ELECTRONIC AND SOLVENT EFFECTS ON MONOSACCHARIDE CONFORMATIONS

A dissertation submitted in accordance with the requirements of the
UNIVERSITY OF CAPE TOWN,

in fulfillment of the requirements for the degree of

MASTER of SCIENCE

by

Christopher Bevan Barnett

Supervisor: Associate Professor Kevin J. Naidoo

September 2007

ABSTRACT

ELECTRONIC AND SOLVENT EFFECTS ON MONOSACCHARIDE CONFORMATIONS

CHRISTOPHER BEVAN BARNETT SEPTEMBER 2007

The hydroxymethyl group rotational preferences of the monosaccharides glucose and galactose are different from each other and non-intuitive (from a steric point of view) in their preferences for gauche conformers. These molecules exhibit very different biological and thermodynamic properties in, for example their binding to glycosides or their liquid crystalline phases in glycolipids. The preference for gauche conformations has been attributed to solvent effects, stereoelectronic effects and hydrogen bonding; yet the experimentally obtained hydroxymethyl rotational populations have not yet been fully rationalised.

In this dissertation, I have used a range of *ab initio*, Molecular Dynamics (MD), Quantum Mechanics/Molecular Mechanics (QM/MM) and free energy computational methods to resolve and explain this observation. The hydroxymethyl free energy surface was calculated using the Potential of Mean Force (PMF), umbrella sampling and Weighted Histogram Analysis Methods (WHAM). The PMF calculations were performed in the canonical (NVT) ensemble in the gaseous and aqueous phase where each monosaccharide was modelled with Parameter Model 3 for Carbohydrates (PM3CARB-1). Density Functional Theory (DFT) calculations were also carried out and Atoms in Molecules (AIM) and Natural Bond Orbital (NBO) analyses were applied.

Gaseous phase simulation results for both glucose and galactose gave hydroxymethyl rotational preferences of gg>tg>gt and gt>gg>tg respectively. These conformational preferences can be rationalised in terms of an intrinsic stereoelectronic effect (found from NBO calculations) and strong intramolecular hydrogen bonding (found in the tg conformer of glucose and the gg conformer of galactose using AIM) in the gaseous phase.

The addition of solvent (water) was found to disrupt the intramolecular hydrogen bonding present in the gaseous phase. Hydroxymethyl rotational preferences in the solution phase were gg>gt>tg for glucose and gt>tg>gg for galactose. The population distributions in solution were also calculated for glucose as gg:gt:tg = 59.21:34.88:0.83 and for galactose as gg:gt:tg = 3.32:79.60:10.15. These populations agree favourably with experimental NMR populations. The solvent conformational preference is dominated by the intrinsic stereoelectronic effect and steric interactions. The gauche effect in monosaccharides has been successfully rationalised.

CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to express sincere thanks to my promoter, Associate Professor Kevin J. Naidoo for providing me with support, enthusiastic comments and ideas throughout the entire M.Sc. process. It has definitely been a challenging one. Perhaps this rough piece of ore is not yet a shiny, perfectly-faceted diamond, but the efforts put into making this a possibility have not gone unnoticed and are genuinely appreciated.

I am thankful for receiving financial aid from the University of Cape Town (UCT); the South African Research Chair Initiative (SARChI) and the Computational Chemistry bursary fund (Thank you again, KJN).

To all of my family, (even if the rain sometimes "washes me in") thank you for always offering me support and guidance. I know I'm often stubborn and that I don't call or accept advice when I should. I am aware that you all care for me and I am very grateful for this. Thanks to Mom and Dad (and Jeannette) for assisting in emergency situations (and in general).

Thanks to all special friends and colleagues for their love and support – Vijandran Murugas, Richard Matthews, Gerhard Venter (Dr.), Jeff Yu-Jen Chen (Dr.), Melissa Petersen (Frau Doktor), Katherine Chen nee de Villiers and Vincent Smith. To those of you who have aided in the dissertation improvement process, thank you for putting up with it and me. I understand that painstakingly long hours were required for proof-reading and even if you missed something (!) I am deeply grateful for all of your insights.

Finally a special thank you to all Computational Chemistry lab colleagues and the Chemistry staff at UCT for making life both easy and difficult but most importantly, fun!

GLOSSARY

Ab initio – "from first principles", this term is applied to quantum mechanics methods to let one know that no experimental data has been used, only fundamental physical constants and the necessary rigorous approximations.

AIM – Atoms in Molecules; A topological analysis of the wavefunction put forward by Bader and implemented into the AIM2000 program by Biegler-König, Schönbohm and Bayles.

AM1 – Austin Model 1, a Semi Empirical model proposed by Dewar.

AMBER – Assisted Model Building and Energy Refinement, a Molecular Mechanics (MM) "engine".

Angstrom or Å - A unit of measure which is equivalent to 10^{-10} m.

Anomer – Refers to the differing directionality of an anomeric substituent, e.g. A pyranose ring can have two forms, one form has the anomeric substituent axial (or alpha) and the other has the anomeric substituent equatorial (or beta), these are termed anomers.

Anomeric carbon – The carbon of pyranose rings which was a ketone or an aldehyde before the ring closing reaction occurred.

Anomeric effect – A stereoelectronic effect which favours the axial over the equatorial anomeric substituent of pyranoses when the anomeric substituent is electronegative.

AO – Atomic Orbital.

au – Atomic Units.

BSSE – Basis Set Superposition Error.

CD – Circular Dichroism.

CHARMM – Chemistry at Harvard Molecular Mechanics. CHARMM is a Molecular Mechanics and dynamics "engine".

CI – Configuration Interaction.

CMO – Canonical Molecular Orbital.

CSFF – Carbohydrate Solution Force Field.

Density - The number (as of particles) per unit of measure.

DFT – Density Functional Theory.

Disaccharide – A dimer consisting of two monosaccharides covalently bonded by a glycosidic linkage.

DNA – Deoxyribonucleic acid.

Enantiomer – Structures that are not identical, but are mirror images of each other.

ESS – Electronic Structure System e.g. GAMESS-UK.

Exo-anomeric effect – Similar to the anomeric effect. It describes the preference of the anomeric substituent OR (alkoxy group) to be gauche to the sugar monosaccharide ring.

FEP – Free energy perturbation method.

fs – femtosecond, a unit of time equivalent to 10^{-15} seconds.

Galactose – A simple monosaccharide, unless otherwise specified or illustrated this should be taken to mean β-D-galactopyranose which has all but the 4-hydroxyl group equatorial.

GAMESS-UK – Generalised Atomic and Molecular Electronic Structure System (from the United Kingdom).

Gauche effect – This is the tendency for the adoption of gauche (synclinal) conformations about a torsional angle C-X-Y-C or X-C-C-Y where X and Y are electronegative atoms or simply the preference for the maximum number of gauche (synclinal) interactions between electron pairs and/or polar bonds.

GHO – Ghost Hybrid Orbital.

Glucose – A simple monosaccharide, unless otherwise specified or illustrated this should be taken to mean β-D-glucopyranose which has all hydroxyl groups equatorial.

GROMOS – Grönigen Molecular Simulation program.

GTO – Gaussian type orbital.

HOMO – Highest occupied molecular orbital.

Hybrid methods – A method that models a system using a combination of other computational methods, e.g. QM/MM.

Hydroxymethyl – The alcohol moiety of a monosaccharide which contains the greatest degree of flexibility (also called the primary alcohol as it is covalently bonded to a carbon with only one other substituent). The torsional rotational angle for this group is usually denoted O5-C5-C6-O6 or C4-C5-C6-O6.

in vacuo – In the gaseous phase.

IR – Infrared spectroscopy.

kcal – kilocalories, a unit of energy measure equivalent to 1000 calories or 4.184 kJ (kilojoules).

LCAO – Linear combination of atomic orbitals.

LDA – Local Density Approximation.

LSDA – Local Spin Density Approximation.

LUMO – Lowest unoccupied molecular orbital.

MC – Monte Carlo.

MD – Molecular Dynamics.

MM – Molecular Mechanics.

MO – Molecular orbital theory.

Monosaccharide – The monomer of higher saccharides. Monosaccharides are of the family of aldoses and ketoses. These can exist in chain or ring form. The ring form (hemi-acetal or acetal) is more prevalent and 5- and 6-membered rings are commonly observed (pyranose and furanose rings respectively).

MPn – Møller Plesset theory with an nth order correction.

μVT – The grand canonical ensemble, a statistical ensemble with μ (the chemical potential), V (volume) and T (temperature) constant.

NAO – Natural Atomic Orbital.

NBO – Natural Bond Orbitals or Natural Bond Orbital analysis.

NHO – Natural Hybridised Orbital.

NLMO – Natural Localised Molecular Orbital.

NMR – Nuclear Magnetic Resonance spectroscopy.

NPT – The isothermal-isobaric ensemble, a statistical ensemble with N (number), P (pressure) and T (temperature) constant.

ns – nanosecond, a unit of time equivalent to 10^{-9} seconds.

NVE – The microcanonical ensemble, a statistical ensemble with N (number), V (volume) and E (energy) constant.

NVT – The canonical ensemble, a statistical ensemble with N (number), V (volume) and T (temperature) constant.

