only proline residues in proteins, but analogous problems involving ribose and deoxyribose in nucleic acids led momentum toward all-atom force fields. Finally, it is difficult with united atom models to make comparisons between computed and observed vibrational frequencies. An extension of the 1984 force field to an all-atom model was published in 1986, as a collaboration between the Kollman and Case groups.[35] Both the 1984 and 1986 parameter sets were primarily developed based on experience with gas phase simulations.

The continued increase in the speed of computers led the Kollman group to decide in the early 1990s that a new round of force-field development was warranted; this came to be known as the ''Cornell et al.'' or ff94 force field.[36] In addition to improvements in the parameters, a more serious attempt was made to explicitly describe the algorithm by which the parameters were

derived, so that consistent extensions could be made to molecules other than proteins.[37] This goal was not really achieved until the development almost a decade later of the antechamber program that completely automates all of the steps in the creation of an Amber-like force field for an arbitrary molecule or fragment.

A key motivation for this development was a desire to produce potentials suitable for condensed phase simulations, since the earlier work had concentrated in large part on gas phase behavior. In particular, the ways in which the optimized potentials for liquid simulations (OPLS) had been parameterized to reproduce the densities and heats of vaporization of neat organic liquids was very influential, along with recognition of the importance of having a balanced description of solute-solvent versus solvent-solvent interactions. A second point arose from the ability to use larger basis sets and fragment sizes to determine atomic charges that mimic the electrostatic potentials outside the molecule found from quantum mechanical calculations. Earlier work had established that fitting charges to the potentials at the Hartree-Fock 6-31G* level tended to overestimate bond-dipoles by amounts comparable to that in empirical water models such as SPC/E or TIP3P; such over polarization is an expected consequence of electronic polarization in liquids. Hence, the use of fitted charges at the HF/6-31G* level appeared to offer a general procedure for quickly developing charges for all twenty amino acids in a way that would be roughly consistent with the water models that were expected to be used. Tests of this idea, with liquid-state simulations of amides and simple hydrocarbons, gave encouraging results.

The actual implementation of this scheme for developing charges had to deal with two complications, which continue to plague force field developers to the present day. First, the effective charges of the more buried atoms are often underdetermined, so that charges for atoms in similar environments in different molecules might vary significantly. In effect, there are many

combinations of atomic charges that will fit the electrostatic potential almost equally well. There are a variety of ways to overcome this problem, often involving statistical techniques based on singular-value decomposition, but Bayly et al.[38,39] chose to use a hyperbolic restraint term to limit the absolute magnitude of charges on non-hydrogen atoms. This is called restrained electrostatic potential (RESP) fit and weakly favors solutions with smaller charges for buried atoms, yielding fairly consistent charge sets with little degradation in the quality of the fit to the electrostatic potential outside the molecule.

A second and more fundamental problem with the RESP procedure is that the resulting charges depend on molecular conformation, often in significant ways. This is a manifestation of electronic polarizability, which can only be described in a very averaged way if fixed atomic charges are to be used. Any real solution to this problem must involve a more complex model. The compromise chosen for the ff94 force field was to fit charges simultaneously to several conformations, in the hopes of achieving optimal averaged behavior.

Once the charges and the internal parameters for bonds and angles were available, the Lennard-Jones parameters could be established primarily by reference to densities and heats of vaporization in liquid-state simulations. Only a small number of sets of 6-12 parameters were necessary to achieve reasonable agreement with experiment. A key expansion from earlier work was the notion that parameters for hydrogens should depend in an important way on the electronegativity of the atoms they are bonded to.[40,41]

As with many other force-field projects, the final parameters to be fit were the ‘‘soft’’ torsional potentials about single bonds. It makes some sense to address these after the charges and Lennard-Jones parameters have been developed, since the energy profile for rotation about torsion angles depends importantly on the non-bonded interactions between the moving groups at the ends,

as well as on whatever intrinsic torsional potential is assigned. The question of how best to partition torsional barriers into bonded versus non-bonded interactions is a thorny one, and many developers of force fields have adopted a strictly empirical approach, fitting k, n, and δ so that the total profile matches some target extracted from quantum mechanics or from experiment.

A key set of torsional parameters are those for the ϕ and ψ backbone angles, since these affect every amino acid residue and heavily influence the relative energies of helices, sheets, and turns in proteins. The ff94 parameters were fit to representative points on the dipeptide maps for glycine and alanine, computed at the MP2 level with a triple-ζ + polarization (TZP) basis set. This is not an unreasonable choice for a target function, but it has a number of intrinsic difficulties. First, the α-helix region near ϕ, ψ = -60, -40 is not a minimum for a gas-phase dipeptide, so fitting just a representative point can lead to errors in the surface as a whole, compared to the full MP2/TZP surface. More importantly, the use of a gas phase dipeptide model as a target ignores both the non-local electronic structure contributions that would be seen in larger fragments[42] and the polarization effects inherent in a condensed phase environment.[43] Some account of the longer-range effects was provided in subsequent parameterizations, referred to as ff96[44] and ff99,[45] in which the ϕ and ψ and potentials were fit to tetrapeptide as well as dipeptide quantum mechanical conformational energies. These later fits provided potential surfaces that were significantly different from those in ff94, but it was hard to tell if physical realism was really being improved.

In recent years it has become computationally feasible to test protein potentials by carrying out converged or nearly converged simulations on short peptides and comparing the resulting conformational populations to those derived from experiment.[46-48] The experimental estimates, obtained mainly from circular dichroism or from NMR, are often only qualitative, but this can be enough to identify obvious errors in computed ensembles. For example, the ff94 parameters appear

to over-stabilize helical peptide conformers in many if not all instances. Computed melting temperatures for polyalanine helices are too high,[49] and helical conformers can predominate in simulations of sequences that experimentally form other structures, such as β-hairpins. At least two modifications of the ff94 ϕ and ψ potentials have been proposed and tested on large-scale peptide simulations.[50,51] It will be of interest to see how these ideas develop as a new generation of long time-scale peptide simulations becomes feasible.

1.5.2 The CHARMM Force Fields

As with Amber, the CHARMM (Chemistry at HARvard using Molecular Mechanics)[52] program was originally developed in the early 1980s and initially used an extended atom force field with no explicit hydrogens. By 1985, this had been replaced by the CHARMM19 parameters, in which hydrogen atoms bonded to nitrogen and oxygen are explicitly represented, while hydrogens bonded to carbon or sulfur are treated as part of extended atoms.[53,54] Key to the parameterization of this model were fits to quantum calculations at the HF/6-31G level of hydrogen bonded complexes between water and the hydrogen bond donors or acceptors of the amino acids or fragments. This involves a series of supermolecular calculations of the model compound, such as formamide or N-methylacetamide and a single water molecule at each of several interaction sites. Before making the fits, the interaction energies are scaled by a factor of 1.16, which is the ratio of the water dimerization energy predicted by the TIP3P model to that predicted at the HF/6-31G level. As in the Amber parameterizations described above, the goal here was to obtain a balanced interaction between solute-water and water-water energies when the latter are represented by TIP3P. For peptides, it was found that fitting the peptide-water interactions in this way led to peptidepeptide hydrogen bonds that were also larger than HF/6-31G values by a

factor very close to 1.16; in other cases, explicit fitting to solute-solute hydrogen bonded dimers may be needed for parameter generation.[55]

As with the contemporaneous Amber 1984 united-atom parameterization, the CHARMM19 values were developed and tested primarily on gas phase simulations. However, the CHARMM19 potential seems to do well in solvated simulations and continues to be used for peptide and protein simulations; this is in contrast to the 1984 Amber force field, which is no longer widely used. In addition, the CHARMM19 values have often been used in conjunction with a distance-dependent dielectric constant as a rough continuum solvation model.

In the early 1990s, the CHARMM development group also recognized the need to refine parameters more explicitly pointed to obtaining a good balance of interaction energies in explicit solvent simulations. The resulting CHARMM22 protein force field was first included in the corresponding version of CHARMM, released in 1992, and was fully described a few years later.[56,57] The key approach from CHARMM19 was carried over by deriving charge models primarily from fits to solute-water dimer energetics. In addition to fitting the dimer interaction energies, charges for model compounds were adjusted to obtain dipole moments somewhat larger than experimental or ab initio values. This has the same goal as the RESP procedure described earlier: bonds are expected to be more polarized in condensed phases than in the gas phase. The use of empirical charges that yield enhanced dipoles both reflects this behavior and allows a reasonably balanced set of interactions with the TIP3P water model, which has a similarly enhanced dipole moment.

Once the charges were determined by these dimer studies, the Lennard-Jones parameters were refined to reproduce densities and heats of vaporization of liquids as well as unit cell parameters and heats of sublimation for crystals. As with the Amber parameterization, generally

only small adjustments from earlier values were required to fit the empirical data. Nevertheless, because of the steep dependence of these forces, such adjustments may be crucial for a well-balanced and successful set of parameters.

As with the Amber ff94 force field, the torsional parameters were finally adjusted to target data derived from vibrational spectra and from ab initio calculations. The torsional potentials for the ϕ and ψ torsions were initially fit to HF/6-31+G* calculations on an analog of the alanine dipeptide in which the terminal methyl groups are replaced by hydrogen. These were then refined in an iterative procedure to improve the agreement with experiment of the backbone angles in simulations of myoglobin. In principle at least, this latter adjustment provides a way of correcting the ab initio dipeptide energy map for effects caused by the protein environment. As with the Amber parameterization, the question of how best to obtain good backbone torsional potentials is a vexing one, and studies are continuing, both at the dipeptide level and with solvated simulations of oligopeptides. Most recently, an extensive reworking of the nucleic acid parameters has resulted in the CHARMM27 force field.[58] However, the CHARMM27 protein parameters are essentially identical to those from the CHARMM22 force field.

One feature of the CHARMM parameterizations is the enforcement of neutral groups, which are small sets of contiguous atoms whose atomic charges are constrained to sum to zero. For example, charges for the C and O atoms of the peptide group form a small neutral group. These groups can be useful when truncating long-range electrostatic interactions: if an entire group is either included or ignored, then there is never any splitting of dipoles. Ignoring charged side chains, each atom would then feel the electrostatic effects of a net neutral environment. The same behavior occurs with solvent molecules, if the interactions of a given water molecule are always treated as a group. Although it was long deemed plausible that such a group-based truncation scheme would

yield better results than an atom based scheme, this is probably not the case for most biomolecular simulations in water.[59-62] In any event, such considerations are now much less important than in earlier times, since many current simulations use Ewald or fast multipole schemes to handle long-range electrostatics, where nothing is gained by having small neutral groups.

1.5.3 The OPLS Force Fields

A third main development in the early 1980s involved potentials developed by Jorgensen and co-workers to simulate liquid state properties, initially for water and for more than fourty organic liquids. These were called OPLS (Optimized Potentials for Liquid Simulations) and placed a strong emphasis on deriving non-bonded interactions by comparison to liquid-state thermodynamics.[63] Indeed, the earliest applications of OPLS potentials were to rigid molecule Monte Carlo simulations of the structure and thermodynamics of liquid hydrogen fluoride.[64] The reproduction of densities and heats of vaporization provides some confidence in both the size of the molecules and in the strengths of their intermolecular interactions. These early models treated hydrogens bonded to aliphatic carbons as part of an extended atom but represented all other hydrogens explicitly.

The initial applications to proteins[65-67] used a polar-hydrogen–only representation, taking the atom types and the valence (bond, angle, dihedral) parameters from the 1984 Amber force field. This was called the AMBER/OPLS force field, and for some time was reasonably popular. As with Amber and CHARMM, an all-atom version (OPLS-AA) was developed later, but with much the same philosophy for derivation of charges and van der Waals parameters from simulations on pure liquids.[68-70] Torsional parameters were developed in a consistent way by fits to HF/6-31G* energy profiles,[71] along with some recent modifications, especially for charged side

chains.[72] Bond stretching and angle bending terms were standardized but were largely taken from the 1986 Amber all-atom force field. The parameter choices were intended to be ‘‘functional group friendly,’’ so that they could be easily transferred to other molecules with similar chemical groupings. Although the parameters were principally derived with reference to condensed phase simulations, comparisons to gas-phase peptide energetics also show good results.[73]

1.5.4 Other Biomolecular Force Fields

There are several other protein potentials that have been widely used. The GROMOS force fields[74] were developed in conjunction with the program package of the same name.[75, 76] The all-atom CEDAR and GROMACS force fields are largely derived from GROMOS. The Merck Molecular Force Field (MMFF) was developed by Halgren, [77-83] and has been aimed more at drug-like organic compounds than at proteins. MMFF was not derived for use in bulk phase simulations and performs poorly when used to model organic liquids.[84] This deficiency is not inherent in the buffered 14-7 function[85] used in MMFF’s van der Waals term, because this same functional form can be reparameterized to fit liquid data.[86] The DISCOVER force field[87] has seen use primarily in conjunction with the commercial INSIGHT modeling package. The MM3 and MM4 potentials for amides[88, 89] are an offshoot of Allinger’s highly respected molecular mechanics parameterizations and have been applied primarily to peptides. These MM methods use atomic charges only at formally charged groups, and rely on bond dipole moments to provide for most electrostatic interactions. A series of potentials refined over many years in Levitt’s group[90-93] are incorporated in the ENCAD (ENergy Calculation And Dynamics) program and have been notably used to study protein folding and unfolding. [94] The ENCAD potential is unique in its use of group-based, rather than atom based, neighbor exclusion of short-range electrostatic interactions.

It also uses pairwise non-bonded potentials shifted to zero energy at short range, and specifically parameterized to reflect these small cutoff distances.

1.6 Weaknesses of the Available Force Fields

Although there are quite a few state-of-the-art force fields, still there are some problems associated with them. A possible avenue for improvement involves the solvation interactions. It is believed that part of the force field inaccuracies can be traced to the approximate treatment of polarization effects using effective partial atomic charges, which leads to an imbalance between the solute-solute and solvent-solvent interactions due to an underestimation of the solute-solvent interactions.[95-99] Most effective charge distributions for molecules are provided by gas phase quantum mechanical calculations, rather than the more appropriate condensed phase calculations which are more expensive. Gas phase calculations only provide the permanent multipole moments with no solvation interactions involved. Unfortunately, the ignored solvation effect can lead to significant changes to the charge distribution which should not be ignored. Hence, most empirical force fields provide only an approximate representation of the molecular polarity in condensed phases. This severely limits the reliability and predictability of molecular properties in biological systems. Therefore, a simple and highly accurate description of the charge distribution in solution is required.

One of the possible developments in force field design is the use of explicit polarization approaches to achieve more accurate results. In principle, this should provide more realistic and accurate results than non-polarizable force fields. However, the additional computational cost and difficulty of finding a unique method to treat pair-wise polarizable interactions has been problematic. [83,100-105] Thus, non-polarizable force fields are still the most popular and widely used

approach. In contrast, non-polarizable force field developers have tried to simply rescale charge distributions in order to distinguish between the gas and condensed phases, but it has been suggested that the electronic rearrangements occurring in the solvation process are far more complicated than provided by simple scaling from the gas phase.

1.7 Kirkwood-Buff (KB) Theory

The Kirkwood-Buff (KB) theory of solutions was proposed by Kirkwood and Buff in 1951.[106] It can be applied to any kind of solutions over the entire range of compositions. It is an exact theory with no approximations, which makes it more valid than other theories.[107] Moreover, the KB theory provides a direct relationship between molecular distributions at the atomic level and bulk thermodynamic properties such as partial molar volume, chemical potential and compressibility. Furthermore, Ben-Naim later developed the inversion procedure of KB theory,[108] providing information about the affinity between a pair of species in the solution mixture from experimental thermodynamic properties. With time, the KB theory has become more popular and it has been widely used by many scientists to a variety of processes.[95, 109-152] In addition, many chemists and physicists are continually developing KB theory and applying it to study solution mixtures.[110, 115, 116, 121-148, 153-169]

The relative distribution of particles in a system can be expressed using radial distribution functions. A radial distribution function (rdf), *g(r)*, provides the probability of finding a particle at a distance *r* around a central particle. It describes how the solution density varies as a function of the distance. In a closed system with *N* particles in a volume *V* and at a temperature *T*, the probability that particle 1 is in dr_1 at r_1 and particle 2 is in dr_2 at r_2 can be expressed using Boltzmann distribution as,[170-172]

$$P(r_1, r_2) = \frac{\iint \dots \int e^{-\beta U_N} dr_3 dr_4 \dots dr_N}{Z_N} \tag{1.10}$$

where β = 1/kT, U_N is the potential energy of N particles, and Z_N is known as the configurational integral. Consequently, the probability that any particle is in dr_1 at r_1, and any particle is in dr_2 at r_2, can be written as,

$$\rho(r_1, r_2) = \frac{N!}{(N-2)!} P(r_1, r_2) \tag{1.11}$$

Moreover, the probability of finding a particle anywhere in the system could be generally expressed as,

$$\frac{1}{V} \int \rho(r_1)\, dr_1 = \rho = \frac{N}{V} \tag{1.12}$$

Therefore, *g(r)* can be introduced as,

$$\rho(r_1, r_2) = \rho^2 g(r_1, r_2) \tag{1.13}$$

which is provided by combining equations (1.10), (1.11), (1.12) and (1.13),

$$g(r_1, r_2) = \frac{V^2 N!}{N^2 (N-2)!} \frac{\iint \dots \int e^{-\beta U_N} dr_3 dr_4 \dots dr_N}{Z_N} \tag{1.14}$$

Figure 1.2 shows a typical radial distribution function (rdf) obtained from a simulation. It starts from zero at short distances due to the strong repulsion between two particles. Then it typically displays a series of fluctuations around *g(r)* = 1, which are generally known as solvation shells. The first peak, which is also the largest one, indicates that one is more likely to find a

particle at this distance, compared to other distances, with respect to that expected for a random bulk solution distribution. As the distance r increases, the distribution of components approaches unity, which indicates a random bulk solution distribution. On the other hand, radial distribution functions can also be obtained from experiments using X-ray diffraction studies of solutions.

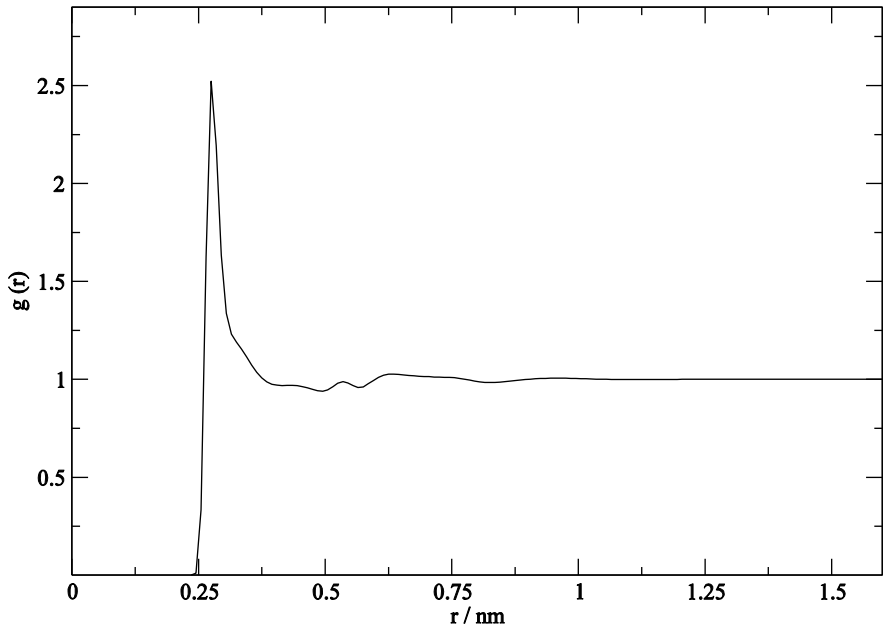


Figure 1.2 An example of a rdf as a function of the distance.

The integration of a radial distribution function between two different species i and j provides a property called the Coordination Number which is given by,

$$CN(i,j) = \rho_j \int_0^R g_{ij}(r) 4\pi r^2 \, dr \tag{1.15}$$

to a distance R from the central molecule.

The radial distribution function provides insight into the liquid structure. The corresponding integrals over g(r), which are called Kirkwood-Buff Integrals (KBIs), are useful to

express thermodynamic properties of solution mixtures, such as compressibilities, partial molar volumes and derivatives of the chemical potentials.[95, 126, 128-130] Hence, combinations of KBIs provide a link between thermodynamic properties and molecular distribution functions for multi-component systems and KBIs are expressed by,

$$G_{ij} = 4\pi \int_0^\infty \left[g_{ij}^{\mu VT}(r) - 1 \right] r^2 dr \tag{1.16}$$

where G_{ij} is the KBI between species i and j, $g_{ij}^{\mu VT}$ is the corresponding radial distribution function in the μVT ensemble, r is the corresponding center of mass-to-center of mass distance. Thus, the theory may be used to compute the thermodynamic quantities of the pair correlation function.

Furthermore, a property called the excess coordination number, N_{ij} can be defined from the KBIs according to,

$$N_{ij} = \rho_j G_{ij} \tag{1.17}$$

where ρ_j is the number density (molar concentration) of species j.

$$\rho_j = \frac{N_j}{V} \tag{1.18}$$

A value of N_{ij} greater than zero indicates an excess of species j in the vicinity of species i over a random distribution, while a negative value corresponds to a depletion of species j surrounding i. In other words, a positive N_{ij} can be interpreted as net favorable (attractive) interactions between species i and j, and a negative N_{ij} is related to net unfavorable (repulsive) interactions. Generic examples of G_{ij} and N_{ij} are illustrated in Figure 1.3 and Figure 1.4, respectively. They provide a sensitive test of the relative distribution of the different species in solution.[130]

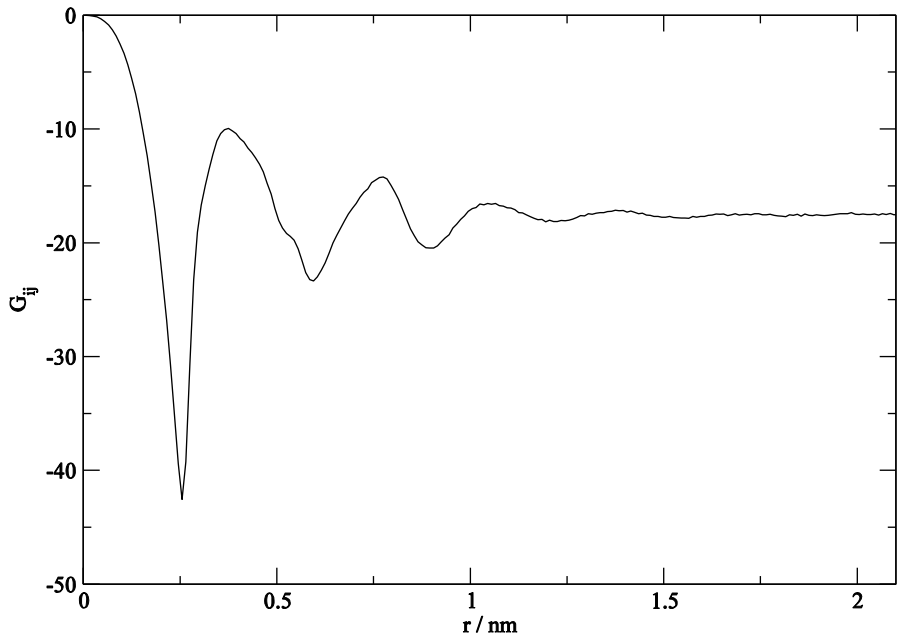


Figure 1.3 An example of a KB integral as a function of integration distance.

In the KB theory of solutions thermodynamic properties of a solution mixture can be derived from radial distribution functions, and vice versa. Hence, KBIs can be determined either from experimental or simulated data. For solution mixtures with water and solute at constant pressure (p) and temperature (T), the chemical potentials (μ_i), partial molar volumes (V_i), and isothermal compressibilities (κ_T) can be obtained experimentally. Then the experimental data can be used to determine KBIs.[144] KB theory can also be applied to biomolecular systems, as well as cosolvent systems to analyze the free energy of molecular binding and characterize the preferential interactions and other thermodynamic properties. In a system of a biomolecule and cosolvent with primary solvent of water (1), the preferential binding parameters can be obtained from equilibrium dialysis experiments and also expressed using KBIs.[123]

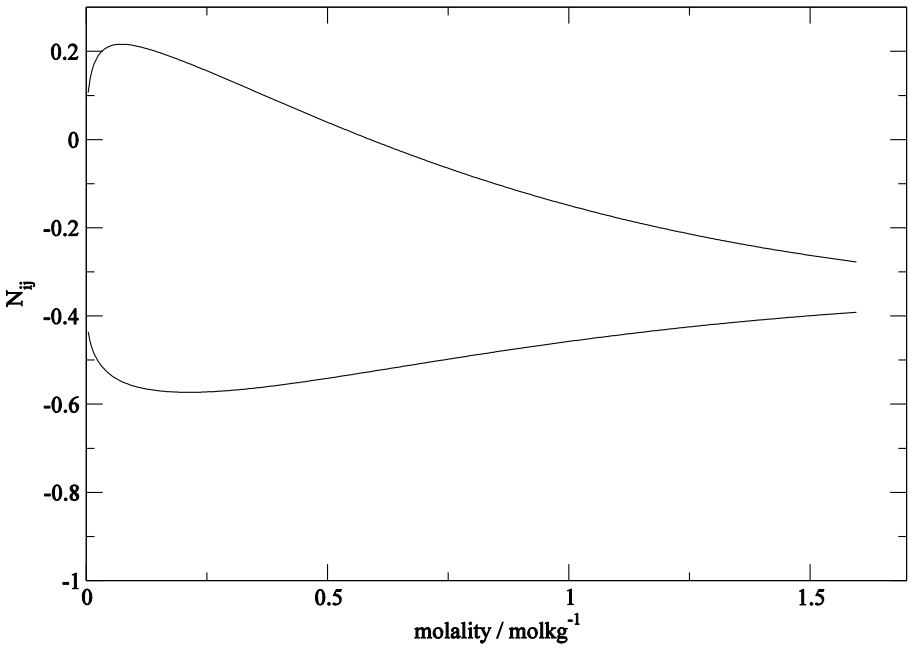


Figure 1.4 An example of excess coordination number as a function of composition.

1.8 Kirkwood-Buff Derived Force Field

The key to an accurate biomolecular simulation is to develop high quality force fields for proteins. It has been observed that currently available force fields tend to over stabilize secondary structure;[173] some are alpha-helix heavy and some may be biased towards beta-sheet. Significant effort to develop and improve force fields has been performed,[174-181] but current force fields can still be improved. In particular, they struggle to reproduce some common physical properties.[95, 96, 126, 129]

KB theory is a powerful tool that can be used to evaluate the ability of a force field to correctly represent relative molecular distribution in solutions. It is an exact theory of solution mixtures and valid for the analysis of both experimental and theoretical solvation quantities with no limitations to the size or character of molecules. The quality of a force field used for simulation can be easily determined by comparing KBIs derived from simulated data to those extracted from

the experimental data. In addition, the KBIs are more sensitive to the parameters from force fields than many other thermodynamic properties, [95, 126, 128-130] which provides a solid basis for judging the accuracy of a particular force field. For instance, the KB integrals are directly related to the molecular affinity information which is a consequence of the interactions among the atoms.

However, many existing force fields perform poorly in their ability to reproduce the experimental KB integrals.[153] This indicates that currently used force fields do not correctly reproduce the solution distributions,[153] and this can lead to inaccurate simulation results. Therefore, it is necessary to develop an improved force field which can truly represent the correct molecular distributions in a solution mixture, and thereby maintain a reasonable balance between solute-solute interactions and solute-solvent interactions. This is the aim of the Kirkwood-Buff derived force field (KBFF) approach. During the past several years the Smith group has been developing Kirkwood-Buff derived force fields as a central aspect of their work. The only major difference to other similar biomolecular force fields is the origin of the effective charge distributions. Other parameters are similar to most common force fields. Moreover, standard bond lengths and bond angles are obtained from experimental data for crystal structures, and other bonded parameters are taken directly from the GROMOS96.[182] The general non-bonded form of the KB force field contains a Lennard-Jones (LJ) 6-12 potential and a Coulomb interaction. The molecular charge is explored thoroughly during the parameterization process, while the van der Waals parameters for hydrocarbons were taken from elsewhere.[182] It has been shown that simulation results from the KBFF models perform fairly well and can be even better than other common force fields with similar computational cost. [183-185]

Below is a list of recent publications regarding our force fields development using the Kirkwood-Buff theory of solutions.

Table 1.1 KBFF models which have been published.

Solute	Solvent	Reference
Acetone	water	Weerasinghe and Smith
Urea	water	Weerasinghe and Smith
NaCl	water	Weerasinghe and Smith
GdmCl	water	Weerasinghe and Smith
Methanol	water	Weerasinghe and Smith
Amides	water	Kang and Smith
Thiols and sulfides	methanol	Bentenitis et al.
Aromatics, Heterocycles	methanol, water	Ploetz and Smith
Alkali halides	water	Gee et al.

Here we continue this research to provide a full force field capable of simulations of peptides and proteins in a variety of solutions.

1.9 Summary and Organization of the Dissertation

Molecular dynamic simulations have played a key role in the study of biological systems and provide information at the atomic level which is not available experimentally. Kirkwood-Buff theory can be used to interpret experimental and computational data and to provide a bridge between them. Here, we use KB theory and computer simulations for a variety of applications.

In Chapter 2 we have provided a rigorous framework for the analysis of experimental and simulation data concerning open and closed multicomponent systems using the KB theory of solutions. The results are illustrated using computer simulations for various concentrations of the

solutes Gly, Gly_2 and Gly_3 in both open and closed systems, and in the absence or presence of NaCl as a cosolvent.

In Chapter 3 we have attempted to quantify the interactions between amino acids in aqueous solutions using the KB theory of solutions. The results are illustrated using computer simulations for various concentrations of the twenty zwitterionic amino acids at ambient temperature and pressure.

In Chapter 4 amino acids were studied at higher temperatures and pressures and the results are discussed in terms of the preferential (solute over solvent) interactions between the amino acids.

In Chapter 5 we have described our most recent efforts towards a complete force field for peptides and proteins. The results are illustrated using molecular dynamic simulations of several tripeptides, selected peptides and selected globular proteins at ambient temperature and pressure followed by replica exchange molecular dynamic simulations of a few selected peptides.

1.10 References

1. Cramer, C. J. Essentials of Computational Chemistry (2nd ed.), Wiley, 2009.

2. Nielsen, S. O.; Lopez, C. F.; Srinivas, G.; Klein, M. L. *J. Phys.: Condens. Mat.* **2004**, 16, R481–R512.

3. Frenkel, D.; Smit, B. Monte Carlo simulations. *In: Understanding Molecular Simulation*: *From Algorithms to Applications* (Frenkel, D., Klein, M., Parrinelo, M., Smit, B., Eds.), Academic Press. San Diego, 2002

4. Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*, Oxford Science Publication, Oxford, 2009.

5. Andersen, H. C. *J. Chem. Phys*. **1980**, 72, 2384–2393.

6. Nose, S. *Mol. Phys*. **1984**, 52, 255–268.

7. Hoover, W. G. *Phys. Rev. A* **1985**, 31, 1695–1697.

8. Martyna, G. J.; Klein, M. L. *J. Chem. Phys*. **1992**, 97, 2635–2644.

9. Hukushima, K.; Nemoto, K. *J. Physical Society of Japan*, **1996**, 65, 1604–1608.

10. Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, 314, 141–151.

11. Swendsen, R. H.; Wang, J. S. *Phys. Rev. Lett.* **1986**, 57, 2607–2609.

12. Hansmann, U. H.; Okamoto, Y. *Curr. Opin. Struct. Biol*. **1999**, 9, 177–83.

13. Earl, D. J.; Deem, M. W. *Phys. Chem. Chem. Phys.* **2005**, 7, 3910–3916.

14. Gallicchio, E.; Levy, R. M.; Parashar, M. *J. Comput. Chem.* **2008**, 29, 788–94.

15. Nymeyer, H.; Gnanakaran, S.; Garcia, A. E. *Methods in Enzymology* **2004**, 383, 119–49.

16. Roitberg, A. E.; Okur, A.; Simmerling, C. *J. Phys. Chem. B*, **2007**, 111, 2415–2418.

17. Zhang, W.; Wu, C.; Duan, Y. *J. Chem. Phys.* **2005**, 123, 54105-9.

18. Momany, F. A.; McGuire, R. F.; Burgess, A. W.; Scheraga, H. A. *J. Phys. Chem.* **1975**, 79, 2361–2381.

19. Nemethy, G.; Pottle, M. S.; Scheraga, H. A. *J. Phys. Chem.* **1983**, 87, 1883–1887.

20. Lifson, S.; Warshel, A. *J. Chem. Phys.* **1969**, 49, 5116–5129.

21. Hagler, A.; Euler, E.; Lifson, S. *J. Am. Chem. Soc.* **1974**, 96, 5319–5327.

22. Hagler, A.; Lifson, S. *J. Am. Chem. Soc.* **1974**, 96, 5327–5335.

23. Niketic, S. R.; Rasmussen, K. *The Consistent Force Field: A Documentation* Springer-Verlag, New York. 1997

24. Allinger, N. L. *Adv. Phys. Org. Chem.* **1976**, 13, 1-82.

25. Burkert, U.; Allinger, N. L. *Molecular Mechanics*, American Chemical Society, Washington, D.C. 1982.

26. Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S. Jr.; Weiner, P. *J. Am. Chem. Soc.* **1984**, 106, 765–784.

27. Weiner, P. K.; Kollman, P. A. *J. Comput. Chem.* **1981**. 2, 287–303.

28. Hagler, A.; Euler, E.; Lifson, S. *J. Am. Chem. Soc.* **1974**, 96, 5319–5327.

29. Hagler, A.; Lifson, S. *J. Am. Chem. Soc.* **1974**, 96, 5327–5335.

30. Jorgensen, W. L. *J. Am. Chem. Soc.* **1981**, 103, 335–340.

31. Williams, D. E. *Rev. Comput. Chem.* **1991**, 2, 219–271.

32. Dougherty, D. A. *Science* **1995**, 271, 163–168.

33. Zacharias, N.; Dougherty, D. A. *Trends Pharm. Sci.* **2002**, 23, 281–287.

34. Tomimoto, M.; Go, N. *J. Phys. Chem.* **1995**, 99, 563–577.

35. Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. *J. Comput. Chem.* **1986**, 7, 230–252.

36. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M. Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, 117, 5179–5197.

37. Fox, T.; Kollman, P. A. *J. Phys. Chem.* B **1998**, 102, 8070–8079.

38. Bayly, C.; Cieplak, P.; Cornell, W.; Kollman, P. A. *J. Phys. Chem.* **1993**, 97, 10269–10280.

39. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Kollman, P. A. *J. Am. Chem. Soc.* **1993**, 115, 9620–9631.

40. Gough, C.; DeBolt, S.; Kollman, P. A. *J. Comput. Chem.* **1992**, 13, 963–970.

41. Veenstra, D.; Ferguson, D.; Kollman, P. A. *J. Comput. Chem.* **1992**, 13, 971–978.

42. Beachy, M. D.; Chapman, D.; Murphy, R. B.; Halgren, T. A.; Friesner, R. A. *J. Am. Chem. Soc.* **1997**, 119, 5908–5920.

43. Osapay, K.; Young, W. S.; Bashford, D.; Brooks, C. L. III; Case, D. A. *J. Phys. Chem.* **1996**, 100, 2698–2705.

44. Kollman, P.; Dixon, R.; Cornell, W.; Fox, T.; Chipot, C.; Pohorille, A. In: *Computer Simulations of Biomolecular Systems,* Vol. 3 (W. F. van Gunsteren, P. K. Weiner, and A. J. Wilkinson, Eds.), 83–96. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997.

45. Wang, J.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, 21, 1049–1074.

46. Damm, W.; van Gunsteren, W. F. *J. Comput. Chem.* **2000**, 21, 774–787.

47. Mitsutake, A.; Sugita, Y.; Okamoto, Y. *Biopolymers* **2001**, 60, 96–123.

48. Garcia, A. E.; Sanbonmatsu, K. Y. *Proteins* **2001**, 42, 345–354.

49. Garcia, A. E.; Sanbonmatsu, K. Y. *Proc. Natl. Acad. Sci. USA* **2002**, 99, 2782–2787.

50. Garcia, A. E.; Sanbonmatsu, K. Y. *Proc. Natl. Acad. Sci. USA* **2002**, 99, 2782–2787.

51. Simmerling, C.; Strockbine, B.; Roitberg, A. E. *J. Am. Chem. Soc.* **2002**, 124, 11258–11259.

52. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* 4, **1983**, 187–217.

53. Reiher, W. E., III. *Theoretical Studies of Hydrogen Bonding*, Ph.D. thesis, Harvard University, 1985.

54. Neria, E.; Fischer, S.; Karplus, M. *J. Chem. Phys.* **1996**, 105, 1902–1921.

55. MacKerell, A. D., Jr. (2001). In: *Computational Biochemistry and Biophysics* (O. Becker, A. D. MacKerell, Jr., B. Roux, and M. Watanabe, Eds.), 7–38. Marcel Dekker, New York. 2001.

56. MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunback, R. L., Jr.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E. III.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wio´rkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, 102, 3586–3616.

57. MacKerell, A. D., Jr. In: *Computational Biochemistry and Biophysics* (O. Becker, A. D. MacKerell, Jr., B. Roux, and M. Watanabe, Eds.), 7–38. Marcel Dekker, New York, 2001.

58. Foloppe, N.; MacKerell, Jr., A. D. *J. Comput. Chem.* **2000**, 1, 86–104.

59. Steinbach, P. J.; Brooks, B. R. *J. Comput. Chem.* **1994**, 15, 667–683.

60. Darden, T.; Pearlman, D.; Pedersen, L. G. *J. Chem. Phys.* **1998**, 109, 10921–10935.

61. Darden, T. A. In: *Computational Biochemistry and Biophysics* (O. Becker, A. D. MacKerell, Jr., B. Roux, and M. Watanabe, Eds.), 91–114. Marcel Dekker, New York, 2001

62. Mark, P.; Nilsson, L. *J. Comput. Chem.* **2002**, 23, 1211–1219.

63. Jorgensen, W. L. In*: Encyclopedia of Computational Chemistry* (P. von R. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, and H. F. Schaefer III, Eds.), 1986–1989. John Wiley and Sons, Chichester, 1998.

64. Jorgensen, W. L. *J. Am. Chem. Soc.* **1981**, 103, 335–340.

65. Jorgensen, W. L.; Swenson, C. J. *J. Am. Chem. Soc.* **1985**, 107, 569–578.

66. Jorgensen, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1988**, 110, 1657–1671.

67. Tirado-Rives, J.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1990**, 112, 2773–2781.

68. Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, 118, 11225–11236.

69. Rizzo, R. C.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1999**, 121, 4827–4836.

70. Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, 105, 6474–6487.

71. Maxwell, D. W.; Tirado-Rives, J.; Jorgensen, W. L. *J. Comput. Chem.* **1995**, 16, 984–1010.

72. Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, 105, 6474–6487.

73. Beachy, M. D.; Chapman, D.; Murphy, R. B.; Halgren, T. A.; Friesner, R. A. *J. Am. Chem. Soc.* **1997**, 119, 5908–5920.

74. van Gunsteren, W. F.; Daura, X.; Mark, A. E. *In: Encyclopedia of Computational Chemistry* (R. von, P. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, and H. F. Schaefer III, Eds.), 1211–1216. John Wiley, Chichester, 1998.

75. van Gunsteren, W. F.; Berendsen, H. J. C. *Groningen Molecular Simulation (GROMOS) Library Manual,* University of Groningen, The Netherlands, 1987.

76. Scott, W. R.; Hu¨nenberger, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennen, J.; Torda, A. E.; Huber, T.; Kru¨ ger, P.; van Gunsteren, W. F. *J. Phys. Chem.* A **1999**, 103, 3596–3607.

77. Halgren, T. A. *J. Comput. Chem.* **1996**, 17, 490–519.

78. Halgren, T. A. *J. Comput. Chem*. **1996**, 17, 520–552.

79. Halgren, T. A. *J. Comput. Chem*. **1996**, 17, 553–586.

80. Halgren, T. A. *J. Comput. Chem*. **1996**, 17, 616–641.

81. Halgren, T. A.; Nachbar, R. B. *J. Comput. Chem*. **1996**, 17, 587–615.

82. Halgren, T. A. *J. Comput. Chem.* **1999**, 20, 720–729.

83. Halgren, T. A. *J. Comput. Chem.* **1999**, 20, 730–748.

84. Kaminski, G.; Jorgensen, W. L. *J. Phys. Chem.* **1996**, 100, 18010–18013.

85. Halgren, T. A. *J. Am. Chem. Soc*. **1992**, 114, 7827–7843.

86. Ren, P.; Ponder, J. W. *J. Phys. Chem. B* **2003**, 107, 5933–5947.

87. Maple, J. R.; Hwang, M.-J.; Jalkanen, K. J.; Stockfisch, T. P.; Hagler, A. T. *Comput. Chem.* **1998**, 18, 430–458.

88. Lii, J. H.; Allinger, N. *J. Comput. Chem*. **1991**, 12, 186–199.

89. Langley, C. H.; Allinger, N. L. *J. Phys. Chem. A* **2002**, 106, 5638–5652.

90. Levitt, M. *J. Mol. Biol*. **1983**, 168, 595–620.

91. Levitt, M.; Sharon, R. *Proc. Natl. Acad. Sci*. **1988**, 85, 7557–7561.

92. Levitt, M.; Hirshberg, M.; Sharon, R.; Daggett, V. *Comp. Phys. Commun*. **1995**, 91, 215–231.

93. Levitt, M.; Hirschberg, M.; Sharon, R.; Laidig, K. E.; Daggett, V. *J. Phys. Chem. B* **1997**, 101, 5051–5061.

94. Daggett, V. *Acc. Chem. Res.* **2002**, 35, 422–429.

95. Weerasinghe, S.; Smith, P. E. *J. Phys. Chem. B* **2003**, 107, 3891-3898.

96. Perera, A.; Sokolic, F. *J. Chem. Phys.* **2004**, 121, 11272-11282.

97. Okur, A.; Strockbine, B.; Hornak, V.; Simmerling, C. *J. Comput. Chem.* **2003**, 24, 21-31.

98. Mu, Y. G.; Kosov, D. S.; Stock, G. *J. Phys. Chem*. B **2003**, 107, 5064-5073.

99. Mazur, A. K. *J. Am. Chem. Soc.* **2003**, 125, 7849-7859.

100. Gao, J. L. *J. Phys. Chem*. B **1997**, 101, 657-663.

101. Xie, W.; Gao, J. *J. Chem. Theory Comput.* **2007**, 3, 1890-1900.

102. Xie, W.; Pu, J.; Mackerell, A. D.; Gao, J. *J Chem. Theory Comput.* **2007**, 3, 1878-1889.

103. Yu, H. B.; Whitfield, T. W.; Harder, E.; Lamoureux, G.; Vorobyov, I.; Anisimov, V. M.; MacKerell, A. D.; Roux, B. *J Chem. Theory Comput* **2010**, 6, 774-786.

104. Piquemal, J. P.; Cisneros, G. A.; Reinhardt, P.; Gresh, N.; Darden, T. A. *J. Chem. Phys.* **2006**, 124, 104101-12.

105. Cisneros, G. A.; Piquemal, J. P.; Darden, T. A. *J. Chem. Phys.* **2006**, 125, 184101-16.

106. Kirkwood, J. G.; Buff, F. P. *J. Chem. Phys.* **1951**, 19, 774.

107. Bennaim, A. *Molecular theory of solutions,* Oxford University Press: New York, 2006.

108. Bennaim, A. *J. Chem. Phys.* **1977**, 67, 4884-4890.

109. Bentenitis, N.; Cox, N. R.; Smith, P. E. *J. Chem. Phys.* B **2009**, 113, 12306.

110. Chitra, R.; Smith, P. E. *J. Chem. Phys. B* **2002**, 106, 1491-1500.

111. Gee, M. B.; Cox, N. R.; Jiao, Y. F.; Bentenitis, N.; Weerasinghe, S.; Smith, P. E. *J. Chem. Theory Comput.* **2011**, 7, 1369.

112. Gee, M. B.; Smith, P. E. Abstracts of Papers of the American Chemical Society **2008**, 236.

113. Gee, M. B.; Smith, P. E. *J. Chem. Phys.* **2009**, 131, 165101.

114. Kang, M.; Smith, P. E. *J. Comput. Chem.* **2006**, 27, 1477-1485.

115. Kang, M.; Smith, P. E. *Fluid Phase Equilibria* **2007**, 256, 14-19.

116. Kang, M.; Smith, P. E. *J. Chem. Phys.* **2008**, 128, 244511.

117. Pierce, V.; Kang, M.; Aburi, M.; Weerasinghe, S.; Smith, P. E. *Cell Biochem. Biophys.* **2008**, 50, 1-22.

118. Ploetz, E. A.; Bentenitis, N.; Smith, P. E. *Fluid Phase Equilibria* **2010**, 290, 43-47.

119. Ploetz, E. A.; Smith, P. E. *Phys. Chem. Chem. Phys.* **2011**, 13, 18154-18167.

120. Ploetz, E. A.; Smith, P. E. *J. Chem. Phys.* **2011**, 135, 044506.

121. Smith, P. E. *J. Phys. Chem. B* **2004**, 108, 18716-18724.

122. Smith, P. E. *J. Phys. Chem. B* **2004**, 108, 16271-16278.

123. Smith, P. E. *J. Phys. Chem. B* **2006**, 110, 2862-2868.

124. Smith, P. E. *Biophys. J.* **2006**, 91, 849-856.

125. Smith, P. E.; Mazo, R. A. *J. Phys. Chem. B* **2008**, 112, 7875-7884.

126. Weerasinghe, S.; Smith, P. E. *J. Chem. Phys.* **2003**, 119, 11342-13349.

127. Weerasinghe, S.; Smith, P. E. *J. Chem. Phys* **2003**, 118, 5901-5910.

128. Weerasinghe, S.; Smith, P. E. *J. Chem. Phys* **2003**, 118, 10663-10670.

129. Weerasinghe, S.; Smith, P. E. *J. Chem. Phys.* **2004**, 121, 2180-2186.

130. Weerasinghe, S.; Smith, P. E. *J. Phys. Chem. B* **2005**, 109, 15080-15086.

131. Marcus, Y. *Monatsh. Fur. Chemie.* **2001**, 132, 1387-1411.

132. Ruckenstein, E.; Shulgin, I. *J. Phys. Chem. B* **1999**, 103, 10266-10271.

133. Ruckenstein, E.; Shulgin, I. *Fluid Phase Equilibria* 2001, 180, 281-297.

134. Shulgin, I.; Ruckenstein, E. *J. Phys. Chem. B* **1999**, 103, 872-877.

135. Shulgin, I.; Ruckenstein, E. *Ind. Eng. Chem. Res.* **2002**, 41, 6279-6283.

136. Shulgin, I.; Ruckenstein, E. *Polymer* **2003**, 44, 901-907.

137. Shulgin, I. L.; Ruckenstein, E. *J. Chem. Phys.* **2005**, 123, 054909.

138. Shulgin, I. L.; Ruckenstein, E. *J. Phys. Chem. B* **2006**, 110, 12707-12713.

139. Shulgin, I. L.; Ruckenstein, E. *J. Phys. Chem. B* **2007**, 111, 3990-3998.

140. Shulgin, I. L.; Ruckenstein, E. *Fluid Phase Equilibria* **2007**, 260, 126-134.

141. Shulgin, I. L.; Ruckenstein, E. *J. Phys. Chem. B* **2008**, 112, 3005-3012.

142. Shimizu, S. *Proceedings of the National Academy of Sciences, USA* **2004**, 101, 1195-1199.

143. Shimizu, S.; Boon, C. L. *J. Chem. Phys.* **2004**, 121, 9147-9155.

144. Shimizu, S.; Matubayasi, N. *Chem. Phys. Lett.* **2006**, 420, 518-522.

145. Shimizu, S.; McLaren, W. M.; Matubayasi, N. *J. Chem. Phys.* **2006**, 124, 234905.

146. Shimizu, S.; Smith, D. J. *J. Chem. Phys.* **2004**, 121, 1148-1154.

147. Hall, D. G. *J. Chem. Soc., Faraday Trans.* **1991**, 87, 3523-3528.

148. Zielenkiewicz, W.; Kulikov, O. V.; Krestov, G. A. Bulletin of the Polish Academy of Sciences-Chemistry **1992**, 40, 293-305.

149. Matteoli, E.; Lepori, L. *J. Chem. Phys.* **1984**, 80, 2856-2863.

150. Matteoli, E.; Lepori, L. *J. Chem. Soc., Faraday Trans.* **1995**, 91, 431-436.

151. Ellegaard, M. D.; Abildskov, J.; O'Connell, J. P. *Fluid Phase Equilibria* **2011**, 302, 93-102.

152. Wedberg, R.; O'Connell, J. P.; Peters, G. H.; Abildskov, *J. Molecular Simulation* **2010**, 36, 1243-1252.

153. Kang, M.; Smith, P. E. *J. Comput. Chem.* **2006**, 27, 1477-1485.

154. Guha, A.; Mukherjee, D. *Journal of the Indian Chemical Society* **1997**, 74, 195.

155. Guha, A.; Ghosh, N. K. *Indian Journal of Chemistry Section a-Inorganic Bio-Inorganic Physical Theoretical & Analytical Chemistry* **1998**, 37, 97.