Oligosaccharide – A carbohydrate chain consisting 2 or more monosaccharides.

ORD – Optical rotatory dispersion.

PES – Potential Energy Surface.

PM3 – Parameterised Model 3, a Semi Empirical model.

PM3CARB-1 – Parameterised Model 3 for carbohydrates, a Semi Empirical model developed by McNamara.

PMF – Potential of Mean Force.

Polysaccharide – A carbohydrate or sugar. A polymer consisting of many monosaccharide monomers.

Primary alcohol – See hydroxymethyl.

ps – picosecond, a unit of time equivalent to 10^{-12} seconds.

QM – Quantum Mechanics.

QM/MM – Quantum Mechanics/Molecular Mechanics. See hybrid methods.

RHF – Restricted Hartree-Fock theory.

ROHF – Spin-Restricted Open-shell Hartree-Fock theory.

RNA – Ribonucleic acid.

Rotamer – A rotational conformer about a torsional angle.

SCF – Self-consistent Field Theory.

Secondary alcohol – An alcohol group attached to a carbon with two other substituents. See primary alcohol.

Stereoelectronic effect – Chemical consequences of the arrangement of orbitals in space.

STO – Slater-type orbital.

Tetrahydropyran – a saturated six-membered ring containing five carbon atoms and one oxygen atom. This can be considered the core of a monosaccharide without any substituents attached.

UHF – Spin-Unrestricted Hartree-Fock Theory.

WHAM – Weighted Histogram Analysis Method.

Chapter 1

INTRODUCTION TO ELECTRONIC EFFECTS IN MONOSACCHARIDES

1.1 Biological role of saccharides

Polysaccharides are polymers formed from monosaccharides which primarily consist of carbon, hydrogen and oxygen. Poly- or monosaccharides are commonly referred to as carbohydrates or sugars. They are often only thought of as an energy resource but these macromolecular entities have a complex conformational phase space and perform a multitude of biological functions.

Phosphorylated sugars are key intermediates in energy generation and biosynthesis while glycogen is used to store energy[1]. Structural support or scaffolding functionality is provided by cellulose and lignin in plants and chitin in arthropods[1,2]. Even the genetic building blocks ribonucleic acid (RNA) and deoxyribonucleic acid (DNA) contain sugars; the sugars ribose and deoxyribose provide a structural framework for RNA and DNA, while the conformational flexibility of these sugars assists in the storage and expression of genetic information[1,2].

Another biologically common and diverse form in which carbohydrates occur are glycoproteins. These biologically ubiquitous entities consist of proteins linked to carbohydrates. The unique oligosaccharide tags on the protein make cell-cell recognition possible which forms an important part of the immune response process[1].

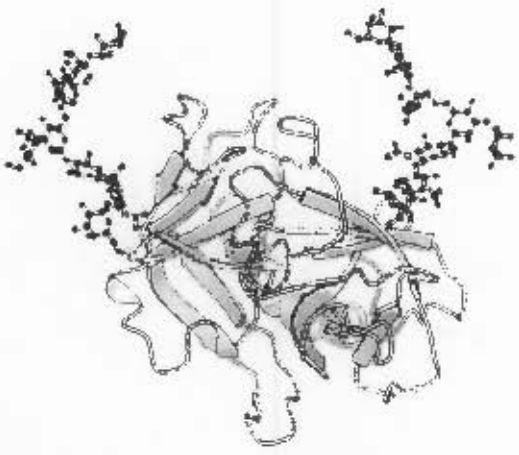

Figure 1.1 – Elastase – a glycoprotein containing the "protein part" (yellow ribbons) and "carbohydrate part" (ball and stick) (extracted from Stryer[2]).

Figure 1.1 gives an example of the glycoprotein elastase, which is a protease (an enzyme that breaks down other proteins). Elastase targets the connective tissue protein elastin and the oligosaccharide chains (indicated by the ball and stick models in Figure 1.1) are vital for molecular recognition[2].

1.2 The role of the monomer

The intricate folding and conformational preferences of polymeric saccharides can be understood by considering the complexity of the monomeric and dimeric units. Even oligosaccharides have a great variety of conformations available to them and this variety can be directly attributed to the multiple bonding sites: primary and secondary alcohol degrees of freedom and ring puckering ability of monosaccharides (to be discussed in the next few pages).

Monosaccharides have at least 3 carbons in a chain and are aldehydes or ketones with multiple hydroxyl groups, where the monosaccharides with 6 carbons are most common. They have a strong tendency to cyclise via hemiacetal or acetal formation[1]. The cyclisation equilibrium process is shown in Figure 1.2.

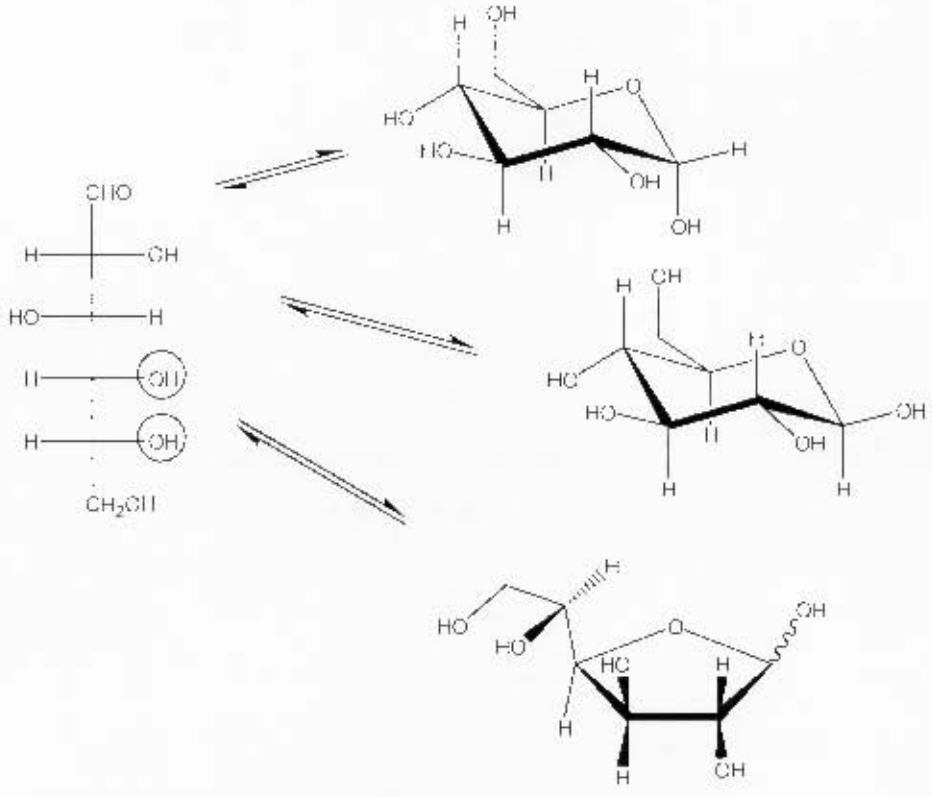

Figure 1.2 – A representation of the different forms of D-glucose. On the left is a Fischer projection of the straight chain form. Clockwise from top right: α-D-glucopyranose; β-D-glucopyranose; D-glucofuranose (the "wiggly" bond means either α or β).

The example in Figure 1.2 is that of D-glucose (D designates the enantiomer under consideration). A Fischer projection of the straight chain form is on the left. This can cyclise to form pyranose or furanose rings when the aldehyde undergoes nucleophilic attack from either of the two circled alcohol groups. Attack can occur from either face of the sp^2 hybridised aldehydic centre yielding two pyranose and two furanose monocycles. The pyranose forms are most abundant and account for up to 99% of D-glucose in solution[1]. The glucopyranose ring with the axial substituent on the anomeric carbon (previously the sp^2 hybridised centre of the aldehyde) is called α-D-glucopyranose while that with the equatorial hydroxyl group is labeled β-D-glucopyranose.

Each pyranose (or hexose) ring has multiple hydroxyl groups and two pyranose rings can join together by a condensation reaction (loss of water – See Figure 1.3) to form

glycosidically linked di-, tri-, oligo-(two, three and a few linked monomers) and polysaccharides (multiple linked monomers)[1,3].

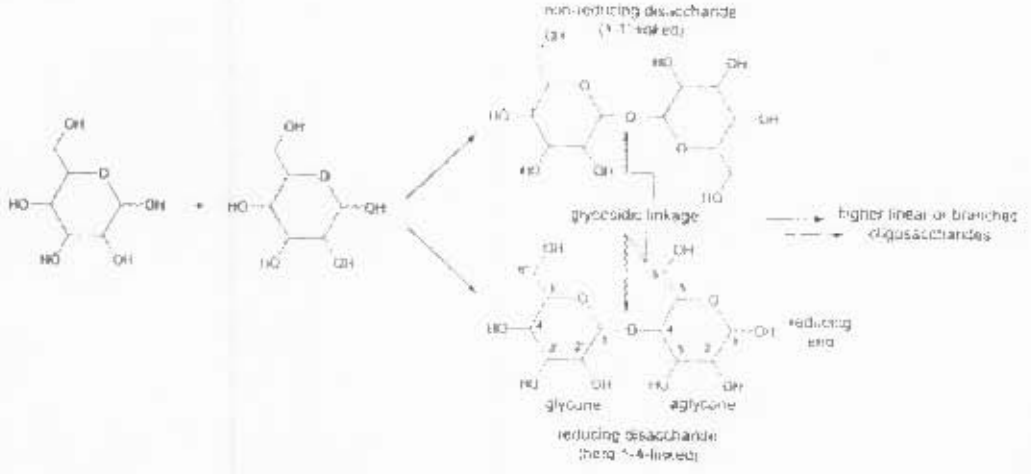

Figure 1.3 – Schematic of two monosaccharides joining (bonding) to form 1→1' and 1→4' glycosidically linked disaccharides. Further formation of glycosidic linkages gives oligo- then polysaccharides (From Lindhorst[4]).