156. Imai, T.; Kinoshita, M.; Hirata, F. *J. Chem. Phys.* **2000**, 112, 9469-9478.

157. Imai, T.; Harano, Y.; Kovalenko, A.; Hirata, F. *Biopolymers* **2001**, 59, 512-519.

158. Imai, T.; Takahiro, T.; Kovalenko, A.; Hirata, F.; Kato, M.; Taniguchi, Y. *Biopolymers* **2005**, 79, 97-105.

159. Lynch, G. C.; Perkyns, J. S.; Pettitt, B. M. *J. Comput. Phys.* **1999**, 151, 135-145.

160. Matteoli, E.; Mansoori, G. A. *J. Chem. Phys.* **1995**, 103, 4672-4677.

161. Matteoli, E. *J. Molecular Liquids* **1999**, 79, 101-121.

162. Nain, A. K. *J. Solution Chemistry* **2008**, 37, 1541-1559.

163. Pandey, J. D.; Verma, R. *Chemical Physics* **2001**, 270, 429-438.

164. Patil, K. J.; Mehta, G. R.; Dhondge, S. S. *Indian Journal of Chemistry Section a-Inorganic Bio-Inorganic Physical Theoretical & Analytical Chemistry* **1994**, 33, 1069.

165. Pjura, P. E.; Paulaitis, M. E.; Lenhoff, A. M. *Aiche Journal* **1995**, 41, 1005-1009.

166. Rosgen, J.; Pettitt, B. M.; Bolen, D. W. *Protein Science* **2007**, 16, 733-743.

167. Rosgen, J. *Osmosensing and Osmosignaling* **2007**, 428, 459-468.

168. Schellman, J. A. *Quarterly Reviews of Biophysics* **2005**, 38, 351-361.

169. Warshavsky, V. B.; Song, X. Y. *Phys. Rev. E* **2008**, 77, 051106.

170. Allen, M. P. T., D. J. *Computer Simulation of Liquids,* Oxford University Press: Oxford, 1989.

171. Widom, B. *Statistical Mechanics: A Concise Introduction for Chemists,* Cambridge University Press: New York, 2002.

172. McQuarrie, D. A. *Statistical Mechanics,* Dover Publications: New York, 1976.

173. Cramer, C. *Essentials of Computational Chemistry: Theories and Models, Second ed.,* John Wiley & Sons Ltd.: West Sussex, 2004.

174. Brooks, C. L. *J. Mol. Biol*. **1992**, 227, 375-380.

175. Tiradorives, J.; Jorgensen, W. L. *Biochemistry* **1993**, 32, 4175-4184.

176. Roccatano, D.; Amadei, A.; Di Nola, A.; Berendsen, H. J. C. *Protein Science* **1999**, 8, 2130-2143.

177. Daggett, V.; Levitt, M. *J. Mol. Biol*. **1993**, 232, 600-619.

178. Mark, A. E.; Vangunsteren, W. F. *Biochemistry* **1992**, 31, 7745-7748.

179. Pande, V. S.; Rokhsar, D. S. *Proceedings of the National Academy of Sciences, USA* **1999**, 96, 9062-9067.

180. Duan, Y.; Kollman, P. A. *Science* **1998**, 282, 740-744.

181. Tsai, J.; Levitt, M.; Baker, D. *J. Mol. Biol.* **1999**, 291, 215-225.

182. Daura, X.; Mark, A. E.; van Gunsteren, W. F. *J. Comput. Chem.* **1998**, 19, 535-547.

183. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, 4, 187-217.

184. Best, R. B.; Buchete, N. V.; Hummer, G. *Biophys J.* **2008**, 95, 4494.

185. Cacace, M. G.; Landau, E. M.; Ramsden, J. *J. Q. Rev. Biophys.***1997**, 30, 241-77.

Chapter 2 - Theory and Simulation of Multicomponent Osmotic Systems

2.1 Abstract

Most cellular processes occur in systems containing a variety of components many of which are open to material exchange. However, computer simulations of biological systems are almost exclusively performed in systems closed to material exchange. In principle, the behavior of biomolecules in open and closed systems will be different. Here, we provide a rigorous framework for the analysis of experimental and simulation data concerning open and closed multicomponent systems using the Kirkwood-Buff (KB) theory of solutions. The results are illustrated using computer simulations for various concentrations of the solutes Gly, Gly_2 and Gly_3 in both open and closed systems, and in the absence or presence of NaCl as a cosolvent. In addition, KB theory is used to help rationalize the aggregation properties of the solutes. Here one observes that the picture of solute association described by the KB integrals, which are directly related to the solution thermodynamics, and that provided by more physical clustering approaches are different. It is argued that the combination of KB theory and simulation data provides a simple and powerful tool for the analysis of complex multicomponent open and closed systems.

2.2 Introduction

Most biological processes occurring under cellular conditions involve systems that are open to some form of matter exchange. In contrast, most *in vitro* experiments study systems closed to matter exchange. It is therefore important to determine any differences in behavior expected under different thermodynamic constraints between otherwise similar systems. While the properties of closed systems have been studied in detail, the study of open systems is less common and yet can provide a wealth of thermodynamic information. Furthermore, the use of computer simulations to help understand biological systems is now common practice. However, simulations of open systems of biological interest remain quite rare. The main aim of the current work is to illustrate how simulation data can be combined with a rigorous theory of solutions (for both open and closed systems) to provide insights into the behavior of biologically relevant solutes and cosolvents.

The thermodynamics of open systems have been studied in detail. [1-6] The usual way to treat binary osmotic systems of a solute (2) in a primary solvent (1) employs a virial expansion for the osmotic pressure (Π) in terms of the solute number density (ρ_2) such that,

$$\beta\Pi = \sum_{n\geq 1} \frac{1}{n} B_n \rho_2^n \tag{2.1}$$

where $\beta=1/RT$, $B_1 = 1$, and several terms (2-5) are typically required in the sum. We note that the above osmotic virial coefficients (B_n) differ slightly from the usual values ($B_n' = B_n/n$) in an effort to simplify some of the results shown below. In the presence of an additional cosolvent (such as NaCl) equilibrium dialysis or isopiestic distillation techniques provide an alternative to the virial expansion approach. [7,8] The above equation can be directly applied to fit the experimental data using the B_n's as fitting constants. Experimental data concerning protein-protein interactions

can be obtained from B_2, however higher order osmotic virial coefficients are not normally required due to the low protein concentrations involved.[9-11] This is not the case for smaller and/or more soluble solutes.

Most statistical thermodynamic theories attempt to relate the virial coefficients to the underlying solute molecular distribution functions.[1,3,12] One of the more versatile approaches is provided by the Kirkwood-Buff (KB) theory of solutions.[12,13] KB theory provides thermodynamic expressions for various properties of both open and closed systems in terms of integrals over molecular distribution functions, commonly referred to as KB integrals (KBIs). In contrast to the traditional McMillan-Mayer (MM) approach, the resulting KB related expressions can easily be applied at any concentration in any multicomponent system. Furthermore, the combined use of KB theory and molecular simulation appears quite natural as the KBIs can be obtained directly from the simulation data at the composition of interest.[14]

The application of KB theory to open systems has been widely recognized.[4,12,15,16] However, only recently have specific applications to evaluate either experimental or simulation data appeared. Kirkwood and Buff recognized the possible uses of their theory for osmotic systems in their original paper.[12] O'Connell and coworkers have since used KB theory to probe the exact relationships between osmotic virial coefficients and other thermodynamic properties of solution mixtures.[15] More recently, KB theory has been used to directly rationalize osmotic pressure data,[16] and to reinterpret light scattering data which can also provide estimates of the second virial coefficient for proteins.[17] Just in the last decade a considerable effort has focused on understanding equilibrium dialysis, and other closely related experimental data, in terms of cosolvent preferential binding.[14,18-22] Finally, KB theory has also been applied to the study of reactive and association equilibria in a variety of ensembles.[23-26] Here, we extend these previous approaches to: i) provide

a simple analysis of experimental osmotic pressure data; ii) indicate how one can obtain valuable information concerning solute-solute distributions; iii) compare and contrast similar properties in both open and closed systems; and iv) illustrate how one can use KB theory to probe association equilibria describing the aggregation of solutes.

The application of KB theory to open systems can be further illustrated using computer simulation data. The simulation of open systems by Monte Carlo methods is quite straight-forward.[27,28] Molecular dynamics simulations of open systems are more problematic due to technical issues surrounding particle creation and annihilation.[29] The simplest methods involve the application of semi-permeable physical boundaries (virtual membranes) between various regions of the system which directly mimic the experimental situation.[30-32] A similar approach is adopted here for the study of small Gly_n (n = 1-3) solutes with and without NaCl as a cosolvent.

2.3 Theory

2.3.1 General Background

In the following sections we will consider solutions containing a principle solvent (1), a solute (2), and in some cases an additional cosolvent (3). The equilibrium concentration of each species is expressed in terms of number densities (molarities), $\rho_i = N_i/V$, or dimensionless molalities, $m_i = \rho_i/\rho_1$, and each species has an associated chemical potential, μ_i. Temperature will be assumed to be constant throughout. The osmotic system(s) of interest involve a central fixed volume of interest (V) which is separated from a large bulk solvent region by a barrier permeable (open) to the solvent, and in some applications the cosolvent, but not to the solute. The bulk solvent is held at a constant chemical potential (μ_1) defined by the solvent at a particular temperature and a fixed outside pressure (P_O). Here, we use pure water at a temperature T = 298.15 K and a pressure

P_O = 1 bar throughout. The pressure generated inside the central fixed volume region (P_I) in the presence of the solute then provides the osmotic pressure via $\Pi = P_I - P_O$. The osmotic pressure, the virial coefficients, and the integrals defined below are then a function of T, $\mu_1(P_O)$, and ρ_2. In the presence of a cosolvent the dependence extends to include ρ_3, when the barrier is impermeable (closed) to the cosolvent, or μ_3 when the barrier is permeable (open) to the cosolvent. However, in the following sections we have not included all of these dependencies in an effort to simplify the notation used.

2.3.2 Kirkwood-Buff Theory of Binary Osmotic Systems

In this section we outline how KB theory can be used to understand osmotic systems. One of the advantages of using KB theory is that the solution thermodynamics can be formulated in terms of integrals which have a well defined physical significance. This is also true of the MM theory of solutions, but there one is restricted to an interpretation in terms of distributions at infinite dilution in the primary solvent.[13] This restriction is not required by KB theory, although MM theory is obtained, as expected, under infinite dilution conditions. The following integrals are required,[12]

$$G_{\alpha\beta} = \frac{1}{V}\iint [g^{(2)}_{\alpha\beta}(r_1, r_2) - 1] dr_1 dr_2 \tag{2.2}$$

$$G_{\alpha\beta\gamma} = \frac{1}{V}\iiint [g^{(3)}_{\alpha\beta\gamma}(r_1, r_2, r_3) - g^{(2)}_{\alpha\beta}(r_1, r_2) - g^{(2)}_{\alpha\gamma}(r_1, r_3) - g^{(2)}_{\beta\gamma}(r_2, r_3) + 2] dr_1 dr_2 dr_3$$

and correspond to integrals over the orientationally averaged two body $g^{(2)}$ and three body $g^{(3)}$ distribution functions between the centers of mass of species α, β and γ, defined in the Grand Canonical ensemble, and integrated over all relative center of mass positions r_1 of particle 1 of

species α, etc. They clearly resemble the integrals appearing in the treatment of imperfect gases or the MM theory of solutions.[33] The key difference is that the solute integrals G_{22} and G_{222} are composition dependent in KB theory. Hence, the distributions (g_{22}, etc) are for pairs of solute molecules after averaging over all other solute and solvent degrees of freedom at the composition of interest. The physical interpretation of the G_{22} integral in open systems is quite simple. A positive value indicates a tendency for the solute to self-associate, while a negative value indicates a preference for solute solvation. We will see that G_{222} provides a measure of triplet solute correlations and determines how G_{22} changes with composition. Alternatively, one can express the above integrals in terms of particle-particle number fluctuations,

$$\rho_2(1 + \rho_2 G_{22}) = \frac{< \delta N_2 \delta N_2 >}{V} = F_{22} \tag{2.3}$$

and,

$$\rho_2(1 + 3\rho_2 G_{22} + \rho_2^2 G_{222}) = \frac{< \delta N_2 \delta N_2 \delta N_2 >}{V} = F_{222} \tag{2.4}$$

where $\delta N_2 = N_2 - < N_2 >$ and the angular brackets denote an ensemble average for a local region within the solution mixture. Here, N_2 is the instantaneous number of solute molecules observed in a small local fixed volume of the solution open to all species. KB theory relates the properties (particle number fluctuations) of systems open to all species, to the properties of semi-open (osmotic) or closed (isothermal isobaric) systems under the same average thermodynamic conditions. We note that one does not have to use the commonly employed superposition approximation for the triplet distributions, or invoke additive potentials, when using KB theory. The evaluation of G_{22} and other G_{ij} values for various solutes represents the major focus of this work.

The application of KB theory to binary osmotic systems provides expressions for derivatives of the osmotic pressure in terms of the above integrals and the solute number density. The first derivative is given by,[12]

$$\beta\left(\frac{\partial\Pi}{\partial\rho_2}\right)_{\mu_1} = \frac{1}{1+\rho_2 G_{22}} \tag{2.5}$$

Clearly, ideal osmotic behavior requires either a small solute concentration or $G_{22} = 0$ for all compositions. A tendency for solute self association ($G_{22} > 0$) would result in a lower than ideal ($\beta\Pi^{id} = \rho_2$) osmotic pressure as the solute concentration is increased, and *vice versa*. An expression for the second derivative has also been provided and can be written,[26]

$$\beta\left(\frac{\partial^2\Pi}{\partial\rho_2^2}\right)_{\mu_1} = -\frac{G_{22}+\rho_2(G_{222}-G_{22}^2)}{(1+\rho_2 G_{22})^3} \tag{2.6}$$

Both derivative expressions apply at any solute concentration. Taking derivatives of the right hand side of Equation (2.5) and equating with the right hand side of Equation (2.6) provides an expression for the derivative of G_{22} with respect to solute concentration at constant T and solvent chemical potential (all such derivatives will be indicated with a prime),

$$G_{22}{}' = \frac{G_{222}-2G_{22}^2}{1+\rho_2 G_{22}} \tag{2.7}$$

Hence, if $G_{222} = 2G_{22}^2$ for all compositions the value of G_{22} will be independent of composition, whereas one requires $G_{222} = 0$ for ideal systems. However, when $G_{222} > 2G_{22}^2$ then G_{22} will tend to increase with composition and *vice versa*. When G_{22} is independent of composition one finds that $\beta\Pi = G_{22}^{-1}\ln(1+\rho_2 G_{22})$.

Given a set of osmotic virial coefficients one can directly express the composition dependence of G_{22} (G_{22}') and G_{222} according to,

$$G_{22} = -\frac{Y_n}{\rho_2(1+Y_n)} \quad G_{222} = \frac{Y_n(1+Y_n)(1+2Y_n) - \rho_2 {Y_n}'}{\rho_2^2(1+Y_n)^3} \quad Y_n = \sum_{n\geq 2} B_n \rho_2^{n-1} \tag{2.8}$$

The above expressions describe the composition dependence of the experimental or simulated solute self-association, and represent the principle quantities of interest in this study. Expansion of the above expressions in a power series in the solute number density leads to,

$$G_{22} \approx -B_2 - [B_3 - B_2^2]\rho_2 - [B_4 - 2B_2B_3 + B_2^3]\rho_2^2 - \cdots \tag{2.9}$$
$$G_{222} \approx -[B_3 - 3B_2^2] - [2B_4 - 9B_2B_3 + 3B_2^3]\rho_2 - \cdots$$

and provide the limiting values of G_{22} and G_{222} for an infinitely dilute solute,

$$G_{22}^{\infty} = -B_2 \qquad G_{222}^{\infty} = 3B_2^2 - B_3 \tag{2.10}$$

together with the derivative of G_{22},

$${G_{22}^{\infty}}' = G_{222}^{\infty} - 2(G_{22}^{\infty})^2 = B_2^2 - B_3 \tag{2.11}$$

The above expressions are necessarily equivalent to those of MM theory, except for the fact that we have not inferred the superposition approximation for the triplet potential of mean force to simplify and evaluate B_3. The above relationships lead to the following osmotic pressure expansion,

$$\beta\Pi = \rho_2 - \tfrac{1}{2}G_{22}^{\infty}\rho_2^2 - \tfrac{1}{3}[G_{222}^{\infty} - 3(G_{22}^{\infty})^2]\rho_2^3 + \cdots \tag{2.12}$$

which provides the B_2 and B_3 coefficients in terms of KB integrals, and is in agreement with previous results.[12] Hence, MM theory is obtained from KB theory when the required derivatives are obtained at infinitely dilute solute concentrations.

The above expressions can be used to analyze experimental or simulated osmotic pressure data for any type of solute. It should be noted, however, that the G_{22} integral diverges ($\propto \rho_2^{-1/2}$)

for low concentration salt solutions.[34] KB theory can still be applied to study salt solutions, but with less interpretive power as provided for non-ionic systems. For both ionic solutes and cosolvents we then distinguish between the traditional salt concentration (ρ_s) and the total ion concentration (ρ_2 or ρ_3).[35] Before leaving this section we note that KB theory can be used to provide an expansion in terms of solute molality,[25,36,37] but the expressions then involve the G_{21} integrals and become somewhat more complicated to interpret.

2.3.3 Kirkwood-Buff Theory of Ternary Osmotic Systems

Ternary osmotic systems are more complicated and yet just as important. In particular, the effects of osmolytes (or molecular crowding) on protein folding and association under cellular (open) conditions requires a detailed knowledge of osmotic systems and their behavior.[38-40] Here, we provide expressions to illustrate the effects of a cosolvent (3) on the osmotic pressure displayed by a solute (2) in a primary solvent (1), which depend on whether the system is open or closed with respect to cosolvent. The following expressions then hold,[25]

$$\begin{aligned} RTd\ln\rho_1 &= (1+N_{11})d\mu_1 + N_{12}d\mu_2 + N_{13}d\mu_3 \\ RTd\ln\rho_2 &= N_{21}d\mu_1 + (1+N_{22})d\mu_2 + N_{23}d\mu_3 \\ RTd\ln\rho_3 &= N_{31}d\mu_1 + N_{32}d\mu_2 + (1+N_{33})d\mu_3 \\ d\mathrm{P} &= \rho_1 d\mu_1 + \rho_2 d\mu_2 + \rho_3 d\mu_3 \end{aligned} \tag{2.13}$$

where we have written $N_{ij} = \rho_j G_{ij}$, and the last expression corresponds to the Gibbs-Duhem equation at constant T. These differentials can be applied toward the analysis of systems in any ensemble where T is held constant. Several different cases will be considered.

If the system is open to both the solvent and the cosolvent then one has $d\mu_1 = d\mu_3 = 0$ and $dP = d\Pi$, which on insertion into the above expressions provide,

$$\beta\left(\frac{\partial\Pi}{\partial\rho_2}\right)_{\mu_1,\mu_3} = \frac{1}{1+N_{22}} \tag{2.14}$$

In this situation there is no explicit dependence of the osmotic pressure on the KB integrals involving either the solvent or cosolvent. However, the value of G_{22} will depend implicitly on the cosolvent concentration. The difference between the G_{22} values in the presence and absence of the cosolvent can be obtained from,

$$\left(\frac{\partial\Pi}{\partial\rho_2}\right)_{\mu_1,\mu_3} - \beta\left(\frac{\partial\Pi}{\partial\rho_2}\right)_{\mu_1} \approx -[N_{22}(\rho_3) - N_{22}(0)] \tag{2.15}$$

which is valid for low solute concentrations. If $G_{22}(\rho_2) > G_{22}(0)$ then the presence of the cosolvent tends to increase the self association of the solute and is characterized by a lower solute osmotic pressure in the presence of the cosolvent compared to that in pure solvent (for the same solute concentration). The above conditions are the same as found in equilibrium dialysis experiments. Here, one can quantify the relative binding of the cosolvent (G_{23}) and solvent (G_{21}) to the solute via the preferential binding parameter,[36,37]

$$\Gamma_{23} = \left(\frac{\partial m_3}{\partial m_2}\right)_{\mu_1,\mu_3} = \frac{N_{23} - m_3 N_{21}}{1 + N_{22} - N_{12}} \tag{2.16}$$

where $m_i = \rho_i/\rho_1$ is the (dimensionless) molality of *i*. This property is particularly useful when describing the effects of cosolvents on molecular association as demonstrated below.

If the system is only open to the solvent and the cosolvent then one has $d\mu_1 = d\rho_3 = 0$ and $dP = d\Pi$, which on insertion into the above expressions provide,

$$\beta\left(\frac{\partial\Pi}{\partial\rho_2}\right)_{\mu_1,\rho_3} = \frac{1 + N_{33} - N_{23}}{(1+N_{22})(1+N_{33}) - N_{23}N_{32}} \tag{2.17}$$

where N_{23} can be considered as a measure of the solute-cosolvent affinity. When species 2 and 3 are both proteins this provides insight into mixed protein-protein interactions that can be extracted from experimental osmotic data. In either case, the above expression reduces to Equation (2.5) when $G_{23} = 0$. Equation (2.17) is much more complicated in comparison to Equation (2.5) or (2.14). However, one can extract information on the cosolvent and solute association via,

$$\beta\left(\frac{\partial\Pi}{\partial\rho_2}\right)_{\mu_1,\rho_3} - \beta\left(\frac{\partial\Pi}{\partial\rho_2}\right)_{\mu_1,\mu_3} \approx -N_{23} \tag{2.18}$$

which is valid for low solute and cosolvent concentrations.

2.3.4 Solute Association Equilibria in Osmotic and Closed Systems

The previous analysis indicates how one can obtain information concerning G_{22} for solutes. It should be noted that this is the most relevant property describing solute-solute association that relates to the thermodynamics of the solution. It involves both the direct binding between solute molecules, together with more subtle and/or long range changes in the solute-solute distribution with respect to a random bulk distribution (see Equation 2.2). Hence, solute-solute association could increase without inferring the formation of well defined dimers, etc. However, a much more physical picture of solute-solute association is provided by spectroscopic studies, where information may be provided concerning the concentration of specific tightly bound dimers. KB theory can also be used to study these types of association equilibria.[23-26] The results for binary and ternary systems are presented here and compared to equivalent results for closed systems.

If we consider a solute which can exist as a monomer (M) and an aggregate (A) consisting of n monomers, then one can define an equilibrium constant for the association reaction $n\mathrm{M} \rightarrow \mathrm{A}$ such that $K = \rho_A/\rho_M^n$ under the equilibrium conditions $\mu_A = n\mu_M$. We note that the equilibrium

constant defined here is not dimensionless. One could include a standard concentration in the definition of the equilibrium to make *K* dimensionless. However, we will only be concerned with changes in the equilibrium constant (KB theory is mute on the value of *K* itself), and hence this factor will disappear. Previous studies indicate that,[25]

$$RTd\ln K = (N_{A1} - nN_{M1})d\mu_1 + (N_{A2} - nN_{M2})d\mu_2 + (N_{A3} - nN_{M3})d\mu_3 \qquad (2.19)$$

for a ternary system. The above differential complements the expressions in Equation (2.13) and involves KB integrals describing the correlation between each solute form and the primary components of the solution. The relationships between the solute KB integrals (independent of solute form) and the integrals for solute specific forms are given by,[25]

$$\delta_{2j} + N_{2j} = f_A N_{Aj} + f_M N_{Mj} \qquad (2.20)$$
$$N_{M2} = 1 + N_{MM} + nN_{MA}$$
$$N_{A2} = n + nN_{AA} + N_{AM}$$

where δ_{ij} is the Kronecker delta function and the monomer and aggregate fractions are given by $f_M = \rho_M/\rho_2$ and $f_A = n\rho_A/\rho_2$, respectively. More details can be found in the original literature.[25,26]

Using Equations (2.13) and (2.19) for binary systems ($\rho_3 = 0$) one finds the following expressions for the effect of increasing solute concentration on the solute association equilibrium in open,

$$RT\left(\frac{\partial \ln K}{\partial \Pi}\right)_{\mu_1} = G_{A2} - nG_{M2} \qquad (2.21)$$
$$\left(\frac{\partial \ln K}{\partial \rho_2}\right)_{\mu_1} = \frac{G_{A2} - nG_{M2}}{1 + N_{22}}$$

and closed,

$$\left(\frac{\partial \ln K}{\partial \rho_2}\right)_P = \frac{(G_{A2} - nG_{M2}) - (G_{A1} - nG_{M1})}{1 + N_{22} - N_{12}} \tag{2.22}$$

systems. After taking derivatives of Equation (2.19) with respect to pressure one can then eliminate the G_{i1} terms to provide,

$$\left(\frac{\partial \ln K}{\partial \rho_2}\right)_P = \frac{G_{A2} - nG_{M2} + \Delta V^*}{1 + N_{22} - \rho_2 RT\kappa_T} \approx \frac{G_{A2} - nG_{M2} + \Delta V^*}{1 + N_{22}} \tag{2.23}$$

where ΔV^* is the change in volume for the process, which can be expressed in terms of KBIs but is simpler to interpret in this form. The above expressions demonstrate that the change in the association equilibrium differs in open and closed systems (possessing the same average thermodynamic properties) by terms in both the numerator and denominator. The compressibility term in the denominator will typically be small (10^{-3}) and can be neglected, and the difference between the ensembles is related to the magnitude of ΔV^*. In open systems an increase in solute concentration increases the equilibrium constant if association of the solute, in any form, is larger to the aggregate than *n* times the monomer. In closed systems the effect of water association is also directly present and can be represented in terms of the volume change associated with the aggregation process. Hence, open systems will resist (compared to closed systems) any processes which result in an increase in volume by a term related to $\Pi\Delta V^*$.

Ternary systems are more complicated and involve additional KB integrals. Furthermore, component 3 may be held at constant chemical potential or fixed concentration, and one can follow the equilibrium by varying either the solute or cosolvent concentration. If the cosolvent concentration is held fixed and the solute concentration varied one finds (in addition to Equation 2.17) that,

$$RT\left(\frac{\partial \ln K}{\partial \Pi}\right)_{\mu_1,\rho_3} = \frac{(G_{A2} - nG_{M2})(1 + N_{33}) - (N_{A3} - nN_{M3})G_{23}}{1 + N_{33} - N_{23}} \tag{2.24}$$

$$\left(\frac{\partial \ln K}{\partial \rho_2}\right)_{\mu_1,\rho_3} = \frac{(G_{A2} - nG_{M2})(1 + N_{33}) - (N_{A3} - nN_{M3})G_{23}}{(1 + N_{22})(1 + N_{33}) - N_{23}N_{32}}$$

These expressions also describe the effect of varying the cosolvent concentration for a fixed solute concentration after a simple index change (2 ↔ 3).

The previous expressions are greatly simplified if we restrict ourselves to situations in which the solute concentration is negligible (a common biological situation) and the cosolvent concentration is varied. Then we find for open systems,

$$RT\left(\frac{\partial \ln K}{\partial \Pi}\right)^{\infty}_{\mu_1,\rho_2} = G_{A3} - nG_{M3} \tag{2.25}$$

$$\left(\frac{\partial \ln K}{\partial \rho_3}\right)^{\infty}_{\mu_1,\rho_2} = \frac{G_{A3} - nG_{M3}}{1 + N_{33}}$$

while for closed systems we have,

$$\left(\frac{\partial \ln K}{\partial \rho_3}\right)^{\infty}_{P,m_2} = \frac{(G_{A3} - nG_{M3}) - (G_{A1} - nG_{M1})}{1 + N_{33} - N_{13}} = \frac{\rho_3^{-1}(\Gamma^{\infty}_{A3} - n\Gamma^{\infty}_{M3})}{1 + N_{33} - N_{13}} \tag{2.26}$$

and provides the KB expression for the *m*-value of protein denaturation when A → D, M → N and $n = 1$. Performing the same manipulation as for binary systems one finds,

$$\left(\frac{\partial \ln K}{\partial \rho_3}\right)^{\infty}_{P,m_2} = \frac{G_{A3} - nG_{M3} + \Delta V^*}{1 + N_{33} - \rho_3 RT\kappa_T} \approx \frac{G_{A3} - nG_{M3} + \Delta V^*}{1 + N_{33}} \tag{2.27}$$

which takes a similar form as before.[41] Hence, the change in the equilibrium constant for association will be larger (more positive) in closed versus open systems when the volume change for association is positive and *vice versa*. The ease with which the above manipulations can be

performed for multicomponent systems in any ensemble represents a particular advantage of the KB approach.

In summary, we have provided a series of expressions which can be applied to understand the behavior of open and closed systems. In particular, expressions describing variations in both the osmotic pressure and association equilibria in terms of a series of KB integrals have been provided. The main difference between equilibria in open and closed systems relates to the volume change accompanying the process. Hence, a significant difference between ensembles would only be expected for large volume changes and/or osmotic pressures. Finally, we want to be clear concerning the exact interpretation of the KBIs. The KBIs are defined in a Grand Canonical ensemble open to all species. Hence, $G_{ij} = G_{ij}(T, V, \mu_1, \mu_2)$ for binary systems. The KBIs obtained from an analysis of the osmotic data correspond to changes in μ_2, ρ_2 or Π, whichever is more convenient to use. The KBIs will not be the same as those obtained from isothermal isobaric (P_O) data, even if the solute and solvent compositions are identical, but will correspond to the KBIs obtained from an isothermal isobaric analysis at the same composition and the higher pressure of $P_O + \Pi$. These differences may or may not be important depending on the exact application.[6,15]

The primary use for the above expressions is two-fold. First, one can apply the expressions provided in Equations (2.5) - (2.7), (2.14), (2.17) to help interpret the experimental data concerning osmotic pressure changes, or Equations (2.21), (2.23), (2.25) and (2.27) to help interpret changes in equilibrium constants, in terms of the distribution functions between molecules provided in Equation (2.2). Hence, one can develop a link between the experimental thermodynamic data and the relative distributions of the various species in solution. Second, one can reverse the whole process and relate the solution distributions, obtained from theory or simulation, to compare with

available experimental data or to make predictions concerning the thermodynamic behavior of the solutions.

2.4 Methods

2.4.1 Molecular Dynamics Simulations

All molecular dynamics simulations were performed using the KBFF models (http://kbff.chem.k-state.edu),[35,42,43] together with the SPC/E water model,[44] as implemented in the GROMACS 4.0.5 package.[45] All simulations were performed at 300 K and the pressure of interest ($P = P_O = 1$ bar or $P = P_O + \Pi$) using the weak coupling technique to modulate the temperature and pressure with relaxation times of 0.1 and 0.5 ps,[46] respectively. A time-step of 2 fs was used and the bond lengths were constrained using the Lincs (solutes) and Settle (water) algorithms.[47,48] The particle mesh Ewald technique was used to evaluate electrostatic interactions with a grid resolution of 0.1 nm.[49] A real space convergence parameter of 3.5 nm^{-1} was used in combination with twin range cutoffs of 1.0 and 1.5 nm, and a nonbonded update frequency of 10 steps. Random initial configurations of molecules in a cubic box were used to study the closed systems. Initial configurations of the different solutions were generated from a cubic box ($L \approx 6.0$ nm) of equilibrated water molecules by randomly replacing waters with solutes until the required concentration was attained. The steepest descent method was then used to perform 100 steps of energy minimization. This was followed by extensive equilibration, which was continued until the rdfs displayed no drift with time (typically 5 ns). Total simulation times were in the 25-50 ns range, and the final 25-30 ns were used for calculating ensemble averages. Configurations were saved every 0.1 ps for the calculation of various properties. Errors ($\pm 1\sigma$) in the simulation data were estimated by using five block averages.

2.4.2 Osmotic Simulations

There are several simulation techniques available to study osmotic systems. Here, we take a very simple physical approach. Simulations of systems extended in the *z* direction (6 x 6 x 24 nm) were performed which included a series of Lennard-Jones (LJ) particles to act as two semi permeable walls separating the bulk solution from a central semi open region of interest. The LJ "walls" were separated by a *z* distance of 12 nm and all solutes were placed in the central region between the two walls. The parameters for the LJ particles were taken to be 0.3 nm and 0.02 kJ/mol, and each wall was constructed of 20x20 particles separated by 0.3 nm in both the *x* and *y* directions. The walls were held fixed during the simulations and all interactions between the LJ particles and between the LJ particles and the solvent were excluded. Periodic boundary conditions were applied in all directions. Anisotropic pressure coupling was used to keep the *x* and *y* box lengths fixed and to maintain a fixed pressure in the *z* direction. All other simulation conditions were the same as for the closed systems. The osmotic pressure was then obtained by determining the pressure on the walls provided by the non-diffusible components.[32]

Several technical issues can arise with such a system setup. First, the presence of the walls could affect the solute distribution and/or the pressure profile for the central region. This issue is discussed in the Results Section. Second, the use of a finite bulk solvent region acting as the chemical potential bath leads to a drop in the pure solvent pressure, as the solvent moves into the central (low μ_1) region, when one simply couples $P_{zz} = P_T$ to a barostat at 1 bar. Hence, the outside pressure displayed by the pure solvent region will be less than 1 bar and therefore the system, while providing the same solvent chemical potential inside and outside the open region, will correspond to different constant solvent chemical potentials for each solute concentration. This makes it difficult to follow the equilibrium line where μ_1 is held constant, at 1 bar for instance, as

performed experimentally. Fortunately, this is easy to correct if we note that the total or reference pressure (P_T), the inside pressure (P_I), and the outside pressure (P_O) are related by,

$$P_T V_T = P_0 V_0 + P_I V_I \tag{2.28}$$

Hence, given the measured osmotic pressure, $\Pi = P_I - P_0$, one can then determine the outside (and thereby inside) pressure according to,

$$P_0 = P_T + \Pi(V_I/V_T) \tag{2.29}$$

To ensure that the solvent chemical potential remains at the same constant value as the solute concentration is increased, one needs to adjust the reference pressure to raise the outside pressure to the desired value of 1 bar. The above equation can be used to predict the value of P_T that is consistent with P_O = 1 bar, assuming the osmotic pressure is independent of P_T, and the whole processes can be iterated (2-3 cycles) to consistency. The same process was performed for the NaCl simulations where μ_1 and μ_3 were held constant, except that the target outside pressure was the osmotic pressure obtained from the NaCl solute simulations. While these adjustments are usually small they can also be important.[6,15]

2.4.3 Analysis of the Simulation Data

The primary analysis involved the determination of the KBIs from the simulations. This was achieved in two ways. The first involved the usual integration of the corresponding rdf. The KBIs are defined in systems open to all components and hence one cannot integrate over the full volume. Hence, the integration was truncated at a distance R from the central particle, where R is the distance at which the rdf approaches unity.[50] This also provides a distance dependent KBI which can be used to determine contributions from various solvation shells,

$$G_{ij}(R) = \int_0^R [g_{ij}^{(2)}(r_{12}) - 1] dr_{12} \tag{2.30}$$

For this work we used a final value of R = 1.5nm. KBIs were only determined from the closed systems at the equivalent state point. The main advantage of this approach is that the rdfs and KBIs provide information concerning the "structure" of the solution surrounding the central *i* particle. The second approach involves the direct application of Equation (2.3) and the determination of the appropriate particle number fluctuations. The closed systems were analyzed by considering a series of reference volumes centered on a randomly chosen origin and then averaging over these volumes. The reference volumes were chosen as cubes of length 3 nm and approximately 10,000 origins were used. The advantage of this approach lies in the large number of origins which can be used, which greatly improves the statistical significance but with the loss of structural information. The determination of G_{222} was performed using the particle number fluctuations and Equation (2.4).

A more physical analysis of solute association was also performed. The most prominent interaction between the Gly_n solutes involved direct association between the N and C terminal groups, as evidenced by the atom based rdfs. Hence, a solute dimer, (and trimer, etc) was defined by considering the contact distance between the nitrogen and the midway point between both oxygens of the carboxylate groups. If this distance was less than 0.5 nm, the first minimum in the rdf between these two groups, then the two solutes were considered to be associated. An iterative procedure was then applied to determine the number of solutes in each solute cluster.

2.5 Results and Discussion

In this section we analyze both the experimental and simulated data for binary mixtures of water containing various concentrations of NaCl, Gly, Gly_2 and Gly_3 as solutes. In addition, simulated data for ternary mixtures of water with a solute and cosolvent are also examined. In performing the osmotic simulations one has little control over the exact concentration of the diffusible components. Hence, while we will constantly refer to systems at 3 m Gly or 6 m NaCl, *etc*, it should be remembered that these are approximate concentrations (to within 10%). The exact concentrations can be found in the various tables. Furthermore, the statistical noise associated with KBIs increases as the concentrations of the components decreases.[51] Therefore, in many situations we have chosen to analyze only the simulations at high solute and cosolvent concentrations, and to use the highest possible concentrations of both solute and cosolvent.

Before continuing with the present analysis it is important to ensure there were no significant artifacts in the osmotic simulations. This is unlikely due to the fact that there is no significant desolvation process for the solutes at the walls, although effects on solute-solute distributions are still possible. The pressure and density profiles for the 6m NaCl osmotic system are displayed in Figure 2.1. The pressure profile, P(z), was determined using the approach outlined in previous work on surface tension.[52] However, here we approximated the pressure contributions using a simple Coulomb plus LJ potential truncated at 1.5 nm for the molecular virial, primarily due to the excessive cost involved with calculating the contributions using the full Ewald potential. Hence, the pressures do not exactly match the pressure determined during the simulation. Nevertheless, it seems clear that neither the density nor pressure profiles indicate any surface effects beyond a few molecular diameters.

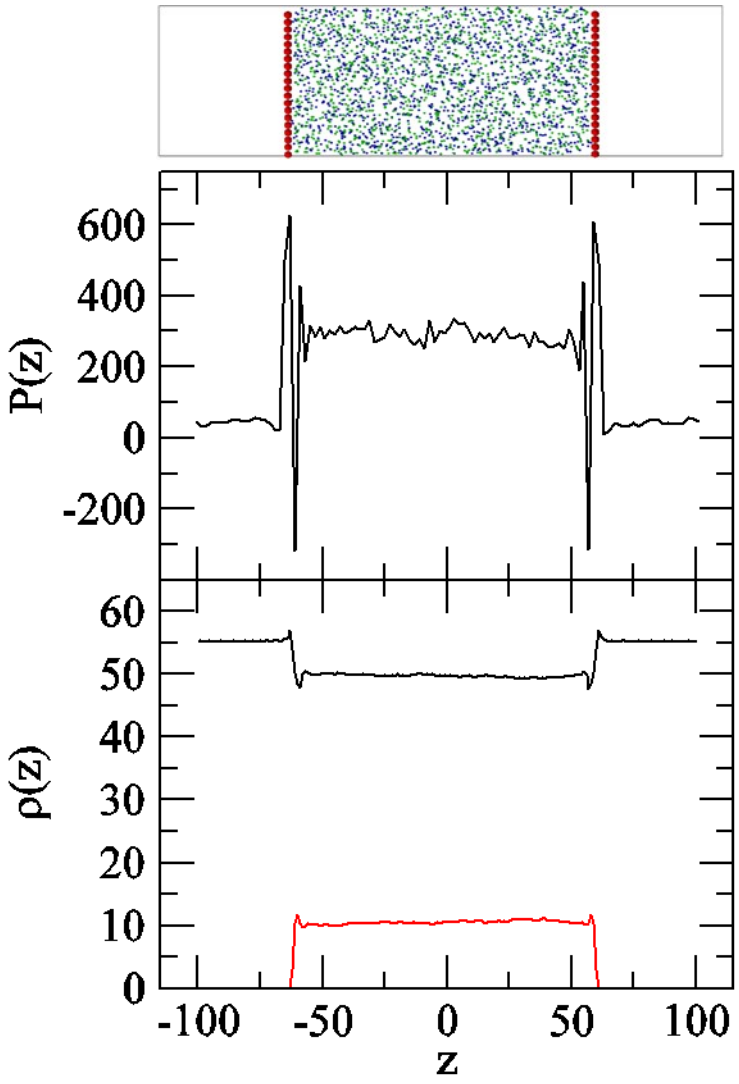


Figure 2.1 Pressure and concentration profiles obtained from the simulation of the 6m NaCl osmotic system at 300 K. The top panel shows a snapshot from the simulation with water molecules removed. The LJ spheres comprising the "walls" are displayed in red. The sodium ions (blue) and chloride ions (green) are confined to the central inside region. The central panel displays the pressure profile in units of bar. The lower panel displays the molar concentrations of water (black) and ions (red).

The experimental and simulated osmotic pressures are displayed in Figure 2.2. The force fields used here performed reasonably well at low solute concentrations, but displayed some deviation from experiment at higher solute concentrations. It also appears that, even if the force fields were perfect, the estimated errors are such that one could not distinguish between the real experimental data and the ideal data provided by the van't Hoff curves, using the current simulation times. This picture changes somewhat when the focus is shifted to the KBIs as we shall see later.

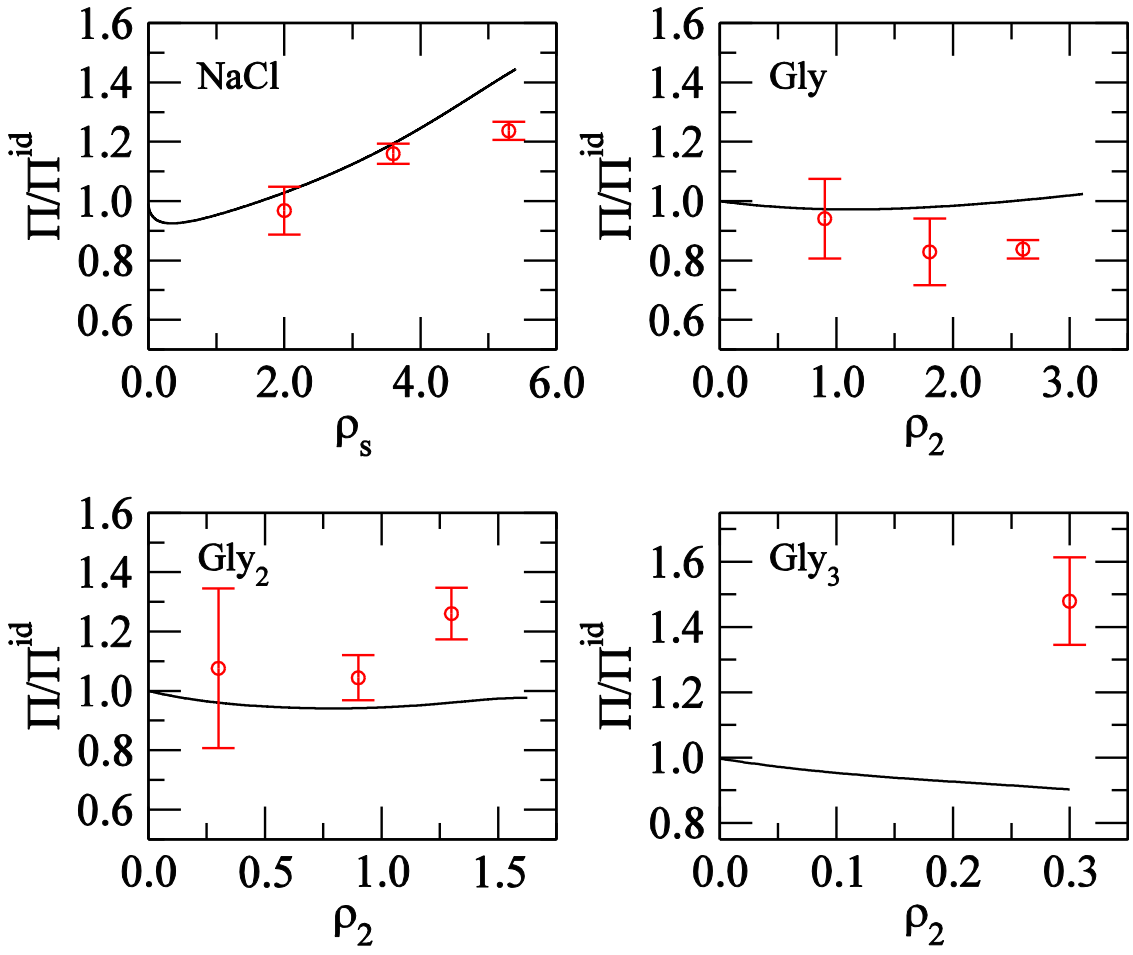


Figure 2.2 Experimental and simulated osmotic pressures at 300 K as a function of solute molarity. Data are displayed as Π/Π^{id} where solid lines correspond to the experimental data and symbols indicate simulated results. Experimental data taken from 54-58.

One of the goals of this work is to investigate the thermodynamics of open (and closed) systems in terms of the KBIs. The presence of the walls and the subsequent loss of periodicity hinder the determination of the KBIs for the inside region. To circumvent this problem, we have performed additional isothermal isobaric simulations at the reference pressure of P_O = 1 bar and also at a pressure of $P_O + \Pi$, using the solute and solvent concentrations obtained for the inside region. The solute-solute rdfs obtained for all three systems are displayed in Figure 2.3.

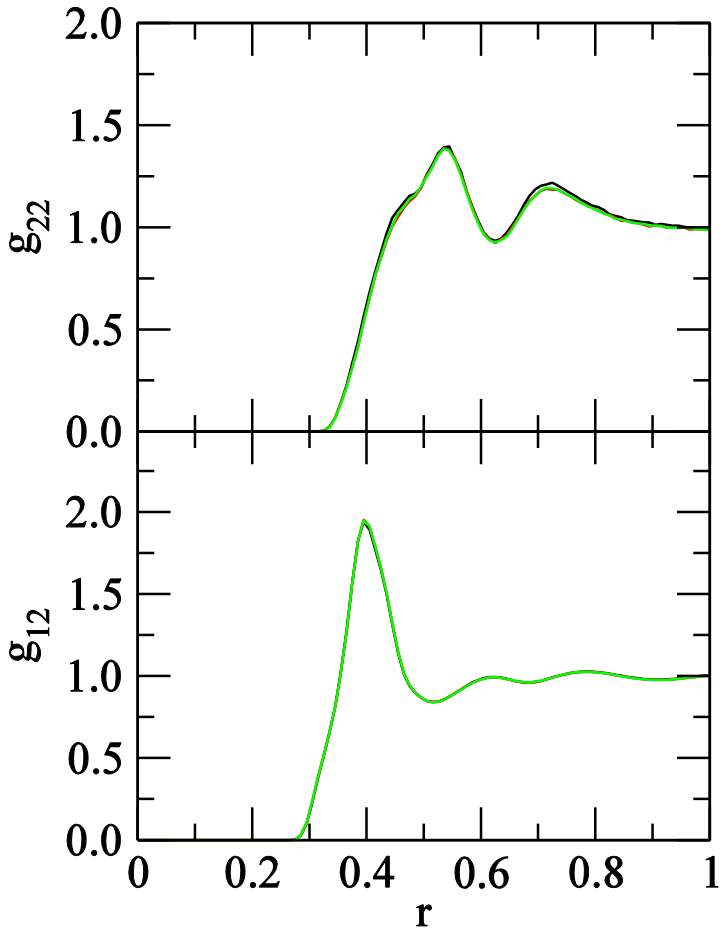


Figure 2.3 Solute-solute (g_{22}) and solute-solvent (g_{12}) rdfs as a function of ensemble and pressure. Data are presented for 3m Gly as a solute, but similar observations are found for the Gly_2 and Gly_3 systems. Curves correspond to the osmotic simulation (black) and closed systems with $P = P_O$ = 1 bar (red) and $P = P_O + \Pi$ = 53 bar (green).

They clearly show that the rdfs are identical within the precision of the simulations. Hence, while the KBIs will vary with composition, they appear to be relatively insensitive to pressure, *i.e*. G_{22} $(T, m_2, P_O) \approx G_{22} (T, m_2, P_O + \Pi)$. This is to be expected for the relatively low pressures exhibited in the current osmotic systems. Consequently, we have obtained all the KBIs presented here from the corresponding closed system simulations.

The experimental and simulated fluctuating quantities are provided in Table 2.1, Table 2.2 and Figure 2.4. The values of G_{22} for all solutes start positive and decrease with increasing solute concentration. Hence, there is a tendency for solute self-association at low solute concentrations which increases as one moves from Gly to Gly_2 to Gly_3. This behavior has been observed before in closed systems where we used the isobaric isothermal results to investigate possible group contributions to the observed association behavior.[53] A comparison of the closed (isothermal isobaric) and open (osmotic) results indicates that the G_{22} values are essentially the same, to within the typical precision of the data, which is to be expected considering the negligible pressure dependence exhibited by the rdfs in Figure 2.3. The simulated values of G_{22} are also provided in Table 2.3 and Figure 2.4. There is not perfect agreement with experiment. The trends in G_{22} with composition appear to be correct and one observes a general agreement in sign. Fortunately, unlike the raw osmotic pressure data, it does appear possible to distinguish the G_{22} values from their ideal values ($G_{22} = 0$). Furthermore, the infinite dilution KBIs obtained from a fit of the simulated osmotic pressures appear to be reasonable (see Table 2.1), which is probably a reflection that the largest disagreement only occurs for high solute concentrations. This is potentially important for applications in force field design as it allows one to determine if one has a correct balance between the solute-solute, solute-solvent and solvent-solvent distributions.