The structural diversity seen in carbohydrates is due to their ability to have multiple points of connection as well as their flexible nature. These "points of connection"- the hydroxyl groups - are either primary or secondary hydroxyls (attached to a primary or secondary carbon). The primary hydroxyl groups have an α (axial) or β (equatorial) configuration allowing for a large variety of branched or straight chain polysaccharides.

Linkages are denoted by the carbon atoms in the attachment as well as the orientation of the linkage (See Figure 1.3 – the orientations are not explicitly shown). The flexible nature of saccharides results not only from the ability of the ring to contort and "flip" (pucker), but also from the rotational flexibility that exists about the glycosidic linkages between monomers.

The free energy surface for cyclohexane illustrated in Figure 1.4 shows the different ring conformations common to 6-membered rings (like pyranoses) as well as the energy required to "flip" between them. It is quite clear that the chair conformer is most stable.

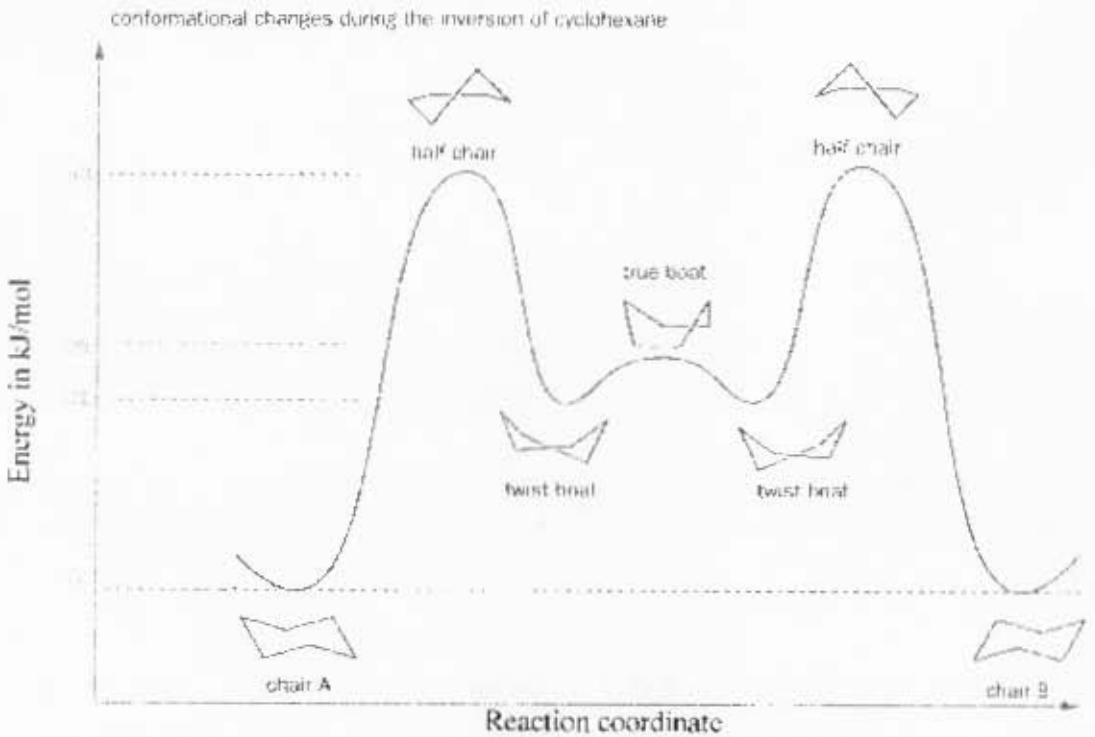

Figure 1.4 – A free energy surface for the inversion of cyclohexane. Note that the chair conformers are most stable (From Clayden[3]).

For pyranose rings the possible ring forms are illustrated in Figure 1.5.

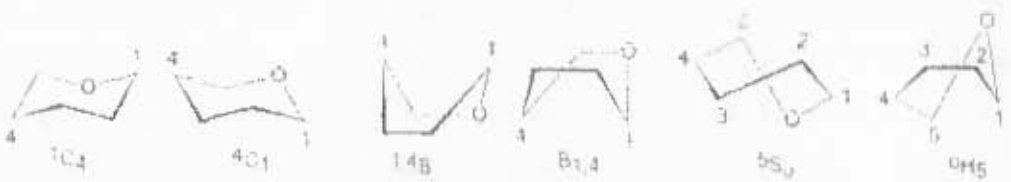

Figure 1.5 – Illustrations of the ring forms of tetrahydropyrans. The capitalised letters denote the conformations and the super/sub scripts denote which atoms are above or below the plane of the ring. C stands for chair, B for boat, S for skew and H for half-chair (From Lindhorst[4]).

When substituted, the ring forms are not equivalent in energy. In monosaccharides such as glucose the $^4C_1$ chair (nomenclature explained in Figure 1.5) is favoured as the $^1C_4$ chair has numerous unfavourable axial interactions (all hydroxyl groups are axial)[4]. This is illustrated in Figure 1.6.

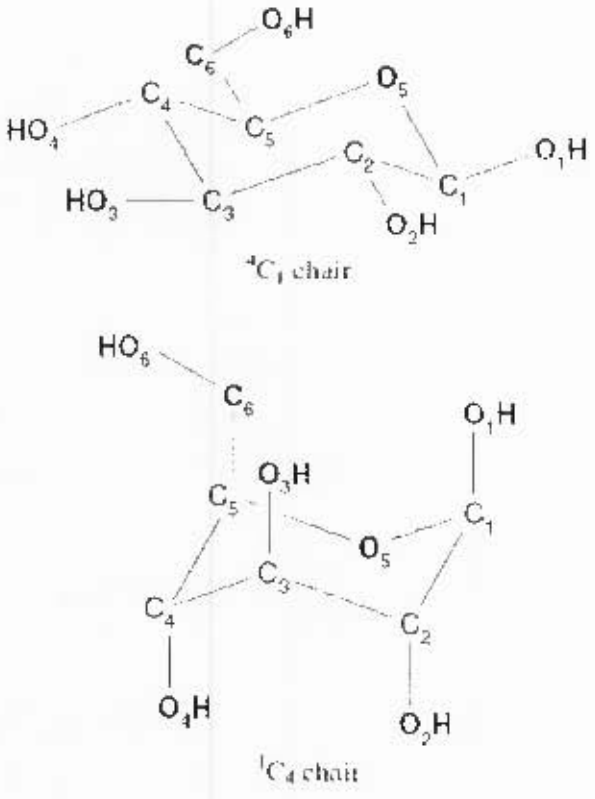

Figure 1.6 – Illustration of the $^4C_1$ chair and the $^1C_4$ chair where subscript labels refer to atom numbering about the ring. The $^4C_1$ is favoured because of the axial group steric clashing in $^1C_4$. (Extracted from McNamara[5])

The hydroxymethyl or primary alcohol moiety provides more flexibility than the secondary alcohols as it has an extra atom which provides an extra degree of rotational freedom. The great flexibility of the primary alcohol is utilised effectively by animals which store glucose in the form of glycogen, a large branched polymer. Most of the glucose units are linked with α-1⟶4 glycosidic bonds but approximately every 10th residue has an α-1⟶6 linkage. These 1⟶6 linkages improve the solubility of glycogen making the glucose stored in glycogen a more readily accessible energy resource[1].

The example of glycogen demonstrates that a linkage involving the hydroxymethyl group provides added flexibility. The hydroxymethyl group has three staggered rotamers or

conformers about the ω-angle (See Figure 1.7) denoted gauche-gauche (gg), gauche-trans (gt) and trans-gauche (tg) (See Figure 1.8).

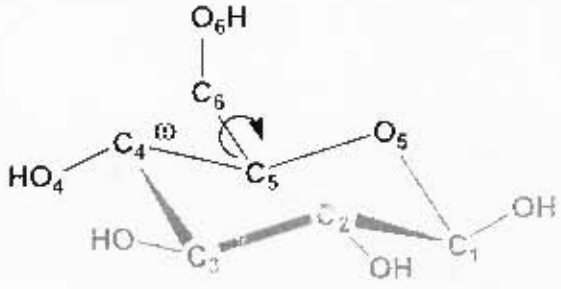

Figure 1.7 – A representation of D-glucose with the hydroxymethyl torsional angle indicated by ω (O5-C5-C6-O6).