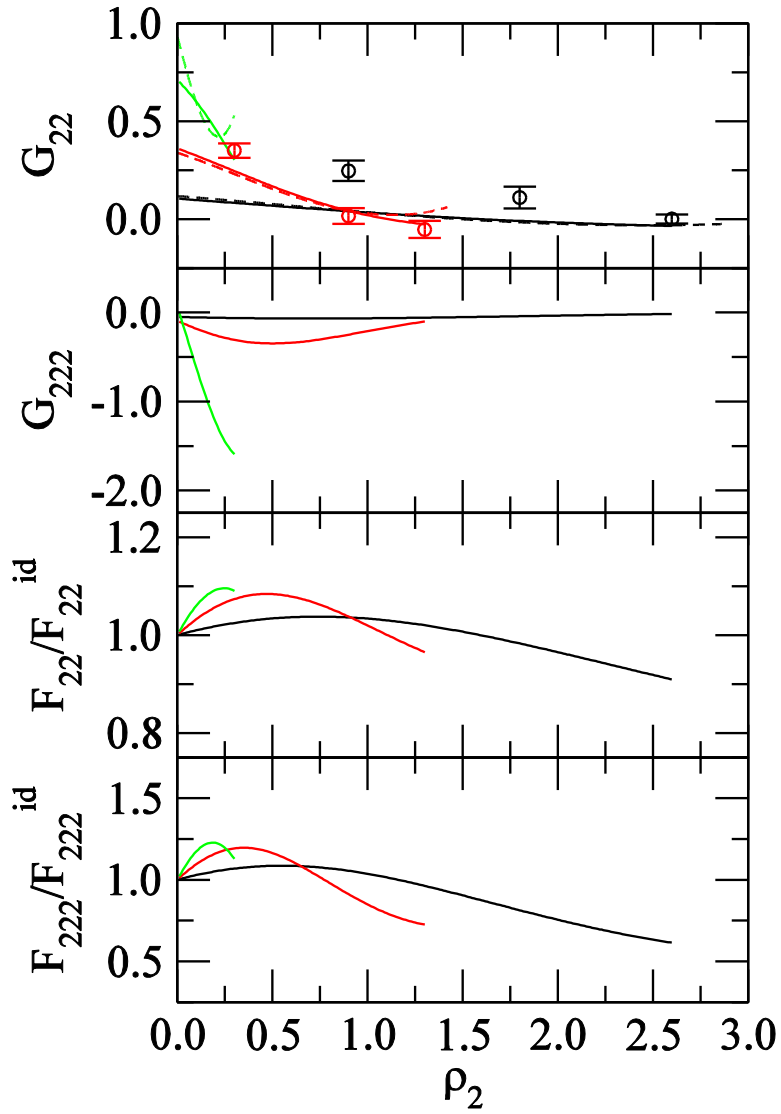


Figure 2.4 Experimental and simulated KBIs and solute fluctuations for Gly (black), Gly_2 (red) and Gly_3 (green) solutes as a function of solute molarity at 300 K. The values of G_{22} and G_{222} are in units of M^{-1} and M^{-2}, respectively. The colors correspond to the different solutes investigated here. Solid lines correspond to the current analysis of the experimental osmotic data, while dashed lines were obtained from an analysis of the corresponding experimental isothermal isobaric data. [54-56] The fluctuating quantities F_{22} and F_{222} are given by Equations 2.3 and 2.4 with ideal values of $F_{22}^{id} = F_{222}^{id} = \rho_2$. Symbols represent the simulated data.

Also displayed in Figure 2.4 are the G_{222} values. The G_{222} values quantify the role of triplet distributions towards the thermodynamic behavior of the mixture and should be zero for ideal solutions. The experimental data suggests that triplet correlations become increasingly important

for Gly_3 and display a strong dependence on concentration. In contrast, the relatively small negative values of G_{222} for Gly and Gly_2 suggest a focus on dimer association for most solute concentrations. The particle number fluctuations (Equation 2.3) are also displayed in Figure 2.4. The data display both positive and negative deviations from ideal behavior with significant deviations even at low solute concentrations. The finite values for F_{222} also indicate that the number fluctuations are not characterized by a symmetric distribution, i.e. they are non-Gaussian. Finally, we attempted to determine F_{222} from our simulations. Even for the highest (most statistically reliable) solute concentration the value of F_{222} was found to be -0.02(30) for 3 m Gly, which is essentially meaningless using the current simulation times of 25 ns or so.

Table 2.1 Experimental and simulated binary osmotic virial coefficients and KB integrals[a].

System		B_2	B_3	B_4	G_{22}^{∞}	$G_{22}^{\infty\prime}$	G_{222}^{∞}
		M^{-1}	M^{-2}	M^{-3}	M^{-1}	M^{-2}	M^{-2}
Gly	Exp	-0.104	0.082	-0.011	0.106	-0.071	-0.050
	MD	-0.260	0.075		0.260	-0.007	0.128
Gly_2	Exp	-0.361	0.484	-0.142	0.330	-0.354	-0.093
	MD	-0.529	1.051		0.529	-0.774	-0.211
Gly_3	Exp	-0.710	1.445		0.973	-0.941	0.067

[a] Obtained from a fit to Equation 2.1 with B1 = 1 at 298.15 K. Experimental data taken from 54-56.

Table 2.2 Summary of the osmotic molecular dynamics simulations[a].

System	Out		In					
	ρ_1	ρ_s	ρ_1	ρ_2	ρ_s	Π	Π^{id}	P_T
			T, μ_1					
2.0m NaCl	55.2		53.7		1.9	96	95	57
4.0m NaCl	55.2		51.9		3.6	207	180	116
6.0m NaCl	55.2		49.7		5.3	325	264	188
1.0m Gly	55.2		53.3	0.9		21	22	10
2.0m Gly	55.2		51.3	1.8		37	45	22
3.0m Gly	55.2		49.3	2.6		52	65	33
0.3m Gly_2	55.2		54.0	0.3		8	7	4
1.0m Gly_2	55.2		51.4	0.9		23	22	10
1.5m Gly_2	55.2		49.8	1.3		41	32	15
0.3m Gly_3	55.2		53.3	0.3		11	7	5
			T, μ_1, μ_3					
3.0m Gly/6.0m NaCl	49.4	5.5	40.4	2.6	5.8	68	65	369
1.5m Gly_2/6.0m NaCl	47.8	5.6	43.7	1.3	5.5	54	32	358
0.3m Gly_3/6.0m NaCl	49.2	5.5	47.6	0.3	5.2	11	7	333
			T, μ_1, ρ_3					
3.0m Gly/6.0m NaCl	55.9		41.1	2.6	5.1	348	319	186
1.5m Gly_2/6.0m NaCl	55.4		42.8	1.3	5.2	399	291	205
0.3m Gly_3/6.0m NaCl	55.3		47.9	0.3	5.2	342	267	185

[a] Mixtures of Gly_n (2) and water (1) in the presence and absence of NaCl (3) at 300 K. Number densities in units of M ($\rho_3 = 2\ \rho_s$). Pressures are in units of bar. Typical standard deviations for the measured osmotic pressures were 4 bar. The ideal osmotic pressure is given by $\beta\Pi^{id} = \rho$, where ρ is the total number density of all non-diffusable species. P_T is the total external pressure applied to each system.

Table 2.3 Simulated KB integrals and preferential interactions[a].

System	In			KBIs				Γ_{23}	Π
	ρ_1	ρ_2	ρ_s	G_{22}	G_{22}^*	G_{23}	G_{21}		
				T, μ_1					
1.0m Gly	53.3	0.9		246(52)	36		-53(2)		21
2.0m Gly	51.3	1.8		110(56)	-33		-56(5)		37
3.0m Gly	49.3	2.6		0(23)	-91		-51(3)		52
0.3m Gly_2	54.0	0.3		350(37)			-83(1)		8
1.0m Gly_2	51.4	0.9		16(41)			-83(3)		23
1.5m Gly_2	49.8	1.3		-53(44)			-80(5)		41
0.3m Gly_3	53.3	0.3		237(376)			-124(13)		11
				T, μ_1, μ_3					
3.0m Gly/6.0m NaCl	40.4	2.6	5.8	-155(2)	38	24(1)	-53(3)	1.22(5)	68
1.5m Gly_2/6.0m NaCl	43.7	1.3	5.5	-261(90)		-2(8)	-80(11)	1.12(12)	54
0.3m Gly_3/6.0m NaCl	47.6	0.3	5.2	-669(214)		-65(9)	-117(8)	0.65(13)	11

[a] Mixtures of Gly_n (2) and water (1) in the presence and absence of NaCl (3) at 300 K. Number densities in units of M ($\rho_3 = 2\ \rho_s$). Pressures in units of bar. KBIs in units of cm^3/mol. G_{22}* corresponds to the integration of G_{22} to the first minimum in the solute-solute rdf (0.62nm for Gly).

The previous analysis has centered upon the KBIs. We have argued that these are the most relevant quantities relating molecular distributions to the corresponding thermodynamics, and can provide an interpretation of solute association. However, it is more typical to analyze simulation results in terms of molecular association defined by some simple distance criteria. This is also likely to be more relevant to spectroscopic data for protein association, for example. To investigate the similarities and differences between these two viewpoints we have analyzed the degree of association of the solutes in our simulations, and investigated the effect of salt on these distributions. A detailed examination of all the solute atom-atom rdfs indicated that the only significant interaction leading to dimer or higher aggregate formation was that between the

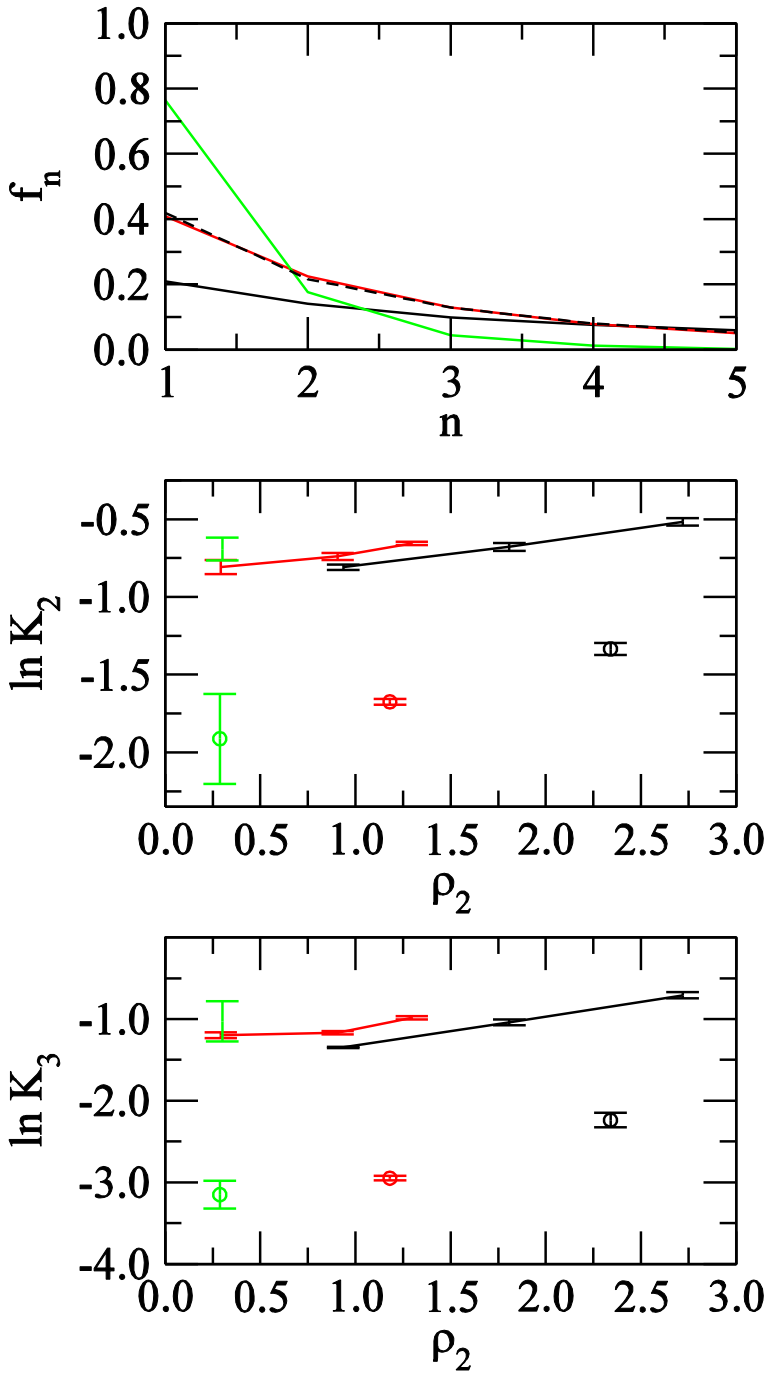


Figure 2.5 The fraction of solute molecules in an aggregate of *n* solute molecules (top) as a function of aggregate size. The equilibrium constants for dimer (middle) and trimer (bottom) formation as a function of solute molarity. See text for definitions. The solid curves correspond to 3.0m Gly (black), 1.5 m Gly_2 (red) and 0.3m Gly_3 (green), while the symbols and dashed curve represents the same solutes in 6.0m NaCl.

zwitterionic N and C terminal groups, and so the first minimum in this rdf was used to define the degree of solute aggregation. The results are presented in Table 2.3 and Figure 2.5 and Figure 2.6.

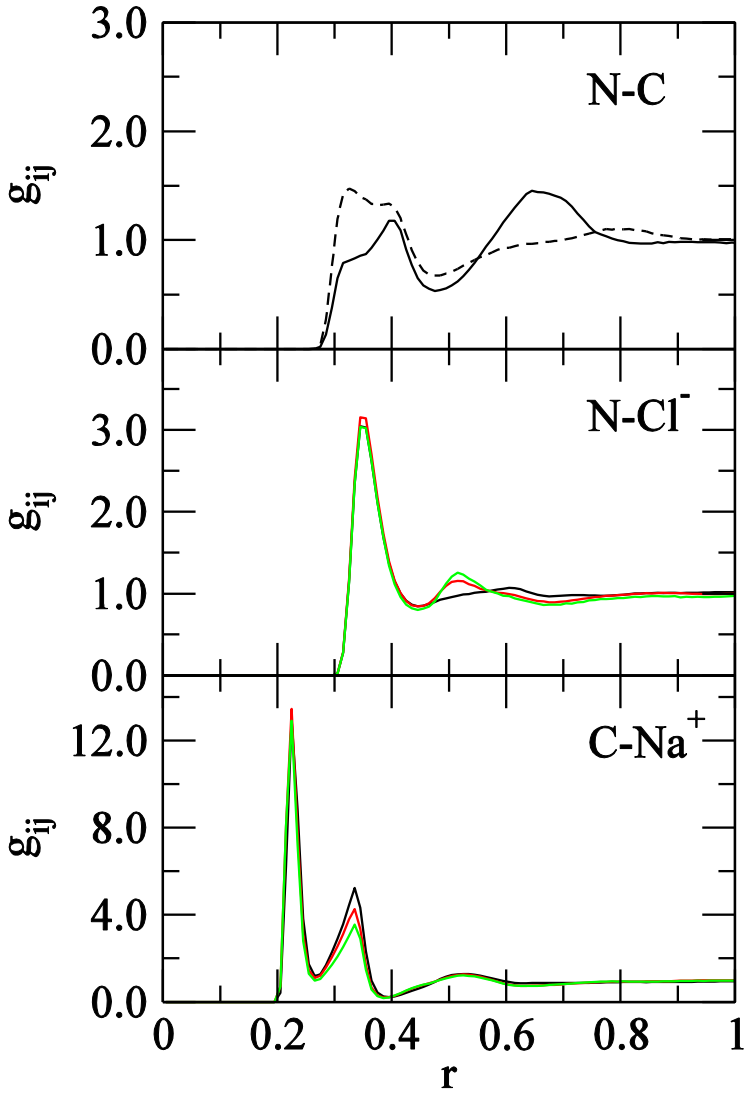


Figure 2.6 Solute-solute and solute-ion atom based rdfs. The N terminus to C terminus rdf for 3 m Gly in the absence and presence of 6m Nacl (top). The N terminus to chloride (center) and the C terminus to sodium (bottom) rdfs for various 3.0 m Gly (black), 1.5 m Gly_2 (red) and 0.3 m Gly_3 (green) concentrations in the presence of 6 m NaCl (bottom).

Figure 2.5 displays the fraction of solute molecules observed in aggregates containing *n* solute molecules during the simulations. The predominant solute form was the monomer for all solutes at all concentrations. However, as solute concentration increases it becomes more difficult to find isolated solute molecules, indicating a potential difficulty one encounters when applying such simple models to concentrated solutions. Figure 2.5 also displays the equilibrium constants for association ($\ln K_n$) as a function of solute concentration. The equilibrium constant data display an increase in solute association with solute concentration for both Gly and Gly_2. This is the opposite trend to that indicated by the previous analysis on the KBIs. However, both approaches agree that solute association (dimer or trimer) increases from Gly to Gly_3. Of course, the difference between the two approaches can be reconciled when one considers that K_n will increase with solute concentration, even if there is no net affinity between the solutes, simply because one has more solutes per unit volume. Comparison with Equation 2.21 indicates that solute association with the dimer or trimer (G_{A2}) must therefore be larger than *n* times the solute association with the monomer (G_{M2}).

The addition of salt had a dramatic effect on the solute association. This was demonstrated by both a significant drop in G_{22} as indicated in Table 2.3, and a drop in the equilibrium constants as shown in Figure 2.5. However, the underlying story was much more complicated. First, the fraction of solute molecules in either the monomer, dimer or trimer form is increased in the presence of salt. This appears to result from a decrease in the number of high *n* aggregates. Second, the equilibrium constant drops in the presence of salt primarily because the monomer concentration increases. Third, as the total concentration is decreased the fraction of monomer will naturally increase. Hence, the monomer fraction is largest for Gly_3 at the concentrations displayed in Figure

2.5, even though Gly_3 displays the largest equilibrium constant for dimer or trimer formation at equivalent solute concentrations.

The change in solute association could be attributable to either a general salt screening of the large dipole-dipole interactions between the solutes, or specific binding of anions and/or cations with the solutes such that solute association is diminished. To investigate further we present the relevant rdfs in Figure 2.6, and have also determined the corresponding preferential interaction coefficients between the solutes and NaCl, which are displayed in Table 2.3. The solute-solute N to C terminal rdf is changed on addition of NaCl. The first peak is decreased and the second peak increased in the presence of NaCl. The increased second solvation shell probability appeared to correspond to the binding of multiple Gly solutes with a shared sodium ion via their carboxylate groups. This also had an effect (-91 to 36 cm^3/mol) on the value of G_{22} truncated after the first solvation shell (denoted as G_{22}*), suggesting an increase in solute-solute contacts at short range, which must be compensated by changes at large distances. The first shell coordination numbers for the N to C termini were 0.74 and 0.48 for 3 m Gly in the absence and presence of 6m NaCl, respectively. The values of Γ_{23} were all positive indicating a net thermodynamic binding of salt ions with the solutes. However, the values of G_{21} were consistently larger than G_{23} suggesting that the greater effect was due to water exclusion from the solutes rather than ion binding. Interestingly, the G_{21} values were the same in the presence and absence of 6m NaCl. The rdfs between the ions and the terminal groups displayed in Figure 2.6 also support a role for ion binding. First shell coordination numbers were found to be 1.14 and 0.67 for the chloride and sodium ions, respectively, and were essentially the same for all three solutes. However, the net ion first shell coordination of 1.81 was significantly higher than that provided by the corresponding

thermodynamic quantities (G_{23} or Γ_{23}). Hence, changes in solute association on the addition of salt appear to be distance dependent.

2.6 Conclusions

Expressions have been provided for the analysis of binary and ternary open and closed systems using the KB theory of solutions and the corresponding KB integrals. The KBIs provide an alternative to the cluster integrals in the MM expressions, which are much easier to determine from simulations of concentrated solutions. The expressions have been illustrated using both experimental and simulation data for small Gly, Gly_2 and Gly_3 zwitterionic peptide solutes in the presence and absence of NaCl. Two measures of solute association were investigated and found to provide different viewpoints of the association process. A thermodynamic measure of solute association is provided by G_{22}, and this is aided by the additional information concerning triplet correlations provided by G_{222}. The experimental and simulation data indicated that solute association prevails at low concentrations and increases within the series Gly < Gly_2 < Gly_3. In addition, solute association decreases with increasing solute concentration for all the solutes. A more physical measure of solute association was investigated and expressed in terms of equilibrium constants for dimer and trimer formation. Here, an increase in the equilibrium constants was observed on increasing the solute concentration, in contrast to the thermodynamic measure of association. The differences arise as the thermodynamic measure includes changes to the solute distribution over all distances, while the physical measure focuses primarily on the first solvation shell. The addition of salt to solutions of Gly_n solutes reduces the values of G_{22} and the equilibrium constants for association. Further analysis of 3m Gly solutions indicated that this was a consequence of the disruption of larger aggregates leading to an increase in the number of monomers, dimers, and trimers. The overall global (long range) effect was clearly solute

disassociation as indicated by the decrease of G_{22} in the presence of NaCl, whereas solute association increased at the local (first shell) level. This suggests an overall salt screening effect that includes a local increase in dimer and trimer formation due to the binding of sodium ions with multiple solute carboxylate groups. It should be noted that, while the value of G_{22} is the most thermodynamically relevant quantity, a clear physical interpretation is often difficult as it probes changes in the solute-solute distribution over multiple solvation shells. In contrast, the physical picture of association is quite clear, but often subjective and not necessarily thermodynamically relevant. The present results therefore illustrate the advantages of a combination of KB theory and computer simulations data provide for the interpretation of complex solution behavior.

2.7 References

1. McMillan Jr, W. G.; Mayer, J. E. *J. Chem. Phys.* **1945***, 13*, 276-305.

2. Hill, T. L. *J. Am. Chem. Soc.* **1957***, 79*, 4885-4890.

3. Hill, T. L. *J. Chem. Phys.* **1959***, 30*, 93-97.

4. O'connell, *J. Mol. Phys.* **1971***, 20*, 27-33.

5. Anderson, C.; Courtenay, E.; Record, M. *J. Phys. Chem. B* **2002***, 106*, 418-433.

6. Curtis, R.; Newman, J.; Blanch, H.; Prausnitz, J. *Fluid Phase Equilib.* **2001***, 192*, 131-153.

7. Timasheff, S. N. *Adv. Protein Chem.* **1998***, 51*, 355-432.

8. Cannon, J. G.; Anderson, C. F.; Record, M. T. *J. Phys. Chem. B* **2007***, 111*, 9675-9685.

9. Lenhoff, A. M. *AIChE J.* **2003***, 49*, 806-812.

10. Moon, Y.; Curtis, R.; Anderson, C.; Blanch, H.; Prausnitz, J. *J. Solution Chem.* **2000***, 29*, 699-718.

11. Moon, Y.; Anderson, C.; Blanch, H.; Prausnitz, J. *Fluid Phase Equilib.* **2000***, 168*, 229-239.

12. Kirkwood, J. G.; Buff, F. P. *J. Chem. Phys.* **1951***, 19*, 774-777.

13. Ben-Naim, A. *Molecular theory of solution*, Oxford University Press: New York 2006.

14. Smith, P. E. *J. Phys. Chem. B* **2004***, 108*, 18716-18724.

15. Cabezas Jr, H.; O'Connell, J. P. *In.d Eng. Chem. Res.* **1993***, 32*, 2892-2904.

16. Terdale, S. S.; Dagade, D. H.; Patil, K. J. *J. Phys. Chem. B* **2006***, 110*, 18583-18593.

17. Blanco, M. A.; Sahin, E.; Li, Y.; Roberts, C. J. *J. Chem. Phys.* **2011***, 134*, 225103.

18. Chitra, R.; Smith, P. E. *J. Phys. Chem. B* **2001***, 105*, 11513-11522.

19. Shimizu, S. *Proc. Natl. Acad. Sci. U. S. A.* **2004***, 101*, 1195-1199.

20. Shimizu, S.; Matubayasi, N. *Chem. Phys. Lett.* **2006***, 420*, 518-522.

21. Shulgin, I. L.; Ruckenstein, E. *J. Chem. Phys.* **2005***, 123*, 054909.

22. Schurr, J. M.; Rangel, D. P.; Aragon, S. R. *Biophys. J.* **2005***, 89*, 2258-2276.

23. Perry, R. L.; O'Connell, J. P. *Mol. Phys.* **1984***, 52*, 137-159.

24. Ben-Naim, A. *J. Chem. Phys.* **1975***, 63*, 2064-2073.

25. Gee, M. B.; Smith, P. E. *J. Chem. Phys.* **2009***, 131*, 165101.

26. Jiao, Y.; Smith, P. E. *J. Chem. Phys.* **2011***, 135*, 014502.

27. Allen, M.; Tildesley, D. *Computer Simulation of Liquids*, Oxford University Press: New York, 1987.

28. Paulsen, M.; Anderson, C.; Record, M. *Biophys. J*. **1988**, *53*, A483.

29. Lynch, G. C.; Perkyns, J. S.; Pettitt, B. M. *J. Comput. Phys.* **1999***, 151*, 135-145.

30. Murad, S.; Powles, J. *J. Chem. Phys.* **1993***, 99*, 7271-7272.

31. Powles, J. G.; Murad, S.; Holtz, B. *Chem. Phys. Lett.* **1995***, 245*, 178-182.

32. Luo, Y.; Roux, B. *J. Phys. Chem. Lett.* **2010***, 1*, 183-189.

33. McQuarrie, D. A. *Statistical Mechanics*, Harper & Row: New York, 1976.

34. Kusalik, P. G.; Patey, G. *J. Chem. Phys.* **1987***, 86*, 5110-5116.

35. Weerasinghe, S.; Smith, P. E. *J. Chem. Phys.* **2003***, 119*, 11342-11349.

36. Smith, P. E. *J. Phys. Chem. B* **2006***, 110*, 2862-2868.

37. Smith, P. E. *Biophys. J.* **2006***, 91*, 849-856.

38. Zhou, H.; Dill, K. A. *Biochemistry (N. Y.)* **2001***, 40*, 11289-11293.

39. Zhou, H. X.; Rivas, G.; Minton, A. P. *Annu. Rev. Biophys.* **2008***, 37*, 375-397.

40. Timasheff, S. N. *Proc. Natl. Acad. Sci. U. S. A.* **2002***, 99*, 9721-9726.

41. Aburi, M.; Smith, P. E. *J. Phys. Chem. B* **2004***, 108*, 7382-7388.

42. Kang, M.; Smith, P. E. *J. Comput. Chem.* **2006***, 27*, 1477-1485.

43. Ploetz, E. A.; Bentenitis, N.; Smith, P. E. *Fluid Phase Equilib.* **2010***, 290*, 43-47.

44. Berendsen, H.; Grigera, J.; Straatsma, T. *J. Phys. Chem.* **1987***, 91*, 6269-6271.

45. Hess, B.; Kutzner, C.; Van Der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008***, 4*, 435-447.

46. Berendsen, H. J.; Postma, J. v.; van Gunsteren, W. F.; DiNola, A.; Haak, J. *J. Chem. Phys.* **1984***, 81*, 3684-3690.

47. Hess, B.; Bekker, H.; Berendsen, H. J.; Fraaije, J. G. *J. Comput. Chem.* **1997***, 18*, 1463-1472.

48. Miyamoto, S.; Kollman, P. A. *J. Comput. Chem.* **1992***, 13*, 952-962.

49. Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993***, 98*, 10089-10092.

50. Weerasinghe, S.; Smith, P. E. *J. Chem. Thermodyn B* **2003***, 107*, 3891-3898.

51. Matteoli, E.; Lepori, L. *J. Chem. Phys.* **1984***, 80*, 2856-2863.

52. Alejandre, J.; Tildesley, D. J.; Chapela, G. A. *J. Chem. Phys.* **1995***, 102*, 4574-4583.

53. Kang, M.; Smith, P. E. *Int. J. Thermophys.* **2010***, 31*, 793-804.

54. Ellerton, H. D.; Reinfelds, G.; Mulcahy, D. E.; Dunlop, P. J. *J. Phys. Chem.* **1964***, 68*, 398-402.

55. Smith, E. R.; Smith, P. K. *J. Biol. Chem.* **1940***, 135*, 273-279.

56. Venkatesu, P.; Lee, M.; Lin, H. *J. Chem. Thermodyn.* **2007***, 39*, 1206-1216.

57. Robinson, R. A.; Stokes, R. H. Electrolyte Solutions, Butterworths: London, 1959.

58. Sohnel, O.; Novotny, P. Densities of Aqueous Solutions of Inorganic Substances, Elsevier: Amsterdam, 1985.

Chapter 3 - Interactions of Amino Acids in Aqueous Solutions

3.1 Abstract

Amino acids are the building blocks of proteins. Solvent mediated interactions between amino acids determine protein structure, protein association, and protein aggregation. Although an understanding of the behavior of proteins in aqueous solutions is important for our understanding, design, and optimization of biological systems, the underlying molecular level interactions are poorly understood. Here, we have attempted to quantify the interactions between amino acids in aqueous solutions using the Kirkwood-Buff (KB) theory of solutions, which provides a link between the molecular interactions and the corresponding solution thermodynamics. The results are illustrated using computer simulations for various concentrations of the twenty zwitterionic amino acids at ambient temperature and pressure. The results are discussed in terms of the preferential (solute over solvent) interactions between the amino acids. Here one can observe that hydrophobic amino acids remain well solvated in the zwitterionic form, but they are observed to associate in the capped form. It is also revealed that the protonation of amino acids with negatively charged polar side chains significantly increases self-association.

3.2 Introduction

Peptide and protein aggregation are directly involved with many age-related diseases and aging itself.[1-4] A better understanding of protein aggregation would hopefully lead to the prediction and even prevention, of these the undesirable conditions. Hence, a number of studies have been pursued to understand and predict the misfolding and subsequent aggregation of proteins.[4,5] However, it is still unclear why certain peptides and proteins tend to aggregate.

In principle, aggregation in a solution mixture results from a shifted balance in the intermolecular interactions between solute and solvent. If the solute-solute interactions are larger than solute-solvent interactions, self-association is likely to occur, and vice versa: i.e. the tendency for aggregation can be predicted using the difference between solute-solute and solute-solvent interactions. Hence, it is reasonable to express the difference between solute-solute and solute-solvent interactions using a quantitative term. The concept of preferential interactions, PI has been introduced previously.[6] Moreover, KB integrals can play an important role in quantifying these PIs.[6] Furthermore, Kang and Smith have developed a pairwise preferential interaction model based on KB integrals to quantify interactions between some of the functional groups commonly observed in peptides.[7] In addition, Karunaweera *et al.* have performed a detailed analysis of experimental and simulation data of glycine monomer, dimer and trimer using the KB theory.[8]

Proteins are usually large molecules consisting of twenty traditional amino acids as building blocks. Therefore, it would be more useful to quantify the interactions between amino acids, or even between functional groups, rather than to deal with the protein as a whole and then maybe we can use the corresponding results to predict the behavior of peptides and proteins. Furthermore, if all twenty traditional amino acids are considered, there are two hundred and ten [(20*19/2) + 20] possible combinations. Hence, initially we are going to focus on just the

self-interactions of individual amino acids in this study. The main aim is to understand the interactions between amino acids in aqueous solutions. In addition, we would also like to determine how amino acid interactions vary with composition. Finally, we will try to quantify the amino acid interactions in terms of PIs.

3.3 Methods

3.3.1 Thermodynamics of Solutions and KB Theory

The notation used here follows the common definition where the subscripts 1 and 2 refer to the solvent (water) and the solute, respectively. The chemical potential, μ plays an important role in thermodynamic changes in a system. Under thermodynamic control, changes in the chemical potential of a species in a system reflects how the species can bring about a change in the system: both physical and chemical changes. According to statistical mechanics, the chemical potential of a species can be expressed as,[9]

$$\mu = W + RT\ ln[\Lambda^3 \rho q^{-1}] = \mu^* + RT\ ln[\Lambda^3 \rho] \qquad (3.1)$$

Here, ρ=N/V is the number density (or molar concentration), N is the number of the species in the system, V is the volume of the system, q is the internal partition function of a molecule and Λ is the thermal de Broglie wavelength of the species. The first term, W quantifies contributions of the interactions among molecules to the chemical potential on the addition of a molecule. If there is no interaction in the system, W = 0 and only the second term, $RT\ ln[\Lambda^3 \rho q^{-1}]$ will be left, simply indicating the chemical potential of an ideal gas at the same temperature and density. Ben-Naim has introduced, μ* = W- RT ln q to represent the pseudo chemical potential.[9] The pseudo chemical potential captures the free energy change for transfer of a molecule from a fixed position in a

vacuum to a fixed position in the solution. This will be the same as the work required for the corresponding cavity formation.[9] Using the pseudo chemical potential, the contribution from the entropy of mixing can be eliminated which is not directly related to the intermolecular interactions.[9] Moreover, expressions for changes or derivatives of the pseudo chemical potential, as well as the total chemical potential required in Equation (3.1), can be easily obtained.

3.3.2 Preferential Interactions

Using KB theory, it is quite easy to show that for any thermodynamically stable mixture of a solute (2) and solvent (1) we can write that,[9]

$$-\beta\left(\frac{\partial\mu_2^*}{\partial\rho_2}\right)_{T,P} = -\left(\frac{\partial\, ln\, y_2}{\partial\rho_2}\right)_{T,P} = \frac{G_{22}-G_{21}}{1+\rho_2(G_{22}-G_{21})} \tag{3.2}$$

where y_2 is the molar activity coefficient of the solute. The above expression reduces to the numerator in the limit of infinite dilution of the solute (2). The value of G_{22}-G_{21} at infinite dilution is the central quantity of interest in this work. It is defined as the preferential interaction (PI) between two infinitely dilute solute molecules in a binary system.[7]

$$PI = G_{22} - G_{21} \tag{3.3}$$

The PI defined here is the same as previous definitions of preferential solvation, PS[10] except for the infinitely dilute solute restriction. However, it will be used in a different manner. The PI at infinite dilution of solute quantifies the interaction between two solute molecules in a large excess of solvent. It results from a balance between solute-solute and solute-solvent interactions. A positive value for PI indicates a favorable solute-solute interaction which tends towards solute association or aggregation where as a negative value indicates a favorable solvation

which tends towards solute hydration and low solute self-association. A value of zero indicates a balance of the interactions, i.e. an ideal solution. The above expression indicates that if the molar activity coefficient decreases with molarity, then the solute must display a tendency towards self-association. The approach therefore provides a way to quantify the degree of molecular association.

3.3.3 Molecular Dynamics Simulations

All molecular dynamics simulations were performed using the KBFF models (http://kbff.chem.k-state.edu)[11-13] together with the SPC/E water model[14] as implemented in the GROMACS 4.0.5 package.[15] All simulations were performed at 300 K and the pressure of 1 bar using the weak coupling technique to modulate the temperature and pressure with relaxation times of 0.1 and 0.5 ps,[16] respectively. A time-step of 2 fs was used and the bond lengths were constrained using the Lincs (solutes) and Settle (water) algorithms.[17,18] The particle mesh Ewald technique was used to evaluate electrostatic interactions with a grid resolution of 0.1 nm.[19] A real space convergence parameter of 3.5 nm^{-1} was used in combination with twin range cutoffs of 1.0 and 1.5 nm, and a nonbonded update frequency of 10 steps. Random initial configurations of molecules in a cubic box were used to study the all systems. Initial configurations of the different solutions were generated from a cubic box ($L \approx 6.0$ nm) of equilibrated water molecules by randomly replacing waters with solutes until the required concentration was attained. The steepest descent method was then used to perform 100 steps of energy minimization. This was followed by extensive equilibration, which was continued until the rdfs displayed no drift with time (typically 5 ns). Total simulation times were in the 25-50 ns range, and the final 25-30 ns were used for

calculating ensemble averages. Configurations were saved every 0.1 ps for the calculation of various properties. Errors (±1σ) in the simulation data were estimated by using five block averages.

3.4 Results and Discussion

Although the experimentally reported solubility values of most of the amino acids are relatively low according to Table 3.1, here we used a concentration of 1.0 m in order to improve the statistics of the results. This is according to the basis that, the lower the concentration, the greater the error in the required integrals. On the other hand, when a concentration of 1.0 m is used many amino acids are well above the experimental solubility limit (Table 3.1). Hence, there are possibilities for phase separations and meta stable states (i.e. an excited state with a longer life time than other excited states). Nevertheless, Pettitt and coworkers have simulated urea solutions at different concentrations above the solubility limit to examine the structures which are responsible for the thermodynamic solution properties and they did not observe nucleation during these relatively short simulation times. [20]

3.4.1 The Effect of Concentration on Amino Acid Interactions

As mentioned above most of the amino acids have relatively low solubility values according to Table 3.1. However, the experimentally reported solubility value for proline is quite high and that of glycine is relatively high too. Therefore, glycine was simulated at several different compositions ranging from 0.5 m to 10.0 m and the resulting radial distribution functions are presented in Figure 3.1. The experimental solubility value of glycine is 3.679 m and according to Figure 3.1, the rdfs are behaving normally well above the experimental solubility limit. Furthermore, as expected the solute-solvent rdf is more prominent than the solute-solute rdf and

Table 3.1 Experimental solubility values of amino acids at 298 K.[21]

Amino acids	Solubility/m
Gly	3.679
Ala	1.972
Val	0.5098
Leu	0.1898
Ile	0.2698
Met	0.4048
Pro	14.8
Phe	0.1945
Trp	0.05989
Ser	4.529
Tyr	0.002964
Cys	0.001058
Asp	0.04463
Glu	0.06927

both rdfs converge to unity after 1.5 nm. Hence, those rdfs can be integrated to obtain the respective KBIs and they can then be used to calculate the preferential interactions which is the central quantity utilized throughout this study. Since rdfs behave normally as the concentration increases, it implies that there is no phase separation and therefore, these results would provide reasonable information concerning interactions at lower concentrations. Moreover, in reality we will encounter such lower concentrations most of the time in biological systems and the insights gained here will be quite significant.

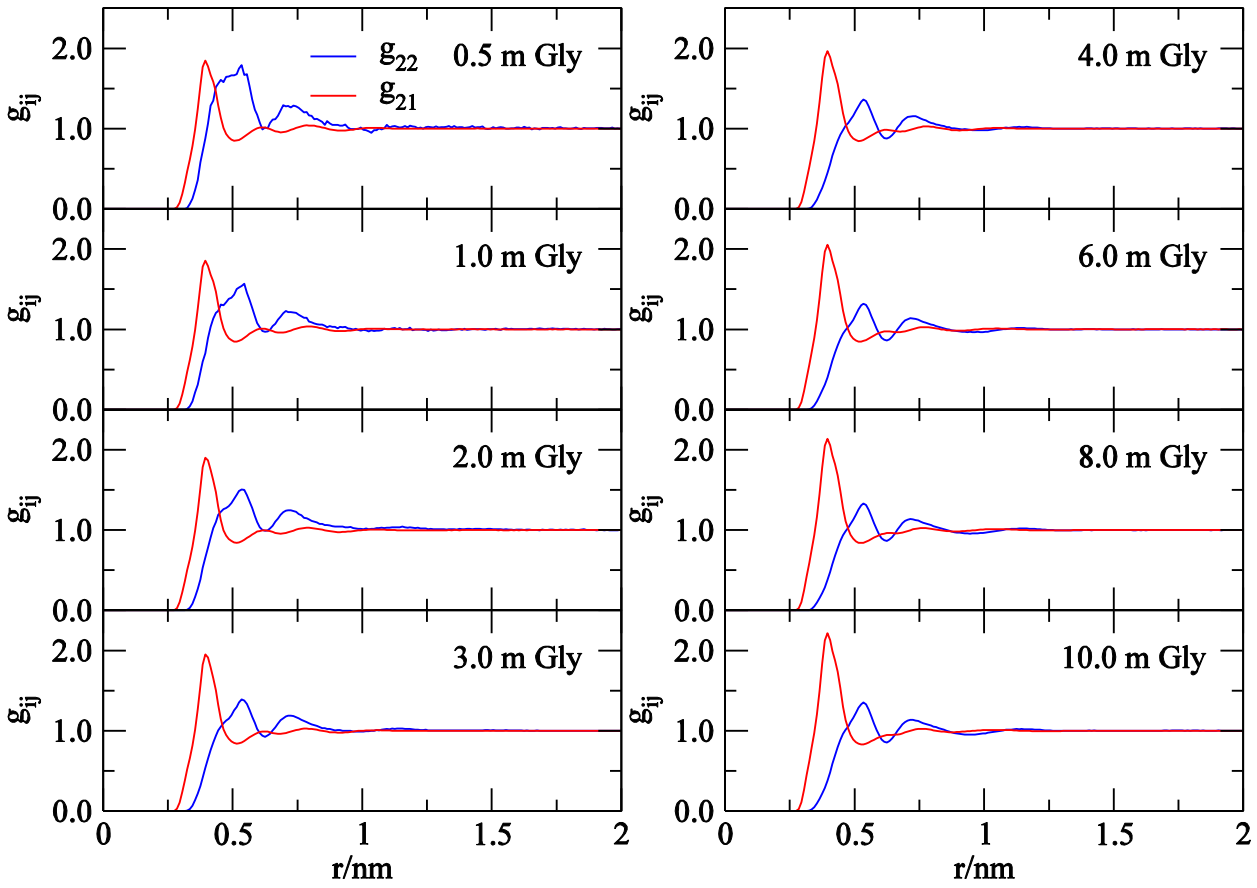


Figure 3.1 The center of mass to center of mass solute-solute (blue) and solute-solvent (red) rdfs of Gly at different compositions at 300 K.

The variation of PIs with the composition are presented in Figure 3.2 and the PIs decrease when the concentration is increased. Although the simulated PIs, represented by the blue symbols in Figure 3.2, are relatively high compared to the experimental PIs, which are shown by the red curve, the simulated PIs follow the same trend as the experimental PIs. Hence, we can claim that our models are in reasonable agreement with the experiments.

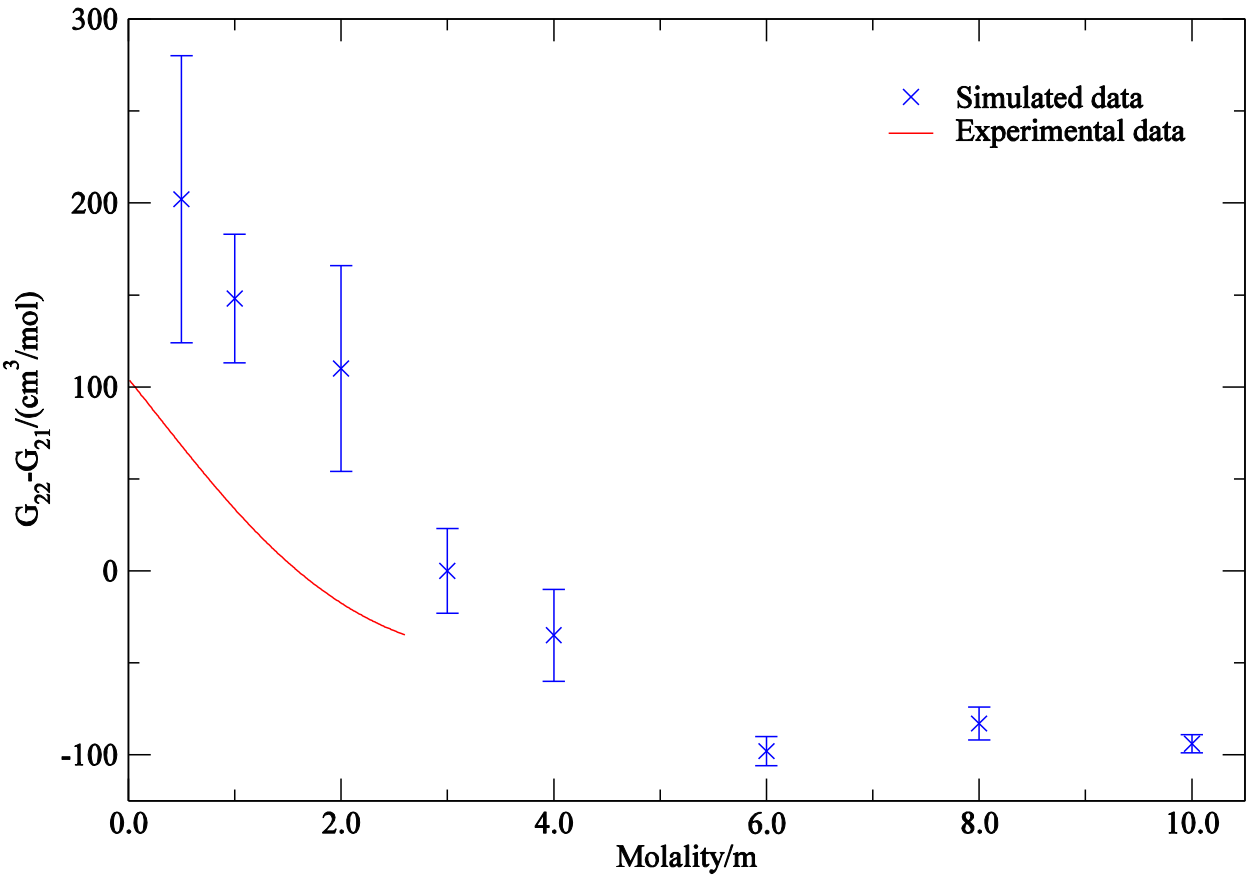


Figure 3.2 PIs of glycine vs composition at 300 K.[22-24]

3.4.2 The Differences in Interactions Among Different Classes of Amino Acids

All twenty traditional amino acids were simulated at a concentration of 1.0 m in their zwitterionic form to study the type of interactions which can exist between different classes of amino acids. The solute-solute and solute-solvent rdfs of the amino acids with nonpolar side chains are presented in Figure 3.3 and all of them converge to unity beyond 1.5 nm except for tryptophan. Neither the solute-solute rdf of tryptophan nor the solute-solvent rdf are converged and this will be discussed further in this section of the chapter. Moreover, the solute-solute and solute-solvent rdfs of the amino acids with uncharged polar side chains are presented in Figure 3.4 and those of amino acids with charged polar side chains are presented in Figure 3.5, respectively. In addition,

all of these rdfs converge to unity beyond 1.5 nm and they can be integrated to calculate the KBIs which then can be used to quantify the interactions between amino acids.

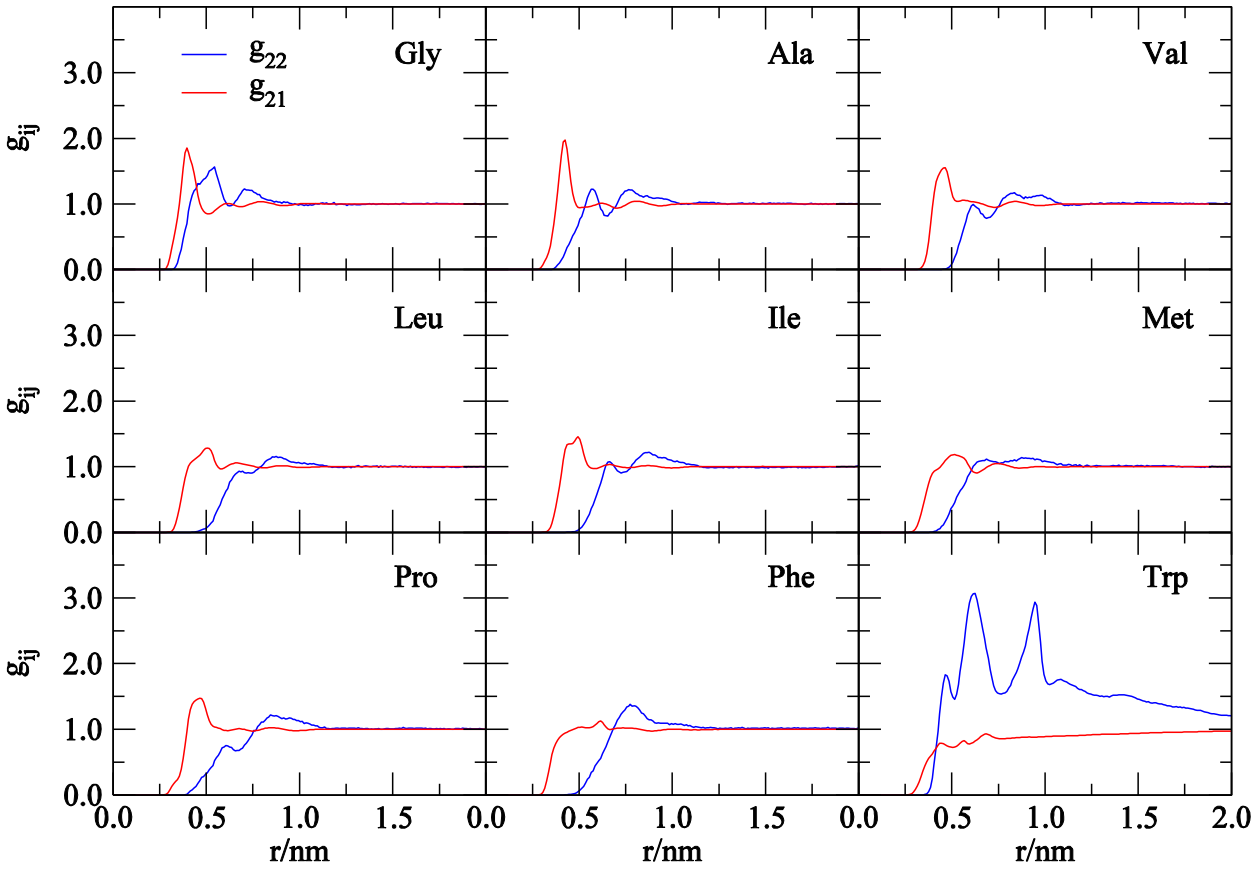


Figure 3.3 The center of mass to center of mass solute-solute (blue) and solute-solvent (red) rdfs of 1.0 m zwitterionic amino acids with nonpolar side chains at 300 K.

The PIs of the twenty traditional amino acids are summarized in Table 3.2 and are expressed as individual PIs and the difference between PI(X) and PI(Gly), ΔPI, where X is the respective amino acid except glycine. As expected of hydrophobic amino acids, glycine has a moderately positive individual PI that indicates it prefers solute-solute interactions which tends

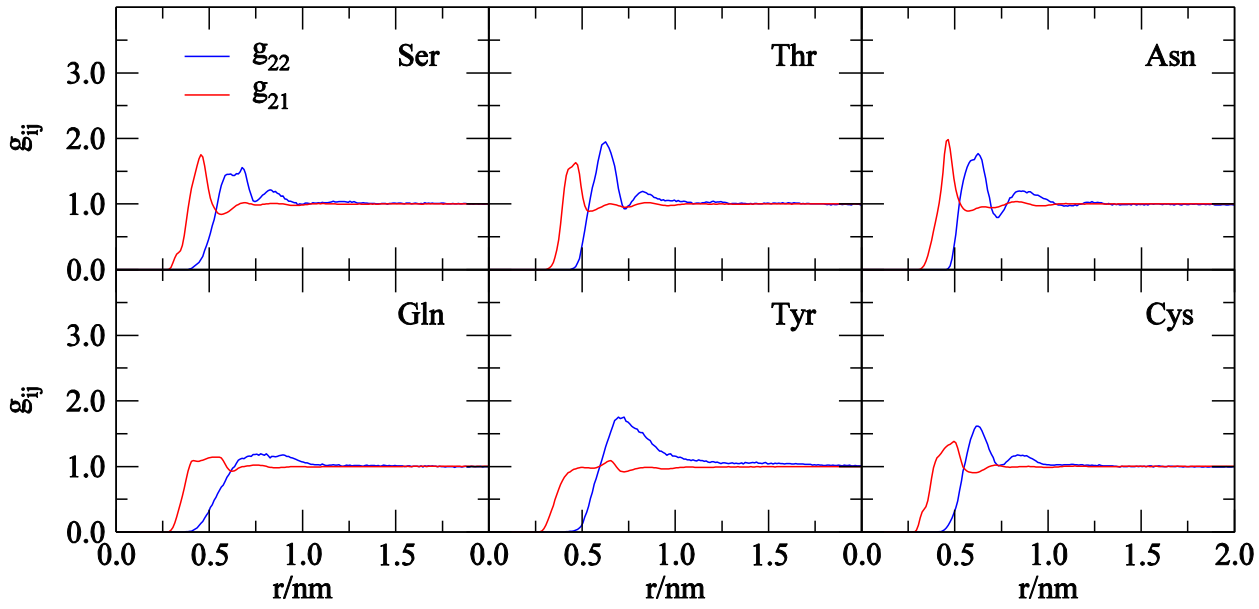


Figure 3.4 The center of mass to center of mass solute-solute (blue) and solute-solvent (red) rdfs of 1.0 m zwitterionic amino acids with uncharged polar side chains at 300 K.

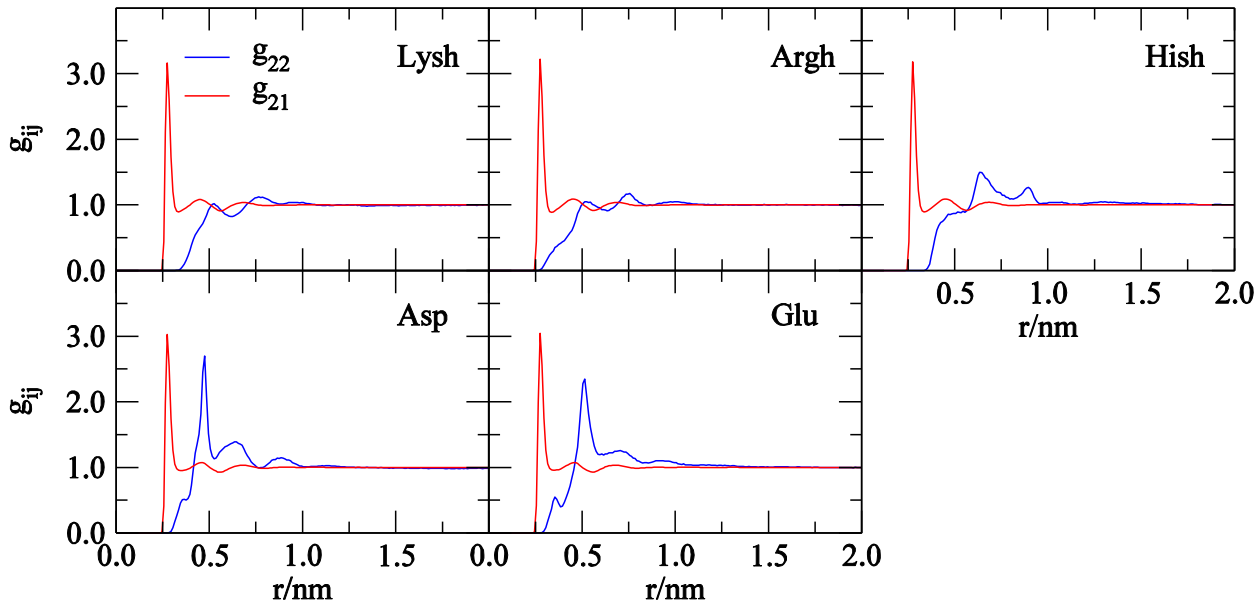


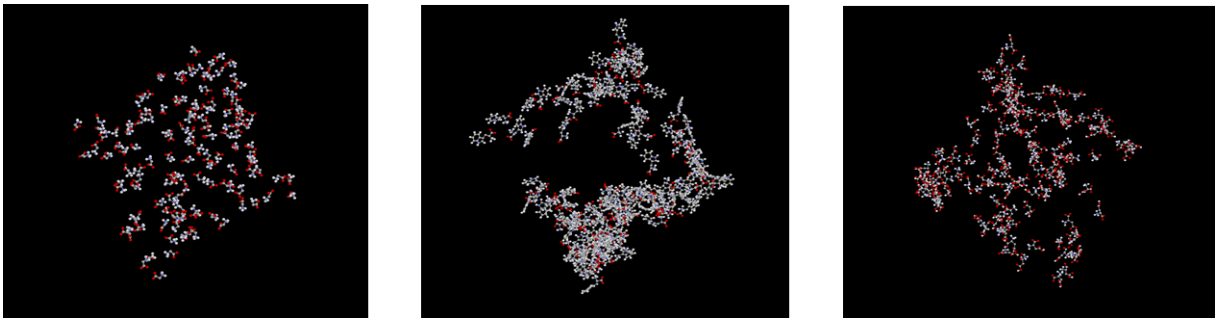
Figure 3.5 The center of mass to center of mass solute-solute (blue) and solute-solvent (red) rdfs of 1.0 m zwitterionic amino acids with charged polar side chains at 300 K.

Table 3.2 PIs (cm^3/mol) of 1.0 m zwitterionic amino acids at 300 K.

Amino acids	PI	ΔPI=PI(X)-PI(Gly)
Gly	148 ± 35	-
Ala	-147 ± 56	-295 ± 66
Val	-447 ± 53	-595 ± 64
Leu	-456 ± 66	-604 ± 75
Ile	-293 ± 17	-441 ± 39
Met	-123 ± 29	-271 ± 45
Pro	-358 ± 45	-506 ± 57
Phe	-87 ± 51	-235 ± 62
Trp	(9282) ± 1503	(9134) ± 1503
Ser	132 ± 119	-16 ± 124
Thr	-23 ± 73	-171 ± 81
Asn	-44 ± 70	-192 ± 78
Gln	-14 ± 86	-162 ± 93
Tyr	851 ± 153	703 ± 157
Cys	159 ± 48	11 ± 59
Lysh	-121 ± 36	-269 ± 50
Argh	-95 ± 38	-243 ± 52
Hish	240 ± 70	92 ± 78
Asp	359 ± 86	211 ± 93
Glu	241 ± 52	93 ± 63

toward solute association or aggregation. However, all the other hydrophobic amino acids which have nonpolar side chains except for tryptophan have moderate to relatively high negative individual PIs. This implies that they are more solvated which tends toward solute hydration and low solute self-association. Moreover, the observed opposite trend is a consequence of the interactions between the side chains and the termini of those amino acids. Although we expected that they would aggregate since they are hydrophobic, in order to do so the hydrophobic side chains

would have to intervene with the zwitterion solvation shells and this is unfavorable. Furthermore, tryptophan has a very large positive individual PI and it indicates that it is close to being phase separated. This abnormal behavior was indicated earlier by the rdfs which were not converged and it is confirmed by Figure 3.6 (b). Also shown in Figure 3.6 are two other examples of the types of associations encountered, one being a moderate association in glycine, Figure 3.6 (a) and the other one being an intermediate association in aspartic acid with an uncharged side chain, Figure 3.6 (c). In addition, among the amino acids with charged polar side chains, the ones with negatively charged side chains seem to aggregate more than the ones with positively charged side chains.



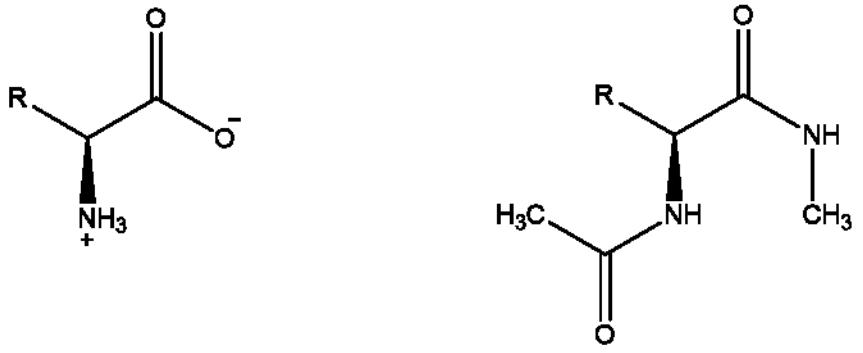
(a) (b) (c)

Figure 3.6 Snap shots of 1.0 m (a) Gly, (b) Trp and (c) Asph during molecular dynamic simulations.

Since glycine is the simplest amino acid with no side chain, when the PI of glycine is subtracted from a PI of another amino acid, the resulting quantity can be used to help interpret the side chain-side chain interactions, although one can argue that there can still be side chain-zwitterion interactions too. The ΔPI values of all hydrophobic amino acids except for glycine and tryptophan reported in Table 3.2 are negative which indicates that the hydrophobic side chains are solvated instead of being aggregated. Furthermore, this observation confirms that hydrophobic amino acids are more solvated in the zwitterionic form.

3.4.3 The Quantification of Amino Acid Interactions in Terms of Zwitterionic and Capped Forms

In order to investigate whether there is an effect from the type termini of the amino acids, i.e. when they are charged (zwitterionic) and uncharged (capped), few amino acids were simulated in both forms at a concentration of 1.0 m. The structures of those two forms are shown in Figure 3.7.



(a) (b)

Figure 3.7 Structures of (a) zwitterionic and (b) capped amino acids. R group represents the side chain of amino acids.

The solute-solute and solute-solvent rdfs for both zwitterionic and capped forms of amino acids with nonpolar side chains are presented in Figure 3.8 and they are converged to unity after 1.5 nm or close to 2.0 nm. Moreover, the rdfs of amino acids with uncharged polar side chains are presented in Figure 3.9 for both zwitterionic and capped forms and all of them are converged to unity after 1.5 nm except for capped version of tyrosine which will be discussed later in this section. Furthermore, the rdfs of amino acids with charged polar side chains are presented in Figure 3.10 for both zwitterionic and capped forms and all of them are converged to unity after 1.5 nm.

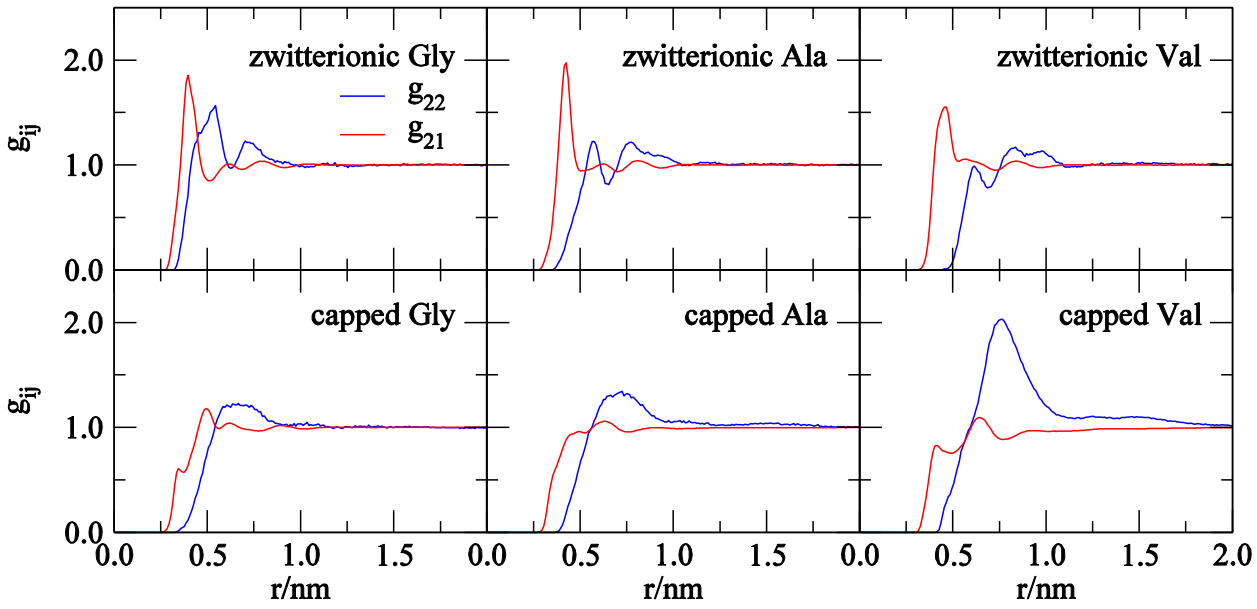


Figure 3.8 The center of mass to center of mass solute-solute (blue) and solute-solvent (red) rdfs of 1.0 m zwitterionic and capped amino acids with nonpolar side chains at 300 K.

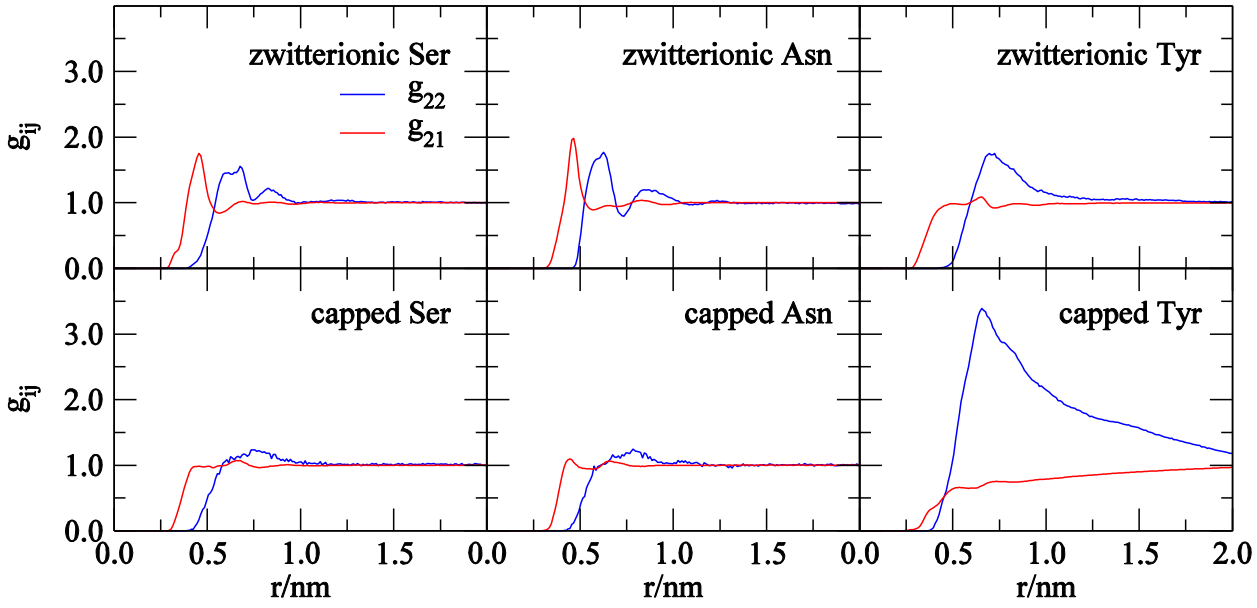


Figure 3.9 The center of mass to center of mass solute-solute (blue) and solute-solvent (red) rdfs of 1.0 m zwitterionic and capped amino acids with uncharged polar side chains at 300 K.

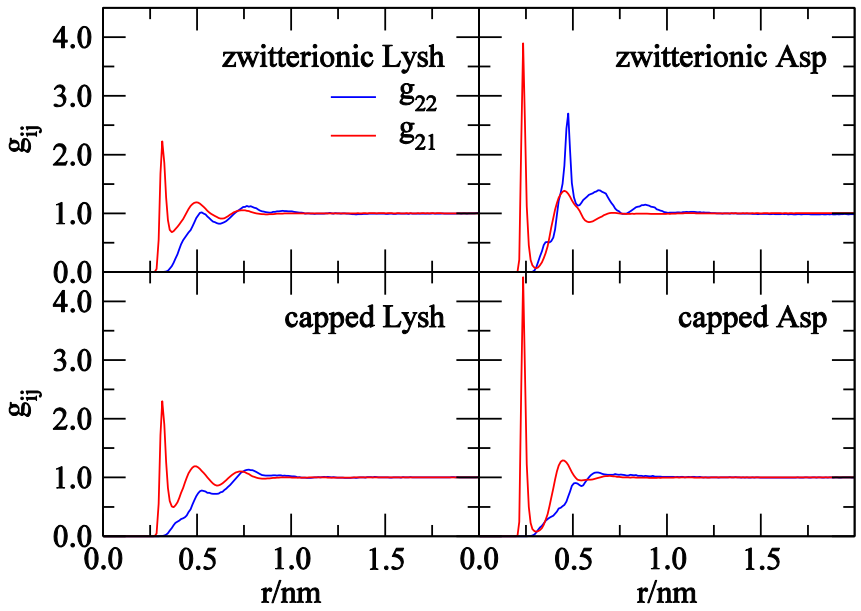


Figure 3.10 The center of mass to center of mass solute-solute (blue) and solute-solvent (red) rdfs of 1.0 m zwitterionic and capped amino acids with charged polar side chains at 300 K.

The PIs for the selected group of amino acids in both zwitterionic and capped forms are presented in Table 3.3. Although there are couple of significant changes in PIs when the type of the termini changes, in most cases the sign of the PI is the same indicating that the type of the interaction will be the same, i.e. it would be an association or a solvation, regardless of the type of the termini. The two of the most significant changes in PIs are in valine and tyrosine. In capped valine solutions there were no apparent side chain-side chain association where as in capped tyrosine solutions there were interactions between the side chain OH and the backbone C terminus (Figure 3.11) and these reasons can be attributed to the significant changes in the observed PIs.

Table 3.3 PIs (cm^3/mol) of 1.0 m zwitterionic vs capped amino acids at 300 K.

Amino acids	Zwitterion PI	Capped PI
Gly	148 ± 35	20 ± 81
Ala	-147 ± 56	-15 ± 105
Val	-447 ± 53	848 ± 274
Ser	132 ± 119	149 ± 47
Asn	-44 ± 70	-163 ± 35
Tyr	851 ± 153	(9820) ± 1558
Lysh	-121 ± 36	-249 ± 12
Asp	359 ± 86	-170 ± 8

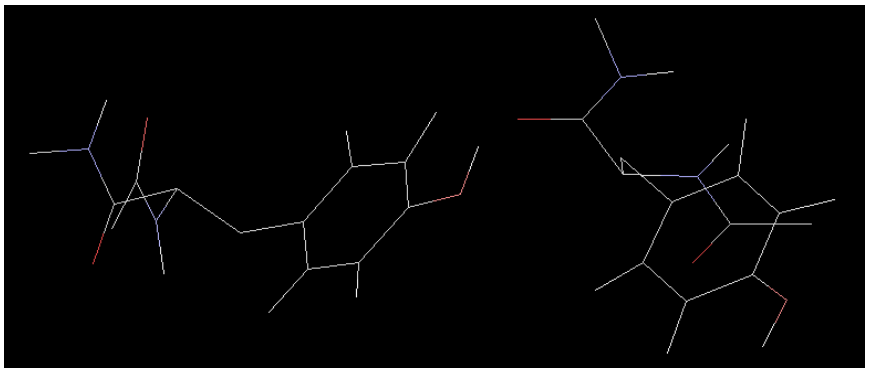


Figure 3.11 A snap shot of capped tyrosine at 300 K.

3.4.4 The Contribution from Uncharged and Charged Polar Side Chains Toward Amino Acid Interactions

Amino acids with charged polar side chains were simulated in two forms, i.e. with the charged side chain which is the normal charged form in which those amino acids exist, and either by removing a hydrogen ion from the amino acids with positively charged side chains (lysine and histidine), or by adding a hydrogen ion to the amino acids with negatively charged side chains

(aspartic acid and glutamic acid) to make the side chain uncharged. The structures of them are presented in Figure 3.12.

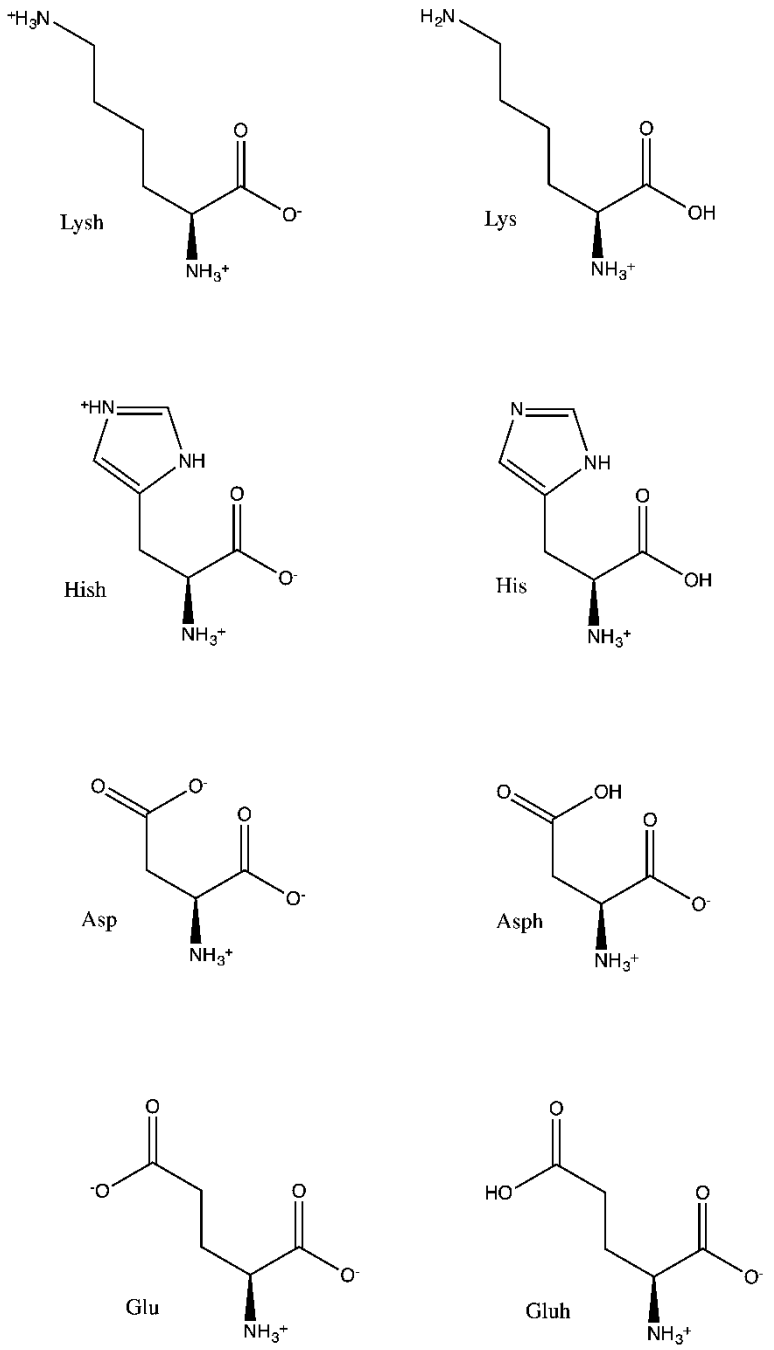


Figure 3.12 Structures of amino acids with charged polar side chains showing their charged state and uncharged state.

The resulting solute-solute and solute-solvent rdfs for both forms of amino acids are presented in Figure 3.13 and all of them converge to unity after 1.5 nm. The PIs for the uncharged and charged forms of the side chain for the respective amino acids are presented in Table 3.4 and the sign of the PI is the same in all four amino acids, indicating that the type of interaction will not change depending on the charge of the side chain. In other words, the charge on the side chain will not be able to change an association to a solvation or vice versa. However, the PI for aspartic acid with an uncharged side chain is relatively large compared to that with a charged side chain. This can be attributed to the interaction between the side chain COOH and the C terminus in the backbone (Figure 3.14) of uncharged aspartic acid.

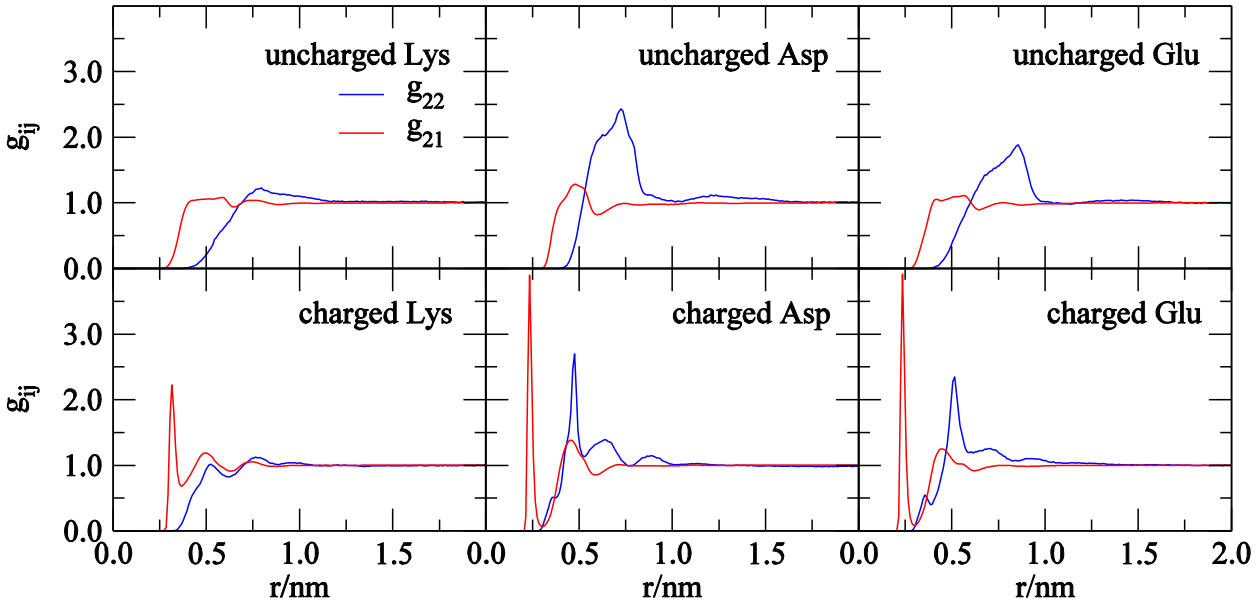


Figure 3.13 The center of mass to center of mass solute-solute (blue) and solute-solvent (red) rdfs of 1.0 m uncharged and charged amino acid side chains.

Table 3.4 PIs (cm^3/mol) of 1.0 m uncharged and charged amino acid side chains at 300 K.

Amino acids	Uncharged PI	Charged PI
Lys	-188 ± 65	-121 ± 36
His	71 ± 96	240 ± 70
Asp	1737 ± 314	359 ± 86
Glu	737 ± 76	241 ± 52

Figure 3.14 A snap shot of unchaged aspartic acid at 300 K.

3.5 Conclusions and Future Directions

In this study we have been able to observe some interesting aspects of amino acid interactions in aqueous solutions and they display a variety of behaviors. Moreover, we have been able to quantify the amino acid interactions in terms of preferential interactions. Hydrophobic amino acids do not associate in the zwitterionic form but they do in the capped form. In addition, the protonation of amino acids with negatively charged polar side chains significantly increases self-association. Furthermore, in future studies we will attempt to study the interactions of mixed amino acids, which will be more representative of the interactions in proteins. In addition, we will try to use a fitting equation to obtain the preferential interactions at infinite dilution, which in turn will resemble the real life situations in biological systems.

3.6 References

1. Truant, R.; Atwal, R. S.; Desmond, C.; Munsie, L.; Tran, T. *Febs Journal* **2008***, 275*, 4252-4262.

2. Batchelor, J. D.; Olteanu, A.; Tripathy, A.; Pielak, G. J. *J. Am. Chem. Soc.* **2004***, 126*, 1958-1961.

3. Stefani, M.; Dobson, C. M. *J. Mo. Med.* **2003***, 81*, 678-699.

4. Cohen, F. E.; Kelly, J. W. *Nature* **2003***, 426*, 905-909.

5. Dobson, C. M. *Nature* **2003***, 426*, 884-890.

6. Kang, M.; Smith, P. E. *Fluid Phase Equilib.* **2007***, 256*, 14-19.

7. Kang, M.; Smith, P. E. *Int. J. Thermophys.* **2010***, 31*, 793-804.

8. Karunaweera, S.; Gee, M. B.; Weerasinghe, S.; Smith, P. E. *J. Chem. Theory Comput.* **2012***, 8*, 3493-3503.

9. Ben-Naim, A. Statistical thermodynamics for chemists and biochemists, Springer Science & Business Media.1992.

10. Matteoli, E.; Mansoori, G. A. Fluctuation theory of mixtures, Taylor & Francis: New York: 1990.

11. Kusalik, P. G.; Patey, G. *J. Chem. Phys.* **1987***, 86*, 5110-5116.

12. Aburi, M.; Smith, P. E. *J. Phys. Chem B* **2004***, 108*, 7382-7388.

13. Kang, M.; Smith, P. E. *J. Comput. Chem.* **2006***, 27*, 1477-1485.

14. Ploetz, E. A.; Bentenitis, N.; Smith, P. E. *Fluid Phase Equilib.* **2010***, 290*, 43-47.

15. Berendsen, H.; Grigera, J.; Straatsma, T. *J. Phys. Chem.* **1987***, 91*, 6269-6271.

16. Hess, B.; Kutzner, C.; Van Der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput*. **2008***, 4*, 435-447.

17. Berendsen, H. J.; Postma, J. v.; van Gunsteren, W. F.; DiNola, A.; Haak, J. *J. Chem. Phys.* **1984***, 81*, 3684-3690.

18. Hess, B.; Bekker, H.; Berendsen, H. J.; Fraaije, J. G. *J. Comput. Chem.* **1997***, 18*, 1463-1472.

19. Miyamoto, S.; Kollman, P. A. *J. Comput. Chem.* **1992***, 13*, 952-962.

20. Kokubo, H.; Pettitt, B. M. *J. Phys. Chem. B* **2007***, 111*, 5233-5242.

21. Fasman, G. D. Handbook of Biochemistry and Molecular Biology, 3rd ed., CRC Press: Ohio 1975.

22. Ellerton, H. D.; Reinfelds, G.; Mulcahy, D. E.; Dunlop, P. J. *J. Phys. Chem.* **1964***, 68*, 398-402.

23. Smith, E. R.; Smith, P. K. *J. Biol. Chem.* **1940***, 135*, 273-279.

24. Venkatesu, P.; Lee, M.; Lin, H. *J. Chem. Thermodyn.* **2007***, 39*, 1206-1216.

Chapter 4 - The Effects of Temperature and Pressure on Amino Acid Interactions in Aqueous Solutions

4.1 Abstract

Studies on protein denaturation play a significant role in understanding the forces that stabilize protein structures and assemblies. Protein denaturation in closed systems occur due to the changes in temperature, pressure and solution composition, while in open systems it is caused by the osmotic pressure or stress. Experimentally, the thermodynamics of protein denaturation is well established and there exist substantial amount of data on protein denaturation. However, it is quite difficult to relate this thermodynamic data to specific interactions with either the native or denatured structures of proteins. Therefore, computer simulations have been extensively used to study protein denaturation and in principle, an atomic level picture of interactions and structural changes can be revealed from them. Here, we have attempted to quantify the effects of temperature and pressure on amino acids interactions and the results are discussed in terms of the preferential (solute over solvent) interactions between the amino acids. If we can classify the variations in these interactions, they might lead to valuable insights toward protein denaturation since amino acids are the building blocks of proteins. It is observed that amino acid association or solvation is residue specific at high temperature whereas amino acid association is always decreased at high pressure.

4.2 Introduction

The mechanisms of protein folding and unfolding have been the subject of intensive experimental and theoretical studies for several decades. However, the details of this process at an atomic level have proved elusive to traditional experimental and theoretical methods. The initiation of molecular dynamics simulations over the last fifteen years has shed some light on this process. In an effort to obtain atomic level information about the protein folding/unfolding process, researchers have been employing computer simulation methods in close collaboration with experiments for couple of decades. The overall results have been in very good agreement with experiment. In order to unfold a protein, some simulations are performed at very high temperatures, typically about 500 K, or 227 ºC. Such drastic measures have been necessary because of the large difference in the experimental timescale for unfolding and that achievable with available computer power. However, as the power of computers increases, the timescale accessible to computer simulations can be extended, allowing the denaturation of proteins to be simulated at much more reasonable temperatures.

In 1895 Royer showed that high pressure kills bacteria and it is one of the initial studies on effects of high pressure on biological systems.[1] Thereafter, in 1914 Bridgman reported the coagulation of egg white at pressures of 10 kbar.[2] There have been rapid development during the last couple of decades in experimental techniques, especially with regards to integration of imaging and spectroscopic methods, such as NMR,[3,4] SANS,[5] and SAXS[6,7] with high pressure. As a consequence, extensive research has been performed on pressure effects on various proteins, protein complexes, and other biomolecules[8,9,10] as well as on viruses,[11,12] bacteria,[13] and cells.[14,15] Experimentally it has been shown that proteins unfold when their aqueous solution is subjected to several kilobars of hydrostatic pressure[5,6,16-24] which leads to water swollen denatured states. This

unfolding behavior of proteins at high pressure, contradicts the fact that the globular folded conformations of proteins are mechanically compressed at high pressures. Moreover, according to the Le Chatelier's principle[25], since unfolding occurs spontaneously at high pressures, the volume change accompanying such a process must be negative. Indeed, experiments show that volume of unfolding of many globular proteins is negative, even though small, usually it is only about 1-3% of the protein volume.[26] Hence, the partial molar volume of unfolded states is lower than that of the folded states, despite the fact that unfolded states are water swollen.[6,27,28] Furthermore, according to experiments, pressure unfolding is a slow process which depends on the protein and the pressure applied and it can take seconds to minutes to hours.[29] It is certain that even with the rapid development of super computers and computing power, direct MD simulations of protein unfolding by pressure in water are computationally impracticable. However, with the efforts to understand how pressure affects fundamental interactions that drive protein folding in water, significant new insights have been obtained.[30-34] For example, Hummer et al.[32] predicted that at room temperature, when the pressure is increased, the hydrophobic interactions between methane like solutes at the pair level are weakened, by using a statistical mechanical theory. Afterwards, this observation was confirmed by extensive simulations of hydrophobic solutes in water.[35-37]

With the above prospects in mind, in this study we try to investigate the effects of temperature and pressure on the interactions of amino acids which are the building blocks of proteins. Moreover, we will try to quantify the interactions of amino acids in terms of preferential interactions.

4.3 Methods

4.3.1 Preferential Interactions

Using KB theory it is quite easy to show that for any thermodynamically stable mixture of a solute (2) and solvent (1) we can write that,[38]

$$-\beta\left(\frac{\partial\mu_2^*}{\partial\rho_2}\right)_{T,P} = -\left(\frac{\partial\ ln\ y_2}{\partial\rho_2}\right)_{T,P} = \frac{G_{22}-G_{21}}{1+\rho_2(G_{22}-G_{21})} \tag{4.1}$$

where y_2 is the molar activity coefficient of the solute. The above expression reduces to the numerator in the limit of infinite dilution of the solute (2). The value of G_{22}-G_{21} at infinite dilution is the central quantity of interest in this work. It is defined as the preferential interaction (PI) between two infinitely dilute solute molecules in a binary system.[39]

$$PI = G_{22} - G_{21} \tag{4.2}$$

A positive value for PI indicates a favorable solute-solute interaction which tends towards solute association or aggregation where as a negative value indicates a favorable solvation which tends towards solute hydration and low solute self-association. A value of zero indicates a balance of the interactions, i.e. an ideal solution.

4.3.2 Molecular Dynamics Simulations

Molecular dynamics simulations were performed according to the same simulation details as described in section 3.3.3. The only differences are that a temperature of 375 K and a pressure of 10000 bar were used to study the effect of temperature and pressure, respectively.

4.4 Results and Discussion

As explained in section 3.4, although the experimentally reported solubility values of most of the amino acids are relatively lower, we used a concentration of 1.0 m in this study as well to improve the statistics of the results.

4.4.1 The Effect of Temperature on Amino Acid Interactions

4.4.1.1 The Quantification of Interactions Among Amino Acids with Nonpolar Side Chains

Amino acids with nonpolar side chains were simulated at a concentration of 1.0 m in their zwitterinic form at 300 K and 375 K to study the effect of temperature on their interactions. The solute-solute rdfs of those amino acids at both temperatures mentioned above are presented in Figure 4.1 and there are slight differences among them, whereas the solute-solvent rdfs are essentially the same and hence they are not shown here. Moreover, all of these rdfs are converged to unity beyond 1.5 nm except for tryptophan and they can be integrated to calculate the KBIs which can then be used to quantify the interactions between amino acids in terms of PIs. Neither the solute-solute rdf of tryptophan nor the solute-solvent rdf are converged and this will be discussed further in this section.

The variation of PIs with temperature for amino acids with nonpolar side chains are summarized in Table 4.1 and they are expressed as individual PIs, the difference of the individual PIs at 300 K and 375 K, $\Delta PI(X)$ and the difference between $\Delta PI(X)$ and $\Delta PI(Gly)$, $\Delta\Delta PI$, where X is the respective amino acid. The individual PIs for all amino acids with nonpolar side chains are increased at 375 K compared to 300 K except for tryptophan. Therefore, it seems like that all amino acids with nonpolar side chains except for tryptophan tend toward more solute association

or aggregation and they become less solvated when the temperature is increased. Moreover, tryptophan is close to being phase separated with PIs of 9282 (±1503) cm^3/mol and 6894 (±1503) cm^3/mol at 300 K and 375 K, respectively. As explained in section 3.4.2, these high PIs are indicative of the non-converged rdfs in Figure 4.1. In addition, the ΔPI(X) for all amino acids with nonpolar side chains except for tryptophan are positive which confirms that they tend toward more solute association or aggregation with increasing temperature.

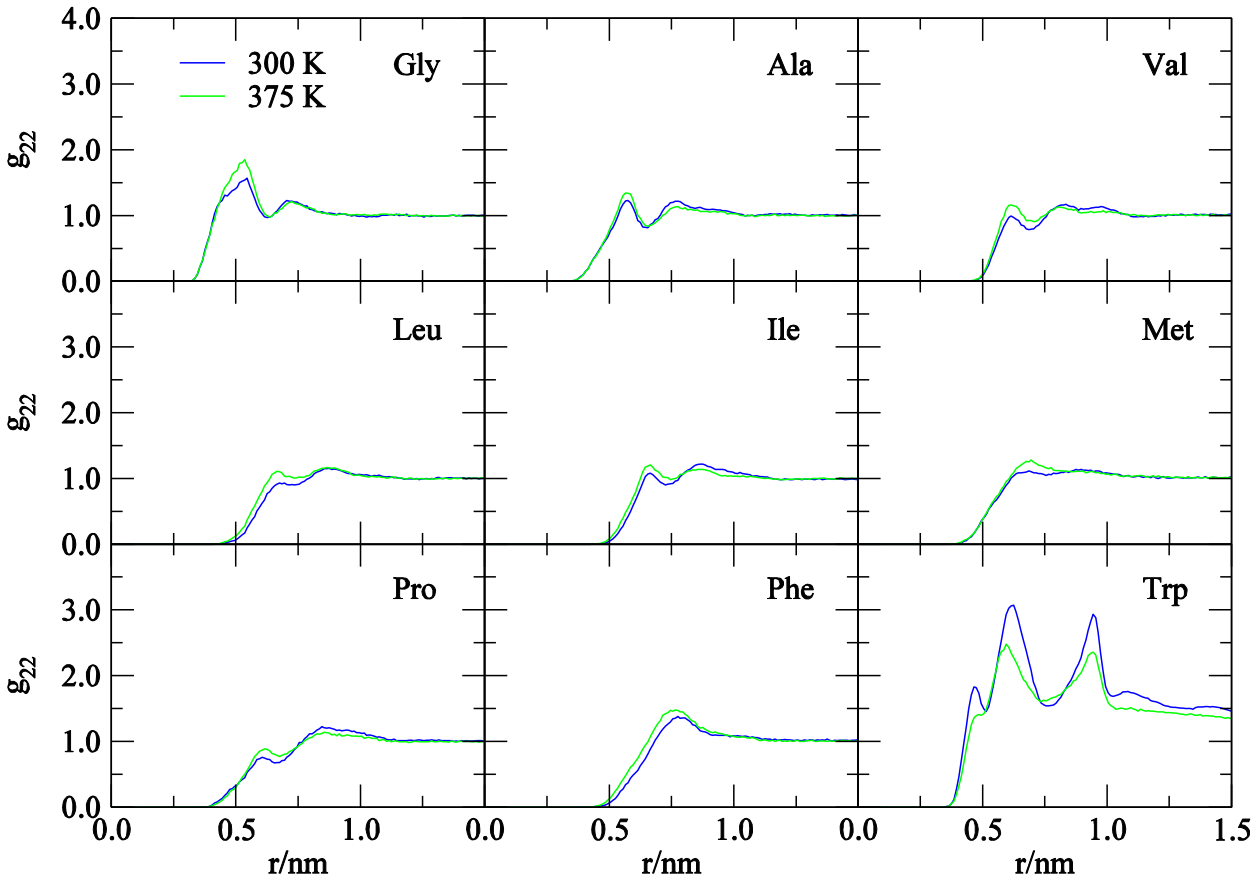


Figure 4.1 Center of mass to center of mass solute-solute rdfs of 1.0 m zwitterionic amino acids with nonpolar side chains at 300 K and 375 K.

Since glycine is the simplest amino acid with no side chain, when the PI of glycine is subtracted from a PI of another amino acid, the resulting quantity can be used to help interpret the

side chain-side chain interactions, although one can argue that still there can be side chain-zwitterion interactions too, as explained in section 3.4.2. The ΔΔPI is positive for most of the amino acids with nonpolar side chains except for alanine, proline and tryptophan. Hence, the side chain-side chain interactions of alanine, proline and tryptophan may be weakened and that of other amino acids with nonpolar side chains may be strengthened with increasing temperature.

Table 4.1 Variation of PIs (cm^3/mol) of 1.0 m zwitterionic amino acids with nonpolar side chains with temperature.

	300 K	375 K		
	PI	PI	$\Delta PI(X)=PI_{375\ K}-PI_{300\ K}$	$\Delta\Delta PI=\Delta PI(X)-\Delta PI(Gly)$
Gly	148 ± 35	177 ± 37	29 ± 51	-
Ala	-147 ± 56	-136 ± 29	11 ± 63	-18 ± 81
Val	-447 ± 53	-375 ± 25	72 ± 59	43 ± 78
Leu	-456 ± 66	-298 ± 29	158 ± 72	129 ± 88
Ile	-293 ± 17	-168 ± 16	125 ± 23	96 ± 56
Met	-123 ± 29	-35 ± 28	88 ± 40	59 ± 65
Pro	-358 ± 45	-330 ± 12	28 ± 47	-1 ± 69
Phe	-87 ± 51	3 ± 25	90 ± 57	61 ± 76
Trp	(9282) ± 1503	(6894) ± 673	-2388 ± 1647	-2417 ± 1648

4.4.1.2 The Quantification of Interactions Among Amino Acids with Uncharged Polar Side Chains

Amino acids with uncharged polar side chains were simulated at a concentration of 1.0 m in their zwitterinic form at 300 K and 375 K to study the effect of temperature on their interactions.

The solute-solute rdfs of those amino acids at both temperatures mentioned above are presented in Figure 4.2 and there are slight differences among them whereas the solute-solvent rdfs are essentially the same and hence they are not shown here. Moreover, all of these rdfs are converged to unity beyond 1.5 nm and they can be integrated to calculate the KBIs which can then be used to quantify the interactions between amino acids in terms of PIs.

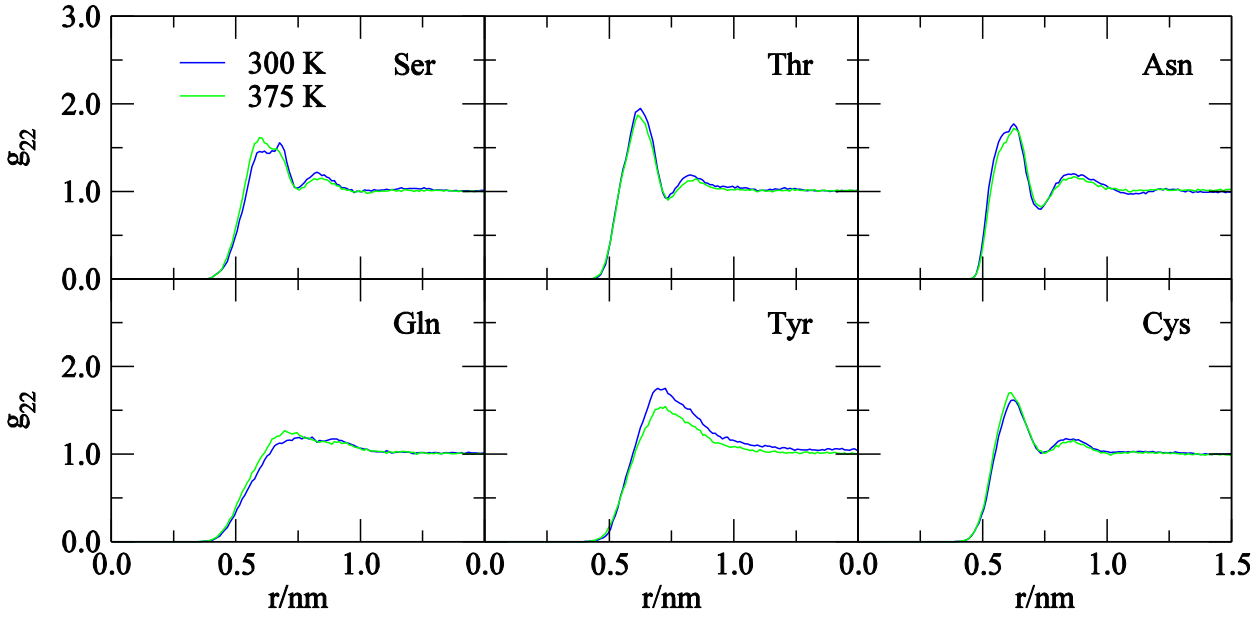


Figure 4.2 Center of mass to center of mass solute-solute rdfs of 1.0 m zwitterionic amino acids with uncharged polar side chains at 300 K and 375 K.