The first description (gauche- or trans-) refers to the relation between the O5 and O6 atoms as given by the O5-C5-C6-O6 torsional angle ω. The second description (-gauche or -trans) refers to the relation between the C4 and O6 atoms as given by the C4-C5-C6-O6 torsional angle. These rotamers are shown in Figure 1.8.

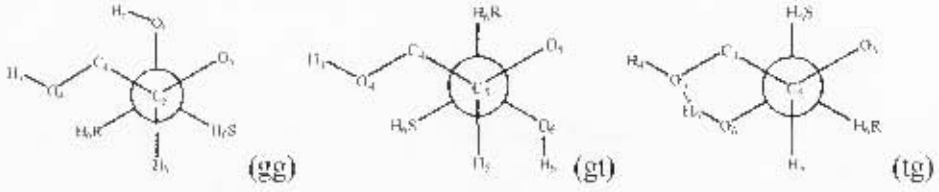

Figure 1.8 - Newman type representations of the hydroxymethyl group rotamers for D-glucose. From left: gg, gt, tg. The dashed lines indicate a potential hydrogen bonding interaction between H6 and O4 in the tg conformer.

The hydroxymethyl group rotation is of interest due to the extra conformational freedom it imparts to the monosaccharide, and due to the importance of the 1→6 linkage in higher saccharides. α, β-D-glucopyranose and α, β-D-galactopyranose are commonly observed hexapyranoses and this research will consider both pyranoses with emphasis on the β form.

1.3 Conformational space and Stereoelectronic effects

The most basic nomenclature and some interesting biological functionality of saccharides have been described. *What about the interactions that affect their conformational space?*

There are a number of interactions at play and many of them arise from the cyclic structure of monosaccharides. These include repulsion of 1, 3 diaxial ring substituents (van der Waals repulsions) and non-bonded interactions i.e. electrostatic interactions and other van der Waals interactions between atoms[4]. Hydrogen bonding can also play a significant role in determining structure, where intermolecular bonds (e.g. with solvent) and intramolecular bonds can occur. The solvent used (in most biological systems this is water) greatly affects the conformational freedom of a carbohydrate. The polarity of the solvent, its ability to form hydrogen bonds as well as the solvent's ability to act as a dielectric (and screen charge) can all affect how a saccharide folds.

Certain configurational and conformational preferences can only be described in terms of stereoelectronic effects. "Stereoelectronic effects are chemical consequences of the arrangement of orbitals in space"[3]. A number of stereoelectronic effects exist in carbohydrates such as the anomeric, exo-anomeric and gauche effect.

1.3.1 The anomeric effect

In most monocyclic systems (cyclohexane derivatives) one expects a substituent to adopt the equatorial position rather than the axial position. This argument is based upon unfavourable steric interactions present in the axial position causing it to be energetically disfavoured when compared to the equatorial position[3]. The chair conformer with the largest number of equatorial substituents and the least synclinal (gauche) interactions should be favoured. This is the case for tetrahydro-2H-pyran-2-ol (Figure 1.9) which favours the equatorial chair in a ratio of $\alpha:\beta = 11:89$[6].

Figure 1.9 – The equatorial (left) and axial (right) tetrahydropyranol chairs conformers (Modified from Kirby[6]).

In polysubstituted tetrahydropyrans this is not always the case. Consider the following two pentasubstituted tetrahydropyrans - (2,3S,4S,5S)-6-hydroxymethyl)tetrahydro-2H-pyran-2,3,4,5,-tetraol and (2,3S,4S,5S)-6-hydroxymethyl)methoxytetrahydro-2H-pyran-3,4,5,-tetraol. These are commonly known as D-glucose and 1-methoxy-D-glucopyranose (where the D designates the enantiomer under consideration).

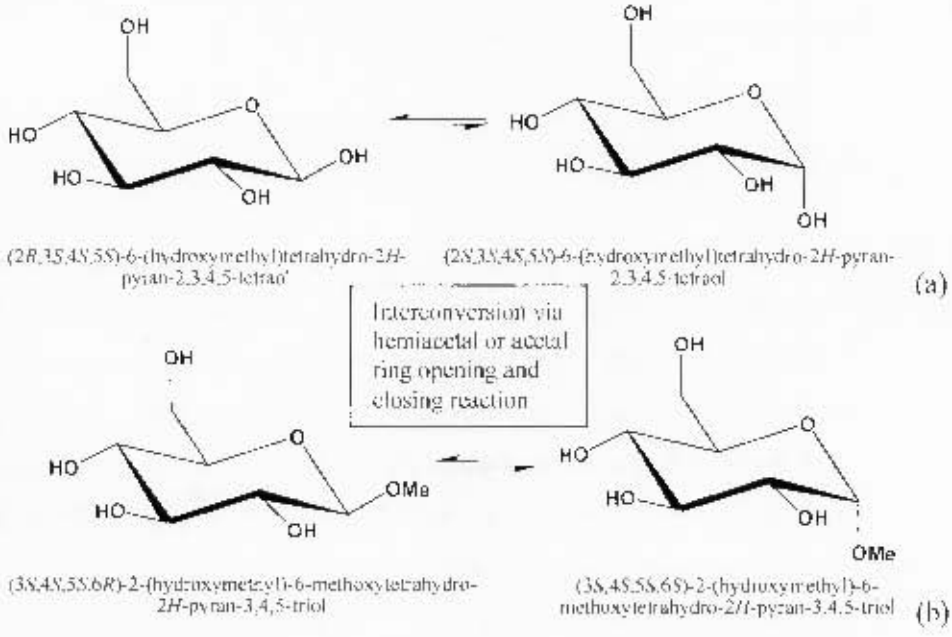

Figure 1.10 – Comparison of the anomeric equilibria for D-glucose (a) and 1-methoxy-D-glucopyranose (b) (Modified from Kirby[6]).

D-glucose is not too dissimilar from tetrahydro-2H-pyran-2-ol and one might expect a similar axial:equatorial equilibrium. The equilibrium still favours the equatorial (β) form

but more α-D-glucose is present at equilibrium (See Figure 1.10a) than predicted by a purely steric argument with a ratio of α:β = 36:64. If a bulkier substituent is attached to the anomeric carbon the equatorial configuration is expected to be more dominant, based purely on steric grounds. Instead the α-anomer becomes more dominant. In fact 1-methoxy-D-glucopyranose at equilibrium favours the more sterically hindered axial configuration (See Figure 1.10b) of the ring in a proportion of α:β = 67:33[1,3,6,7].

Anomeric substitution by other electronegative substituents such as bromide and chloride also yields a favoured axial configuration. In these cases the conversion between anomers cannot occur via hemiacetal or acetal opening and closure. Instead if the electronegative substituent exhibits a strong enough effect the $^4C_1$ chair can flip to become the $^1C_4$ chair.[8].

This preference for the axial configuration at the 2-position of tetrahydropyrans with electronegative substituents is termed the anomeric effect[6]. Alternatively, the anomeric effect is the effect which provides stabilisation of the axial substituent such that the inherent steric bias of the substituent is overcome[9].

An explanation for the anomeric effect has been given in terms of dipole-dipole interactions where the conformation in which the dipoles partially cancel out is favoured over the one in which the dipoles strengthen each other[4] (See Figure 1.11).

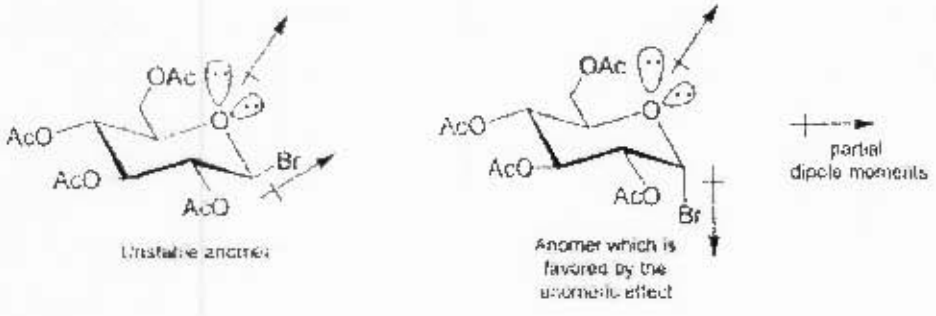

Figure 1.11 – Dipole moments for an acetylated and brominated derivative of glucose. The equatorial conformer is on the left and the axial conformer on the right (From Lindhorst[4]).

An alternative explanation for this preference is a stereoelectronic one. Highest occupied molecular orbital (HOMO) to lowest occupied molecular orbital (LUMO) interactions for

each of the axial and equatorial conformers of 1-methoxy-D-glucopyranose are possible. The potential HOMO-LUMO interaction between the oxygen lone pair and antibonding orbitals are shown in Figure 1.12.

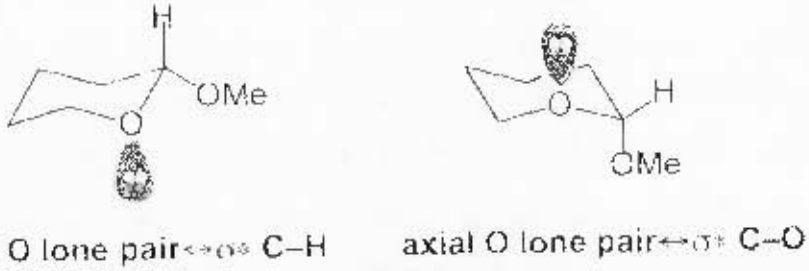

Figure 1.12 Potential orbital interactions between the axial lone pair of the ring oxygen and the antibonding orbitals (σ* - not drawn) of C-H (in the equatorial conformer) and C-O (in the axial conformer) (From Evans[5]).

Electronegative atoms (X, Y) other than oxygen, exhibit this effect and a more generic pyranose derivative is shown in Figure 1.13, in which the overlap and orbital diagram for bonding and anti-bonding orbital overlap (HOMO-LUMO overlap) are illustrated. This is termed negative hyperconjugation and leads to an energy lowering[4].

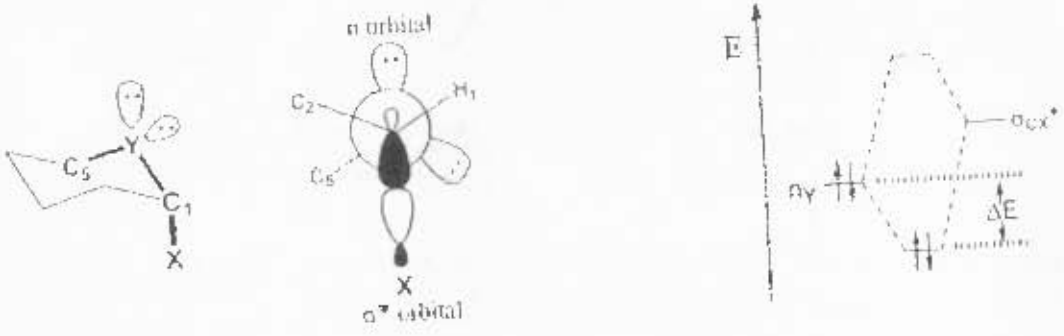

Figure 1.13 An axial anomer of a substituted pyranose derivative (left). The non-bonding electrons of Y and the antibonding orbital of the C-X bond are shown (middle). The filled nonbonding and empty antibonding orbitals can mix to form a filled lower energy orbital and an empty higher energy orbital (right) (From Lindhorst[4]).

The anomeric effect can also be described as a preference for gauche (synclinal) interactions. If one considers the interactions occurring across the X-C-Y-C unit (See Figure 1.13), the axial conformer shown has the C-X bond gauche to the C5-Y bond whereas in the equatorial form this would be in the trans orientation. Thus, the gauche form is preferred.

1.3.2 The exo-anomeric effect

The exo-anomeric effect is similar to the anomeric effect in that it specifies the preference for the gauche conformers about O-C-O-R bonds. Specifically, it specifies the preference for the anomeric substituents (usually O-R (alkoxy) groups) of tetrahydropyrans to be in the gauche form. Figure 1.14 depicts three axial and three equatorial forms of 2-alkoxytetrahydropyran[6].

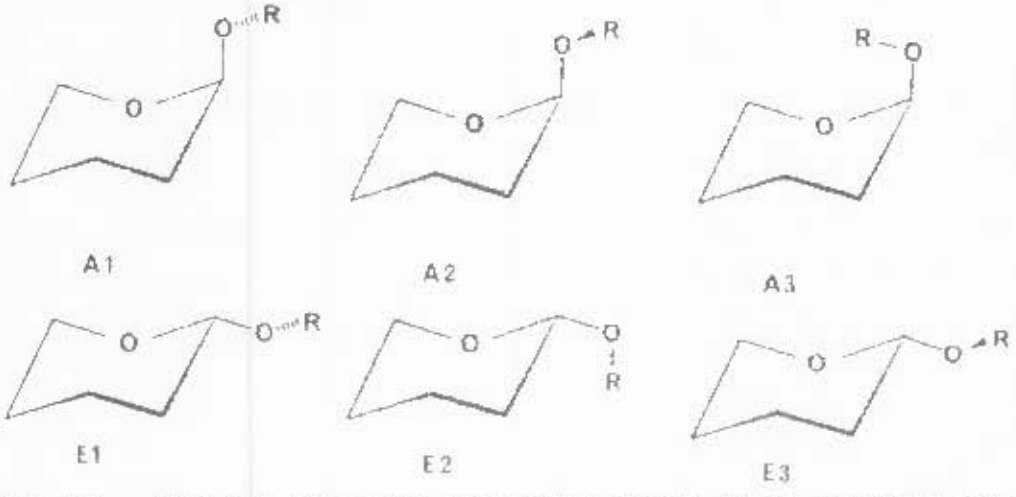

Figure 1.14 Illustrations of the axial and equatorial positions for the anomeric substituent of tetrahydropyrans (Marked A for axial and E for equatorial) (Extracted from Kirby[8]).

The conformers that minimise steric clashing and maximise the hyperconjugative interactions are favoured. These are A1 and E1 (See Figure 1.14). Figure 1.15 illustrates that A1 is favoured over A3 because stereoelectronic interactions are maximised and steric interactions minimised (See Figures 1.14 and 1.15).

Figure 1.15 – The anomeric carbon to oxygen bond is indicated (left). The two C-O bond rotamers (middle and right) are both gauche rotamers. The favoured rotamer (indicated) has favourable stereoelectronic interactions while minimising unfavourable steric clashes (Extracted from Evans[9]).

1.3.3 The gauche effect

The gauche effect is the tendency for the adoption of gauche (synclinal) conformations about a torsional angle C-X-Y-C or X-C-C-Y where X and Y are electronegative atoms or simply the preference for the maximum number of gauche (synclinal) interactions between electron pairs and/or polar bonds[6,10,11].

An improved definition proposed by Kirby is based upon the stereoelectronic (orbital-based) arguments discussed earlier; "There is a stereoelectronic preference for conformations in which the best donor lone pair or bond is antiperiplanar to the best acceptor bond"[11].

The donors and acceptors in decreasing ability are

Donors: $n_N > n_O > \sigma_{C-C}, \sigma_{C-H} > \sigma_{C-X}$ (X=N>O>S>Halogen)

Acceptors: $\pi^*_{C=O} > \sigma^*_{C-Hal} > \sigma^*_{C=O} > \sigma^*_{C=N} > \sigma^*_{C-C} > \sigma^*_{C-H}$

where n is a non-bonding orbital, σ is an occupied Lewis (or bonding) orbital and π^* and σ^* are the unoccupied non-Lewis (or in this case, antibonding) orbitals.

Simple substituted alkanes in which the gauche effect occurs include 1,2-difluoroethane[12] and 1,2-ethane-diol[13]. In 1,2-difluoroethane the gauche effect can be neatly described in terms of acceptor-donor capability[8]. The favourable $\sigma_{C-H} \rightarrow \sigma^*_{C-F}$ interactions outweigh the steric preference for the two fluorines to be anti-periplanar (trans). Although C-H bonds are not the most convincing donor orbitals they are better than the C-F orbitals. It is often

easier to think of these systems avoiding less preferred interactions (i.e. trans C-X donor/acceptor interactions) rather than having a preference for gauche interactions[8]. However, for 1,2-ethane-diol the gauche preference can be understood by taking hydrogen bonding into account rather than the gauche effect[13,14].

This gauche stereoelectronic effect is of particular interest as it is directly relevant in determining the conformational freedom about the hydroxymethyl (primary alcohol) of monosaccharides and 1→6 linkages of higher saccharides (The hydroxymethyl group is an O-C-C-O fragment with two vicinal oxygen atoms).

1.4 Experimental techniques applied to carbohydrates

Carbohydrates are difficult to study experimentally because of their highly flexible nature. The wide variety of experimental methods available includes crystallographic techniques, Nuclear Magnetic Resonance (NMR), optical rotatory dispersion (ORD), ultrasonic spectroscopy, circular dichroism and neutron diffraction.

Solid state techniques such as X-ray crystallography do not necessarily reproduce the correct conformational ratios that would be observed in solution. Furthermore, polysaccharides can be difficult to crystallise and the method of crystallisation (solvent from which we crystallise, crystal inducement method) can affect the conformation that is observed in the crystal[15].

NMR techniques have been widely applied to carbohydrates in a number of solvents and can give insight into conformational aspects of these molecules[16-20]. The effect of a solvent can also be considered[21]. Sometimes experimental results can be ambiguous depending on how proton signals are assigned (Specifically in the case of negative tg populations[21])

Optical rotatory dispersion experiments show whether polarised light is rotated in a dextro- or lævo-rotatory manner (right or left rotation from a reference). These techniques have been particularly useful in calculating equilibrium anomeric (α:β) ratios[7].

Ultrasonic spectroscopy has recently been applied to aqueous carbohydrate solutions to observe the kinetics of ring puckering and hydroxymethyl rotation[22].

The spectra from circular dichroism have been used to gain insight into the rotational dependence of the hydroxymethyl group and to analyse oligosaccharides[23,24]. Recently the circular dichroic exciton chirality method has been applied to chromophorically substituted gluco- and galactopyranosides[25,26].

Neutron diffraction methods as applied to carbohydrates have also received a lot of exposure due to the work of Brady et al.[27,28]. Water structuring and solution dynamics[29] as well as hydroxymethyl rotations[27] have been investigated using this method.