The variation of PIs with temperature for amino acids with uncharged polar side chains are summarized in Table 4.2 and they are expressed as individual PIs, the difference of the individual PIs at 300 K and 375 K, ΔPI(X) and the difference between ΔPI(X) and ΔPI(Gly), ΔΔPI, where X is the respective amino acid. The individual PIs for threonine and asparagine are increased at 375 K compared to 300 K whereas that of serine, glutamine, tyrosine and cysteine are decreased with increasing temperature. Therefore, it seems that threonine and asparagine tend toward more

solute association or aggregation and they become less solvated when the temperature is increased. On the other hand, serine, glutamine, tyrosine and cysteine seems to be more solvated with increasing temperature which tends toward solute hydration and low solute self-association.

Furthermore, the ΔPI(X) for threonine and asparagine are positive while it is negative for serine, glutamine, tyrosine and cysteine. Thus, it is confirmed that threonine and asparagine tend toward more solute association or aggregation whereas serine, glutamine, tyrosine and cysteine tend toward solute hydration and low solute self-association with increasing temperature. Moreover, the ΔΔPI for threonine and asparagine are positive while it is negative for serine, glutamine, tyrosine and cysteine. Hence, the side chain-side chain interactions of threonine and asparagine may be strengthened and that of serine, glutamine, tyrosine and cysteine may be weakened with increasing temperature.

Table 4.2 Variation of PIs (cm^3/mol) of 1.0 m zwitterionic amino acids with uncharged polar side chains with temperature.

	300 K	375 K		
	PI	PI	$\Delta PI(X)=PI_{375\ K}-PI_{300\ K}$	$\Delta\Delta PI=\Delta PI(X)-\Delta PI(Gly)$
Ser	132 ± 119	30 ± 37	-102 ± 125	-131 ± 135
Thr	-23 ± 73	87 ± 31	110 ± 79	81 ± 94
Asn	-44 ± 70	50 ± 37	94 ± 79	65 ± 94
Gln	-14 ± 86	-128 ± 30	-114 ± 91	-143 ± 104
Tyr	851 ± 153	413 ± 64	-438 ± 166	-467 ± 173
Cys	159 ± 48	34 ± 35	-125 ± 59	-154 ± 78

4.4.1.3 The Quantification of Interactions Among Amino Acids with Charged Polar Side Chains

Amino acids with charged polar side chains were simulated at a concentration of 1.0 m in their zwitterinic form at 300 K and 375 K to study the effect of temperature on their interactions. The solute-solute rdfs of those amino acids at both temperatures mentioned above are presented in Figure 4.3 and there are slight differences among them whereas the solute-solvent rdfs are essentially the same and hence they are not shown here. Moreover, all of these rdfs are converged to unity beyond 1.5 nm and they can be integrated to calculate the KBIs which can then be used to quantify the interactions between amino acids in terms of PIs.

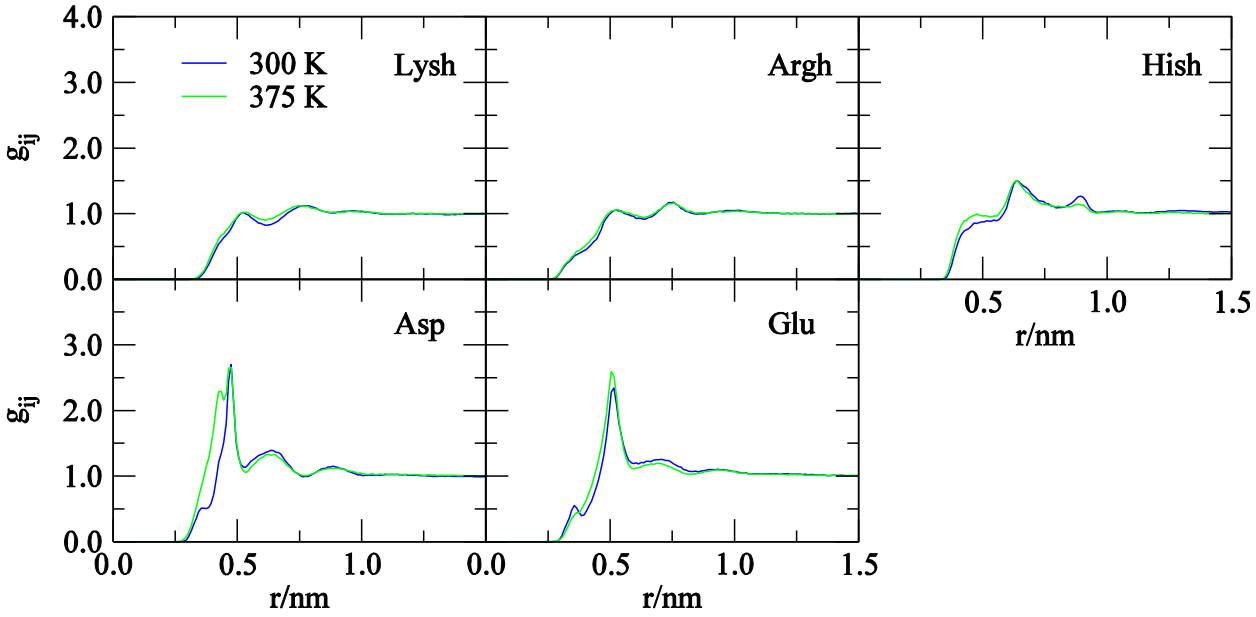


Figure 4.3 Center of mass to center of mass solute-solute rdfs of 1.0 m zwitterionic amino acids with charged polar side chains at 300 K and 375 K.

The variation of PIs with temperature for amino acids with charged polar side chains are summarized in Table 4.3 and they are expressed as individual PIs, the difference of the individual

PIs at 300 K and 375 K, ΔPI(X) and the difference between ΔPI(X) and ΔPI(Gly), ΔΔPI, where X is the respective amino acid. The individual PIs for all amino acids with charged polar side chains are increased at 375 K compared to 300 K. Hence, it seems that they tend toward more solute association or aggregation and become less solvated when the temperature is increased. In addition, the ΔPI(X) for all amino acids with charged polar side chains are positive which confirms that they tend toward more solute association or aggregation with increasing temperature. Moreover, the ΔΔPI for arginine, aspartic acid and glutamic acid are positive while it is negative for lysine and histidine. Hence, the side chain-side chain interactions of arginine, aspartic acid and glutamic acid may be strengthened and that of lysine and histidine may be weakened with increasing temperature.

Table 4.3 Variation of PIs (cm^3/mol) of 1.0 m zwitterionic amino acids with charged polar side chains with temperature.

	300 K	375 K		
	PI	PI	$\Delta PI(X)=PI_{375\ K}-PI_{300\ K}$	$\Delta\Delta PI=\Delta PI(X)-\Delta PI(Gly)$
Lysh	-121 ± 36	-95 ± 16	26 ± 39	-3 ± 64
Argh	-95 ± 38	-1 ± 17	94 ± 42	65 ± 66
Hish	240 ± 70	254 ± 43	14 ± 82	-15 ± 97
Asp	359 ± 86	444 ± 13	85 ± 87	56 ± 101
Glu	241 ± 52	405 ± 47	164 ± 70	135 ± 87

4.4.1.4 The Quantification of Amino Acids Interactions in Terms of Zwitterionic and Capped Forms

In order to investigate the effect of temperature on the type of termini, glycine and valine were simulated in the zwitterionic form as well as in the capped form at a concentration of 1.0 m at 300 K and 375 K. The solute-solute rdfs of the two types of termini at both temperatures mentioned above are presented in Figure 4.4 and there are slight differences among them whereas the solute-solvent rdfs are essentially the same and hence they are not shown here. Moreover, all of these rdfs are converged to unity beyond 1.5 nm and they can be integrated to calculate the KBIs which can then be used to quantify the amino acid interactions in terms of PIs.

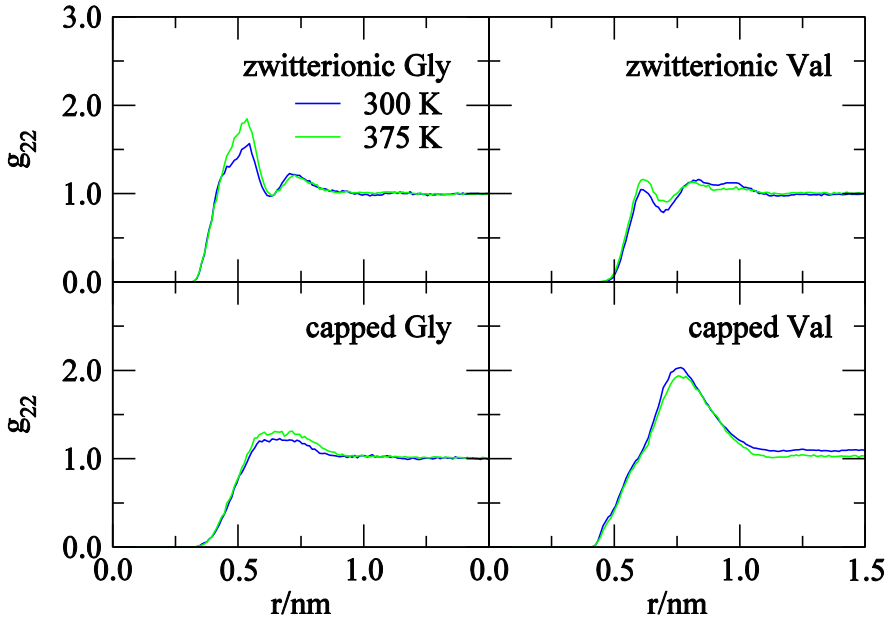


Figure 4.4 Center of mass to center of mass solute-solute rdfs of 1.0 m zwitterionic and capped glycine and valine at 370 K and 375 K.

The variation of PIs with temperature for both types of termini are presented in Table 4.4 and the PIs of glycine and valine for both types of termini are increased with increasing

temperature. Hence, it seems that both types of termini of glycine and valine tend toward more solute association or aggregation and become less solvated when the temperature is increased. Moreover, the changes in valine are significantly larger than the changes in glycine. All PIs of glycine are positive which indicates solute association or aggregation. On the other hand, the zwitterionic form of valine has negative PIs whereas the capped form has positive PIs which indicates solute hydration and solute association, respectively.

Table 4.4 Variation of PIs (cm^3/mol) of 1.0 m zwitterionic and capped glycine and valine with temperature.

	300 K		375 K	
	Zwitterion	Capped	Zwitterion	Capped
	PI	PI	PI	PI
Gly	148 ± 35	20 ± 81	177 ± 37	156 ± 27
Val	-447 ± 53	848 ± 274	-375 ± 25	1653 ± 138

4.4.1.5 The Contribution from Charged and Uncharged Polar Side Chains Toward Amino Acid Interactions

In order to investigate the effect of temperature on the charged and uncharged side chains, the amino acids with charged polar side chains were simulated at a concentration of 1.0 m at 300 K and 375 K. The solute-solute rdfs of the amino acids with two types of side chains at both temperatures mentioned above are presented in Figure 4.5 and there are slight differences among them where as the solute-solvent rdfs are essentially the same and hence they are not shown here. Furthermore, all of these rdfs are converged to unity beyond 1.5 nm and they can be integrated to calculate the KBIs which can then be used to quantify the amino acid interactions in terms of PIs.

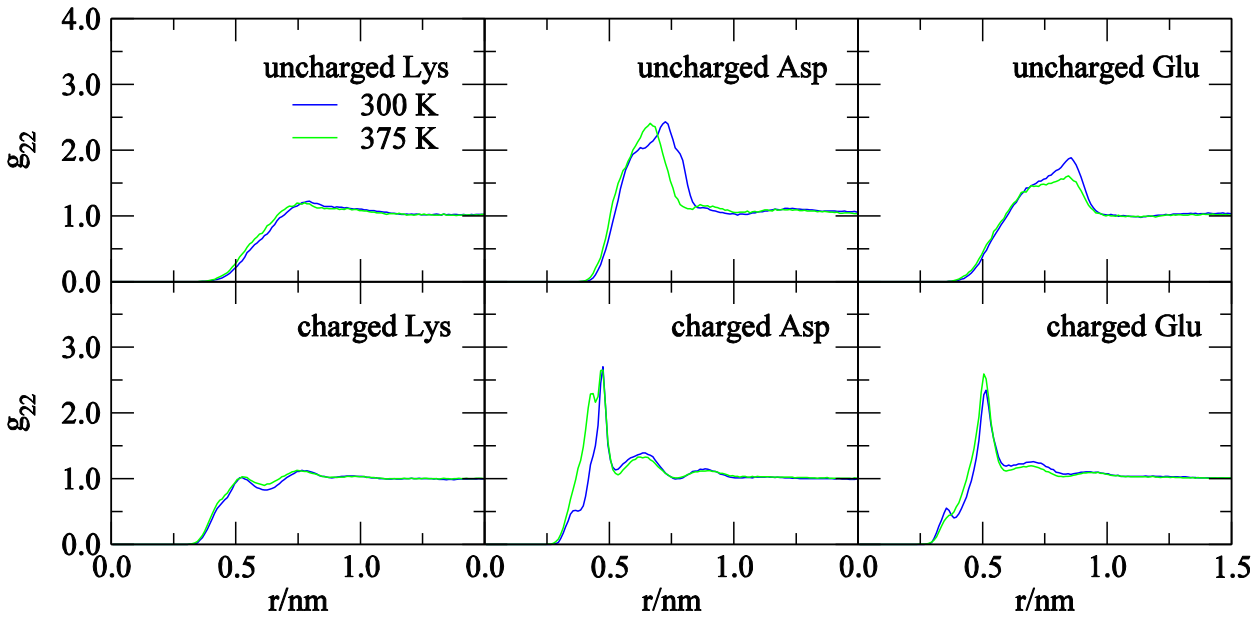


Figure 4.5 Center of mass to center of mass solute-solute rdfs of 1.0 m uncharged and charged amino acid side chains at 300 K and 375 K.

The variation of PIs with temperature for both types of side chains are presented in Table 4.5. The PIs of lysine, histidine, aspartic acid and glutamic acid with a charged side chain are increased with increasing temperature. Therefore, it seems that charged polar amino acids with charged side chains tend toward more solute association or aggregation and become less solvated when the temperature is increased. Furthermore, the PI of lysine with an uncharged side chain is increased whereas that of histidine, aspartic acid and glutamic acid are decreased with increasing temperature. Hence, it seems that lysine with an uncharged side chain tends toward more solute association or aggregation and becomes less solvated when the temperature is increased whereas histidine, aspartic acid and glutamic acid with an uncharged side chain seems to be more solvated with increasing temperature which tends toward solute hydration and low solute self-association.

Table 4.5 Variation of PIs (cm^3/mol) of 1.0 m uncharged and charged amino acid side chains with temperature.

	300 K		375 K	
	Charged	Uncharged	Charged	Uncharged
	PI	PI	PI	PI
Lys	-121 ± 36	-188 ± 65	-95 ± 16	-158 ± 27
His	240 ± 70	71 ± 96	254 ± 43	10 ± 59
Asp	359 ± 86	1737 ± 314	444 ± 13	1031 ± 100
Glu	241 ± 52	737 ± 76	405 ± 47	630± 88

4.4.2 The Effect of Pressure on Amino Acid Interactions

4.4.2.1 The Quantification of Interactions Among Amino Acids with Nonpolar Side Chains

Amino acids with nonpolar side chains were simulated at a concentration of 1.0 m in their zwitterinic form at 1 bar and 10000 bar to study the effect of pressure on their interactions. The solute-solute rdfs of those amino acids at both pressures mentioned above are presented in Figure 4.6 and they indicate general loss of structure when the pressure is increased whereas the solute-solvent rdfs are essentially the same and hence they are not shown here. Moreover, all of these rdfs are converged to unity beyond 1.5 nm except for tryptophan and they can be integrated to calculate the KBIs which can then be used to quantify the interactions between amino acids in terms of PIs. Neither the solute-solute rdf of tryptophan nor the solute-solvent rdf are converged and this will be discussed further in this section.

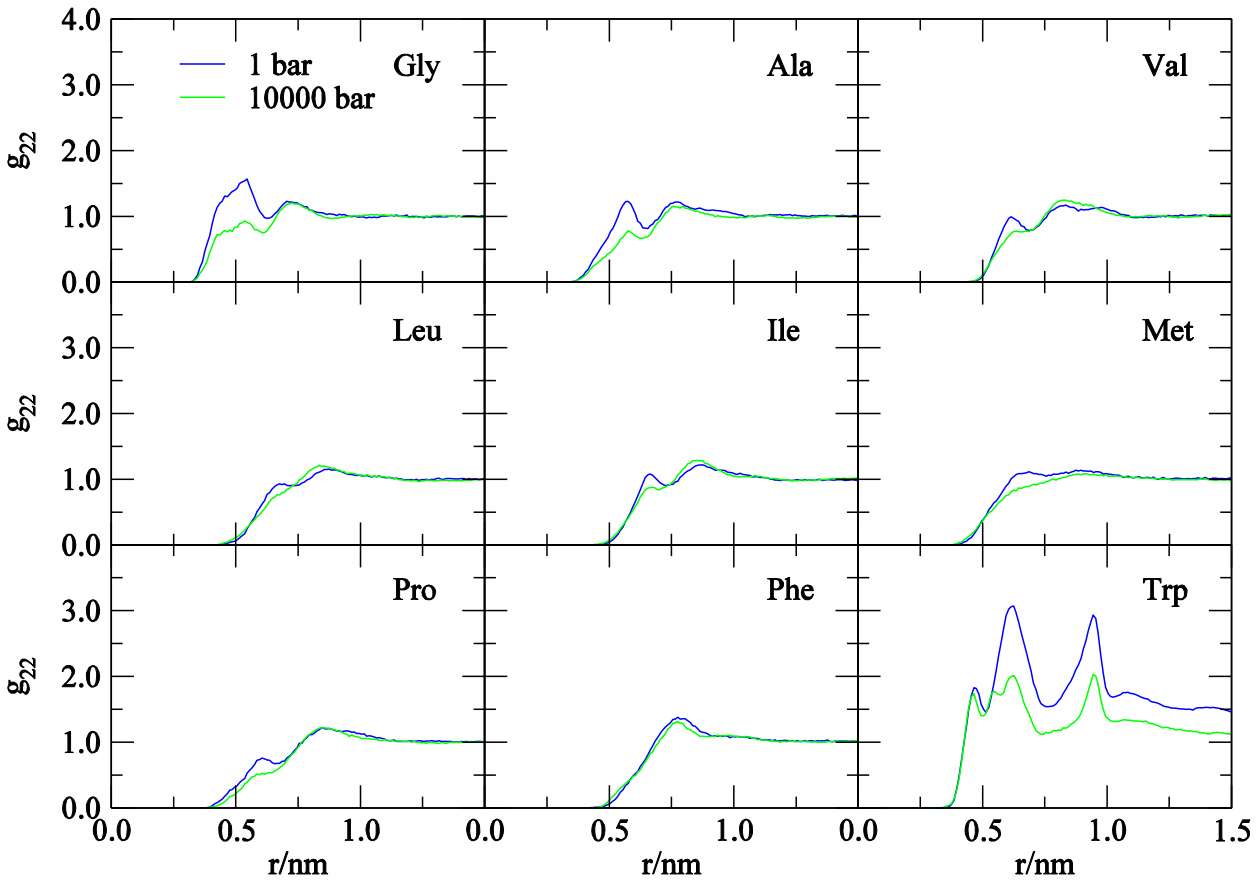


Figure 4.6 Center of mass to center of mass solute-solute rdfs of 1.0 m zwitterionic amino acids with nonpolar side chains at 1 bar and 10000 bar.

The variation of PIs with pressure for amino acids with nonpolar side chains are summarized in Table 4.6 and they are expressed as individual PIs, the difference of the individual PIs at 1 bar and 10000 bar, ΔPI(X) and the difference between ΔPI(X) and ΔPI(Gly), ΔΔPI, where X is the respective amino acid. The individual PIs for all amino acids with nonpolar side chains are decreased at 10000 bar compared to 1 bar except for proline. Therefore, all amino acids with nonpolar side chains except for proline seem to be more solvated with increasing pressure which tends toward solute hydration and low solute self-association. However, proline seems to tend toward more solute association or aggregation and it becomes less solvated when the pressure is increased. Moreover, tryptophan is close to being phase separated with PIs of

9282 (±1503) cm^3/mol and 5174 (±598) cm^3/mol at 1 bar and 10000 bar, respectively. As explained earlier these high PIs are indicative of the non-converged rdfs in Figure 4.6.

Furthermore, the ΔPI(X) for all amino acids with nonpolar side chains are negative except for proline, which confirms that they tend toward solute hydration and low solute self-association with increasing pressure. On the other hand, it is confirmed that proline tend toward more solute association or aggregation and it becomes less solvated when the pressure is increased. Moreover, the ΔΔPI for all amino acids with nonpolar side chains are positive except for tryptophan. Hence, the side chain-side chain interactions of them may be strengthened and that of tryptophan may be weakened with increasing pressure.

Table 4.6 Variation of PIs (cm^3/mol) of 1.0 m zwitterionic amino acids with nonpolar side chains with pressure.

	1 bar	10000 bar		
	PI	PI	$\Delta PI(X)=PI_{10000\ bar}-PI_{1\ bar}$	$\Delta\Delta PI=\Delta PI(X)-\Delta PI(Gly)$
Gly	148 ± 35	-193 ± 42	-341 ± 55	-
Ala	-147 ± 56	-392 ± 25	-245 ± 61	96 ± 82
Val	-447 ± 53	-481 ± 42	-34 ± 68	307 ± 87
Leu	-456 ± 66	-482 ± 28	-26 ± 72	315 ± 90
Ile	-293 ± 17	-306 ± 23	-13 ± 29	328 ± 62
Met	-123 ± 29	-333 ± 30	-210 ± 42	131 ± 69
Pro	-358 ± 45	-334 ± 48	24 ± 66	365 ± 86
Phe	-87 ± 51	-325 ± 29	-238 ± 59	103 ± 80
Trp	(9282) ± 1503	(5174) ± 598	-4108 ± 1618	-3767 ± 1619

4.4.2.2 The Auantification of Interactions Among Amino Acids with Uncharged Polar Side Chains

Amino acids with uncharged polar side chains were simulated at a concentration of 1.0 m in their zwitterinic form at 1 bar and 10000 bar to study the effect of pressure on their interactions. The solute-solute rdfs of those amino acids at both pressures mentioned above are presented in Figure 4.7 and they indicate general loss of structure when the pressure is increased whereas the solute-solvent rdfs are essentially the same and hence they are not shown here. Moreover, all of these rdfs are converged to unity beyond 1.5 nm and they can be integrated to calculate the KBIs which can then be used to quantify the interactions between amino acids in terms of PIs.

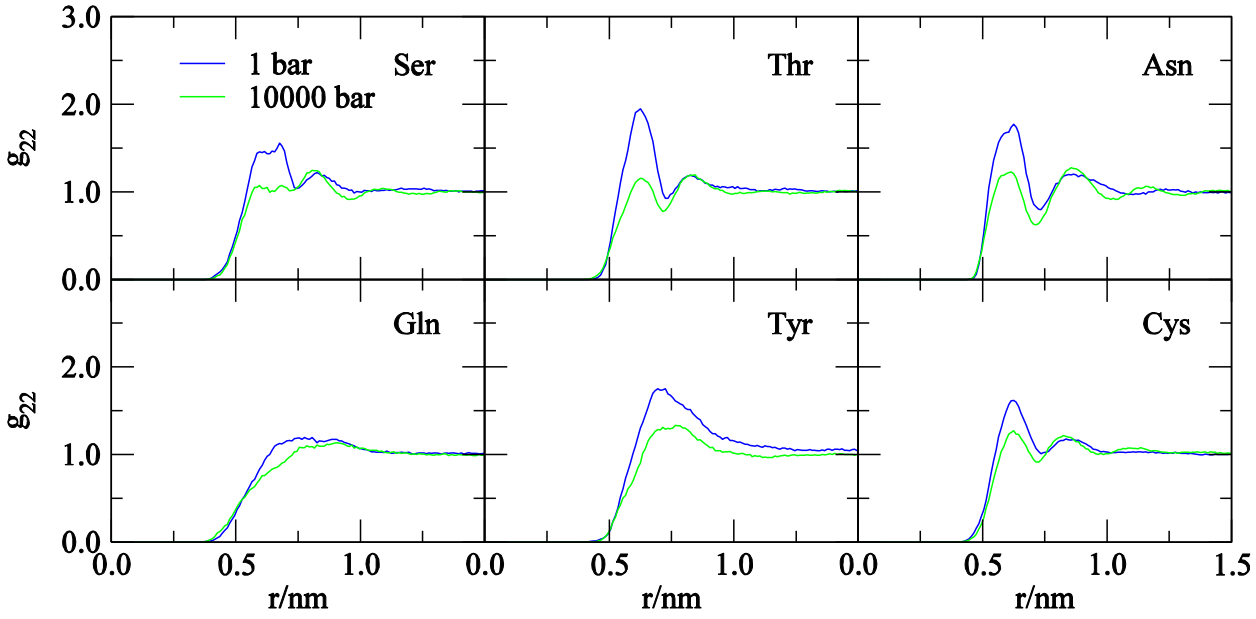


Figure 4.7 Center of mass to center of mass solute-solute rdfs of 1.0 m zwitterionic amino acids with uncharged polar side chains at 1 bar and 10000 bar.

The variation of PIs with pressure for amino acids with uncharged polar side chains are summarized in Table 4.7 and they are expressed as individual PIs, the difference of the individual

PIs at 1 bar and 10000 bar, ΔPI(X) and the difference between ΔPI(X) and ΔPI(Gly), ΔΔPI, where X is the respective amino acid. The individual PIs for all amino acids with uncharged polar side chains are decreased at 10000 bar compared to 1 bar. As a result, they seem to be more solvated with increasing pressure which tends toward solute hydration and low solute self-association. In addition, the ΔPI(X) for all amino acids with uncharged polar side chains are negative which confirms that they tend toward solute hydration and low solute self-association with increasing pressure. Moreover, the ΔΔPI for threonine, asparagine and glutamine are positive while it is negative for serine, tyrosine and cysteine. Hence, the side chain-side chain interactions of threonine, asparagine and glutamine may be strengthened whereas that of serine, tyrosine and cysteine may be weakened with increasing pressure.

Table 4.7 Variation of PIs (cm^3/mol) of 1.0 m zwitterionic amino acids with uncharged polar side chains with pressure.

	1 bar	10000 bar		
	PI	PI	$\Delta PI(X)=PI_{10000\ bar}-PI_{1\ bar}$	$\Delta\Delta PI=\Delta PI(X)-\Delta PI(Gly)$
Ser	132 ± 119	-281 ± 69	-413 ± 138	-72 ± 148
Thr	-23 ± 73	-323 ± 21	-300 ± 76	41 ± 94
Asn	-44 ± 70	-267 ± 37	-223 ± 79	118 ± 96
Gln	-14 ± 86	-312 ± 22	-298 ± 89	43 ± 104
Tyr	851 ± 153	13 ± 82	-838 ± 174	-497 ± 182
Cys	159 ± 48	-184 ± 46	-343 ± 66	-2 ± 86

4.4.2.3 The Quantification of Interactions Among Amino Acids with Charged Polar Side Chains

Amino acids with charged polar side chains were simulated at a concentration of 1.0 m in their zwitterinic form at 1 bar and 10000 bar to study the effect of pressure on their interactions. The solute-solute rdfs of those amino acids at both pressures mentioned above are presented in Figure 4.8 and they indicate general loss of structure when the pressure is increased whereas the solute-solvent rdfs are essentially the same and hence they are not shown here. Moreover, all of these rdfs are converged to unity beyond 1.5 nm and they can be integrated to calculate the KBIs which can then be used to quantify the interactions between amino acids in terms of PIs.

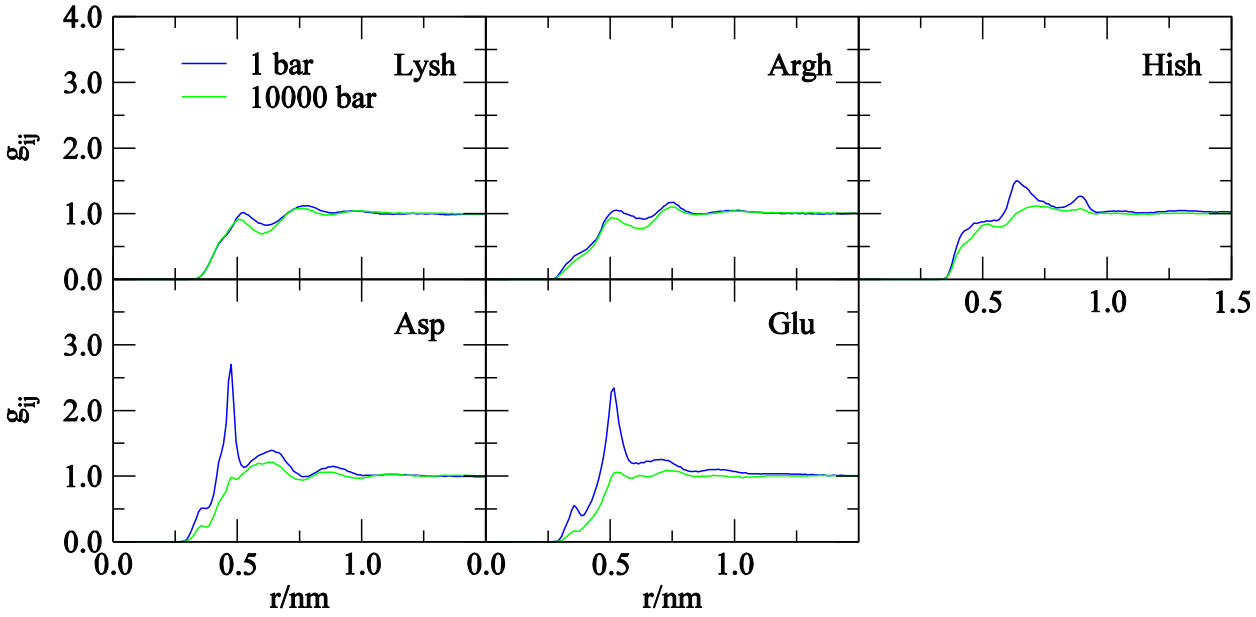


Figure 4.8 Center of mass to center of mass solute-solute rdfs of 1.0 m zwitterionic amino acids with charged polar side chains at 1 bar and 10000 bar.

The variation of PIs with pressure for amino acids with charged polar side chains are summarized in Table 4.8 and they are expressed as individual PIs, the difference of the individual

PIs at 1 bar and 10000 bar, ΔPI(X) and the difference between ΔPI(X) and ΔPI(Gly), ΔΔPI, where X is the respective amino acid. The individual PIs for all amino acids with charged polar side chains are decreased at 10000 bar compared to 1 bar. Hence, they seem to be more solvated with increasing pressure which tends toward solute hydration and low solute self-association. Moreover, the ΔPI(X) for all amino acids with charged polar side chains are negative which confirms that they tend toward solute hydration and low solute self-association with increasing pressure. Furthermore, the ΔΔPI for lysine and arginine are positive while it is negative for histidine, aspartic acid and glutamic acid. Hence, the side chain-side chain interactions of lysine and arginine may be strengthened whereas that of histidine, aspartic acid and glutamic acid may be weakened with increasing pressure.

Table 4.8 Variation of PIs (cm^3/mol) of 1.0 m zwitterionic amino acids with charged polar side chains with pressure.

	1 bar	10000 bar		
	PI	PI	$\Delta PI(X)=PI_{10000\ bar}-PI_{1\ bar}$	$\Delta\Delta PI=\Delta PI(X)-\Delta PI(Gly)$
Lysh	-121 ± 36	-252 ± 10	-131 ± 37	210 ± 66
Argh	-95 ± 38	-233 ± 12	-138 ± 40	203 ± 68
Hish	240 ± 70	-162 ± 29	-402 ± 76	-61 ± 93
Asp	359 ± 86	-96 ± 19	-455 ± 88	-114 ± 104
Glu	241 ± 52	-117 ± 13	-358 ± 54	-17 ± 77

4.4.2.4 The Quantification of Amino Acids Interactions in Terms of Zwitterionic and Capped Forms

In order to investigate the effect of pressure on the type of termini, glycine and valine were simulated in the zwitterionic form as well as in the capped form at a concentration of 1.0 m at 1 bar and 10000 bar. The solute-solute rdfs of the two types of termini at both pressures mentioned above are presented in Figure 4.9 and there are slight differences among them where as the solute-solvent rdfs are essentially the same and hence they are not shown here. Moreover, all of these rdfs are converged to unity beyond 1.5 nm and they can be integrated to calculate the KBIs which can then be used to quantify the amino acid interactions in terms of PIs.

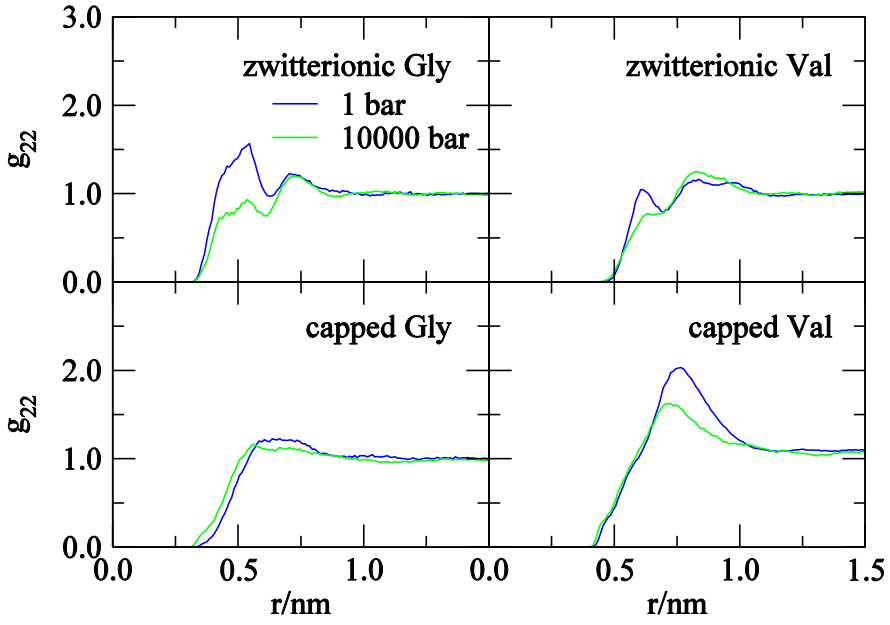


Figure 4.9 Center of mass to center of mass solute-solute rdfs of 1.0 m zwitterionic and capped glycine and valine at 1 bar and 10000 bar.

The variation of PIs with pressure for both types of termini are presented in Table 4.9 and the PIs of glycine and valine for the zwiterionic form are decreased whereas that for the capped form are increased with increasing pressure. Therefore, it indicates that the zwitterionic form seem

to be more solvated with increasing pressure which tends toward solute hydration and low solute self-association. On the other hand, the capped form tends toward more solute association or aggregation when the pressure is increased. Moreover, the most significant change is observed in the zwitterionic form of glycine where it changes from a moderate positive value at 1 bar to a moderate negative value at 10000 bar indicating a transformation from a solute association or aggregation to a solute hydration and low solute self- association.

Table 4.9 Variation of PIs (cm^3/mol) of 1.0 m zwitterionic and capped glycine and valine with pressure.

	1 bar		10000 bar	
	Zwitterion	Capped	Zwitterion	Capped
	PI	PI	PI	PI
Gly	148 ± 35	20 ± 81	-193 ± 42	62 ± 74
Val	-447 ± 53	848 ± 274	-481 ± 42	1080 ± 177

4.4.2.5 The Contribution from Uncharged and Charged Polar Side Chains Toward Amino Acid Interactions

In order to investigate the effect of pressure on the charged and uncharged side chains, the amino acids with charged polar side chains were simulated at a concentration of 1.0 m at 1 bar and 10000 bar. The solute-solute rdfs of the amino acids with two types of side chains at both pressures mentioned above are presented in Figure 4.10 and there are slight differences among them where as the solute-solvent rdfs are essentially the same and hence they are not shown here. Furthermore, all of these rdfs are converged to unity beyond 1.5 nm and they can be integrated to calculate the KBIs which can then be used to quantify the amino acid interactions in terms of PIs.

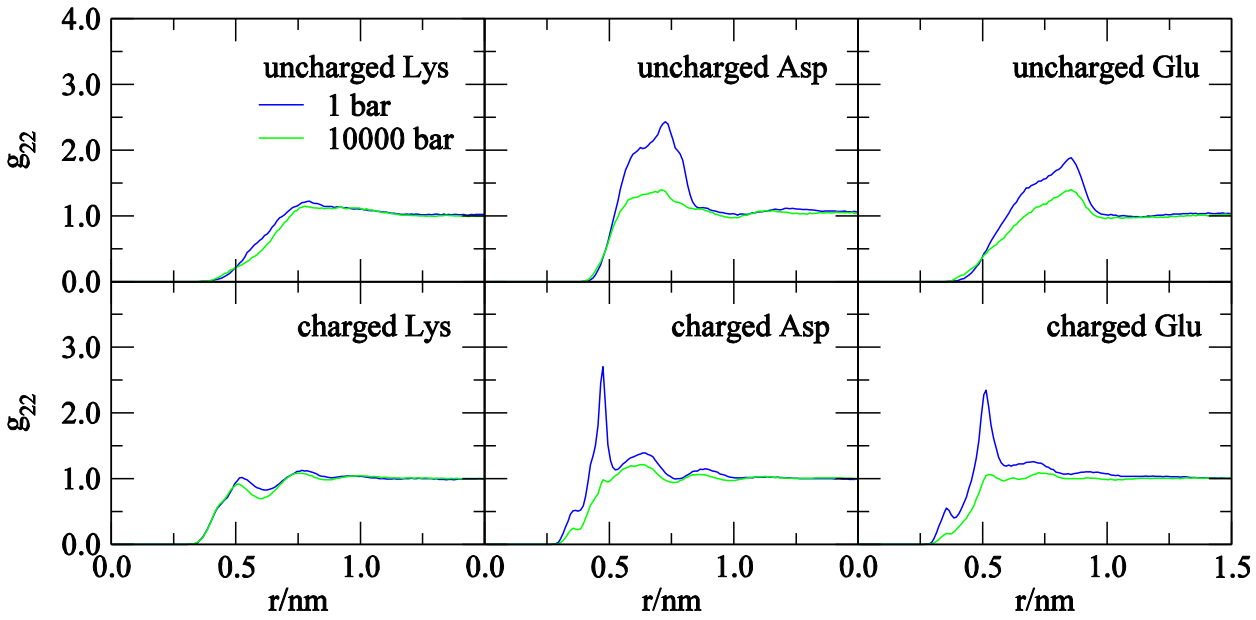


Figure 4.10 Center of mass to center of mass solute-solute rdfs of 1.0 m uncharged and charged amino acid side chains at 10000 bar.

The variation of PIs with pressure for both types of side chains are presented in Table 4.10. Not only the PIs of lysine, histidine, aspartic acid and glutamic acid with charged side chains but also the PIs of them with uncharged side chains, are decreased with increasing pressure. Therefore, it seems that charged polar amino acids with charged side chains as well as uncharged side chains seems to be more solvated with increasing pressure which tends toward solute hydration and low solute self-association.

Table 4.10 Variation of PIs (cm^3/mol) of 1.0 m uncharged and charged amino acid side chains with pressure.

	1 bar		10000 bar	
	Charged	Uncharged	Charged	Uncharged
	PI/(cm^3/mol)	PI/(cm^3/mol)	PI/(cm^3/mol)	PI/(cm^3/mol)
Lys	-121 ± 36	-188 ± 65	-252 ± 10	-418 ± 56
His	240 ± 70	71 ± 96	-162 ± 29	-224 ± 31
Asp	359 ± 86	1737 ± 314	-96 ± 19	-146 ± 69
Glu	241 ± 52	737 ± 76	-117 ± 13	-80 ± 65

4.5 Conclusions

In this study we have been able to quantify the amino acid interactions in terms of PIs at relatively higher temperatures and pressures. The amino acids with nonpolar side chains except for tryptophan and all the amino acids with charged polar side chains tend toward more solute association or aggregation whereas amino acids with uncharged polar side chains except threonine and asparagine tends toward solute hydration and low solute self-association when the temperature is increased. On the other hand, all the amino acids except for proline tends toward solute hydration and low solute self-association when the pressure is increased. Thus, at a higher pressure a general increase in solvation was observed which agrees with the swollen structures of experiments. The amino acids used in this study were in the zwitterionic form in most of the instances. However, in proteins they are connected to each other via peptide bonds and do not exist as zwitterions. Therefore, at this point we will not be able to provide any insights into temperature and pressure denaturation of proteins. Nevertheless, as pointed out in chapter 3 when we study the mixed amino

acids in the future we will be able to provide some insights toward temperature and pressure denaturation of proteins.

4.6 References

1. Royer, H. Arch Physiol Normale Pathol. **1895,** 7-12.
2. Bridgman, P.W. J Biol Chem. **1914**, 19, 511–512.
3. Jonas, J.; Jonas, A. Annu Rev Biophys Biomol Struct. **1994**, 23, 287–318.
4. Akasaka, K. Chem. Rev. **2006**, 106, 1814–1835.
5. Paliwal, A.; Asthagiri. D; Bossev, D. P.; Paulaitis, M. E. Biophys. J. **2004**, 87, 3479–3492.
6. Panick, G.; Malessa, R.; Winter, R.; Rapp. G.; Frye, K. J.; Royer, C. A. *J. Mol. Biol.* **1998**, 275, 389–402.
7. Woenckhaus, J.; Kohling, R.; Winter, R.; Thiyagarajan, P.; Finet, S. *Rev. Sci. Instrum.* **2000**, 71, 3895–3899.
8. Kim, Y. S.; Randolph, T. W.; Seefeldt, M. B.; Carpenter, J. F. *Methods Enzymol.* **2006**, 413, 237–253.
9. Macgregor, R. B. *Biopolymers* **1998**, 48, 253–263.
10. Chalikian, T. V.; Macgregor, R. B. *Phys. Life. Rev*. **2007**, 4, 91–115.
11. Silva, J. L.; Luan, P.; Glaser, M.; Voss, E. W.; Weber, G. *J. Virol.* **1992**, 66, 2111–2117.
12. Kingsley, D. H.; Hoover, D. G.; Papafragkou, E.; Richards, G. P. Inactivation of hepatitis A virus and calicivirus by high hydrostatic pressure. *J. Food. Prot.* **2002**, 65, 1605–1609.
13. Bartlett, D. H. *Biochim. Biophys. Acta.* **2002**, 1595, 367–381.
14. Frey, B.; Franz, S.; Sheriff A.; Korn. A.; Bluemelhuber, G.; Gaipl, U. S.; Voll, R. E.; Meyer-Pittroff, R.; Herrmann, M. Cell. Mol. Biol. **2004**, 50, 459–467.
15. Frey, B.; Hartmann, M.; Herrmann, M.; Meyer-Pittroff, R.; Sommer, K,; Bluemelhuber, G. *Microsc. Res. Tech*. **2006**, 69, 65–72.
16. Brandts, J. F.; Oliveira, R. J.; Westort, C. *Biochemistry* **1970**, 9, 1038–1047.
17. Samarasinghe, S. D.; Campbell, D. M.; Jonas, A.; Jonas, J. *Biochemistry* **1992**, 31, 7773–7778.
18. Silva, J. L.; Weber, G. *Annu. Rev. Phys. Chem.* **1993**, 44, 89–113.
19. Gross, M.; Jaenicke, R. *Eur. J. Biochem.* **1994**, 221, 617–630.
20. Ruan, K. C.; Lange, R.; Bec, N.; Balny, C. *Biochem. Biophys. Res. Comm.* **1997**, 239, 150–154.

21. Ruan, K. C.; Xu, C. H.; Yu, Y.; Li, J.; Lange, R.; Bec, N.; Balny, C. *Eur. J. Biochem*. **2001**, 268, 2742–2750.

22. Herberhold, H.; Marchal, S.; Lange, R.; Scheyhing, C. H.; Vogel, R. F.; Winter, R. *J. Mol. Biol.* **2003**, 330, 1153–1164.

23. Kitahara, R.; Yokoyama, S.; Akasaka, K. *J. Mol. Biol.* **2005**, 347, 277–285.

24. Wiedersich, J.; Kohler, S.; Skerra, A.; Friedrich, *J. Proc. Natl. Acad. Sci. USA* **2008**, 105, 5756–5761.

25. Atkins, P.; Paula, J. D. Physical chemistry. Oxford: Oxford University Press; 2006.

26. Royer, C. A. *Biochim. Biophys. Acta.* **2002**, 1595, 201–209.

27. Peng, X. D.; Jonas, J.; Silva, J. L. *Proc. Natl. Acad. Sci. USA* **1993**, 90, 1776–1780.

28. Lassaslle, M. L.; Yamada, H.; Akasaka, K. *J. Mol. Biol.* **2000,** 298, 93–302.

29. Woenckhaus, J.; Kohling, R.; Thiyagarajan, P.; Littrell, K. C.; Seifert, S.; Royer, C. A.; Winter, R. *Biophys. J.* **2001**, 80, 1518–1523.

30. Wallqvist, A. *J. Phys. Chem.* **1991,** 95, 8921–8927.

31. Payne, V. A.; Matubayasi, N.; Murphy, L. R.; Levy, R. M. *J. Phys. Chem. B* **1997**, 101, 2054–2060.

32. Hummer, G.; Garde, S.; Garcia, A. E.; Paulaitis, M. E.; Pratt, L. R. *Proc. Natl. Acad. Sci. USA* **1998**, 95, 1552–1555.

33. Cheung, J. K.; Shah, P.; Truskett, T. M. *Biophys. J.* **2006**, 91, 2427–2435.

34. Patel, B. A.; Debenedetti, P. G.; Stillinger, F. H.; Rossky, P. J. *Biophys. J.* **2007**, 93, 4116–4127.

35. Ghosh, T.; Garcia, A. E.; Garde, S. *J. Am. Chem. Soc.* **2001**, 123, 10997–11003.

36. Ghosh, T.; Garcia, A. E.; Garde, S. *J. Chem. Phys.* **2002**, 116, 2480–2486.

37. Rick, S. W. *J. Phys. Chem. B* **2000**, 104, 6884–6888.

38. Ben-Naim, A. Statistical Thermodynamics for Chemists and Biochemists, Plenum Press: New York, 1992.

39. Kang, M.; Smith, P. E. *Int J Thermophys* **2010**, 31, 793-804.

Chapter 5 - Development of Torsional Potentials for the KBFF Model of Peptides and Proteins

5.1 Abstract

Computer simulations have become a significant tool for studying the structure and dynamics of biological macromolecules since experimental methods cannot reveal those properties under most circumstances. However, the accuracy of simulation data is determined by the quality of the force field which is being used. Although there are many state of the art force fields which are widely used, it is common to find discrepancies in the conformational preferences of different amino acid residues. Therefore, not only there is room for improvement of those established force fields, but also there are opportunities to develop new force fields as well. Recently, we have developed a series of force fields with the intention of simulating biological systems by attempting to accurately reproduce experimental Kirkwood-Buff integrals which are observed for solution mixtures. Here, we describe our most recent efforts towards a complete force field for peptides and proteins. The results are illustrated using molecular dynamic simulations of several tripeptides, selected peptides and globular proteins at ambient temperature and pressure followed by replica exchange molecular dynamic simulations of a few selected peptides. It is observed that the side chain torsional potentials of KBFF are in good agreement with the experimental data whereas the backbone potentials need further improvement.

5.2 Introduction

Accurate empirical force fields are required to study the behavior of proteins and peptides via computer simulations.[1,2] There are several state of the art force fields which are currently available such as CHARMM19[3] and 22,[4] OPLS,[5] AMBER,[6] and GROMOS.[7] Moreover, all of them are specifically designed to the study biological systems. Although these force fields have been extensively used to study a wide variety of biological systems in some detail, still there are several inherent shortcomings which reduce their accuracy. Mainly there are two issues and one of them is related to the degree of sampling achieved during a simulation. The simulation time should be long enough to enable the sampling of all relevant molecular conformations and this essentially determines the precision of the simulation results. The other issue is the accuracy of the force field. An inaccurate energy function may bias the simulation towards incorrect behavior and hence it will significantly affect the accuracy of the data.[8]

Furthermore, larger molecules such as proteins, which have many potential conformations, are more susceptible to sampling problems compared to smaller molecules which may not be severely affected by sampling limitations. Many approaches to improve the degree of sampling in molecular simulations have been developed and they consist of techniques related not only to software issues but also to hardware issues as well.[8-11] With current approaches and computers one can perform simulations of reasonably large systems on the microsecond timescale when applying enhanced sampling.[8-11] Unfortunately, many of these longer MD simulations are not accurate when existing force fields are used.[9,12] Therefore, continuous improvements in the accuracy of force field are still required. Some studies have focused on developing more accurate polarizable force fields,[13] but those are computationally expensive. Thus, there is still room to improve existing non-polarizable force fields.