1.5 Computational and theoretical methods applied to carbohydrates

Computational studies of carbohydrates have included Molecular Mechanics (MM) and Molecular Dynamics (MD) calculations (covered in Chapter 2 and 3), Quantum Mechanics (QM) techniques (Chapter 2) and hybrid QM/MM methods (Chapter 2).

MM and MD approaches have been used to examine the behaviour of glycosidic linkages[30-32], as well as the hydroxymethyl group rotations[27] . The effect of water structuring around sugars has also been investigated with MD simulations[33].

Static QM calculations have been used to describe the torsional rotation energies of carbohydrate mimics and derivatives[34-36]. Furthermore, analysis of static QM calculations has been used to consider the relationship between intramolecular hydrogen bond strength and electron density[37,38]. *Ab initio* calculations have been used to examine the anomeric effect in pyranoses and glucose[39].

Free energy calculations have also been used to probe the free energy surface of 1→4 glycosidic linkages using MD[40] and hybrid QM/MM potentials[41,42].

1.6 Computational, theoretical and experimental methods applied to glucose and galactose

The primary intention of this research is to investigate the hydroxymethyl preference of glucose and galactose. *What other methods have been applied to this problem?*

1.6.1 Glucose

Experimental observation of the glucose hydroxymethyl group (See Figure 1.16) in aqueous solutions reveals a preference for gauche conformers in the order gg>gt>tg[43,44].

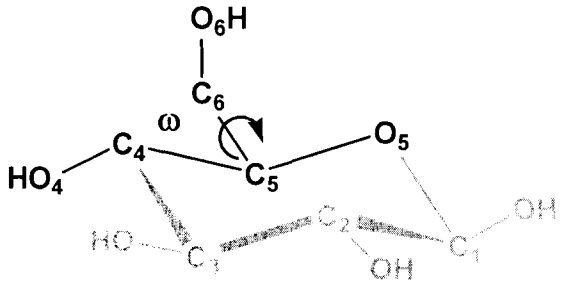

Figure 1.16 – Representation of D-glucose with the hydroxymethyl torsional angle indicated by ω (O5-C5-C6-O6).

NMR studies of the hydroxymethyl group conformation[19] showed that the gg conformer population changes according to the solvent polarity leading to postulates that the conformational preference may be governed by solvent effects. However, Bock and Duus[19] alluded to the existence of an intrinsic stereoelectronic effect being responsible for the gg, gt preference in Methyl-(α and β)-4-deoxy-D-glucose[19]. In this compound no 1, 3 diaxial interactions destabilise the tg conformer yet gg and gt rotamers are still favoured.

A Semi Empirical study of D-glucopyranose using continuum solvent models was performed by Cramer and Truhlar[45] where they showed that only after solvent models were applied did the relative hydroxymethyl ordering agree with experimental NMR values. They ascribed the unusual conformational preference of the primary alcohol in glucose to intramolecular hydrogen bonding and the solvent, where it was identified as the reason for this preference. Analysis of a series of *ab initio* calculations on D-glycero-hexopyranosides led Tvaroška[46] to agree with the gg>gt>tg preference. Once again, intramolecular hydrogen bonding, in particular between O5 and the hydroxyl of the C6 carbon (6-hydroxyl group), was cited as the reason for stabilisation of the gg and gt conformers[46].

1.6.2 Galactose

Galactose (See Figure 1.17) differs from glucose through the change of the hydroxyl from equatorial to axial on the C4 carbon.

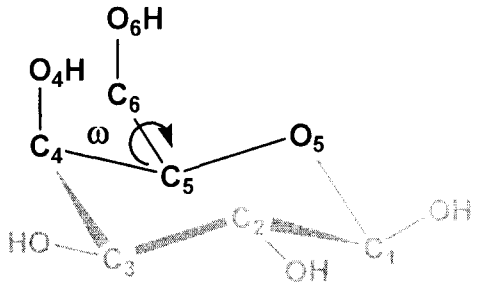

Figure 1.17 – Representation of D-galactose with the hydroxymethyl torsional angle indicated by ω (O5-C5-C6-O6).

A gauche preference is also observed, but along with the positional change of the 4-hydroxyl group (OH4), the conformational preference changes from gg>gt>tg for glucose to gt>gg>tg[43] for galactose. The effect of the OH4 configurational change in the 4-hydroxyl group configuration was assessed using NMR, IR spectroscopy and optical rotation experiments[47] in a variety of solvents and intramolecular hydrogen bond motifs, with no conclusive rational put forward.

1.6.3 The gauche effect?

A major reason for this unusual preference in both glucose and galactose has been ascribed to the gauche effect. While this preference is attributed to favourable bonding orbital (σ) to antibonding orbital (σ*) interactions, it does not convincingly explain the ordering of the conformations or the differences that arise from a hydroxyl configurational change at the C4 position. There is no general agreement and it is debated whether the reasons are due to solvation or hydrogen bonding[45,46].

1.7 Objectives

Previous studies have postulated that the major reason for the conformational preferences of the hydroxymethyl group in glucose and galactose is due to the gauche effect. For these reasons the vacuum and solution free energy profiles for both glucose and galactose have been calculated, and their thermodynamic and electronic structures in the gg, gt and tg conformations have been investigated.

Glucose and galactose, as well as the staggered rotamers for the hydroxymethyl group, are shown in Figures 1.18-1.20.

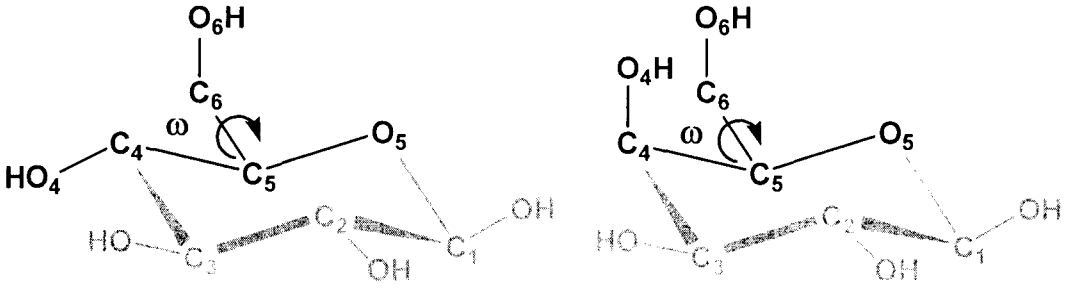

Figure 1.18 – Representations of D-glucose (left) and D-galactose (right) with the hydroxymethyl torsional angle indicated by ω (O5-C5-C6-O6).

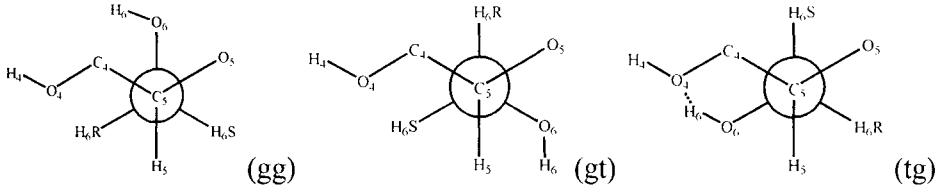

Figure 1.19 – Newman type representations of the hydroxymethyl group rotamers for D-glucose. From left: gg, gt, tg. The dashed lines indicate a potential hydrogen bonding interaction between H6 and O4 in the tg conformer.

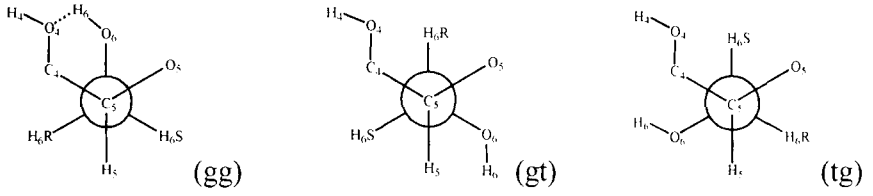

Figure 1.20 – Newman type representations of the hydroxymethyl group rotamers for D-galactose. From left: gg, gt, tg. The dashed lines indicate a potential hydrogen bonding interaction between H6 and O4 in the tg conformer.

The free energy methodology employs the Potential of Mean Force (PMF) to give the Helmholtz free energy surface for hydroxymethyl rotation. This iterative process produces an energy surface which is dependent on thorough sampling of hydroxymethyl phase space, especially the high energy regions. Other methods that have been employed include iterative umbrella sampling (to direct hydroxymethyl phase space sampling) and the Weighted Histogram Analysis Method (WHAM) which allows the probability histograms of multiple simulations to be merged, thereby increasing the rate of convergence of the PMF. Upon characterising the free energy surface the preferred conformers, transition states and population distributions can be calculated.

The dynamics simulations make use of the hybrid QM/MM methodology within the CHARMM[48] program which is a MM/MD "engine". GAMESS-UK[49], an electronic structure programme has been used for *ab initio* calculations and electron population analysis (using Atoms in Molecules[50,51] and Natural Bond Orbital analysis[52,53]).

From these calculations we aim to describe the nature of the gauche effect in monosaccharides.

1.8 Overview

Chapter 2 introduces computational approaches for describing molecular systems. Important concepts that are covered include Molecular Mechanics (MM) and Force Fields, where atoms are treated as deformable balls and *ab initio* Quantum Mechanics (QM) as well as Density Functional Theory (DFT) methods where the electrons of atoms are explicitly described. Electron population analysis techniques such as AIM and NBO analysis are also discussed. Approximate Semi Empirical approaches for treating electrons are also covered followed by a short discourse on hybrid QM/MM potentials and optimisation techniques.

A statistical description of molecular systems is introduced in Chapter 3 by way of ensembles and MD simulation methods. This chapter also highlights the important Potential of Mean Force (PMF) method for generating the Helmholtz free energy surface for a degree of freedom. The iterative PMF approach is described and important theories which enhance the PMF application are also discussed. Specifically, umbrella sampling and the Weighted Histogram Analysis Method (WHAM). Umbrella sampling directs the simulation into uncharted or poorly sampled regions of phase space. This improves "adaptively" as the PMF calculations progress. WHAM provides one with improved convergence to the final PMF by adding the probability histograms of multiple simulations using weighting coefficients.

Chapter 4 reports specifics of the methodologies used and the results of the hybrid QM/MM PMF gaseous and solution phase dynamics simulations of glucose and galactose. DFT calculations, Natural Bond Orbital (NBO) and Atoms in Molecules (AIM) analysis were also carried out and these results are discussed in depth.

Conclusions complete this dissertation

Chapter 2

BACKGROUND TO COMPUTATIONAL METHODS – DESCRIBING MOLECULAR SYSTEMS

The theoretical approach used to model physical properties of a system falls into three broad categories: Empirical Force Field models, Quantum Mechanical models or a combined hybrid Quantum Mechanical/ Molecular Mechanical (QM/MM) approach. [54]

A diverse array of methods is available and a great deal of time goes into deciding upon the method(s) to be used. To make this decision one must consider the size of a problem, the accuracy required as well as time and funds available. Certain modelling techniques suit certain problems. In the end the researcher must justify his/her decision based on which method willl describe the problem most adequately.

The empirical or Molecular Mechanics (MM) description of a system will be described first followed by a relatively detailed discussion of electronic structure methods. This will include Hartree-Fock theory, Density Functional Theory (DFT), the Semi Empirical approach and electron population analysis. Hybrid QM/MM methods and a brief overview of optimisation techniques close off the chapter.

2.1 Empirical Force Fields and Molecular Mechanics

Molecular Mechanics (MM) methods treat the system using the laws of classical physics. MM is often referred to as the ball and spring model because the bonds between atoms are represented as springs and obey Hooke's law. The Born-Oppenheimer approximation is applied, this is the postulate that the potential energy of the molecule can be described as a function of the nuclear coordinates of each atom[54]. Electronic contributions are ignored in

this approach[55] (electrons are not treated as separate particles , although several Force Fields have attempted to remedy this by moving charge away from the nucleus and adding lone pair terms) and because of this properties/processes bases on electron interactions (e.g. bond breaking or forming) cannot be studied[54] (this statement is not strictly accurate – a new reactive Force Field[56] has been developed for hydrocarbons that allows reactions to occur.).

Atoms are modelled as deformable balls using mathematical potentials. Each MM method is defined by a Force Field, which treats "connected" and "not-connected" atoms using different potentials. This Force Field contains the following information[57]:

- Equations describing how the potential energy is affected by the location of atoms.
- Atom types which define the behaviour of an atom in a specific chemical environment (e.g. the effect of charge, hybridisation).
- Parameter sets which define force constants relating atomic character to energy and structural data (e.g. bond lengths, angles).

Note that because these methods rely on a set of experimentally derived parameter sets they are termed empirical methods. They are also very quick – definitely the fastest of all the methods in this chapter. The price paid for this speed is a loss of accuracy. If a good Force Field doesn't exist, results will generally be poor. Conversely, a good Force Field can reproduce experimental data within in a large degree of accuracy.

2.1.1 How is the potential energy defined?

The kinetic energy of the stationary nuclei is zero, thus the total potential energy of a system is the sum of bonded interactions and non-bonded interactions as shown in Equation 2.1[54].

$$U = U_{bonds} + U_{angles} + U_{torsional\ angles} + U_{improper\ torsional\ angles} + U_{vanderWaals} + U_{electrostatic} \qquad \text{Eq 2.1}$$

The bonded interactions are bond stretching, angle bending and torsional angle rotation. Improper torsional angle interactions (the angle used to keep four atoms in a plane) also contribute to the overall potential. Non-bonded interactions are defined between parts of a molecule that are not directly bonded or between two different molecules, these include electrostatic and van der Waals interactions and are shown in Figure 2.1.

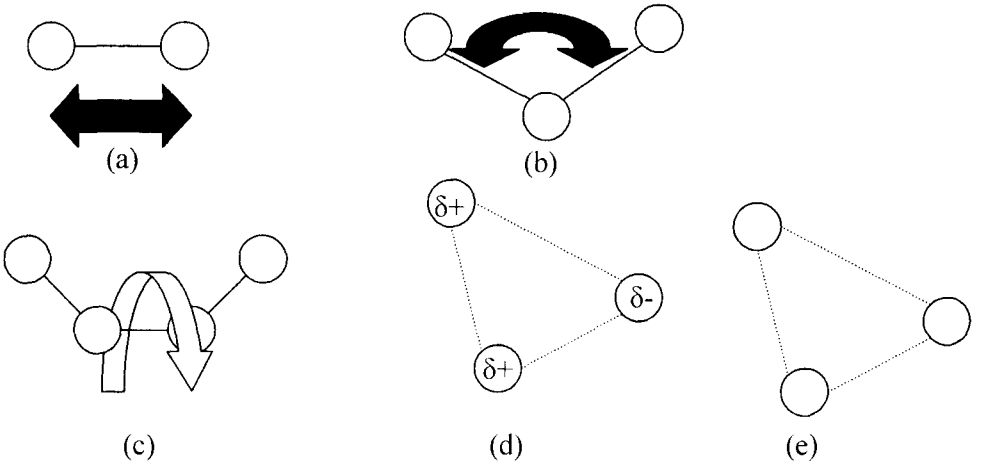

Figure 2.1 – Schematic representation of the different types of bonding and non-bonding interactions: these being (a) bond stretching, (b) angle bending, (c) torsional angle rotation, (d) electrostatic interactions and (e) van der Waals interactions (adapted from Leach[54] and Lewars[58]).

Note that all bonded terms are described by Hooke's law type functions, i.e. a simple quadratic form. The non-bonded van der Waals term is described by a Lennard-Jones[54] potential while the non-bonded electrostatic term is described by the standard coulomb charge expression (See Equation 2.2).

The energy function implemented in the CHARMM[48] program fits the general force field description[48] but also includes a Urey-Bradley term which models angle-bending using the non-bonded interactions between atoms separated by two covalent bonds.

$$U(R^N) = \sum_{Bonds} k_b (b - b_0)^2 + \sum_{Urey-Bradley} k_{UB} (S - S_0)^2 + \sum_{Angles} k_\theta (\theta - \theta_0)^2 +$$
$$\sum_{Dihedrals} k_\chi [1 + \cos(n\chi - \gamma)] + \sum_{\text{Im } propers} k_{\text{Im } p} (\varphi - \varphi_0)^2 + \qquad \text{Eq 2.2}$$
$$\sum_{Non-bonded} \varepsilon_{ij} [(\frac{R_{\min ij}}{r_{ij}})^{12} - (\frac{R_{\min ij}}{r_{ij}})^{6}] + \frac{q_i q_j}{\in_l r_{ij}}$$

R^N is the set of all molecular coordinates upon which the potential (U) depends. The first term of Equation 2.2 models deformations from the equilibrium bond length b_0 for a bond b where K_b is the bond force constant. The Urey-Bradley deformation term contains S, the distance between the two atoms separated by two covalent bonds, S_0 the equilibrium distance and K_{UB} as the force constant. This term is often purposefully not implemented by setting K_{UB} to zero. Similarly for the angle and improper torsional angle terms, θ is the bond angle, θ_0 is the equilibrium bond angle, K_θ is the angle force constant, φ is the improper dihedral angle, φ_0 is the equilibrium value of the improper dihedral and K_{imp} is the planar deviation force constant. The dihedral term describes the energy associated with rotation about bond torsional angles, where χ is the torsional angle, γ the phase angle, n the multiplicity and K_χ the dihedral force constant. The final term is the non-bonded term which includes van der Waals interactions modelled by the Lennard-Jones potential and electrostatic interactions modelled by the Coulomb potential[54]. Pairs of atoms in the same molecule that are at least three bonds away and those in different molecules are included in this non-bonded sum. $R_{mini,j}$ is the radius for two atoms i and j at which the interaction energy is at a minimum, r_{ij} is the distance between atoms i and j and ε_{ij} is the Lennard-Jones well depth. For the Coulomb potential, q_i and q_j are the charges on atoms i and j where $\in_l$ is the dielectric constant of the medium the simulation is carried out in.

Force Fields only take pairwise interactions into account. Three molecules A, B and C are only described by the pairwise interactions A-B, A-C and B-C. The effect of the three-way A-B-C interaction is not explicitly calculated and instead it is parameterised into the pair potential. This parameterisation also applies for many-body interactions, thus the term effective pair potential is used. More sophisticated Force Field definitions exist, the reader is directed to *Molecular Modelling*[54] for further information on these Class II force fields.