Recently, there have been notable efforts to improve the accuracy of the existing force fields in terms of the re-parameterization of the torsional potentials. These potentials which are defined by Equation 5.1, determine the conformational preferences observed for the backbone (alpha versus beta) of amino acid residues.

$$U_{\substack{proper \\ dihedral}} = \Sigma k_{\phi}[1 + \cos(n\phi - \phi_s)] \quad (5.1)$$

Modifications to the protein backbone and sidechain torsional potentials of the Amber and CHARMM force fields have been performed and they have resulted in significant improvements in the accuracy of the results.[14] Moreover, the Amber99SB force field[15] and the CMAP correction for CHARMM22[16] focused on the backbone torsional potentials. On the other hand, the recently developed Amber 99SB ILDN force field[17] included improved sidechain dihedral potentials.

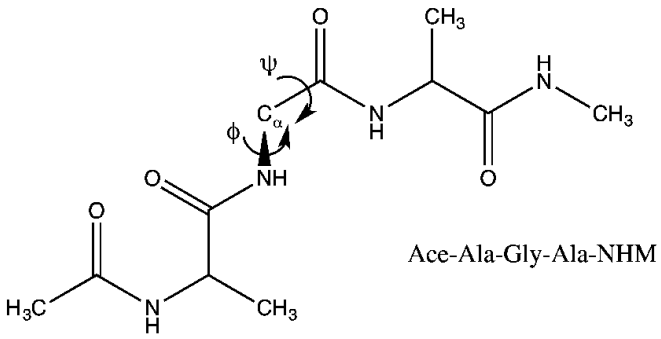
In this study, we present our most recent efforts to develop the torsional potentials of peptides and proteins for the KB derived force field, KBFF. Furthermore, in order to improve the backbone torsional potentials, we have used the approach used in the implementation of CMAP correction for CHARMM22 force field.[16]

5.3 Methods

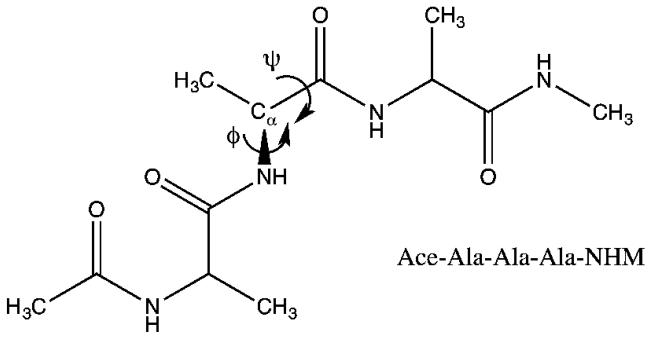
5.3.1 Model Systems for Peptides and Proteins

As mentioned above most of the recent efforts to improve biomolecular force fields have centered around the critically important ϕ and ψ degrees of freedom. Moreover, the typical model systems which were used in those studies are glycine, alanine and proline dipeptides, which are actually single capped amino acids, Ace-X-NMH, where X represents the respective amino acid Gly, Ala, Pro, etc. However, a more appropriate model system which is representative of the peptides and proteins would be the tripeptides of the form, Ace-AXA-NHM, where A represents

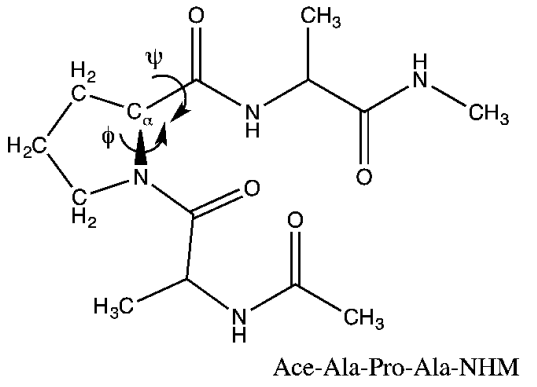
Ala and X represents Gly, Ala, Pro, etc. As shown in Figure 5.1, there are two peptide bonds surrounding the central ϕ (C-N-C_α-C) and ψ (N-C_α-C-N) dihedral angles in each tripeptide.



Ace-Ala-Gly-Ala-NHM



Ace-Ala-Ala-Ala-NHM



Ace-Ala-Pro-Ala-NHM

Figure 5.1 Capped glycine, capped alanine and capped proline tripeptides.

5.3.2 Regular Molecular Dynamics Simulations

All molecular dynamics simulations were performed using the KBFF models (http://kbff.chem.k-state.edu)[18,19,20] together with the SPC/E water model[21] as implemented in the GROMACS 4.0.5 package.[22] All simulations were performed at 300 K and the pressure of 1 bar using the weak coupling technique to modulate the temperature and pressure with relaxation times of 0.1 and 0.5 ps,[23] respectively. A time-step of 2 fs was used and the bond lengths were constrained using the Lincs (solutes) and Settle (water) algorithms.[24,25] The particle mesh Ewald technique was used to evaluate electrostatic interactions with a grid resolution of 0.1 nm.[26] A real space convergence parameter of 3.5 nm^{-1} was used in combination with twin range cutoffs of 1.0 and 1.5 nm, and a nonbonded update frequency of 5 steps. All tripeptides were solvated in a cubic box of 4.0 nm.

5.3.2.1 Small Peptides

Several small peptides with well characterized secondary structure were selected for validation and testing of the KBFF and the full list of peptides is shown in Table 5.1. MD simulations were performed using the KBFF models. Each peptide was initially solvated in 7.5 nm cubic water boxes containing about 13750 water molecules. The net charges of the peptide systems were neutralized by adding sodium or chloride ions as required. Each system was initially subject to energy minimization, followed by 1 ns of MD simulation in the NPT ensemble, with position restraints on the backbone atoms. After this initial equilibration, each system was simulated for 100ns in the NPT ensemble. The trajectories obtained from these 100ns runs were used for subsequent data analysis.

Table 5.1 Selected small peptides for the validation and testing of KBFF.

Peptide	Sequence	Length	T
AAQAA	Ace-AAQAAAAQAAAAQAA-NHM	15	300 K
pepIII	Ace-AETAAAKFLRAHA-NH_2	13	300 K
CLN025	YYDPETGTWY	10	300 K
Trpzip1	SWTWEGNKWTWK-NH_2	12	300 K
GB1 hairpin	GEWTYDDATKTFTVTE	16	300 K
GB1m3	KKWTYNPATGKFTVQE	16	300 K

5.3.2.2 Globular Proteins

MD simulations of globular proteins such as ubiquitin, lysozyme and RNaseA were also performed using the KBFF models and the details of the systems are summarized in Table 5.2. The net charge of the proteins was neutralized by adding sodium or chloride ions as required. Each system was initially subject to energy minimization which was followed by three 1 ns periods of MD simulation in the NPT ensemble where the temperature was at 100K, 200 K and 300K while applying position restraints to the backbone atoms. After this initial equilibration, each system was simulated for 100ns in the NPT ensemble and the trajectories obtained from these 100ns runs were used for subsequent data analysis.

Table 5.2 Selected globular proteins for the validation of KBFF.

Protein	PDB ID	Number of residues	Length of the cubic box	Number of water molecules
BPTI	5PTI	58	6.9 nm	10376
CI-2	2CI2	83	7.6 nm	13428
GA98	2LHC	56	8.2 nm	18211
GB98	2LHD	56	6.7 nm	9738
Lysozyme	4LZT	129	7.9 nm	15982
NTL9	2HBB	51	7.2 nm	12235
RNaseA	2AAS	124	7.7 nm	14852
Ubiquitin	1UBQ	76	7.4 nm	13352

5.3.3 Replica Exchange Molecular Dynamics Simulations of Small Peptides

Replica exchange molecular dynamics simulations of the selected small peptides listed in Table 5.1 were performed according to the same simulation details as described in section 5.3.2. Each peptide was initially solvated in 6.0 nm cubic water boxes containing about 7100 water molecules. The net charges of the peptide systems were neutralized by adding sodium or chloride ions as required. Each system was initially subject to energy minimization which was followed by three 1 ns periods of MD simulation in the NPT ensemble where the temperature was at 100K, 200 K and 270K while applying position restraints to the backbone atoms. After this initial equilibration, 48 replicas of each peptide were generated using a script and those replicas were distributed in a temperature range of 270 K to 370 K. Then, each system was simulated for about 100ns in the NPT ensemble while attempting to exchange between replicas after every 500 steps

periodically. The trajectories obtained from these production runs were used for subsequent data analysis.

5.3.4 Crystal Structure Data Bases

In this study we have considered two crystallographic data bases to obtain the experimental torsional (backbone and side chain) angles. These two data bases are the protein data bank and the protein coil library.[27] The protein data bank is a crystallographic data base which consists of three dimensional structural data of large biological molecules such as proteins and nucleic acids. Approximately half of the structure of folded proteins is either alpha helix or beta strand. Thus, this data base contains more structured conformations and it is dominated by alpha helices. On the other hand, the protein coil library is developed by removing alpha helices and beta strands from the reported structures in the PDB. Hence, it is composed of non-alpha helix and non-beta stand fragments and tends to favor extended conformations.

Since the PDB contains more structured conformations it seems to be an extreme and hence, we decided to use the protein coil library in this study. Recently, a couple of state of the art force fields have been improved by using the protein coil library[28, 29] and our implementation of it will be explained in section 5.4.2.

5.4 Results and Discussion

5.4.1 Side Chain Torsional Potentials

Our first goal was to establish a set of side chain torsional potentials which would be in agreement with the protein coil library and the PDB. The side chain torsional potentials for all the residues which have a side chain were fitted to Equation 5.1 and they are presented in Figure 5.2,

Figure 5.3 and Figure 5.4. According to those figures there are no major differences between the side chain torsional potentials predicted by the protein coil library and the PDB. Moreover, with already developed side chain torsional potentials of KBFF model[30] we are able to match them quite well in some residues while the other residues are in good agreement too.

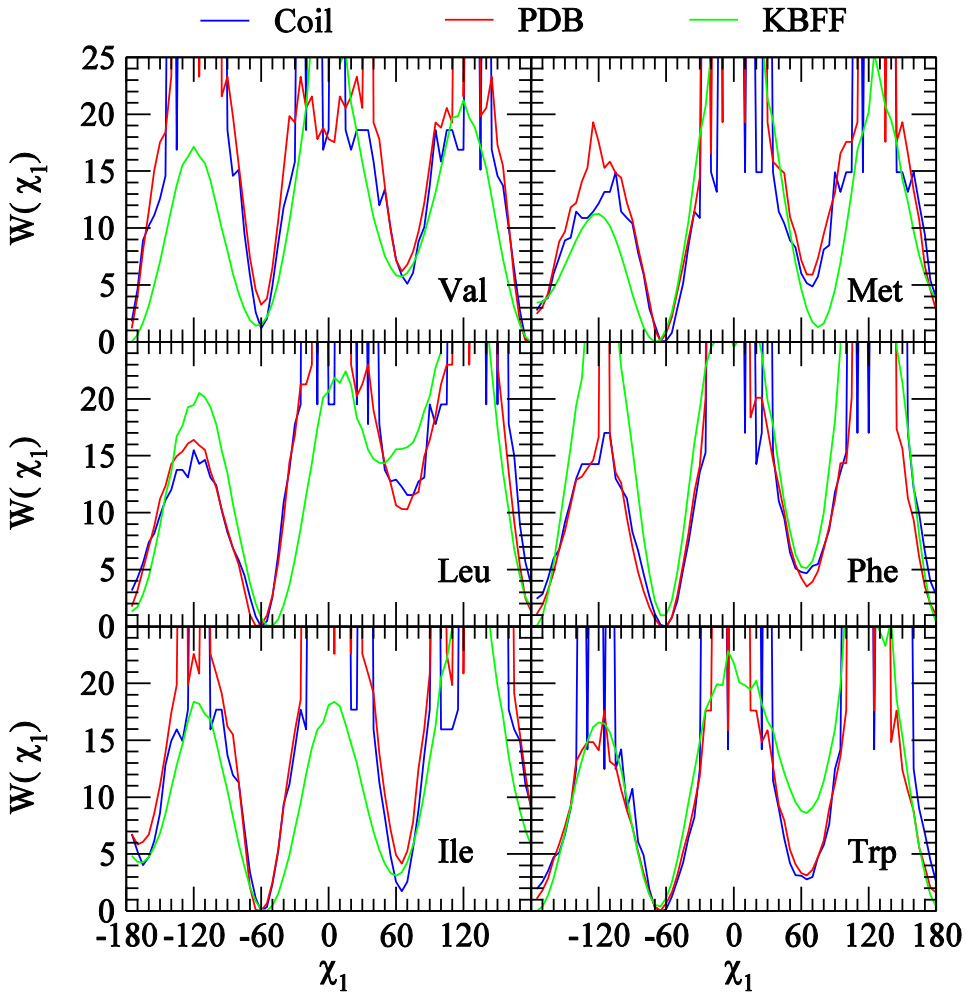


Figure 5.2 Side chain torsional potentials of residues with nonpolar side chains.

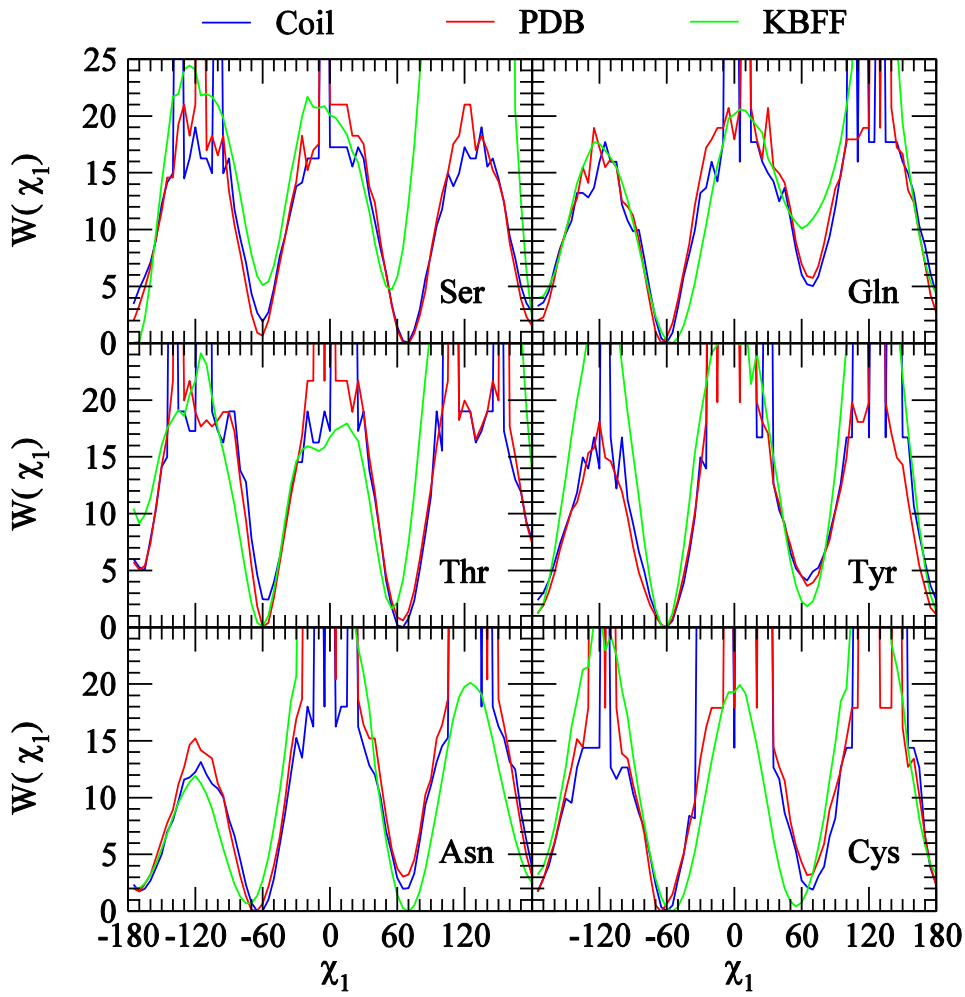


Figure 5.3 Side chain torsional potentials of residues with uncharged polar side chains.

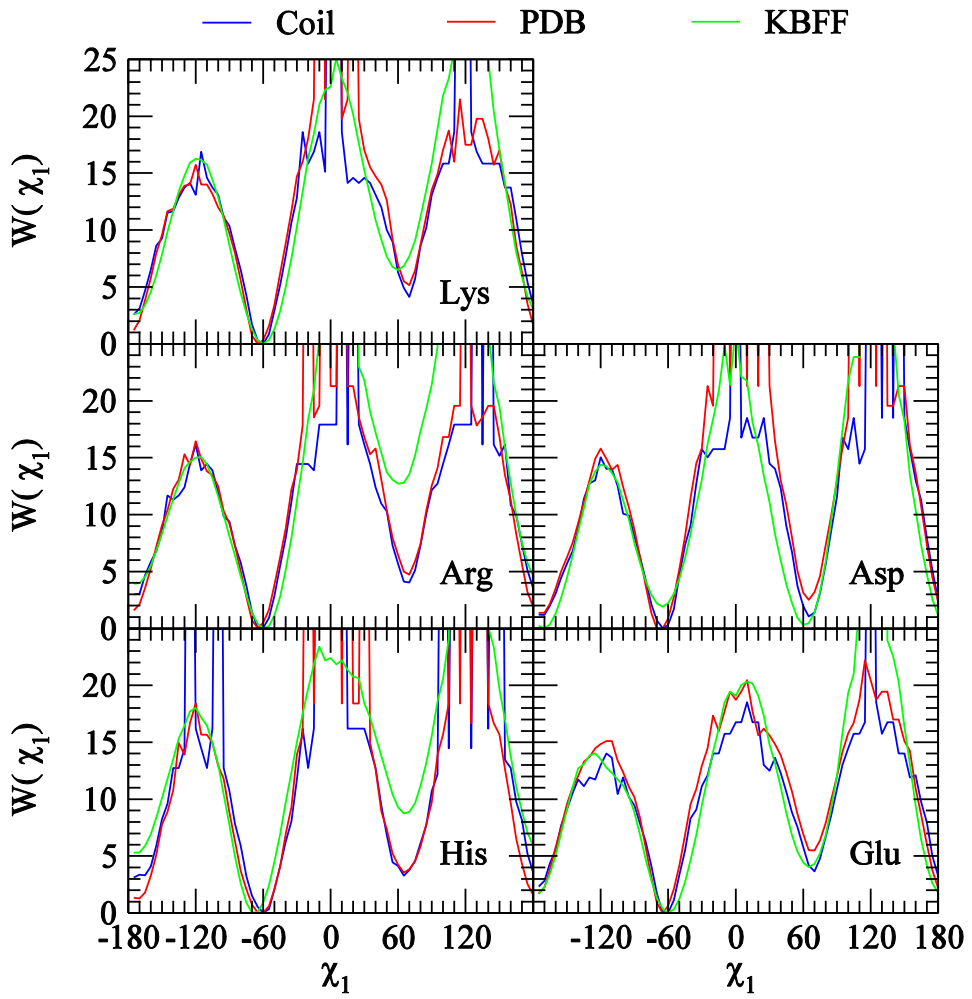


Figure 5.4 Side chain torsional potentials of residues with charged polar side chains.

5.4.2 Backbone Torsional Potentials

Once the side chain torsional potentials were finalized our next challenge was to develop a set of backbone torsional potentials. The model compounds used in here were the relevant tripeptides as described in section 5.3.1. Moreover, glycine, alanine and proline were treated separately and all the other amino acid residues were treated together. Glycine was treated separately since there are two hydrogen atoms attached to the C_α, alanine was treated separately since it has only a methyl group as the side chain and all the other residues except for glycine, alanine and proline were treated together since all of them have side chains of moderate sizes. In

addition, the CMAP correction used in CHARMM22[16] was implemented on all those residues except for proline. Since proline has restrictions on phi-psi space which arises due to the five membered ring, its phi is restricted to -60º. Hence, a simple one-dimension potential which is defined by Equation 5.1 was applied on psi of proline.

When implementing the CMAP correction on KBFF, the relevant tripeptide was simulated without any phi-psi potentials and the resulting base energy map was subtracted from the respective energy map produced by the corresponding residue of the protein coil library. Then, that difference in energy was fitted to a polynomial to generate the corresponding energy grid. Finally, the selected tripeptide was simulated with that CMAP correction. Furthermore, when the fitting was done glycine was fitted to glycine, alanine was fitted to alanine and all the other amino acid residues except for proline were fitted together to all the amino acid residues from the protein coil library except for glycine, alanine and proline.

All of the twenty tripeptides were simulated for 500 ns at 300 K and the percentage conformations are summarized in the third column of Table 5.3. Moreover, the percentage conformations of amino acid residues from the PDB and the coil data base are summarized in the first and second columns, respectively. As explained earlier the alpha helix percentages of alanine, glutamine and glutamic acid are relatively higher in the PDB than in the coil library. Furthermore, the percentage conformations obtained using KBFF seems to be in pretty good agreement with the coil library over the PDB. Therefore, our decision to use the coil library is justified by these results.

Table 5.3 The conformational populations (%) of the tripeptides at 300 K.

	PDB				Coil Library				KBFF			
	α^a	$\beta_1{}^a$	$\beta_2{}^a$	$\phi>0$	α	β_1	β_2	$\phi>0$	α	β_1	β_2	$\phi>0$
Gly	61	23	16	0	60	21	19	0	61	22	18	0
Ala	67	19	14	0	45	16	35	4	53	26	22	0
Pro	41	1	58	0	34	1	66	0	29	1	70	0
Val	38	52	11	0	36	35	28	1	24	34	41	2
Leu	56	30	13	1	47	20	30	3	39	24	34	3
Ile	43	47	10	0	38	34	27	1	29	34	37	0
Met	56	31	12	1	41	25	28	6	40	28	30	2
Phe	45	42	11	2	42	28	24	5	22	36	41	1
Trp	49	36	14	1	45	23	27	4	17	35	45	4
Ser	47	33	19	2	44	24	27	4	23	44	33	1
Thr	45	41	13	0	48	30	21	1	42	23	35	0
Asn	46	28	16	10	40	21	18	21	48	23	28	0
Gln	60	26	12	3	47	23	22	8	36	31	33	0
Tyr	45	42	12	2	44	28	23	5	45	25	28	2
Cys	43	41	14	2	35	32	27	6	44	23	25	8
Lys	58	26	13	3	48	21	22	8	29	31	40	0
Arg	56	28	13	2	45	23	24	8	32	33	34	1
His	47	35	14	4	45	26	19	10	30	39	29	3
Asp	51	24	20	5	45	20	25	10	38	20	41	1
Glu	64	22	12	2	53	17	24	6	47	19	33	0

[a] α: $\phi<0^o$ and $-150^o<\psi<30^o$, β_1: $\phi<-90^o$ and $\psi>30^o$, β_2: $\phi>-90^o$ and $\psi>30^o$

5.4.3 Regular Molecular Dynamic Simulations

5.4.3.1 Small Peptides

The variation of root mean square deviation (RMSD) values with time for the selected peptides simulated with regular MD are presented in Figure 5.5. It seems like that the selected hairpins behaved better than the two helices. Moreover, among the hairpins, CLN025 seems to perform quite well with a RMSD value of less than 2 Aº. In addition, the other three hairpins are

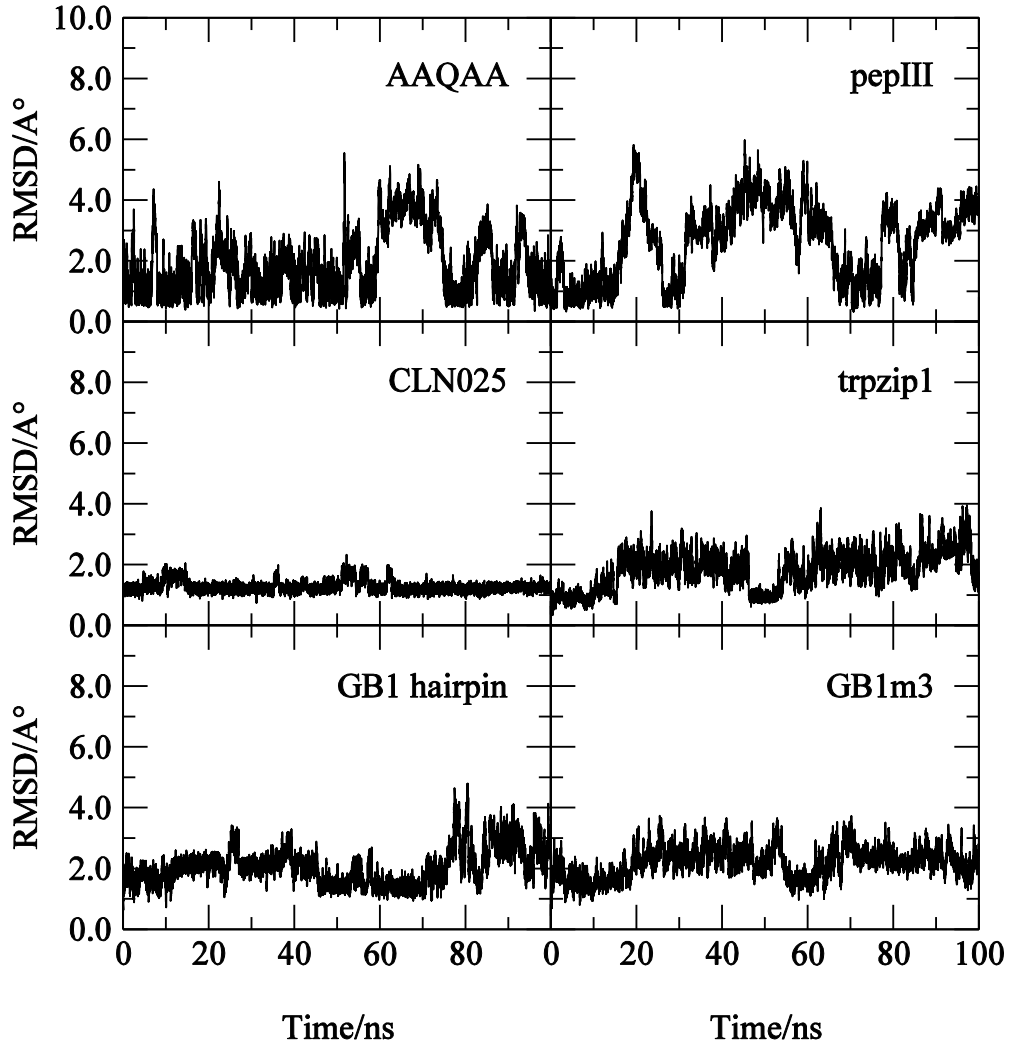


Figure 5.5 The RMSDs of helices and hairpins at 300 K.

not as stable as CLN025. On the other hand, AAQAA was performing quite well up to about 50 ns and afterwards seems like that it is deviated a bit. Also, pepIII does not look good at the end of 100 ns. However, according to the experiments the percentage of helices should be about 50% around 280 K. Consequently, the helices may not be that bad since these RMSDs were calculated from simulations performed at 300 K. Hopefully, those RMSDs of helices will be improved along with the other peptides when they are simulated for longer, which would be helpful in validating the KBFF.

5.4.3.2 Globular Proteins

The variation of RMSD values with time for the selected globular proteins simulated with regular MD are presented in Figure 5.6 and GB98 and CI-2 seem to be the best out of this series of proteins with RMSDs of about 2 Aº. Furthermore, BPTI, ubiquitin, lysozyme and NTL9 seems to be behaving reasonably up to 100 ns. In addition, GA98 seems to be the most deviated one as of now and also RNaseA is not perfect either. Moreover, both GA98 and GB98 have identical sequences consisting of fifty-six residues each and they only differ by the forty-fifth residue. Still they are structurally different and that might be a reason for the extremes observed in RMSDs, i.e. GB98 having a better RMSD whereas GB98 having the most deviated RMSD. Hopefully, all of these globular proteins will improve when simulated for another couple of hundred nanoseconds, which would be helpful in validating the KBFF.

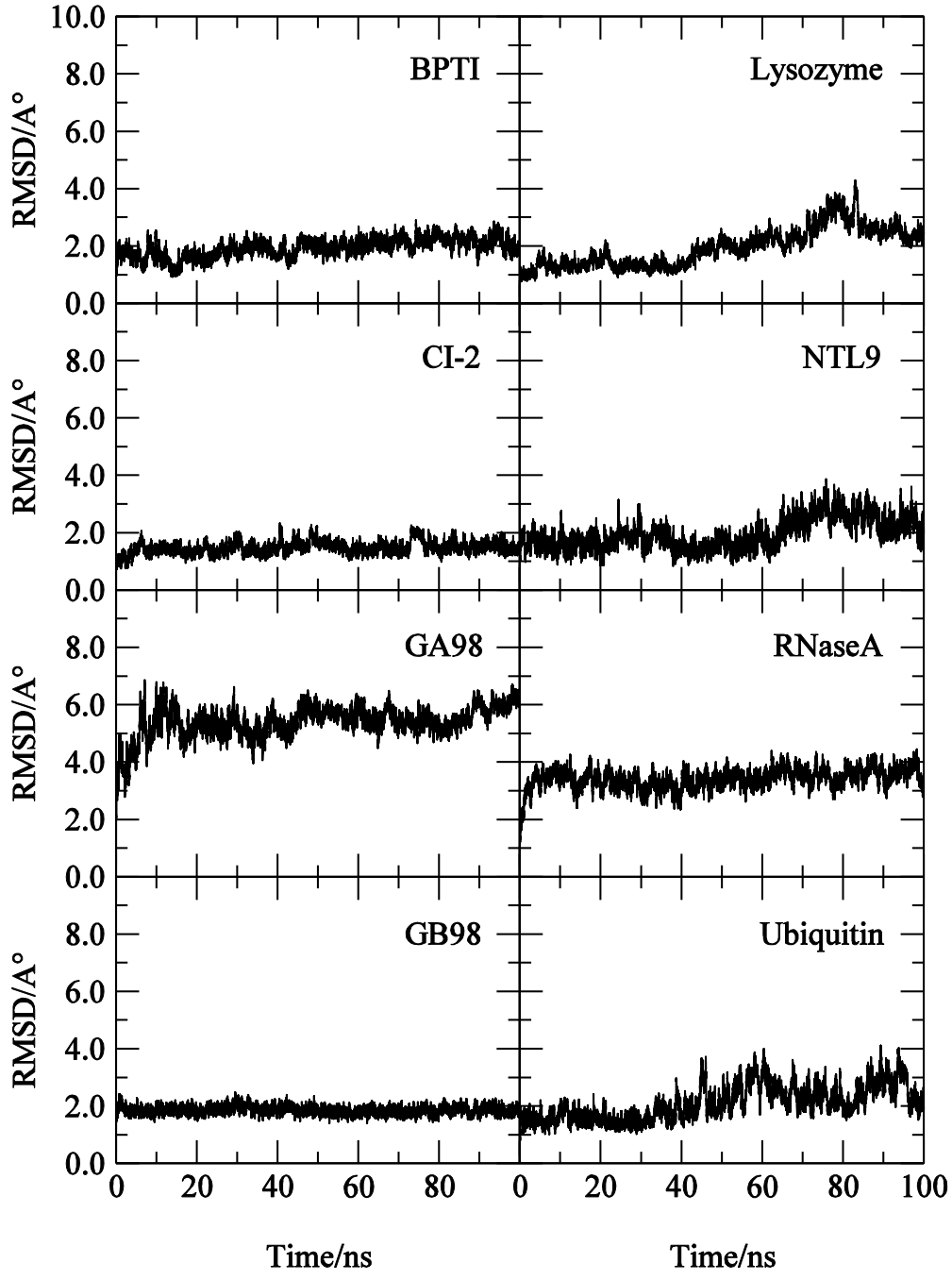


Figure 5.6 The RMSDs of globular proteins at 300 K.

5.4.4 Replica Exchange Molecular Dynamic Simulations of Small Peptides

In order to achieve better sampling replica exchange molecular dynamic simulations of the selected small peptides were performed and the trajectories were analyzed to determine the melting curves. They are presented in Figure 5.7 and helices are relatively more stable than the hairpins.

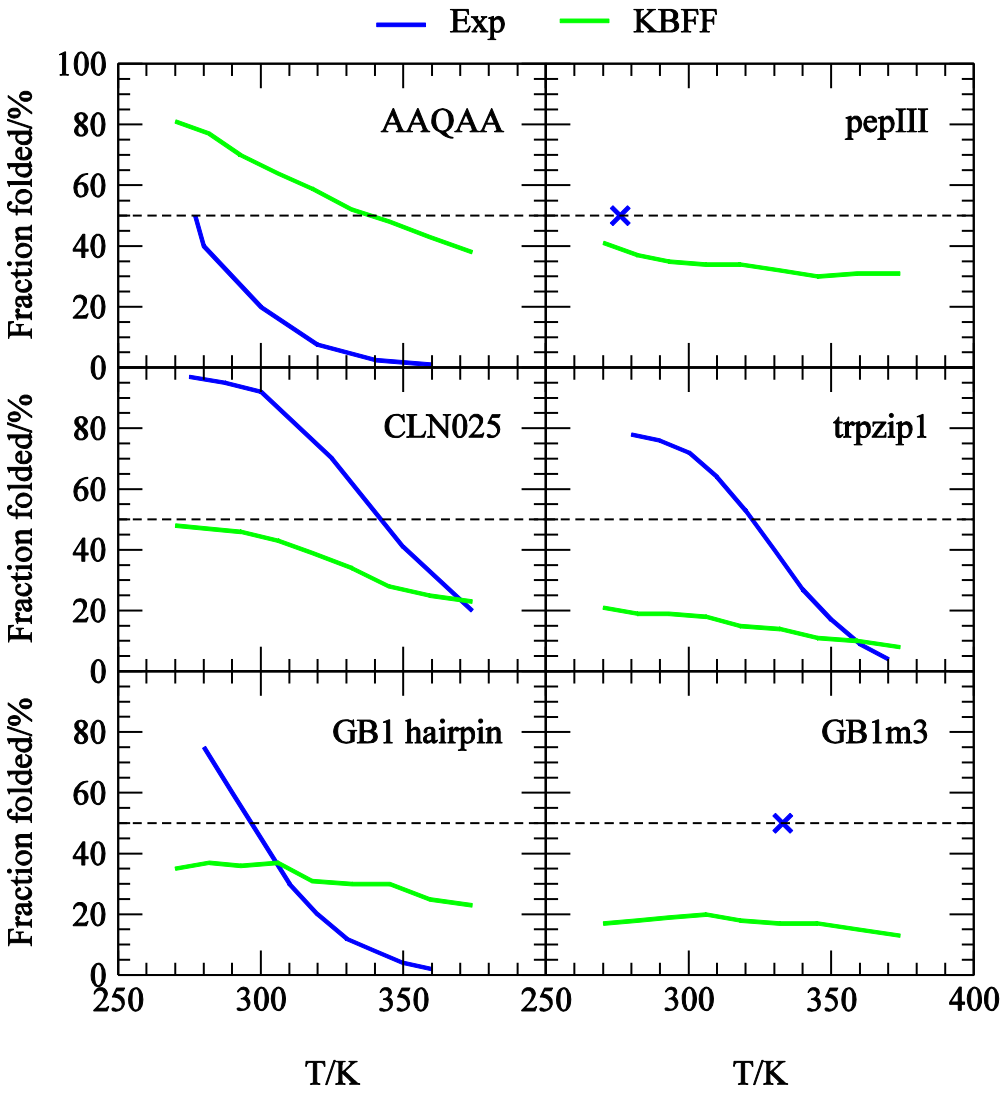


Figure 5.7 The melting curves of the selected small peptides.[31-36]

Moreover, out of the selected two helices AAQAA is overly stable compared to the experiments whereas the fraction folded of pepIII is about 40% at the melting temperature. Furthermore,

AAQAA is mainly consisting of alanine and the over stability of AAQAA may be due to the high helical nature of the current alanine map.

On the other hand, all the four selected hairpins are relatively unstable compared to the experiments and the melting curve of CLN025 is the closest to the experimental melting curve. When peptides fold and unfold the degrees of freedom in a hairpin is greater compared to that in helices and this may be one of the reasons for hairpins to be unstable. Also, there may be unnecessary hydrogen bonds which are created between the OH of the side chains and the C terminus of the backbone which causes the unfolded structures of hairpins to be more favored. In addition, the instability of hairpins may be due to the competition between alpha turn residues and beta non-turn residues. Moreover, the hairpins were relatively unstable with most of the state of the art force fields in their early stages.

Recently, Jiang and coworkers developed a new strategy for protein force field parametrization where they have used the backbone and side chain conformational distributions of all twenty amino acid residues obtained from protein coil library were used as the target data.[28] In their study they modified the torsion potentials and some local non-bonded interactions in OPLS-AA/L force field and the new force field was named as Residue Specific Force Field 1 (RSFF1). RSFF1 gave a good balance between alpha helical and beta sheet secondary structures and successfully folded a set of alpha helix proteins and beta hairpins. However, it overestimates the melting temperature and the stability of native state of these peptides/proteins.

Moreover, in another study Zhou and coworkers modified the Amber ff99SB force field and it was named as Residue Specific Force Field 2 (RSFF2).[29] RSFF2 gave melting curves of alpha helical peptides and Trp-cage in good agreement with experimental data whereas it overestimated the melting temperature and the stability of beta hairpins.

As discussed above although there have been improvements still the hairpins are not in perfect agreement with experiments. Furthermore, Mercadante and coworkers have used the KB derived force field to reproduce the correct conformational ensemble of intrinsically disordered proteins and new route for tackling the deficiencies of current protein force fields in describing protein solvation.[37] Hence, it seems that there are promising signs of KBFF being more reliable.

5.5 Conclusions and Future Directions

The side chain torsional potentials of KBFF are in good agreement with both the PDB and the protein coil library. Hence, it can be concluded that the side chain torsional potentials are finalized for KB derived force field. Moreover, the percentage conformations of most of the tripeptides are reproducing the percentage conformations of both the PDB and the protein coil library or at least that of the protein coil library. Therefore, it seems that it is more appropriate to use the protein coil library to obtain the CMAP corrections of KBFF. Furthermore, the RMSDs of most of the selected small peptides and globular proteins are well behaved during regular MD simulations.

However, with REMD simulations, AAQAA is overly stable and pepIII and the selected hairpins are relatively unstable with KBFF as of now. The causes for the instability of hairpins may be the unnecessary hydrogen bonds created between the OH of the side chains and the C terminus of the backbone and the competition between alpha turn residues and beta non-turn residues. In the future we are planning to treat each residue separately and attempt to develop a residue specific KBFF which will be able to improve folding of peptides and proteins.

5.6 References

1. Brooks, C. L., III; Karplus, M.; Pettitt, B. M. Proteins: A Theoretical Perspective of Dynamics, Structures, and Thermodynamics; John Wiley & Sons: New York, **1988**.

2. Becker, O. M.; MacKerell, J.; A. D.; Roux, B.; Watanabe, M., Eds. Computational Biochemistry and Biophysics; Marcel-Dekker, Inc.: New York, 2001.

3. Neria, E.; Fischer, S.; Karplus, M. *J. Chem. Phys.* **1996**, 105, 1902-1921.

4. MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem.* B **1998**, 102, 3586-3616.

5. Jorgensen, W. L.; Tiradorives, J. *J. Am. Chem. Soc.* **1988**, 110, 1657.

6. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. J. *Am. Chem. Soc.* **1995**, 117, 5179-5197.

7. van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hunenberger, P. H.; Kruger, P.; Mark, A. E.; Scott, W. R. P. Biomolecular Simulation: The GROMOS96 Manual and User Guide; BIOMOS b.v.: Zurich, 1996.

8. Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. *Curr. Opin. Struct. Biol.* **2009**, 19, 120-127.

9. Freddolino, P. L.; Liu, F.; Gruebele, M.; Schulten, K. *Biophys. J.* **2008**, 94, L75-L77.

10. Shaw, D. E.; Dror, R. O.; K., S. J.; P., G. J.; Mackenzie, K. M.; Bank, J. A.; Young, C.; M., D. M.; B., B.; Bowers, K. J.; Chow, E.; P., E. M.; Ierardi, D. J.; Klepeis, J. L.; S., K. J.; Larson, R. H.; Lindorff-Larsen, K.; Maragakis, P.; Moraes, M. A.; Piana, S.; Shan, Y.; Towles, B. Millisecond-scale molecular dynamics simulations on Anton. In: Proceedings of the 2009 ACM/IEEE Conference on Supercomputing (SC09). ACM Press: Washington, DC, **2009**.

11. Liwo, A.; Czaplewski, C.; Oldziej, S.; Scheraga, H. A. *Curr. Opin. Struct. Biol.* **2008**, 18, 134.

12. Perez, A.; Marchan, I.; Svozil, D.; Sponer, J.; Cheatham, T. E.; Laughton, C. A.; Orozco, M. *Biophys. J.* **2007**, 92, 3817-3829.

13. Friesner, R. A. *Adv. Protein. Chem*. **2006**, 72, 79-104.

14. Best, R. B.; Buchete, N. V.; Hummer, G. *Biophys. J.* **2008**, 95, L7-L9.

15. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins-Structure Function and Bioinformatics* **2006**, 65, 712-725.

16. Mackerell, A. D.; Feig, M.; Brooks, C. L. *J. Comput. Chem.* **2004**, 25, 1400-1415.

17. Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. *Proteins-Structure Function and Bioinformatics* **2010**, 78, 1950-1958.

18. Kusalik, P. G.; Patey, G. N. *J. Chem. Phys.* **1987**, *86*, 5110-5116.

19. Aburi, M.; Smith, P. E. *J. Phys. Chem. B* **2004**, 108, 7382-7388.

20. Kang, M.; Smith, P. E. *J. Comput. Chem*. **2006**, 27, 1477-1485.

21. Ploetz, E. A.; Bentenitis, N.; Smith, P. E. *Fluid Phase Equilib*. **2010**, 290, 43-47.

22. Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem*. **1987**, 91, 6269-6271.

23. Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput*. **2008**, 4, 435-447.

24. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Dinola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, 81, 3684-3690.

25. Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem*. **1997**, 18, 1463-1472.

26. Miyamoto, S.; Kollman, P. A. *J. Comput. Chem.* **1992**, 13, 952-962.

27. Fitzkee, N. C.; Fleming, P. J.; Rose, G. D. *Poteins: Structure, Function, and Bioinformatics* **2005**, 58, 852-854.

28. Jiang, F.; Zhou, C.; Wu, Y. *J. Phys. Chem. B* **2014**, 118, 6983-6998.

29. Zhou, C. Y.*;* Jiang, F.; Wu, Y.D. *J. Phys. Chem. B* **2015**, 119, 1035-1047.

30. Jiao, Y. Dissertation 2012.

31. Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. *PLoS ONE* **2012**, 7, e32131.

32. Shoemaker, K. R. Kim, P. S.; York, E.; Stewart, J. M.; Baldwin, R. L. *Nature* **1987**, 326, 563-567.

33. Honda, S.; Akiba, T.; Kato, Y. S.; Sekijima, M.; Ishimura, K.; Ooishi, A.; Watanabe, H.; Odahara, T.; Harata, K. *J. Am. Chem. Soc.* **2008**, 130, 15327-15331.

34. Cochran, A. G.; Skelton, N. J.; Starovanski, M. A. *Proceedings of the National Academy of Sciences of the United States of America* **2001**, 98, 5578-5583.

35. Munoz, V.; Tompson, P. A.; Hofrichter, J.; Eaton, W. A. *Nature* **1997**, 390, 196-199.

36. Fesinmeyer, R. M.; Hudson, F. M.; Andersen, N. H. *J. Am. Chem. Soc.* **2004**, 126, 7238−7243

37. Mercadante, D.; Milles, S.; Fuertes, G.; Svergun, D. I.; Lemke, E. A.; Gräter, F. *J. Phys. Chem. B* **2015**, 119, 7975-7984.

Appendix A - Experimental Triplet and Quadruplet Fluctuation Densities and Spatial Distribution Function Integrals for Pure Liquids

A.1 Abstract

Fluctuation Solution Theory has provided an alternative view of many liquid mixture properties in terms of particle number fluctuations. The particle number fluctuations can also be related to integrals of the corresponding two body distribution functions between molecular pairs in order to provide a more physical picture of solution behavior and molecule affinities. Here, we extend this type of approach to provide expressions for higher order triplet and quadruplet fluctuations, and thereby integrals over the corresponding distribution functions, all of which can be obtained from available experimental thermodynamic data. The fluctuations and integrals are then determined using the International Association for the Properties of Water and Steam Formulation 1995 (IAPWS-95) equation of state for the liquid phase of pure water. The results indicate small, but significant, deviations from a Gaussian distribution for the molecules in this system. The pressure and temperature dependence of the fluctuations and integrals, as well as the limiting behavior as one approaches both the triple point and the critical point, are also examined.

A.2 Introduction

From a theoretical point of view, liquids and liquid mixtures are commonly characterized in terms of probability distribution functions. These distribution functions provide a way to describe liquid structure, and can be further used to relate this structure to the corresponding thermodynamics.[1] The Kirkwood-Buff (KB) theory of solutions provides such a link between

integrals over the spatial pair distribution functions and the thermodynamic properties of any stable multicomponent system.[2] These integrals can also be expressed in terms of particle fluctuation densities and both quantities can be considered to characterize a liquid or liquid mixtures.[3] Consequently, KB theory, also more generally known as Fluctuation Solution Theory (FST), has provided an alternative view of many solution properties in terms of particle number fluctuations and distributions.[2,4-5]

Many theoretical approaches also employ distribution functions beyond the simple pair distribution. The role of triplet and higher distribution functions in liquids is well established.[6-9] However, quantitative information concerning these distributions remains quite limited, especially from experimental sources.[7-10] In particular, there are no systematic studies of triplet correlations over a wide range of temperature and pressure for complex liquids using experimental data. Previous work has been primarily restricted to scattering studies that provide the pair distribution function, and thereby partial information concerning triplet distributions, via studies of the pressure and temperature dependence of the pair distribution. Unfortunately, scattering studies are usually limited to molecules of low complexity.

Here we describe an extension of traditional FST to generate triplet and quadruplet particle number fluctuations, together with the corresponding integrals over the triplet and quadruplet spatial distribution functions, in an effort to provide experimental data concerning higher distributions in pure liquids at any density. The new expressions are then combined with existing relationships to systematically investigate these higher correlations in pure water over a range of pressure and temperature. Finally, we compare and contrast this approach with currently available experimental methods that attempt to access similar liquid state correlations. The results can therefore be used to provide rigorous tests of current theories of solutions, to guide the

development of accurate models for computer simulation, and to further relate solution structure to solution thermodynamics.

A.3 Theory

As is traditional with Fluctuation Solution Theory, one starts with the equations of the Grand Canonical Ensemble (GCE) and then uses various thermodynamic transformations to provide properties corresponding to either semi-open osmotic systems, or fully closed isothermal isobaric systems.[11] Hence, all ensemble averages and distribution functions in this type of approach correspond to the GCE. The fluctuating quantities are then related to integrals over distribution functions. Usually the application is to liquids, but there is no reason the same approach cannot be applied to pure gases or even solids.

The thermodynamic potential and partition function in the GCE - where the set of chemical potentials ($\{\mu\}$), the volume (V) and the absolute temperature (T) are the independent variables - can be written,[1]

$$\beta p = V^{-1} \ln \Xi$$
$$\Xi = \sum_{\{N\}=0}^{\infty} Q(\{N\}, V, T) e^{\beta \boldsymbol{\mu} \cdot \mathbf{N}} \tag{A.1}$$

where $\beta = (k_B T)^{-1}$, p is the pressure, $\boldsymbol{\mu} \cdot \mathbf{N} = \mu_1 N_1 + \mu_2 N_2 \cdots$, and N_α is the number of molecules of species α. The sum is over the full permutation of molecule numbers, and $Q(\{N\}, V, T)$ is the canonical partition function for each system of $\{N\}$ molecules in the same fixed volume. The most useful value for the Boltzmann constant (k_B) in this work is 0.083143714 bar L mol^{-1} K^{-1}. The above partition function applies for any system where Boltzmann statistics are obeyed. Our

main focus will be single component liquids. We will, however, retain the full multicomponent expressions during the initial derivation and then simplify to a single component later.

The corresponding differential for the GCE can be written in terms of just the intensive variables,

$$d\beta p = -\frac{U}{V} d\beta + \sum_{\alpha} \frac{N_\alpha}{V} d\beta\mu_\alpha \tag{A.2}$$

where U is the internal energy. The ensemble average of a property (X) in the GCE is given by,

$$\langle X \rangle = e^{-\beta pV} \sum_{\{N\}} \sum_{i} X e^{-\beta E_i} e^{\beta \boldsymbol{\mu}\cdot\mathbf{N}} \tag{A.3}$$

and Equation A.1 and A.2 lead directly to the following relationships for the internal energy density and the particle number densities of each species,

$$\begin{aligned} \frac{U}{V} &= -\left(\frac{\partial V^{-1} \ln \Xi}{\partial \beta} \right)_{\{\beta\mu\}} = \frac{\langle E \rangle}{V} \\ \frac{N_\alpha}{V} &= \left(\frac{\partial V^{-1} \ln \Xi}{\partial \beta\mu_\alpha} \right)_{\beta,\{\beta\mu\}'} = \frac{\langle N_\alpha \rangle}{V} = \rho_\alpha \end{aligned} \tag{A.4}$$

for any multicomponent system.[1] Here, the prime indicates that all chemical potentials except for the one of interest are held constant.

One can continue to take derivatives with respect to the chemical potentials in the GCE to provide,[6]

$$\left(\frac{\partial \rho_\alpha}{\partial \beta\mu_\beta}\right)_{\beta,\{\beta\mu\}'} = V^{-1}\left[\left\langle \delta N_\alpha \delta N_\beta \right\rangle\right] \equiv B_{\alpha\beta}$$

$$\left(\frac{\partial B_{\alpha\beta}}{\partial \beta\mu_\gamma}\right)_{\beta,\{\beta\mu\}'} = V^{-1}\left[\left\langle \delta N_\alpha \delta N_\beta \delta N_\gamma \right\rangle\right] \equiv C_{\alpha\beta\gamma} \tag{A.5}$$

$$\left(\frac{\partial C_{\alpha\beta\gamma}}{\partial \beta\mu_\delta}\right)_{\beta,\{\beta\mu\}'} = V^{-1}\begin{bmatrix}\left\langle \delta N_\alpha \delta N_\beta \delta N_\gamma \delta N_\delta \right\rangle - \left\langle \delta N_\alpha \delta N_\beta \right\rangle \left\langle \delta N_\gamma \delta N_\delta \right\rangle - \\ \left\langle \delta N_\alpha \delta N_\gamma \right\rangle \left\langle \delta N_\beta \delta N_\delta \right\rangle - \left\langle \delta N_\alpha \delta N_\delta \right\rangle \left\langle \delta N_\beta \delta N_\gamma \right\rangle\end{bmatrix} \equiv D_{\alpha\beta\gamma\delta}$$

where $\delta X = X - \langle X \rangle$ denotes a fluctuation in the value of *X*. The particle number fluctuations can also be expressed in terms of density fluctuations, but this becomes increasingly more awkward for the higher moments of the distribution. The most appropriate intensive properties are the fluctuations per unit volume, or fluctuation densities, as displayed above. The fluctuation densities provide quantitative measures of the correlation between particles in an open system.

The above expressions are restricted to open systems. The next step is to provide a connection to equivalent closed systems, which are of more common interest. The intensive GCE averages are a function of the intensive thermodynamic variables associated with the GCE and therefore one can write the following general differential,

$$d\langle X \rangle = \left(\frac{\partial \langle X \rangle}{\partial \beta}\right)_{\{\beta\mu\}} d\beta + \sum\nolimits_\alpha \left(\frac{\partial \langle X \rangle}{\partial \beta\mu_\alpha}\right)_{\beta,\{\beta\mu\}'} d\beta\mu_\alpha \tag{A.6}$$

where *X* is an intensive property. It should be noted that there is no volume derivative in the above expressions as it can be shown that this derivative is zero when *X* is intensive, i.e. intensive properties only depend on the intensive variables.[6] If we restrict ourselves to the study of isothermal changes, the results relate to particle number fluctuations only. Energy fluctuations will

be included fully at a later date. However, some preliminary results are invoked at the end of the Results section (see A.3.6).

Taking derivatives of Equation A.6 with respect to pressure provides a series of useful relationships. When $X = \beta p = V^{-1} \ln \Xi$ one obtains the common relationship between the partial molar volumes ($\bar{V}_\alpha$),

$$\left(\frac{\partial V^{-1} \ln \Xi}{\partial p} \right)_{\beta,\{N\}} = \beta \sum\nolimits_\alpha \rho_\alpha \bar{V}_\alpha = \beta \tag{A.7}$$

Using $\langle X \rangle = \rho_\alpha$ in Equation A.6 provides,

$$\left(\frac{\partial \rho_\alpha}{\partial p} \right)_{\beta,\{N\}} = \beta \sum\nolimits_\beta B_{\alpha\beta} \bar{V}_\beta = \rho_\alpha \kappa_T \tag{A.8}$$

and corresponds to the Kirkwood-Buff theory of solutions expression for the isothermal compressibility,[2]

$$\kappa_T \equiv -\left(\frac{\partial \ln V}{\partial p} \right)_{\beta,\{N\}} \tag{A.9}$$

This well-known relationship can be used to extract particle-particle number (or density) fluctuations in pure liquids. Here, we wish to go beyond these pair correlations to examine the triplet and higher fluctuations (correlations). Finally, using $\langle X \rangle = B_{\alpha\beta}$ and $\langle X \rangle = C_{\alpha\beta\gamma}$ in Equation A.6 provides,

$$\begin{aligned} \left(\frac{\partial B_{\alpha\beta}}{\partial p} \right)_{\beta,\{N\}} &= \beta \sum\nolimits_\gamma C_{\alpha\beta\gamma} \bar{V}_\gamma \\ \left(\frac{\partial C_{\alpha\beta\gamma}}{\partial p} \right)_{\beta,\{N\}} &= \beta \sum\nolimits_\delta D_{\alpha\beta\gamma\delta} \bar{V}_\delta \end{aligned} \tag{A.10}$$

The above expressions describe the pressure dependence of the two and three body particle number fluctuations. Obviously, one could continue indefinitely. However, it is unclear if reliable estimates for higher distributions can be obtained from experiment. This represents a particular aim of the current work.

It is clear from the above expressions that if the multivariate particle number probability distribution for the liquid was simply Gaussian in nature, the *C*'s and *D*'s would then be zero and the *B*'s would therefore be independent of pressure. This is clearly not the case, as noted previously.[12] The expressions in Equation A.5 correspond to the cumulants of the multivariate particle number probability distribution expressed in terms of the central moments. Alternatively, they can be viewed as the mean, covariance, coskewness, and excess cokurtosis of the same distribution. We note that there are several different definitions of skewness and excess kurtosis in the literature. The definition of skewness and excess kurtosis referred to here are those provided by $VC_{\alpha\beta\gamma}$ and $VD_{\alpha\beta\gamma\delta}$ expressions, respectively.

Before discussing a real system of common interest, we note that the fluctuating quantities can be related to a series of corresponding distribution functions (see section A.7.1). For single component systems the relationships between the fluctuating quantities and the corresponding distribution functions are provided by,

$$B_{11} = \frac{\left\langle (\delta N_1)^2 \right\rangle}{V} = \rho_1 (1 + \rho_1 G_{11})$$

$$C_{111} = \frac{\left\langle (\delta N_1)^3 \right\rangle}{V} = \rho_1 (1 + 3\rho_1 G_{11} + \rho_1^2 G_{111}) \qquad \text{(A.11)}$$

$$D_{1111} = \frac{\left\langle (\delta N_1)^4 \right\rangle - 3\left\langle (\delta N_1)^2 \right\rangle^2}{V} = \rho_1 (1 + 7\rho_1 G_{11} + 6\rho_1^2 G_{111} + \rho_1^3 G_{1111})$$

which involve integrals over the *n*-body spatial distribution functions $g^{(n)}_{\alpha\beta\ldots}(r_1, r_2, \ldots)$ that are similar in form to the integrals appearing in the theory of imperfect gases or the McMillan-Mayer theory of (osmotic) solutions,[1]

$$G_{11} = V^{-1}\int\left[g^{(2)}_{11} - 1\right]dr_1 dr_2$$

$$G_{111} = V^{-1}\int\left[g^{(3)}_{111} - 1 - 3(g^{(2)}_{11} - 1)\right]dr_1 dr_2 dr_3 \tag{A.12}$$

$$G_{1111} = V^{-1}\int\left[g^{(4)}_{1111} - 1 - 4(g^{(3)}_{111} - 1) - 3(g^{(2)}_{11} - 1)(g^{(2)}_{11} - 1) + 6(g^{(2)}_{11} - 1)\right]dr_1 dr_2 dr_3 dr_4$$

The integrals can also be expressed in terms of particle number fluctuations,

$$\rho_1^2 V G_{11} = \left\langle (\delta N_1)^2 \right\rangle - \left\langle N_1 \right\rangle$$

$$\rho_1^3 V G_{111} = \left\langle (\delta N_1)^3 \right\rangle - 3\left\langle (\delta N_1)^2 \right\rangle + 2\left\langle N_1 \right\rangle \tag{A.13}$$

$$\rho_1^4 V G_{1111} = \left\langle (\delta N_1)^4 \right\rangle - 3\left\langle (\delta N_1)^2 \right\rangle^2 - 6\left\langle (\delta N_1)^3 \right\rangle + 11\left\langle (\delta N_1)^2 \right\rangle - 6\left\langle N_1 \right\rangle$$

if desired. Hence, if one can obtain the fluctuating quantities in terms of experimental data then the corresponding integrals over the center of mass based two, three, and four body distribution functions can also be obtained.

The integrals over the spatial distribution functions are valid for any liquid density and are obtained after averaging over the positions (and orientations) of all the other molecules in the system. Hence, they can be used for regions of the phase diagram where many expansions do not usually apply. Furthermore, any orientational effects of the molecules do not appear (directly) in the associated integrals. This means the integrals for molecular systems adopt a much simpler form than observed for many low density expansions. They cannot be used to probe the detailed nature of the interaction energy between molecules, as required in many integral equation theories, but they are valid for both pairwise additive and non-additive potentials.

The fluctuation densities and corresponding integrals provide alternative, but complimentary, descriptions of the correlation between particles within the system. The above integrals and probability distributions are often used to provide insight into the "structure" of liquids and liquid mixtures.[4] It should be noted that, because the distribution functions are defined for the GCE, the corresponding integrals are not those expected for closed systems (where $\rho_\beta G_{\alpha\beta} = -\delta_{\alpha\beta}$), and the distribution functions tend towards unity in an exact manner when all molecules become widely separated.

The pressure dependence of the *G*'s can be obtained by taking derivatives of Equation A.11 and comparing with Equation A.10. One finds,

$$G'_{11} = \beta(G_{111} - 2G_{11}^2)$$
$$G'_{111} = \beta(G_{1111} - 3G_{111}G_{11}) \qquad \text{(A.14)}$$

where the prime indicates an isothermal derivative with respect to pressure (a notation that will be used throughout this article). The first relationship in Equation A.14 is the integrated form of the well-known expression for the pressure dependence of the pair correlation function.[13-14]

In summary, we have provided an extension of the traditional FST approach to investigate the fluctuations and distribution integrals for real solutions using experimental data. In doing so, we are not attempting to provide a low density expansion valid for solutions. Furthermore, no attempt is made to link the results to the underlying pair interactions from which they came. We are providing access to the experimental fluctuations and integrals over distribution functions that characterize the liquid and give rise to the thermodynamic properties at a particular state point, regardless of the density. The relationship between the present approach and existing previous studies of higher distribution functions will be discussed further in Section IV.

A.3 Results

Some of the expressions presented here for pure systems have appeared previously. For instance, the expression for the compressibility given by Equations A.8 and A.11 is well known.[2] It is also known that the pressure dependence of the pair or radial distribution function (rdf) is related to the triplet distribution.[13] However, we have found no relevant quantitative experimental data in the literature concerning the fluctuations (beyond the compressibility) for molecular liquids over a range of pressures and temperatures.

Using the results from Equations A.8 and A.10 the fluctuating quantities for pure gases and liquids can be expressed in terms of pressure derivatives of the density and are given by,

$$B_{11} = \beta^{-1}\rho_1\rho_1'$$

$$C_{111} = \beta^{-2}\rho_1[\rho_1\rho_1'' + (\rho_1')^2] \qquad \text{(A.15)}$$

$$D_{1111} = \beta^{-3}\rho_1[\rho_1^2\rho_1''' + 4\rho_1\rho_1'\rho_1'' + (\rho_1')^3]$$

Expressions for the fluctuations in terms of derivatives of the molar volume or the isothermal compressibility are given in section A.7.2. The above expressions essentially correspond to the familiar Kirkwood-Buff inversion procedure,[15] that provides fluctuating quantities in terms of experimental observables. Once the fluctuating quantities have been obtained, the integrals over the distribution functions can also be extracted using the relationships outlined in Equation A.11. However, to obtain reliable values for the derivatives an accurate equation of state (EOS) is required.

The results obtained for pure water as a function of pressure and temperature are determined here using the IAPWS-95 EOS as implemented in the National Institute of Standards and Technology (NIST) Standard Reference Database 10: NIST/American Society of Mechanical

Engineers Steam Properties Database version 2.22.[16-17] The source code provides a series of thermodynamic properties as a function of pressure (or density) and temperature via a subroutine call. First and second derivatives of the density are provided directly by the EOS. The third derivatives were obtained numerically via a finite difference approach using the second derivatives and a value of $dp = \pm 10^{-20}$ bar. Calculations were performed in quadruple precision.

A.3.1 Density and Pressure Expansions

Before presenting the results for water, it is informative to clarify the uses and exact meaning of the integrals presented here, especially as similar quantities are also found in the literature. In Section II we provided expressions that relate a series of fluctuating quantities to a series of corresponding pressure derivatives. Consequently, one of the most obvious uses for the fluctuations is to rationalize changes in the density as a function of pressure along a particular isotherm. A simple Taylor series expansion provides,

$$\rho_1(p) = \rho_1(p_o) + \rho_1'(p_o)\Delta p + \tfrac{1}{2}\rho_1''(p_o)\Delta p^2 + \tfrac{1}{6}\rho_1'''(p_o)\Delta p^3 + O(\Delta p^4) \qquad \text{(A.16)}$$

in terms of the pressure change $\Delta p = p - p_o$ from a reference pressure p_o. The pressure derivatives appearing in the above expression are given by the fluctuating quantities appearing in Equation A.15,

$$\rho_1' = \frac{\beta B_{11}}{\rho_1}$$

$$\rho_1'' = \frac{\beta^2}{\rho_1^3}\left[\rho_1 C_{111} - B_{11}^2\right] \qquad \text{(A.17)}$$

$$\rho_1''' = \frac{\beta^3}{\rho_1^5}\left[\rho_1^2 D_{1111} - 4\rho_1 C_{111} B_{11} + 3B_{11}^3\right]$$

or via integrals over the distribution function according to,

$$\rho_1' = \beta\left[1+\rho_1 G_{11}\right]$$

$$\rho_1'' = \beta^2\left[G_{11}+\rho_1(G_{111}-G_{11}^2)\right] \tag{A.18}$$

$$\rho_1''' = \beta^3\left[2G_{111}-3G_{11}^2+\rho_1(G_{1111}-4G_{111}G_{11}+3G_{11}^3)\right]$$

The above derivative expressions are valid at any liquid or gas density, but should be evaluated at the reference pressure/density for use in Equation A.16. In the case of gases, where ρ_1 will be very small, the derivatives may simplify further.

Expressions for the equivalent virial coefficients (B_n) can also be obtained from Equation A.16 using a series reversion approach – although to obtain the fourth virial coefficient one requires an additional density derivative. This provides the following expansion,

$$\beta\Delta p = \Delta\rho_1\left[B_1+B_2\Delta\rho_1+B_3\Delta\rho_1^2+\cdots\right] \tag{A.19}$$

for which the B's are given by,

$$B_1(T,p_o) = \left[1+\rho_1 G_{11}\right]^{-1}$$

$$B_2(T,p_o) = -\tfrac{1}{2}\left[1+\rho_1 G_{11}\right]^{-3}\left[G_{11}+\rho_1(G_{111}-G_{11}^2)\right] \tag{A.20}$$

$$B_3(T,p_o) = -\tfrac{1}{6}\left[1+\rho_1 G_{11}\right]^{-5}\begin{bmatrix}2G_{111}-6G_{11}^2+\rho_1(G_{1111}+6G_{11}^3-8G_{111}G_{11})\\+\rho_1^2(G_{1111}G_{11}+2G_{111}G_{11}^2-3G_{111}^2)\end{bmatrix}$$

Again, the virial coefficient expressions in Equation A.20 appear more complicated than the traditional expressions as they are valid for any reference pressure (away from a first order transition). The traditional virial EOS is provided when p_o and ρ_1 are zero. Hence, the more common virial EOS is actually a limiting case of FST. Clearly, the expansion provided in Equation A.16 and A.18 is simpler in form for finite reference pressures (densities).

One can also develop expansions for G_{11} using Equation A.14. This further illustrates how the structure of the liquid or gas changes with pressure or density. The pressure derivatives are then,

$$G'_{11} = \beta(G_{111} - 2G_{11}^2)$$
$$G''_{11} = \beta^2(G_{1111} - 7G_{111}G_{11} + 8G_{11}^3) \tag{A.21}$$

while the corresponding density derivatives are given by,

$$\left(\frac{\partial G_{11}}{\partial \rho_1}\right)_\beta = \left[1 + \rho_1 G_{11}\right]^{-1}\left[G_{111} - 2G_{11}^2\right]$$
$$\left(\frac{\partial^2 G_{11}}{\partial \rho_1^2}\right)_\beta = \left[1 + \rho_1 G_{11}\right]^{-3}\begin{bmatrix} G_{1111} - 8G_{111}G_{11} + 10G_{11}^3 \\ +\rho_1(G_{1111}G_{11} - G_{111}^2 - 4G_{111}G_{11}^2 + 6G_{11}^4) \end{bmatrix} \tag{A.22}$$

where we have used the chain rule to write $\partial G_{11} / \partial \rho_1 = G'_{11} / \rho'_1$, etc.

A.3.2 Gas Phase Fluctuations and Distribution Function Integrals

While our primary focus is the liquid phase – as this has traditionally been the more difficult system to study – a brief discussion of the results for the gas region are in order (the gas phase diagrams containing results up to 1250 K are provided in the Supplemental Materials). The two body fluctuations (B_{11}/ρ_1) generally increase with decreasing temperature and increasing pressure. The skewness of the distributions (C_{111}/ρ_1) is always positive (an excess of particles in the volume is favored over a depletion) and the excess kurtosis (D_{1111}/ρ_1) is always positive (the actual distribution is more peaked than a normal distribution). The magnitude of C_{111}/ρ_1 and D_{1111}/ρ_1 follow the same pressure and temperature trends as B_{11}/ρ_1. The pair distribution integral $\rho_1 G_{11}$ is positive and generally increases with decreasing temperature and increasing pressure, while the triplet distribution integral $\rho_1^2 G_{111}$ is positive over most of the gas phase region investigated here,

but can take on negative values at high pressure and high temperature near the supercritical region. The quadruplet distribution integral $\rho_1^3 G_{1111}$ can adopt either positive (as one approaches the critical point) or negative (for low pressures and higher temperature) values. As expected, the gas phase approaches ideal behavior at low pressure and high temperature, where $B_{11} = C_{111} = D_{1111} = \rho_1$, which corresponds to the known Poisson distribution for the particle number fluctuations.[18] All the G's would be zero for a perfect gas, but are finite for real gases even at low pressures.

Figure A.1 displays the two virial coefficient forms described above as obtained at 298.15 K. In addition, we have included the following related virial coefficients,

$$B_1^* = 1$$

$$B_2^* = -\tfrac{1}{2} G_{11} \qquad \text{(A.23)}$$

$$B_3^* = -\tfrac{1}{3}\left[G_{111} - 3G_{11}^2 \right]$$

These correspond to the traditional (low density) forms of the virial coefficients. However, the integrals used in Equation A.23 can be applied to finite densities and do not form part of an expansion. They are, however, useful as integrals that correspond to particle correlations at a finite density, after averaging over all other molecules in the system, where one can apply the Kirkwood Superposition Approximation to the (orientationally averaged) potential of mean force between the molecules at that particular finite density. The first three virial coefficients decrease with an increase in pressure (or density). This is in agreement with the fact that G_{11} and G_{111} are positive over the pressures studied. The data for B_i* also indicate that the presence of additional water molecules in the gas phase serves to decrease the effective pair and triplet correlations – as would be expected.

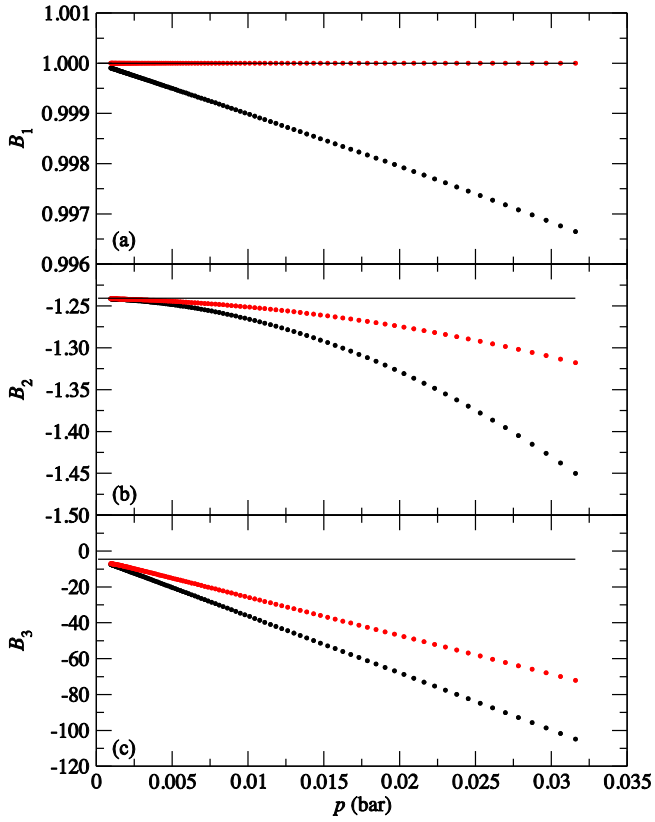


Figure A.1 The (a) first, (b) second, and (c) third virial coefficients for water vapor at 298.15 K for pressures up to the saturation pressure. Black dotted lines: B_i from Equation A.20. Red dotted lines: B_i^* from Equation A.23. Black solid lines: the traditional (zero pressure and density) virial coefficient values provided by the IAPWS-95 EOS. Units for the B_2 coefficients are in M^{-1} while the B_3 coefficients are in M^{-2}.

A.3.3. Liquid Phase Fluctuations and Distribution Function Integrals

The results for liquid water are displayed in Figure A.2 and Figure A.3 as dimensionless quantities. In the Supplemental Materials we also provide Figure A.2 and Figure A.3 as the raw quantities, but they essentially exhibit the same overall trends. Figure A.2 indicates that the fluctuation cumulants alternate in sign for the liquid region. As expected, the two body fluctuations (B_{11}/ρ_1) generally increase with increasing temperature and decreasing pressure. The skewness of the distributions (C_{111}/ρ_1) is always negative (a depletion of particles in the volume is favored over an excess) and the excess kurtosis (D_{1111}/ρ_1) is always positive (the actual distribution is more

peaked than a normal distribution). The underlying distributions tend to a normal distribution (B_{11} is constant, $C_{111} = D_{1111} = 0$) as the pressure increases and the temperature decreases. Figure A.3 indicates that the integrals ($G_{\alpha\beta..}$) alternate in sign and increase in magnitude as the pressure decreases and/or temperature increases. As expected, B_{11}/ρ_1 and $\rho_1 G_{11}$ tend to large positive values as the critical point is approached, as do the values of D_{1111}/ρ_1 and $\rho_1^3 G_{1111}$, whereas the values of C_{111}/ρ_1 and $\rho_1^2 G_{111}$ become very large and negative in this region.

The alternating signs of the fluctuation quantities appears to be an inherent characteristic of liquids and is determined by the expressions found in Equations A.8 and A.10. Hence, one observes $\rho_1' > 0$ as $B_{11} > 0$, $B_{11}' < 0$ as $C_{111} < 0$ and $C_{111}' > 0$ as $D_{1111} > 0$. These patterns indicate that all the fluctuating quantities decrease in magnitude as the pressure increases. The fluctuating quantities and corresponding integrals are displayed in Figure A.4 and Table A.1 for selected isobars, isotherms and state points.

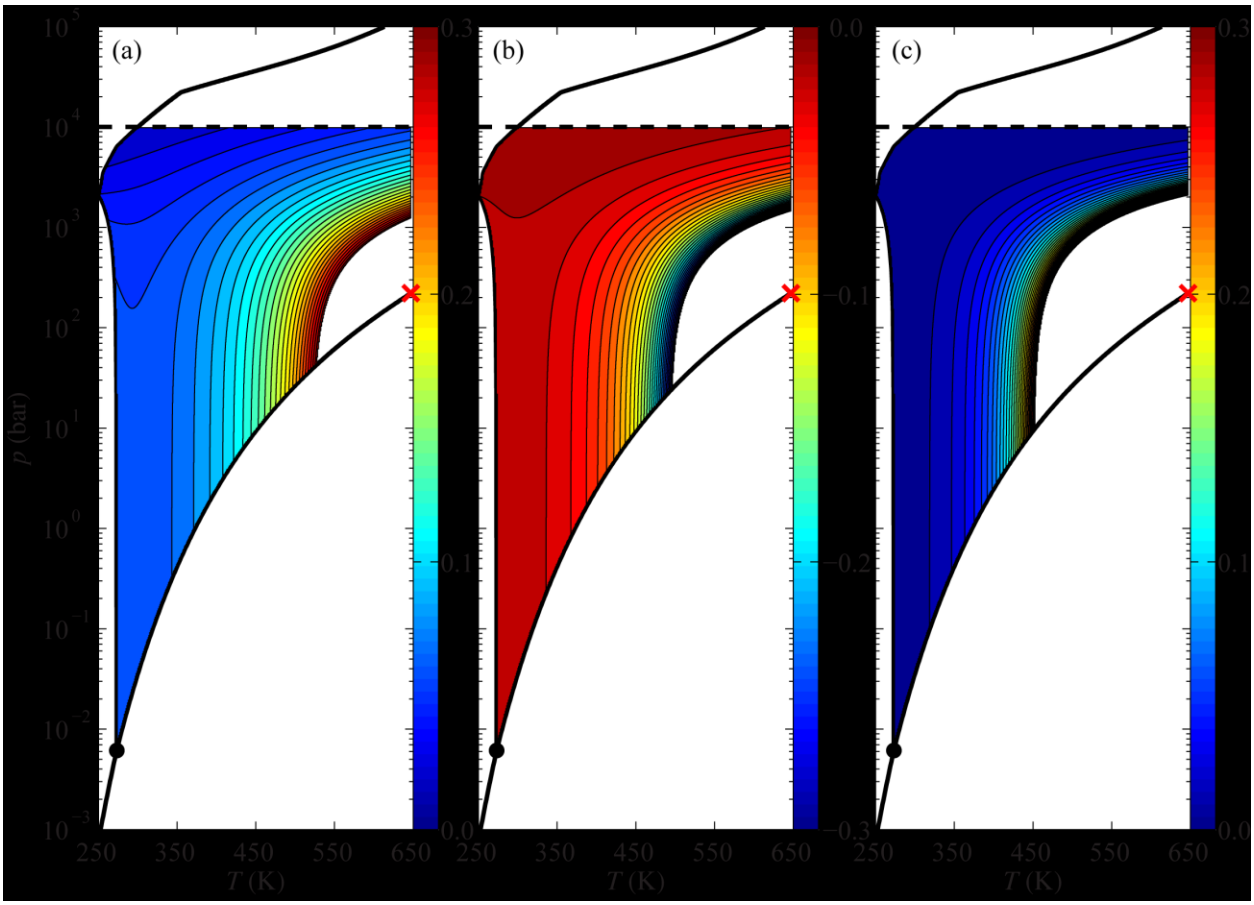


Figure A.2 Liquid phase fluctuation cumulants (a) B_{11}/ρ_1, (b) C_{111}/ρ_1, and (c) D_{1111}/ρ_1. The triple point is indicated by a black dot and the critical point by a red "×." The horizontal dashed line is the maximum valid pressure for the IAPWS-95 Equation of State. Only the liquid phase was contoured. Data outside of the ranges depicted on the color bars were removed, due to the divergence of these properties at the critical point.

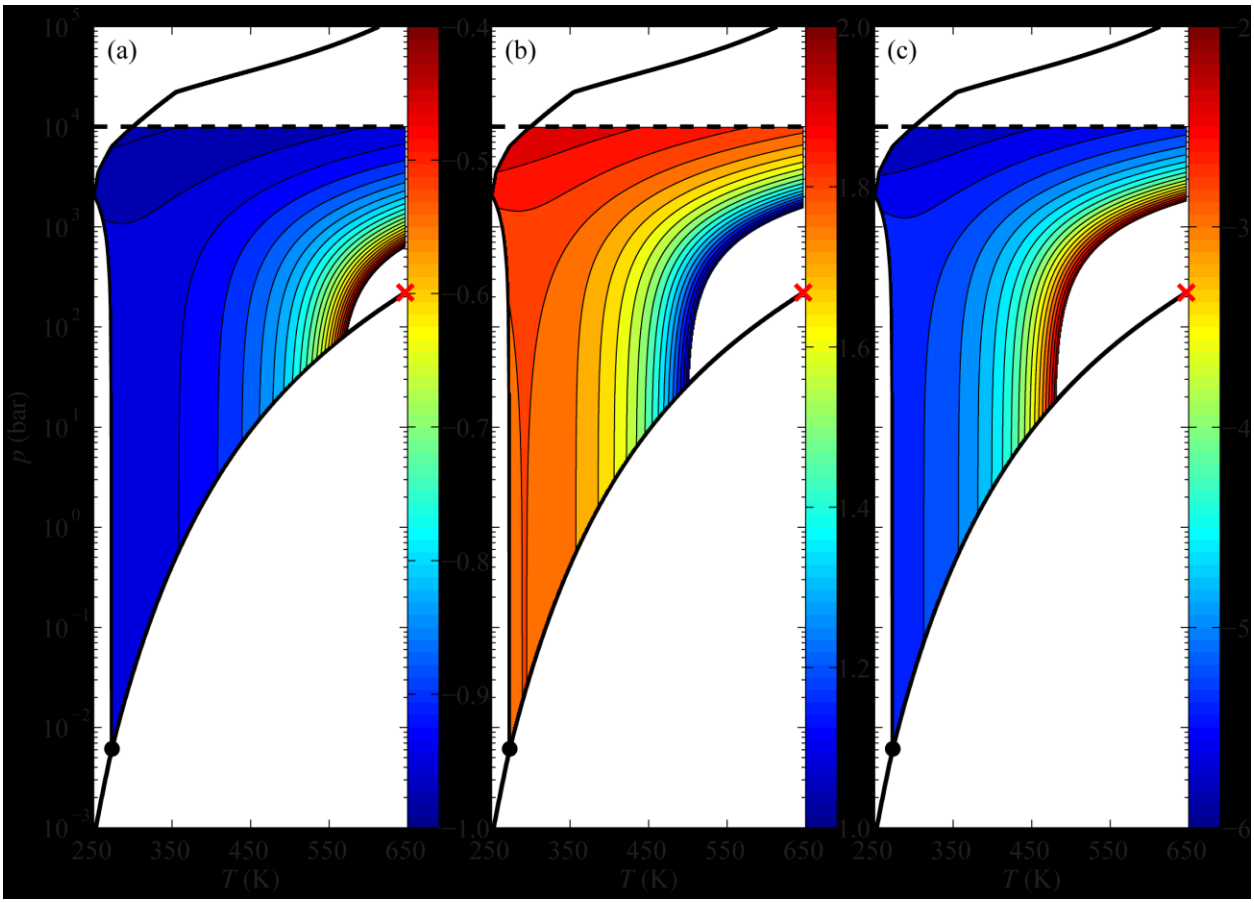


Figure A.3 Liquid phase distribution function integrals (a) $\rho_1 G_{11}$, (b) $\rho_1^2 G_{111}$, and (c) $\rho_1^3 G_{1111}$. The triple point is indicated by a black dot and the critical point by a red "×." The horizontal dashed line is the maximum valid pressure for the IAPWS-95 Equation of State. Only the liquid phase was contoured. Data outside of the ranges depicted on the color bars were removed, due to the divergence of these properties at the critical point.

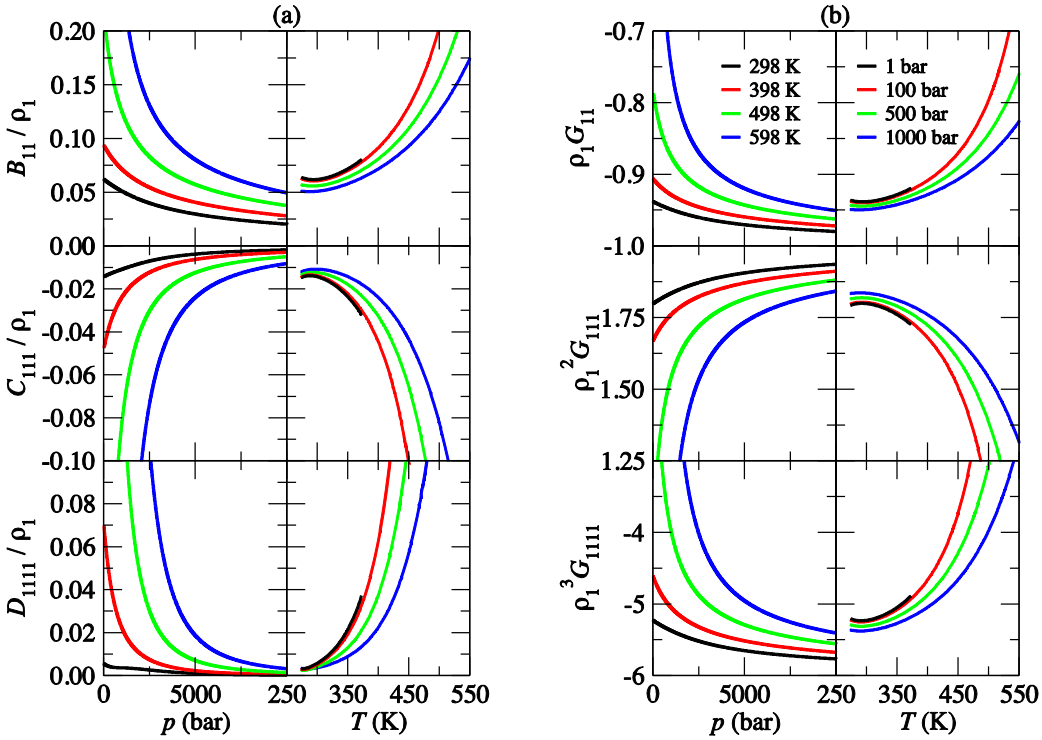


Figure A.4 Liquid phase (a) fluctuation cumulants and (b) distribution function integrals for selected isotherms [left column of panels (a) and (b)] and isobars [right column of panels (a) and (b)].

Table A.1 Fluctuations and integrals for water at various state points.

Property	Ambient $T = 298.15$ K $p = 1$ bar	Triple Point [(l) approach, along T_{tp}]	Triple Point [(g) approach, along p_{tp}]	Triple Point [(g) approach, along T_{tp}]
ρ_1	55.34456	55.49695	2.694692×10^{-4}	2.694697×10^{-4}
B_{11}/ρ_1	0.062076	0.064150	1.0012	1.0012
C_{111}/ρ_1	-0.01419	-0.01483	1.004	1.004
D_{1111}/ρ_1	0.00561	0.00327	1.01	1.01
$\rho_1 G_{11}$	-0.937924	-0.935850	0.001218	0.001218
$\rho_1^2 G_{111}$	1.7996	1.7927	0.000252	0.000252
$\rho_1^3 G_{1111}$	-5.226	-5.202	0.00025	0.00025

According to the IAPWS-95 EOS.

Triple point (tp) values are estimated by approaching the tp from three directions.

Units: ρ_1 is in M and all other properties are dimensionless.

Triple Point: = 273.16 K, 0.00611655 bar, 0.0180190 M^{-1} (liquid), 3,710.98 M^{-1} (vapor)

A.3.4 Moelwyn-Hughes Isotherms

The IAPWS-95 EOS for water, while very accurate, is quite complicated and similar quality expressions for other liquids are relatively few in number. In an effort to provide a more accessible analysis of pure liquids, while maintaining a significant degree of accuracy, we have investigated a simple relationship accredited to Moelwyn-Hughes.[19] The Moelwyn-Hughes isotherm can be developed from the semi-empirical observation that the bulk modulus is proportional to pressure for a variety of substances. Hence,

$$\left(\frac{\partial \kappa_T^{-1}}{\partial p}\right)_T \equiv \mu(T) \qquad \text{(A.24)}$$

where μ is a constant for a fixed temperature. The relationship holds over a reasonable range of temperatures and pressures. The value of μ can also provide details regarding the intermolecular

potential. For instance, it can be shown that μ =1 for ideal gases, μ = 8 for a Lennard-Jones 6-12 potential, and μ = 6-11 for typical real liquids.[19-20]

Assuming the Moelwyn-Hughes isotherm is obeyed one can integrate to obtain an expression for the compressibility as a function of a pressure change,

$$\frac{\kappa_T(p,T)}{\kappa_T(p_o,T)} = \left[1+\Delta p\mu(T)\kappa_T(p_o,T)\right]^{-1} \tag{A.25}$$

and integrate again to obtain the density as a function of a pressure change,

$$\frac{\rho_1(p,T)}{\rho_1(p_o,T)} = \left[1+\Delta p\mu(T)\kappa_T(p_o,T)\right]^{1/\mu(T)} \tag{A.26}$$

A further integration provides the change in chemical potential as a function of a pressure change, but that is not needed in the present study. Here, we use the above expressions to provide the fluctuating quantities and integrals. First, we note that Equation A.8 implies,

$$B_{11}(p,T) = \beta^{-1}\rho_1(p,T)^2\kappa_T(p,T) \tag{A.27}$$

and hence Equation A.25 and Equation A.26 provide the value of B_{11} and G_{11} anywhere along the isotherm. Using the expression provided in Equation A.24 and then comparing with the derivative of Equation A.27 obtained using the expressions given in Equation A.8 and Equation A.10 provides,

$$\begin{aligned} C_{111} &= (2-\mu)\frac{B_{11}^2}{\rho_1} \\ \rho_1^2 G_{111} &= (1-\mu)+(1-2\mu)\rho_1 G_{11}+(2-\mu)\rho_1^2 G_{11}^2 \end{aligned} \tag{A.28}$$

where we have dropped the explicit dependencies on pressure and temperature for clarity. A further pressure derivative, assuming μ is constant, then provides,

$$D_{1111} = (2-\mu)(3-2\mu)\frac{B_{11}^3}{\rho_1^2}$$

$$\rho_1^3 G_{1111} = -(1-\mu)(1+2\mu) + (5-9\mu+6\mu^2)\rho_1 G_{11} + 3(2-\mu)(1-2\mu)\rho_1^2 G_{11}^2 + (2-\mu)(3-2\mu)\rho_1^3 G_{11}^3 \quad \text{(A.29)}$$

Hence, using this approach all the cumulants and integrals over distribution functions can be related to B_{11} and/or G_{11} through a single constant, μ.

The relatively simple forms for the fluctuations shown above allow us to characterize a series of possible situations. First, when μ = 1 we obtain results consistent with the Poisson distribution observed for ideal gases. When μ = 2 the moments describe a Gaussian distribution for the particle number fluctuations where B_{11} is independent of pressure and C_{111} = D_{1111} = 0 – the G's being non zero. An intermediate case where μ = 3/2 would result in $B_{11} > 0$, $C_{111} < 0$, and $D_{1111} = 0$. For real liquids, where μ = 6 – 11,[20] we find that $B_{11} > 0$, $C_{111} < 0$, and $D_{1111} > 0$. This is the pattern observed in Figure A.2. Unfortunately, a clear pattern does not emerge for the G's from the above equations, although one is observed experimentally (see Figure A.3). The optimal value of μ does depend slightly on temperature. Therefore, to account for isothermal changes at temperatures greater ($^+$) or less ($^-$) than T_o = 298.15 K, we have fitted the dependence of μ on temperature according to the relationship,

$$\mu(T^+) = \mu(T_o) + a^+\Delta T + b^+\Delta T^2$$

$$\mu(T^-) = \mu(T_o) + a^-\Delta T + b^-\Delta T^2 + c^-\Delta T^3 \quad \text{(A.30)}$$

where $\Delta T = T - T_o$. The parameters for Equation A.30 were obtained from the IAPWS-95 EOS by first fitting a series of bulk modulus versus pressure ($p_{sat} < p \leq 5$ bar) isotherms using Equation A.25 with p_o = 1 bar, and then fitting the resulting μ(T) values to Equation A.30. The parameters are then μ(T_o) = 5.68, $a^+ = 1.58\times10^{-2}$, $b^+ = 2.1\times10^{-5}$, $a^- = 1.27\times10^{-2}$, $b^- = 7.3\times10^{-5}$, and

$c^{-} = -1.2\times10^{-5}$. The results from this approach are compared to the more exact results provided by the IAPWS-95 EOS in Figure A.5. The results at 298.15 K and 1 bar are in very good agreement with those reported in Table A.1, with the exception of the value for D_{1111}/ρ_1.

Figure A.5 indicates that the distribution function integrals are very well reproduced by the Moelwyn-Hughes approximation over a range of pressures and isotherms, even though the slope of the bulk modulus vs pressure (μ) is changing over this range. The fluctuation cumulants are more problematic. The pair fluctuations are well reproduced, while the triplet fluctuations are well reproduced at low pressures and start to deviate from reality as the pressure increases. The quadruplet fluctuations are poorly reproduced at low pressures, primarily due to the assumption that μ is independent of pressure, but are very well reproduced at higher pressures. All the data is well reproduced for the 318 K isotherm as the value of μ is independent of pressure for this temperature. Interestingly, this isotherm is very close to the compressibility minimum observed for liquid water at relatively low pressures.

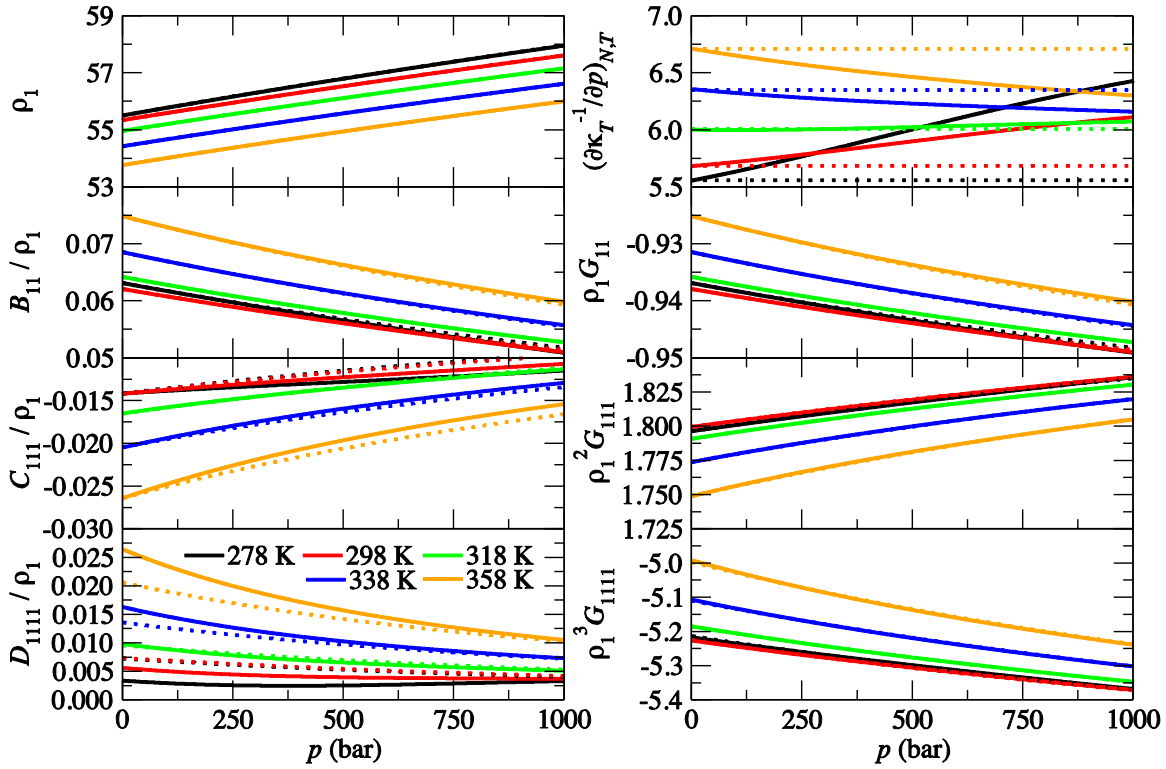


Figure A.5 Properties of liquid water according to the Moelwyn-Hughes isotherms (dotted lines) provided by Equations A.24-A.30 compared to the values given by the IAPWS-95 EOS (solid lines). The density (ρ_1) is displayed in units of *M*.

A.3.5 Linear Density Approximation

A further simplification can be achieved when the density varies linearly with pressure. This is often observed or assumed. For instance, the simulated compressibility of a solution often involves a simple finite difference density calculation.[21] In this case, one can use Equation A.26 with $\mu = 1$. However, linear behavior of any kind (slope) will satisfy Equation A.24 with $\mu = 1$. To distinguish these possibilities, and to simplify the resulting expressions, we define a new positive constant μ_L by,

$$\beta\mu_L(T) \equiv \rho_1(p_o,T)\kappa_T(p_o,T)\mu(T) \qquad \text{(A.31)}$$

such that $\rho_1(p) = \rho_1(p_o) + \beta\mu_L\Delta p$ for linear behavior. Hence, $\mu_L = \mu = 1$ for an ideal gas ($\rho_1 = \beta p$ and $\kappa_T = p^{-1}$ for any p and T), but not necessarily for a liquid where μ_L will typically be much smaller than unity (0.06 for water at 298 K and 1 bar). It is relatively easy to show, using Equation A.15 with $\rho_1'' = \rho_1''' = 0$, that the fluctuating quantities are then given by,

$$\begin{aligned} B_{11} &= \rho_1\mu_L \\ C_{111} &= \rho_1\mu_L^2 \\ D_{1111} &= \rho_1\mu_L^3 \end{aligned} \tag{A.32}$$

which are all positive quantities. The corresponding integrals are provided by,

$$\begin{aligned} \rho_1 G_{11} &= \mu_L - 1 \\ \rho_1^2 G_{111} &= (\mu_L - 1)(\mu_L - 2) \\ \rho_1^3 G_{1111} &= (\mu_L - 1)(\mu_L - 2)(\mu_L - 3) \end{aligned} \tag{A.33}$$

and will alternate in sign if $\mu_L < 1$. Both of these expressions suggest the general relationships,

$$\begin{aligned} \kappa_{n+1} &= \mu_L^n \kappa_1 \\ \rho_1^n G_{1(n+1)} &= \prod_{k=1}^{n} (\mu_L - k) \end{aligned} \tag{A.34}$$

where κ_i is the ith cumulant of the particle number distribution.

Unfortunately, it is immediately clear that this is a poor approximation for the fluctuating quantities. The value of B_{11} is reasonably well reproduced for small deviations from the reference state. However, this provides the wrong sign and magnitude for C_{111} indicating that the higher fluctuations are sensitive probes of the density variations. Thus, a linear density approximation,

while adequate for obtaining the compressibility (B_{11}), is probably insufficient to obtain the higher order fluctuations in most cases.

A.3.6 Temperature Related Effects

The previous discussion has focused on pressure effects at constant T. However, many interesting observations occur as a function of T. FST can also be extended to include derivatives with respect to T.[11,22-24] Temperature effects naturally introduce energy fluctuations. The density maximum observed for water indicates a value of zero for the thermal expansion. FST provides the following expression for the thermal expansion coefficient (α_P) of a pure liquid,[24]

$$T\alpha_P = T\left(\frac{\partial \ln V}{\partial T}\right)_p = -\beta \frac{\langle \delta N_1 \delta \varepsilon \rangle}{\langle N_1 \rangle} = -\beta \frac{B_{1\varepsilon}}{\rho_1} \qquad \text{(A.35)}$$

in terms of the fluctuations in an excess energy $\varepsilon = E - N_1 H_1$, where E is the instantaneous internal energy of the volume of interest and H_1 is the average molar enthalpy of the solution. The subscript ε indicates a substitution of N_1 by ε in the previous expressions for B_{11} (and later C_{111}). When the thermal expansion is zero the following condition must hold,

$$\langle \delta N_1 \delta E \rangle = \langle \delta N_1 \delta N_1 \rangle H_1 \approx \langle \delta N_1 \delta N_1 \rangle U_1 \qquad \text{(A.36)}$$

where the approximation should be reasonable for liquids under ambient conditions. Hence, the density maximum is characterized by the absence of a correlation between the particle number and the internal energy, i.e. $E = N_1 H_1 \approx N_1 U_1$.

The condition for the well-known minimum in the compressibility of water, located at 315±5 K between 1 and 8 bar,[25-30] can also be phrased in terms of fluctuations. This requires the temperature derivatives developed in our previous work and provided in section A.7.3.[11] The simplest result obtained from Equation A.8 is,

$$T\left(\frac{\partial[\rho_1\kappa_T/\beta]}{\partial T}\right)_p = \frac{\beta}{\rho_1^2}\left[\rho_1 C_{11\varepsilon} - B_{11}B_{1\varepsilon}\right] \tag{A.37}$$

Therefore, a minimum in the compressibility term is characterized by,

$$\langle N_1\rangle\langle\delta N_1\delta N_1\delta\varepsilon\rangle = \langle\delta N_1\delta N_1\rangle\langle\delta N_1\delta\varepsilon\rangle \tag{A.38}$$

which indicates that the triplet correlation is then simply related to the corresponding pair correlations. This can also be expressed in terms of density and energy fluctuations by dividing throughout by V^3.

A.3.7. Behavior of the Fluctuations Approaching the Critical Point

As mentioned previously, the fluctuation densities tend to ±∞ at the critical point for this second order transition. The critical point is characterized by the fact that,[31]

$$\frac{\partial p}{\partial V} \to 0 \qquad\qquad \frac{\partial^2 p}{\partial V^2} \to 0 \tag{A.39}$$

Hence, many of the derivatives required in order to obtain the fluctuations from Equation A.15 become very large in this region. The correlation length is therefore very long and the integrals described here also become large. However, pressure varies smoothly as a function of T along the critical isochore. Using the thermodynamic identity,

$$\left(\frac{\partial p}{\partial T}\right)_{\rho_1}\left(\frac{\partial T}{\partial \rho_1}\right)_p\left(\frac{\partial \rho_1}{\partial p}\right)_T = -1$$

$$\left(\frac{\partial p}{\partial T}\right)_{\rho_1} = \frac{\alpha_P}{\kappa_T} = -\frac{\rho_1 B_{1\varepsilon}}{TB_{11}} \tag{A.40}$$

the above expressions indicate that, while the thermal expansion and compressibility diverge as one approaches the critical point, their ratio remains finite. It also indicates that the same behavior

with respect to T or p will be exhibited along the critical isochore. Furthermore, the approach to the critical point along the critical isotherm or isobar will also be the same.

Even though the fluctuations appear to diverge at the critical point, one can still investigate this divergence in terms of the traditional scaling laws and also obtain relationships between the triplet and pair correlations under these circumstances. To do this we examine the behavior of B_{11}, which is closely related to the isothermal compressibility and bulk modulus, and tends to infinity at the critical point. Analysis of the derivative of B_{11} with respect to either T or p provides expressions in terms of both the particle-particle and particle-energy fluctuations (see A.7.3 for the isochoric expressions),

$$\left(\frac{\partial \ln B_{11}}{\partial \ln T}\right)_{\rho_1} = \beta\left[\frac{C_{11\varepsilon}}{B_{11}} - \frac{C_{111}B_{1\varepsilon}}{B_{11}^2}\right]$$

$$\left(\frac{\partial \ln B_{11}}{\partial \ln p}\right)_{T} = \frac{\beta p}{\rho_1}\frac{C_{111}}{B_{11}} \qquad \text{(A.41)}$$

$$\left(\frac{\partial \ln B_{11}}{\partial \ln T}\right)_{p} = \beta\frac{C_{11\varepsilon}}{B_{11}}$$

The expressions found in Equations A.41 are valid anywhere away from a first order phase boundary. These derivatives tend to infinity at the critical point, but they do so in a well-defined manner. To see this we need to examine the critical exponent associated with the limiting behavior of B_{11}. A series of related critical exponents can be defined by,

$$\left(\frac{\partial \ln B_{11}}{\partial \ln|\Delta T|}\right)_{\rho_{1,c},\Delta T\to 0} \equiv \gamma$$

$$\left(\frac{\partial \ln B_{11}}{\partial \ln|\Delta T|}\right)_{p_c,\Delta T\to 0} \equiv \gamma_p \qquad \text{(A.42)}$$

$$\left(\frac{\partial \ln B_{11}}{\partial \ln|\Delta p|}\right)_{T_c,\Delta p\to 0} \equiv \gamma_T$$

where $\Delta T = T - T_c$ and $\Delta p = p - p_c$. We note that γ is the traditional exponent describing the divergence of the compressibility along the critical isochore.[32] Then, we can relate the two sets of derivatives via,

$$\left(\frac{\partial \ln B_{11}}{\partial \ln T}\right)_{\rho_{1,c},\Delta T\to 0} = -\frac{\gamma T}{|\Delta T|}$$

$$\left(\frac{\partial \ln B_{11}}{\partial \ln T}\right)_{p_c,\Delta T\to 0} = -\frac{\gamma_p T}{|\Delta T|} \qquad \text{(A.43)}$$

$$\left(\frac{\partial \ln B_{11}}{\partial \ln p}\right)_{T_c,\Delta p\to 0} = -\frac{\gamma_T p}{|\Delta p|}$$

This strongly suggests that the divergence of the fluctuating quantities is related in a simple manner – as one moves from the pair fluctuations to the triplet and quadruplet, the divergence increases by a factor of ΔT^{-1} or Δp^{-1} each time. Hence, the ratio of C_{111}/B_{11} and D_{1111}/C_{111} quantities diverge in the same manner, as do the ratios of integrals G_{111}/G_{11} and G_{1111}/G_{111}. The following limiting behavior is therefore observed along the critical isochore,

$$G_{11}, B_{11}, B_{1\varepsilon} \propto (p - p_c)^{-\gamma} \propto (T_c - T)^{-\gamma}$$

$$G_{111}, C_{111}, C_{11\varepsilon} \propto (p - p_c)^{-\gamma-1} \propto (T_c - T)^{-\gamma-1} \qquad \text{(A.44)}$$

$$G_{1111}, D_{1111} \propto (p - p_c)^{-\gamma-2} \propto (T_c - T)^{-\gamma-2}$$

while along the critical isotherm or isobar one finds,

$$G_{11}, B_{11}, B_{1\varepsilon} \propto (p - p_c)^{-\gamma_T} \propto (T_c - T)^{-\gamma_p}$$

$$G_{111}, C_{111}, C_{11\varepsilon} \propto (p - p_c)^{-\gamma_T-1} \propto (T_c - T)^{-\gamma_p-1} \qquad \text{(A.45)}$$

$$G_{1111}, D_{1111} \propto (p - p_c)^{-\gamma_T-2} \propto (T_c - T)^{-\gamma_p-2}$$

From the relationship provided in Equation A.40, it is clear that $\gamma_T = \gamma_p$. The value of both constants can be obtained from the Taylor expansion provided in Equation A.16. If we rewrite the expansion for the critical isotherm using $p_o = p_c$ as,

$$\Delta\rho_1 = \rho_1'(p_c)\Delta p\left[1 + \frac{\frac{1}{2}\rho_1''(p_c)\Delta p}{\rho_1'(p_c)} + \frac{\frac{1}{6}\rho_1'''(p_c)\Delta p^2}{\rho_1'(p_c)} + O(\Delta p^3)\right] \qquad \text{(A.46)}$$

Then the relationships provided in Equations A.17 and A.45 indicate that all the terms in the square brackets are finite and constant when approaching the critical point. In fact, the general relationship,

$$\left(\frac{\partial^n \rho_1}{\partial p^n}\right)_\beta \left(\frac{\partial p}{\partial \rho_1}\right)_\beta^{-1} \propto \Delta p^{1-n} \qquad \text{(A.47)}$$

would then hold where the derivatives are dominated by the first term in the square brackets of Equation A.17. It is known that $|\Delta\rho_1|^\delta \propto |\Delta p|$ as one approaches the critical point and so $\rho_1'(p_c) \propto \Delta p^{\delta^{-1}-1}$, and consequently $B_{11} \propto \Delta p^{\delta^{-1}-1}$, or $\gamma_T = \gamma_p = 1 - \delta^{-1}$. The IAPWS-95 EOS state is a classical EOS for water that provides a value of $\gamma = 1$. The IAPWS-95 EOS has nonanalytic

terms and the value of β is set at 0.3,[17] close to the renormalization group theory value (0.326),[32] and hence a value of δ close to the renormalization group theory value (4.8)[32] would be expected. We have examined the limiting behavior of the fluctuations in Table A.2. The results are in agreement with the EOS and a value of $\delta \approx 5$. It should be noted that the analysis of critical exponents provided in Table A.2 does not shed any new light on the experimental data, as these exponents result from the EOS, but they do provide support for the results presented in Equations A.41 and A.43 to A.45. It should also be noted that the uncertainties in the properties generally increase as the critical region is approached and that the IAPWS-95 isothermal compressibility has an unphysical indentation in a region from T_c to T_c + 2 K for densities ±0.5% from ρ_c. Hence, the properties in Table A.2 may not be entirely representative of real experimental data in this region.[17]

The scaling relations illustrated in Equations A.44 and A.45 also suggest quantities for which limiting fluctuation ratio values can be obtained at the critical point and can therefore be used to characterize the distribution. Specifically, these quantities are ratios of particle and/or energy fluctuations of the same order (pair/pair, triplet/triplet, etc.). It should also be noted that the terms preceded by delta functions in Equation S.5 become negligible as we approach the critical point and hence the behavior of B_{11}, C_{111}, and D_{1111} is determined by the behavior of G_{11}, G_{111}, and G_{1111}, respectively. The estimated values for these quantities are provided in Table A.2. Examination of the critical exponents in Table A.2 also suggests that the following ratio of moments should be constant at the critical point,

$$\frac{B_{11}D_{1111}}{C_{111}^2} = \frac{G_{11}G_{1111}}{G_{111}^2} = \frac{\left\langle (\delta\rho)^2 \right\rangle \left[\left\langle (\delta\rho)^4 \right\rangle - 3\left\langle (\delta\rho)^2 \right\rangle^2 \right]}{\left\langle (\delta\rho)^3 \right\rangle^2} \qquad \text{(A.48)}$$

where $\rho = N_1/V$ is an instantaneous density. The above ratio is also the ratio of the fourth standardized central moment to the square of the third standardized central moment.

Table A.2 Critical point behavior

Estimated finite critical point quantities for water

	$\frac{\beta B_{1\varepsilon}}{B_{11}}$	$\frac{C_{111}}{B_{11}^2}$	$\frac{D_{1111}}{B_{11}^3}$	$\frac{\beta C_{11\varepsilon}}{C_{111}}$	$\frac{B_{11}D_{1111}}{C_{111}^2}$
		M^{-1}	M^{-2}		
$\rho_{1,c}$, $\partial p/\partial T$	-1.80				
$\rho_{1,c}$, $T > T_c$	-1.8	0.149	ND	-2.0	ND
T_c, $p > p_c$	-1.7			-1.7	2.2
p_c, $T < T_c$	-1.7			-1.7	2.2

Critical exponents from the IAPWS-95 EOS

	B_{11}^{-1}	$\lvert C_{111}\rvert^{-1}$	D_{1111}^{-1}	$\lvert B_{1\varepsilon}\rvert^{-1}$	$\lvert C_{11\varepsilon}\rvert^{-1}$
$\rho_{1,c}$, $T > T_c$	1.01	2.02	ND	1.00	2.00
T_c, $p > p_c$	0.81	1.81	2.80	0.81	1.83
p_c, $T < T_c$	0.83	1.84	2.83	0.84	1.85

T_c = 647.096 K, p_c = 220.64 bar, $\rho_{1,c}$ = 17.87 M, β_c = 0.186 mol/kJ, and $\beta_c p_c/\rho_{1,c}$ = 0.229.

Values of $\Delta p \approx$ m bar and $\Delta T \approx$ m K were used to determine the critical exponenets.

ND, could not be determined accurately.

A.4 Discussion

The fluctuations investigated here using FST are the same properties that cause radiation to scatter when it impinges on a liquid.[33] The measured distribution of scattered radiation intensity, $I(Q)$, provides information on the distribution of the atomic positions in the liquid.[34-35] Several steps and corrections must be taken to go from $I(Q)$ to the structure factor, $S(Q)$.[36-38] The resulting $S(Q)$ is a sum of weighted averages of $m(m+1)/2$ partial structure factors, $S_{\alpha\beta}(Q)$, where m is the number of distinct atomic species.[36,39] For water there are three partial structure factors, $S_{OO}(Q)$, $S_{OH}(Q)$, and $S_{HH}(Q)$.[36,38] If each $S_{\alpha\beta}(Q)$ can be de-convoluted, the site-site rdfs can be obtained by Fourier transformation.[36,38,40-41] So far, it has only been possible to obtain the complete set of site-site rdfs for a very small set of all the molecules that make up chemical space.[10,37,39,42]

Even knowing $g_{\alpha\beta}^{(2)}$ for a liquid does not "close the book" on its structure.[43] The structure of monatomic liquids is fully defined by the relative probability of finding $n = 1, 2, \ldots N$, of the N molecules in the system at various separation radii.[8,43-50] These relative probabilities are trivial for $n = 1$, $g_{\alpha}^{(1)} = 1$, and obtainable for $n = 2$, $g_{\alpha\beta}^{(2)}$, for monatomic liquids as described above. For molecular systems, a complete description of the structure would additionally require knowledge of the relative probability of finding, triplets, *etc.*, of the molecules' constituent atoms at various separation radii and the angular relative probability distributions that describe the molecular orientations.[39,41,43,51] Experimental studies that provide $g_{\alpha\beta}^{(2)}$ do not provide the complete angular distributions.[41,43,52-53] As a workaround, Soper and coworkers obtain angular distributions from computer simulations.[41,51,54] Indeed, the full set of site-site and angular dependent distribution functions are required for integral equation studies of molecules, but FST does not require this exhaustive level of detail regardless of the type of molecule under study. All that is needed in the

FST approach is the center of mass based, not the site-site and/or the angular dependent, distributions for any type of molecule. From this input information, a thermodynamic and microstructural description of the system can be obtained.

In addition to the sequence of positional distribution functions, the sequence of interatomic potentials is also relevant to this discussion.[7] In most studies, all that is considered is the two-body correlations and/or potential.[13,43,46,50] Largely, this is because there has been no experimental determination of three-body distribution functions for (three-dimensional) fluids.[8,39,42,47-49,55] However, several examples where knowledge of the triplet correlations are important have been discussed by von Grunberg,[47-48] Abascal,[56] Winter,[49] and Rice.[42] Higher-order potential terms are included in some atomic and homonuclear diatomic fluid theories, where accurate pair potentials can be obtained.[13,43,53,57] In contrast, molecular potentials are generally "effective" potentials, meaning that a pair potential is adjusted to reproduce target data to circumvent the need for the correct combinations of pair plus higher potentials.[43]

Attempts to measure higher distribution functions directly, despite the experimental limitations with the conventional scattering methods, have led to interesting video microscopy studies of quasi two-dimensional, colloidal systems.[42,47-48] Additionally, indirect experimental measurements of integrals over the triplet distribution function may be obtained for fluids for which $g_{\alpha\beta}^{(2)}$ is known. However, this approach is not applicable to molecular fluids without making approximations.[8] The isothermal pressure derivative of $S(Q)$ involves the pressure or density derivative of $g_{\alpha\beta}^{(2)}$, which can be written as an integral over $g_{\alpha\beta\gamma}^{(3)}$.[14] For example, the pressure derivative of $g_{11}^{(2)}$ is given by,[8]

$$\rho_1 k_B T \left(\frac{\partial \rho_1^2 g_{11}^{(2)}(r)}{\partial p} \right)_T = \rho_1^3 \int \left[g_{111}^{(3)}(\mathbf{r},\mathbf{s}) - g_{11}^{(2)}(r) \right] d\mathbf{s} + 2\rho_1^2 g_{11}^{(2)}(r) \qquad \text{(A.49)}$$

where **r** and **s** are interatomic vectors connecting atoms at positions 1 and 2 and positions 1 and 3, respectively. The above expression is consistent with Equation A.14. It should be noted that neither $g_{111}^{(3)}$ itself nor an integral over $g_{111}^{(3)}$ by itself are obtained, but instead an integral over $g_{111}^{(3)} - g_{11}^{(2)}$. Nevertheless, the integral has been useful for the few systems for which the rdfs are obtainable. These relationships, first provided by Buff and coworkers and by Schofield,[22,58-61] and have been used extensively by Egelstaff and others to test theories and models for $g_{\alpha\beta\gamma}^{(3)}$.[8-9,13-14,22,49,55,58-62] and are described below.

Our work is similar to the $S(Q\rightarrow 0)$ limit of Egelstaff's (and others') work, which is focused on $S(Q)$ and therefore only technically valid for monatomic liquids.[8] Egelstaff did make very brief mention of this $S(0)$ limit. For example, most similar to our work, he compared the pressure derivative of the bulk modulus for argon at its triple point from experiment with various models for $g^{(3)}$.[13] Similar analysis was performed for gaseous krypton at a state point near its critical point.[9] Ram and Egelstaff assessed the pressure of room temperature krypton versus density.[53,63] They compared experiment and simulation using an accurate pair potential and attributed deviations in the agreement to higher order interactions.[53,63]

Gray, Gubbins, and Egelstaff have derived general expressions for thermodynamic derivatives of properties that are a function of the phase variables averaged over the GCE to derive the thermodynamic derivatives of dynamic correlation functions in a systematic fashion.[14] They discussed using derivatives of radiation scattering functions to study higher-order correlations and to test models or theories of fluids and used examples taken from other studies of the derivative of the Van Hove self-scattering function for hydrogen gas at 85 K and 120 atm and of the pressure derivative of the distinct Van Hove scattering function for nitrogen gas at room temperature and

200 atm.[14] They briefly noted that their equations could also be extended to mixtures and would then be considered generalizations of KB theory.[14]

Several other works are also directly dependent upon scattering data. Buff and Brout[22] and Schofield[58] developed recursion relationships between the density derivatives of correlation functions of all orders and integrals over higher order correlation functions, but no real applications were provided. In a later study, Buff and Brout used the concentration derivative of the rdf for argon at 91.8 K and 2 atm to assess the quality of the superposition approximation and also obtained G_{111} from a similar route to that described here.[59] Egelstaff, Page, and Heard assessed the isothermal pressure derivative of $S(Q)$ for liquid rubidium near its triple point and argon near its critical point, and also assessed the isothermal pressure derivative of the centers-$S(Q)$ for liquid carbon tetrachloride (reliant upon approximations since this is polyatomic) at an intermediate state point at 296 K.[8] They compared their findings to integral equation theory results using various models for $g^{(3)}$.[8] Gläser and coworkers presented a neutron diffraction investigation of the $S(Q)$ of liquid cesium close to its liquid-vapor critical point to obtain the isothermal density derivative of $S(Q)$.[62] Soper and coworkers measured the isothermal density derivative of $S(Q)$ of dense fluid helium by neutron diffraction to test a model for $g^{(3)}$.[55] Finally, Egelstaff[64] additionally assessed the second derivative of $S(Q)$ with respect to pressure for liquid neon at 35 K and $\rho = 1.119$ g/cm^3, which is related to integrals over the quadruplet correlation function, and Ballentine[65] has further considered the long wavelength limit.

Most notably, Gorbaty and coworkers determined the pair correlation function of liquid water at pressures up to 7.7 kbar.[66] No attempt to connect to three body correlations was reported. Soper also measured the rdf in water and ice over a range of temperatures and pressures, but connections to the integrals over the three body correlation function were, again, not reported.[37]

A.5 Conclusions

We have illustrated how triplet and quadruplet fluctuation densities, or integrals over triplet and quadruplet distribution functions, can be obtained from existing experimental data for pure liquids. The results should help to provide a deeper understanding of liquids and liquid mixtures. The only other experimental technique that we know of which provides such data for solutions is that of solution scattering studies.[13] However, these are limited to triplet correlations and to relatively simple pure liquids and very simple mixtures. No such limitation is found with the current approach. It appears that the experimental extraction of these correlations is viable for the triplet distributions, while the quadruplet distributions are somewhat more problematic, but also seem obtainable. Fortunately, while the current approach does not provide the experimental values of $g^{(3)}_{\alpha\beta\gamma}$ or $g^{(4)}_{\alpha\beta\gamma\delta}$ as a function of intermolecular distance (no currently available approach does), the thermodynamics of the solution are directly related to integrals over these distribution functions, which are provided by the current approach. For simple liquids, it has already been shown that this type of integral is useful.[8-9,13-14,22,49,55,58-62] Using FST, integrals over $g^{(3)}_{\alpha\beta\gamma}$ and $g^{(4)}_{\alpha\beta\gamma\delta}$ may now be obtained without the need for scattering experiments, albeit with a loss of spatial resolution. They may be obtained for any liquid (or solution) where the required bulk thermodynamic data has been determined. Thus, we believe the field may now begin to assess how important the $g^{(3)}_{\alpha\beta\gamma}$ and $g^{(4)}_{\alpha\beta\gamma\delta}$ distributions are for describing the thermodynamic properties of any system of interest.

A.6 Supplementary Information

A.6.1 Distribution functions and fluctuation densities

In the grand canonical ensemble the probability that any n_α molecules of species α, and n_β molecules of species β, *etc.* are within $d\{r\}$ at $\{r\}$ is given by $\rho^{(n)}(\{r\})d\{r\}$ where,[1]

$$\int \rho^{(n)}\left(\{r\}\right)d\{r\} = \left\langle \prod_s \frac{N_s!}{\left(N_s - n_s\right)!} \right\rangle \tag{S.1}$$

Here, the product involves the different species (s) present in the solution, while n_s is the number of molecules of each species in the (n) particle distribution. We require integrals up to and including the four body distribution for a general mixture of any number of components. These involve integrals over the following spatial probability density distributions,

$$\begin{aligned}
&\int \rho_\alpha^{(1)}\left(r_1\right)dr_1 = \left\langle N_\alpha \right\rangle \\
&\int \rho_{\alpha\beta}^{(2)}\left(r_1, r_2\right)dr_1 dr_2 = \left\langle N_\alpha (N_\beta - \delta_{\alpha\beta}) \right\rangle \\
&\int \rho_{\alpha\beta\gamma}^{(3)}\left(r_1, r_2, r_3\right)dr_1 dr_2 dr_3 = \left\langle N_\alpha (N_\beta - \delta_{\alpha\beta})(N_\gamma - \delta_{\alpha\gamma} - \delta_{\beta\gamma}) \right\rangle \\
&\int \rho_{\alpha\beta\gamma\delta}^{(4)}\left(r_1, r_2, r_3, r_4\right)dr_1 dr_2 dr_3 dr_4 = \left\langle N_\alpha (N_\beta - \delta_{\alpha\beta})(N_\gamma - \delta_{\alpha\gamma} - \delta_{\beta\gamma})(N_\delta - \delta_{\alpha\delta} - \delta_{\beta\delta} - \delta_{\gamma\delta}) \right\rangle
\end{aligned} \tag{S.2}$$

By analogy with the theory of imperfect gases we define the following integrals,[1-2]

$$\rho_\alpha \rho_\beta VG_{\alpha\beta} \equiv \int \left[\rho_{\alpha\beta}^{(2)} - \rho_\alpha^{(1)} \rho_\beta^{(1)} \right] dr_1 dr_2$$

$$\rho_\alpha \rho_\beta \rho_\gamma VG_{\alpha\beta\gamma} \equiv \int \left[\rho_{\alpha\beta\gamma}^{(3)} - \rho_{\alpha\beta}^{(2)} \rho_\gamma^{(1)} - \rho_{\alpha\gamma}^{(2)} \rho_\beta^{(1)} - \rho_{\beta\gamma}^{(2)} \rho_\alpha^{(1)} + 2\rho_\alpha^{(1)} \rho_\beta^{(1)} \rho_\gamma^{(1)} \right] dr_1 dr_2 dr_3$$

$$\rho_\alpha \rho_\beta \rho_\gamma \rho_\delta VG_{\alpha\beta\gamma\delta} \equiv \int \begin{bmatrix} \rho_{\alpha\beta\gamma\delta}^{(4)} - \rho_{\alpha\beta\gamma}^{(3)} \rho_\delta^{(1)} - \rho_{\alpha\beta\delta}^{(3)} \rho_\gamma^{(1)} - \rho_{\alpha\gamma\delta}^{(3)} \rho_\beta^{(1)} - \rho_{\beta\gamma\delta}^{(3)} \rho_\alpha^{(1)} \\ -\rho_{\alpha\beta}^{(2)} \rho_{\gamma\delta}^{(2)} - \rho_{\alpha\gamma}^{(2)} \rho_{\beta\delta}^{(2)} - \rho_{\alpha\delta}^{(2)} \rho_{\beta\gamma}^{(2)} \\ +2\rho_{\alpha\beta}^{(2)} \rho_\gamma^{(1)} \rho_\delta^{(1)} + 2\rho_{\alpha\gamma}^{(2)} \rho_\gamma^{(1)} \rho_\delta^{(1)} + 2\rho_{\alpha\delta}^{(2)} \rho_\beta^{(1)} \rho_\gamma^{(1)} \\ +2\rho_{\beta\gamma}^{(2)} \rho_\alpha^{(1)} \rho_\delta^{(1)} + 2\rho_{\beta\delta}^{(2)} \rho_\alpha^{(1)} \rho_\gamma^{(1)} + 2\rho_{\gamma\delta}^{(2)} \rho_\alpha^{(1)} \rho_\beta^{(1)} \\ -6\rho_\alpha^{(1)} \rho_\beta^{(1)} \rho_\gamma^{(1)} \rho_\delta^{(1)} \end{bmatrix} dr_1 dr_2 dr_3 dr_4 \qquad \text{(S.3)}$$

where we have removed the spatial dependencies for clarity. Similar integrals arise during the expansion of the partition function (see Equation A.1) employed in gas and solution theory. The first two integrals also appeared in the original KB paper.[2] The above integrals can be expressed in terms of spatial probability distributions via their definitions,

$$g_\alpha^{(1)}(r_1) \equiv \frac{\rho_\alpha^{(1)}(r_1)}{\rho_\alpha} = 1 \qquad g_{\alpha\beta}^{(2)}(r_1, r_2) \equiv \frac{\rho_{\alpha\beta}^{(2)}(r_1, r_2)}{\rho_\alpha \rho_\beta}$$

$$g_{\alpha\beta\gamma}^{(3)}(r_1, r_2, r_3) \equiv \frac{\rho_{\alpha\beta\gamma}^{(3)}(r_1, r_2, r_3)}{\rho_\alpha \rho_\beta \rho_\gamma} \qquad g_{\alpha\beta\gamma\delta}^{(4)}(r_1, r_2, r_3, r_4) \equiv \frac{\rho_{\alpha\beta\gamma\delta}^{(4)}(r_1, r_2, r_3, r_4)}{\rho_\alpha \rho_\beta \rho_\gamma \rho_\delta} \qquad \text{(S.4)}$$

A combination of Equations S.3 and S.4, followed by some minor rearrangement, provides the integrals given in the main text as Equation A.12. Furthermore, a combination of Equations S.2 and S.3 provides the *G*'s in terms of fluctuating quantities,

$$\rho_\alpha \rho_\beta VG_{\alpha\beta} = \left\langle \delta N_\alpha \delta N_\beta \right\rangle - \delta_{\alpha\beta} \left\langle N_\alpha \right\rangle$$

$$\begin{aligned}\rho_\alpha \rho_\beta \rho_\gamma VG_{\alpha\beta\gamma} = &\left\langle \delta N_\alpha \delta N_\beta \delta N_\gamma \right\rangle - \delta_{\beta\gamma} \left\langle \delta N_\alpha \delta N_\beta \right\rangle - \delta_{\alpha\beta} \left\langle \delta N_\alpha \delta N_\gamma \right\rangle \\ &- \delta_{\alpha\gamma} \left\langle \delta N_\beta \delta N_\gamma \right\rangle + 2\delta_{\alpha\beta}\delta_{\alpha\gamma} \left\langle N_\alpha \right\rangle\end{aligned}$$

$$\begin{aligned}\rho_\alpha \rho_\beta \rho_\gamma \rho_\delta VG_{\alpha\beta\gamma\delta} = &\left\langle \delta N_\alpha \delta N_\beta \delta N_\gamma \delta N_\delta \right\rangle - \left\langle \delta N_\alpha \delta N_\beta \right\rangle \left\langle \delta N_\gamma \delta N_\delta \right\rangle \\ &- \left\langle \delta N_\alpha \delta N_\gamma \right\rangle \left\langle \delta N_\beta \delta N_\delta \right\rangle - \left\langle \delta N_\alpha \delta N_\delta \right\rangle \left\langle \delta N_\beta \delta N_\gamma \right\rangle \\ &- (\delta_{\alpha\delta} + \delta_{\gamma\delta}) \left\langle \delta N_\alpha \delta N_\beta \delta N_\gamma \right\rangle - (\delta_{\alpha\gamma} + \delta_{\beta\gamma}) \left\langle \delta N_\alpha \delta N_\beta \delta N_\delta \right\rangle \\ &- (\delta_{\alpha\beta} + \delta_{\beta\delta}) \left\langle \delta N_\alpha \delta N_\gamma \delta N_\delta \right\rangle \\ &+ (\delta_{\alpha\gamma}\delta_{\alpha\delta} + \delta_{\alpha\delta}\delta_{\gamma\delta} + \delta_{\beta\delta}\delta_{\gamma\delta} + \delta_{\beta\gamma}\delta_{\gamma\delta}) \left\langle \delta N_\alpha \delta N_\beta \right\rangle \\ &+ (\delta_{\alpha\beta}\delta_{\alpha\gamma} + \delta_{\alpha\beta}\delta_{\gamma\delta} + \delta_{\alpha\gamma}\delta_{\beta\gamma} + \delta_{\alpha\gamma}\delta_{\beta\delta}) \left\langle \delta N_\alpha \delta N_\delta \right\rangle \\ &+ (\delta_{\alpha\beta}\delta_{\alpha\delta} + \delta_{\alpha\delta}\delta_{\beta\gamma} + \delta_{\alpha\delta}\delta_{\beta\delta}) \left\langle \delta N_\alpha \delta N_\gamma \right\rangle - 6\delta_{\alpha\beta}\delta_{\alpha\gamma}\delta_{\alpha\delta} \left\langle N_\alpha \right\rangle\end{aligned} \tag{S.5}$$

These expressions can then be rearranged to provide the equivalent fluctuating quantities given in the main text. It should be noted that when all the particles are of a different type (all delta functions are zero) then the integrals are simply the cumulants of the multivariate particle distribution.

A.6.2 Fluctuating quantities and distribution function integrals from experimental data

The fluctuating quantities were expressed in terms of density derivatives in the main text. They can also be expressed in terms of derivatives of the molar volume ($V_1 = 1 / \rho_1$) and are then given by,

$$B_{11} = -\frac{1}{\beta V_1^3} V_1'$$

$$C_{111} = -\frac{1}{\beta^2 V_1^5} [V_1 V_1'' - 3(V_1')^2] \tag{S.6}$$

$$D_{1111} = -\frac{1}{\beta^3 V_1^7} [V_1^2 V_1''' - 10 V_1 V_1' V_1'' + 15(V_1')^3]$$

where the prime again indicates an isothermal derivative with respect to pressure. The corresponding fluctuations can also be expressed in terms of pressure derivatives of the compressibility and are then given by the expressions,

$$B_{11} = \beta^{-1}\rho_1^2\kappa_T$$

$$C_{111} = \beta^{-2}\rho_1^3[\kappa_T' + 2\kappa_T^2] \tag{S.7}$$

$$D_{1111} = \beta^{-3}\rho_1^4[\kappa_T'' + 7\kappa_T\kappa_T' + 6\kappa_T^3]$$

The most convenient choice will depend on the EOS used to fit the experimental or simulated data.

A.6.3 Energy fluctuations

Here, we provide expressions for the triplet fluctuations involving the excess energy of the region of interest. Previously we have shown that,[11,67]

$$C_{11\varepsilon} = -\left(\frac{\partial B_{11}}{\partial \beta}\right)_p \tag{S.8}$$

which is also evident from Equation A.6. Using the expression for B_{11} provided in Equation A.15 allows one to derive an expression for the particle-particle-excess energy triplet fluctuation density in terms of experimental data for pure components. The result is,

$$\frac{\beta C_{11\varepsilon}}{\rho_1} = \frac{\rho_1\kappa_T}{\beta}(1 - T\alpha_P) + \frac{T}{\beta}\left(\frac{\partial^2\rho_1}{\partial p\partial T}\right) \tag{S.9}$$

whereas the FST expression for the second density derivative is,

$$\frac{T}{\beta}\left(\frac{\partial^2\rho_1}{\partial p\partial T}\right) = \frac{\beta C_{11\varepsilon}}{\rho_1} - \frac{B_{11}}{\rho_1}\left[1 + \frac{\beta B_{1\varepsilon}}{\rho_1}\right] \tag{S.10}$$

These expressions are used to obtain $C_{11\varepsilon}$ from the EOS.

The same type of approach can also be used to provide derivatives along a particular isochore. From Equation A.6 with $\langle X \rangle = B_{11}$ we find,

$$\left(\frac{\partial B_{11}}{\partial \beta}\right)_{\rho_1} = -C_{11E} + C_{111}\left(\frac{\partial \beta\mu_1}{\partial \beta}\right)_{\rho_1} \tag{S.11}$$

Using the fact that,

$$\left(\frac{\partial \beta\mu_1}{\partial \beta}\right)_{\rho_1} = \left(\frac{\partial \beta\mu_1}{\partial \beta}\right)_{p} + \left(\frac{\partial \beta\mu_1}{\partial p}\right)_{\beta}\left(\frac{\partial p}{\partial \beta}\right)_{\rho_1} = H_1 - \frac{T\alpha_P}{\rho_1 \kappa_T} \tag{S.12}$$

where we have also used Equation A.40 leads to,

$$\left(\frac{\partial B_{11}}{\partial \beta}\right)_{\rho_1} = -C_{11\varepsilon} + C_{111}\frac{B_{1\varepsilon}}{B_{11}} \tag{S.13}$$

which provides the first expression in Equation A.41.

A.7 References

1. T. L. Hill, *Statistical Mechanics: Principles and Selected Applications*. McGraw-Hill Book Company, Inc.: New York (1956).
2. J. G. Kirkwood and F. P. Buff, *J. Chem. Phys.* **19**, 774 (1951).
3. A. Ben-Naim, In *Fluctuation Theory of Solutions: Applications in Chemistry, Chemical Engineering and Biophysics*, P. E. Smith; E. Matteoli; J. P. O' Connell, Eds. Taylor & Francis: Boca Raton (2013) pp 35-63.
4. A. Ben-Naim, *Molecular Theory of Solutions*. Oxford University Press: New York (2006).
5. P. E. Smith, E. Matteoli, and J. P. O' Connell, *Fluctuation Theory of Solutions: Applications in Chemistry, Chemical Engineering and Biophysics*. Taylor & Francis: Boca Raton (2013).
6. C. G. Gray and K. E. Gubbins, *Theory of Molecular Fluids. Vol. 1: Fundamentals*. Oxford University Press: New York (1984).
7. P. A. Egelstaff, *J. Phys. (Paris)* **46**, 1 (1985).
8. P. A. Egelstaff, D. I. Page, and C. R. T. Heard, *J. Phys. C: Solid State Phys.* **4**, 1453 (1971).
9. D. J. Winfield and P. A. Egelstaff, *Can. J. Phys.* **51**, 1965 (1973).
10. P. A. Egelstaff, In *Molecular Liquids: New Perspectives in Physics and Chemistry*, J. J. C. Teixeira-Dias, Ed. Springer: Netherlands **Vol. 379** (1992) pp 29-44.
11. E. A. Ploetz and P. E. Smith, In *Adv. Chem. Phys.*, John Wiley & Sons, Inc.: (2013) pp 311-372.
12. R. F. Greene and H. B. Callen, *Phys. Rev.* **83**, 1231 (1951).
13. P. A. Egelstaff, *Annu. Rev. Phys. Chem.* **24**, 159 (1973).
14. K. E. Gubbins, C. G. Gray, and P. A. Egelstaff, *Mol. Phys.* **35**, 315 (1978).
15. A. Ben-Naim, *J. Chem. Phys.* **67**, 4884 (1977).
16. A. H. Harvey, A. P. Peskin, and S. A. Klein *NIST Standard Reference Database 10: NIST/ASME Steam Properties*, Version 2.22; U.S. Department of Commerce: Gaithersburg (2008).
17. W. Wagner and A. Pruss, *J. Phys. Chem. Ref. Data* **31**, 387 (2002).

18. L. D. Landau and E. M. Lifshitz, *Statistical Physics, Part I*. 3rd ed.; Pergamon Press: Oxford. Translated from the Russian by J. B. Sykes and M. J. Kearsley **Vol. 5 of Course of Theoretical Physics** (1980).
19. E. A. Moelwyn-Hughes, *Physical Chemistry. 2nd Ed*. Pergamon Press: Oxford (1961).
20. D. P. Kharakoz, *Biophys. J.* **79**, 511 (2000).
21. H. B. Yu and W. F. van Gunsteren, *J. Chem. Phys.* **121**, 9549 (2004).
22. F. P. Buff and R. Brout, *J. Chem. Phys.* **23**, 458 (1955).
23. P. G. Debenedetti, *J. Chem. Phys.* **86**, 7126 (1987).
24. E. A. Ploetz and P. E. Smith, *J. Chem. Phys.* **135**, 044506 (2011).
25. H. L. Pi, J. L. Aragones, C. Vega, E. G. Noya, J. L. F. Abascal, M. A. Gonzalez, and C. McBride, *Mol. Phys.* **107**, 365 (2009).
26. G. S. Kell, *J. Chem. Eng. Data* **20**, 97 (1975).
27. R. J. Speedy and C. A. Angell, *J. Chem. Phys.* **65**, 851 (1976).
28. T. S. Carlton, *J. Phys. Chem. B* **111**, 13398 (2007).
29. F. Mallamace, C. Corsaro, and H. E. Stanley, *Sci. Rep.* **2**, 993 (2012).
30. M. Vedamuthu, S. Singh, and G. W. Robinson, *J. Phys. Chem.* **99**, 9263 (1995).
31. A. Munster, *Classical Thermodynamics*. Stonebridge Press: Bristol (1970).
32. W. Wagner, (Gibbs Award Lecture) From the Beginning to this Day: My First Naive Ideas and the Research Results Achieved. In *15th International Conference on the Properties of Water and Steam*, R. Span; I. Weber, Eds. VDI - The Association of German Engineers and GET - Society for Energy Technology: Berlin, (2008).
33. P. A. Egelstaff, *Brookhaven Symposia in Biology*, 126 (1976).
34. G. Allen and J. S. Higgins, *Rep. Prog. Phys.* **36**, 1073 (1973).
35. P. A. Egelstaff, In *Molecular Liquids: New Perspectives in Physics and Chemistry*, J. J. C. Teixeira-Dias, Ed. Springer: Netherlands **Vol. 379** (1992) pp 1-27.
36. B. Tomberli, C. J. Benmore, P. A. Egelstaff, J. Neuefeind, and V. Honkimaki, *J. Phys.: Condens. Matter* **12**, 2597 (2000).
37. A. K. Soper, *Chem. Phys.* **258**, 121 (2000).
38. P. A. Egelstaff, *Phys. Chem. Liq.* **40**, 203 (2002).
39. G. W. Neilson and A. K. Adya, *Annu. Rep. Prog. Chem. Sect. C: Phys. Chem.* **93**, 101 (1997).

40. P. A. Egelstaff, In *Methods in the Determination of Partial Structure Factors of Disordered Matter by Neutron and Anomalous X-ray Diffraction*, J. B. Suck; P. Chieux; D. Raoux; C. Riekel, Eds. World Scientific Publishing Co: Singapore (Singapore) (1993) pp 1-15.
41. A. K. Soper, *J. Chem. Phys.* **101**, 6888 (1994).
42. H. M. Ho, B. H. Lin, and S. A. Rice, *J. Chem. Phys.* **125**, 184715 (2006).
43. P. A. Egelstaff, *Adv. Chem. Phys.* **53**, 1 (1983).
44. P. Linse, *J. Chem. Phys.* **94**, 8227 (1991).
45. W. J. McNeil, W. G. Madden, A. D. J. Haymet, and S. A. Rice, *J. Chem. Phys.* **78**, 388 (1983).
46. J. A. Krumhansl and S. S. Wang, *J. Chem. Phys.* **56**, 2034 (1972).
47. K. Zahn, G. Maret, C. Russ, and H. H. von Grunberg, *Phys. Rev. Lett.* **91**, 115502 (2003).
48. C. Russ, K. Zahn, and H.-H. von Grunberg, *J. Phys.: Condens. Matter* **15**, S3509 (2003).
49. M. A. Schroer, M. Tolan, and R. Winter, *Phys. Chem. Chem. Phys.* **14**, 9486 (2012).
50. P. A. Egelstaff, *Pure Appl. Chem.* **51**, 2131 (1979).
51. A. K. Soper, *J. Mol. Liq.* **78**, 179 (1998).
52. L. Van Hove, *Phys. Rev.* **95**, 249 (1954).
53. P. A. Egelstaff, *Methods Exp. Phys.* **23**, 405 (1987).
54. A. K. Soper and C. J. Benmore, *Phys. Rev. Lett.* **101**, 065502 (2008).
55. W. Montfrooij, L. A. de Graaf, P. J. van den Bosch, A. K. Soper, and W. S. Howells, *J. Phys.: Condens. Matter* **3**, 4089 (1991).
56. S. Jorge, E. Lomba, and J. L. F. Abascal, *J. Chem. Phys.* **116**, 730 (2002).
57. J. A. Barker, R. A. Fisher, and R. O. Watts, *Mol. Phys.* **21**, 657 (1971).
58. P. Schofield, *Proc. Phys. Soc.* **88**, 149 (1966).
59. F. P. Buff and R. Brout, *J. Chem. Phys.* **33**, 1417 (1960).
60. F. P. Buff and F. M. Schindler, *J. Chem. Phys.* **29**, 1075 (1958).
61. F. P. Buff, *J. Chem. Phys.* **23**, 419 (1955).
62. R. Winter, F. Hensel, T. Bodensteiner, and W. Glaser, *J. Phys. Chem.* **92**, 7171 (1988).
63. J. Ram and P. A. Egelstaff, *Phys. Chem. Liq.* **14**, 29 (1984).
64. P. A. Egelstaff and S. S. Wang, *Can. J. Phys.* **50**, 684 (1972).
65. L. E. Ballentine and A. Lakshmi, *Can. J. Phys.* **53**, 372 (1975).
66. A. V. Okhulkov, Y. N. Demianets, and Y. E. Gorbaty, *J. Chem. Phys.* **100**, 1578 (1994).

67. Y. F. Jiao and P. E. Smith, *J. Chem. Phys.* **135**, 014502 (2011).

Appendix B - Copyright Clearance

RightsLink®

Home | Account Info | Help

Logged in as:
Sadish Karunaweera
LOGOUT

Title: Theory and Simulation of Multicomponent Osmotic Systems

Author: Sadish Karunaweera, Moon Bae Gee, Samantha Weerasinghe, et al

Publication: Journal of Chemical Theory and Computation

Publisher: American Chemical Society

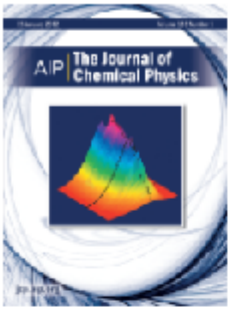
Date: Oct 1, 2012

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms & Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from (COMPLETE REFERENCE CITATION). Copyright (YEAR) American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your request. No additional uses are granted (such as derivative works or other editions). For any other uses, please submit a new request.

BACK | CLOSE WINDOW



Title:	Experimental triplet and quadruplet fluctuation densities and spatial distribution function integrals for pure liquids
Author:	Elizabeth A. Ploetz,Sadish Karunaweera,Paul E. Smith
Publication:	Journal of Chemical Physics
Volume/Issue	142/4
Publisher:	AIP Publishing LLC
Date:	Jan 26, 2015
Page Count:	14

Order Completed

Thank you for your order.

This Agreement between Sadish Karunaweera ("You") and AIP Publishing LLC ("AIP Publishing LLC") consists of your license details and the terms and conditions provided by AIP Publishing LLC and Copyright Clearance Center.

Your confirmation email will contain your order number for future reference.

Printable details.

License Number	4000771109414
License date	Dec 02, 2016
Licensed Content Publisher	AIP Publishing LLC
Licensed Content Publication	Journal of Chemical Physics
Licensed Content Title	Experimental triplet and quadruplet fluctuation densities and spatial distribution function integrals for pure liquids
Licensed Content Author	Elizabeth A. Ploetz,Sadish Karunaweera,Paul E. Smith
Licensed Content Date	Jan 26, 2015
Licensed Content Volume	142
Licensed Content Issue	4
Requestor type	Author (original article)
Format	Print and electronic
Portion	Excerpt (> 800 words)
Requestor Location	Sadish Karunaweera 2020 Tunstall Circle Apt S-07 MANHATTAN, KS 66502 United States Attn: Sadish Karunaweera
Billing Type	Invoice
Billing address	Sadish Karunaweera 2020 Tunstall Circle Apt S-07 MANHATTAN, KS 66502 United States Attn: Sadish Karunaweera
Total	0.00 USD

ORDER MORE | CLOSE WINDOW

AIP PUBLISHING LLC LICENSE TERMS AND CONDITIONS

Dec 02, 2016

This Agreement between Sadish Karunaweera ("You") and AIP Publishing LLC ("AIP Publishing LLC") consists of your license details and the terms and conditions provided by AIP Publishing LLC and Copyright Clearance Center.

License Number	4000771109414
License date	Dec 02, 2016
Licensed Content Publisher	AIP Publishing LLC
Licensed Content Publication	Journal of Chemical Physics
Licensed Content Title	Experimental triplet and quadruplet fluctuation densities and spatial distribution function integrals for pure liquids
Licensed Content Author	Elizabeth A. Ploetz,Sadish Karunaweera,Paul E. Smith
Licensed Content Date	Jan 26, 2015
Licensed Content Volume Number	142
Licensed Content Issue Number	4
Type of Use	Thesis/Dissertation
Requestor type	Author (original article)
Format	Print and electronic
Portion	Excerpt (> 800 words)
Will you be translating?	No
Title of your thesis / dissertation	Theory and Simulation of Intermolecular Interactions in Biological Systems
Expected completion date	Dec 2016
Estimated size (number of pages)	200
Requestor Location	Sadish Karunaweera 2020 Tunstall Circle Apt S-07 MANHATTAN, KS 66502 United States Attn: Sadish Karunaweera
Billing Type	Invoice
Billing Address	Sadish Karunaweera 2020 Tunstall Circle Apt S-07 MANHATTAN, KS 66502 United States Attn: Sadish Karunaweera
Total	0.00 USD
Terms and Conditions	

AIP Publishing LLC -- Terms and Conditions: Permissions Uses

AIP Publishing hereby grants to you the non-exclusive right and license to use and/or distribute the Material according to the use specified in your order, on a one-time basis, for the specified term, with a maximum distribution equal to the number that you have ordered. Any links or other content accompanying the Material are not the subject of this license.

1. You agree to include the following copyright and permission notice with the reproduction of the Material:"Reprinted from [FULL CITATION], with the permission of AIP Publishing." For an article, the credit line and permission notice must be printed on the first page of the article or book chapter. For photographs, covers, or tables, the notice may appear with the Material, in a footnote, or in the reference list.
2. If you have licensed reuse of a figure, photograph, cover, or table, it is your responsibility to ensure that the material is original to AIP Publishing and does not contain the copyright of another entity, and that the copyright notice of the figure, photograph, cover, or table does not indicate that it was reprinted by AIP Publishing, with permission, from another source. Under no circumstances does AIP Publishing purport or intend to grant permission to reuse material to which it does not hold appropriate rights.
   You may not alter or modify the Material in any manner. You may translate the Material into another language only if you have licensed translation rights. You may not use the Material for promotional purposes.
3. The foregoing license shall not take effect unless and until AIP Publishing or its agent, Copyright Clearance Center, receives the Payment in accordance with Copyright Clearance Center Billing and Payment Terms and Conditions, which are incorporated herein by reference.
4. AIP Publishing or Copyright Clearance Center may, within two business days of granting this license, revoke the license for any reason whatsoever, with a full refund payable to you. Should you violate the terms of this license at any time, AIP Publishing, or Copyright Clearance Center may revoke the license with no refund to you. Notice of such revocation will be made using the contact information provided by you. Failure to receive such notice will not nullify the revocation.
5. AIP Publishing makes no representations or warranties with respect to the Material. You agree to indemnify and hold harmless AIP Publishing, and their officers, directors, employees or agents from and against any and all claims arising out of your use of the Material other than as specifically authorized herein.
6. The permission granted herein is personal to you and is not transferable or assignable without the prior written permission of AIP Publishing. This license may not be amended except in a writing signed by the party to be charged.
7. If purchase orders, acknowledgments or check endorsements are issued on any forms containing terms and conditions which are inconsistent with these provisions, such inconsistent terms and conditions shall be of no force and effect. This document, including the CCC Billing and Payment Terms and Conditions, shall be the entire agreement between the parties relating to the subject matter hereof.

This Agreement shall be governed by and construed in accordance with the laws of the State of New York. Both parties hereby submit to the jurisdiction of the courts of New York County for purposes of resolving any disputes that may arise hereunder.

V1.1

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.