Calculating the energy of a static system configuration with MM methods is a quick procedure but choosing the "correct" Force Field is a vital decision. Each of the many MM packages available (AMBER[59], CHARMM[48], GROMOS[60]) tends to have its own generally applicable Force Field. Force Fields are often parameterised for a certain subset of molecules and thus a Force Field applicable to the system of interest must be chosen.

2.1.2 Carbohydrate models

Available carbohydrate Force Fields include the AMBER[61-64], CHARMM[65-67] and GROMOS[68-70] Force Fields. These Force Fields reproduce experimental data in varying degrees of accuracy. They can be exceptionally poor in molecules which deviate drastically from the "training set" and where electronic effects may dominate, e.g. addition of resonance containing or highly charged groups into carbohydrates such as the N-acetyl group. The carbohydrate solution Force Field (CSFF) was recently developed[71] to improve upon the CHARMM carbohydrate Force Field[65,66]. The new CSFF improved the primary alcohol rotation frequency to agree more accurately with experimental NMR data[16,43,44]. This is important for studies of oligo- or polysaccharides with 1→6 linkages. The CSFF was implemented in preliminary studies leading up to the work in this dissertation.

2.1.3 Water models

Since chemical reactions often occur in solvent, it is sensible to properly describe the solvent using an appropriate model. Other than being the most abundant solvent on earth, water is also prevalent in biological systems where much computational and experimental research is focused. Sensibly it has many models attributed to describing it. These range from simple models such as TIP3P[72,73] and SPC/E[74] to polarisable and *ab initio* water models[54].

2.1.3.1 Explicit models - 'Simple' rigid water models

Here individual molecules are considered explicitly. The 'simple' water models are most commonly used as they are computationally efficient and provide a reasonable level of accuracy. Figure 2.2 illustrates some of the most commonly used rigid water models.

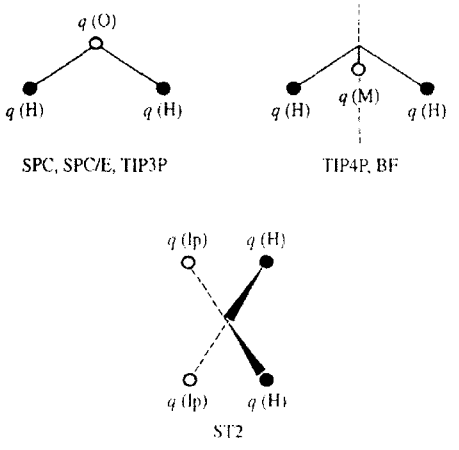

Figure 2.2 – Illustrations of popular water models with charges indicated (Extracted from Leach[54,73])

Common features in these models include a rigid water geometry and 3 (or 5) interaction sites. The TIP3P[72,73] water (available in CHARMM[48]) used in this study has 3 electrostatic interaction sites. A partial negative charge on the oxygen(q(O)) is balanced by the two positive charges on the hydrogen atoms (q(H)). Van der Waals interactions in this model are only accounted for by a Lennard-Jones potential on the oxygen atoms[54].

Table 2.1 compares water model parameters. Values compared include the charges (q), bond lengths (r) and the HOH angle as well as A and C Lennard-Jones parameters for water's oxygens that have been optimised to yield reasonable structural and enegetic results for a water dimer[73]. All of these models produce reasonable results[54].

Table 2.1: A comparison of various water models parameters[54,73]

	SPC	SPC/E	TIP3P	BF	TIP4P	ST2
r(OH), Å	1.0	1.0	0.9572	0.96	0.9572	1.0
HOH, deg	109.47	109.47	104.52	105.7	104.52	109.47
$A \times 10^{-3}$, kcal Å^{12}/mol	629.4	629.4	582.0	560.4	600.0	238.7
C, kcal Å^6/mol	625.5	625.5	595.0	837.0	610.0	268.9
q(O)	−0.82	−0.8472	−0.834	0.0	0.0	0.0
q(H)	0.41	0.4238	0.417	0.49	0.52	0.2375
q(M)	0.0	0.0	0.0	−0.98	−1.04	−0.2375
r(OM), Å	0.0	0.0	0.0	0.15	0.15	0.8

These rigid models are quick and sufficient for modelling most systems. More complex models such as polarisable models and *ab initio* models are also available. More information on these topics is available in the *Encyclopaedia of Computational Chemistry*[15].

2.1.3.2 Implicit solvent models

The final solvent model to be considered is that in which individual solvent molecules are not explicitly modelled. Rather, a field represents the screening of the solvent. For solutes in which specific interactions (like hydrogen bonding) to solvent are expected to occur, implicit models are unlikely to represent this interaction very well. Implicit solvent models are not expected to perform particularly well for carbohydrates although they have been utilised[45].

2.2 General background to Quantum Mechanics

Quantum Mechanics (QM) or electronic structure methods[45] are considered the most accurate for describing chemical systems. They explicitly model electrons in atomic and molecular systems using an implementation of the Schrödinger equation. Electronic

structure programs that can be used include Gaussian[75] and GAMESS-UK[49]. The latter[49] was used exclusively in this project.

The time-dependent or relativistic Schrödinger equation can be written as:

$$\left\{\frac{-\hbar^2}{2m}\nabla^2 + \hat{V}\right\}\Psi(\vec{r},t) = \frac{i\hbar\partial\Psi(\vec{r},t)}{\partial t} \qquad \text{Eq 2.3}$$

where ∇^2 (del squared) is:

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \qquad \text{Eq 2.4}$$

and $\vec{r} = (r, R)$ (i.e. is a function of nuclear and electronic coordinates).

A number of postulates are necessary to solve the Schrödinger equation[55]. If the potential energy V does not depend on time we can consider the time independent equation. Using the separation of variables technique:

$$\Psi(\vec{r},t) = \Psi(\vec{r})\tau(t) \qquad \text{Eq 2.5}$$

and for time independent systems we can solve the spatial Schrödinger equation

$$\left\{\frac{-\hbar^2}{2m}\nabla^2 + V\right\}\Psi(\vec{r}) = E\Psi(\vec{r}) \qquad \text{Eq 2.6}$$

which only depends on the spatial terms (i.e. coordinates).

Equation 2.6 can be further abbreviated to

$$\hat{H}\Psi = E\Psi \qquad \text{Eq 2.7}$$

if we denote the Hamiltonian operator $\hat{H}$ as :

$$\hat{H} = \left\{ \frac{-\hbar^2}{2m} \nabla^2 + \hat{V} \right\}. \qquad \text{Eq 2.8}$$

The Hamiltonian operator contains a kinetic energy and potential energy contribution. The kinetic energy operator $\hat{T}$ is

$$\hat{T} = \frac{-\hbar^2}{2m} \nabla^2 \qquad \text{Eq 2.9}$$

, while the potential energy operator $\hat{V}$ is given by:

$$\hat{V} = \frac{1}{4\pi\varepsilon_0} \left(- \sum_{i}^{electrons} \sum_{I}^{nuclei} \left(\frac{Z_I e^2}{|R_I - r_i|} \right) + \sum_{i}^{electrons} \sum_{j<i}^{electrons} \left(\frac{e^2}{|r_i - r_j|} \right) + \sum_{I}^{nuclei} \sum_{J<I}^{nuclei} \left(\frac{Z_I Z_J e^2}{|R_I - R_j|} \right) \right) \qquad \text{Eq 2.10}$$

where e is the charge and Z the atomic number. The summations in Equation 2.10 represent electron-nuclei attraction, electron-electron repulsion and nuclear-nuclear repulsion. The potential expression consists of electrostatic interactions (Coulomb interactions) between nuclei, between electrons as well as between electrons and nuclei.

If the Hamiltonian operator is applied to the wavefunction (Ψ) of a molecule an energy eigenvalue is obtained, i.e. a solution to the wave equation. However, when solving anything more than the most simple of systems, e.g. particle in a box, harmonic oscillator, particle on a ring, particle on a sphere and the hydrogen atom it is only possible to approximate the correct answer. Common simplifications in these simple systems are restrictions known as boundary conditions[55]. To solve the wave equation for large systems a number of postulates must be made[55].

2.2.1 The Born-Oppenheimer Approximation

The first postulate is the neglect of nuclear kinetics or the Born-Oppenheimer approximation[76]. This is reasonable as electronic motion is much greater than nuclear

motion and the electrons are of lighter mass than the protons. We can expect the electrons to 'instantaneously' adjust to the nuclear coordinates of a molecule[54,58,77]. The effective energy is thus:

$$\hat{H}^{elec}\Psi^{elec}(r,R)=E^{eff}(R)\Psi^{elec}(r,R) \qquad \text{Eq 2.11}$$

where the wavefunction depends on the positions of the nuclei (R) and electrons (r). The electronic Hamiltonian is as before a sum of kinetic and potential contributions:

$$\hat{H}^{elec}=\hat{T}^{elec}+\hat{V} \qquad \text{Eq 2.12}$$

where the kinetic energy is

$$\hat{T}^{elec}=-\left(\frac{h^2}{8\pi^2 m}\right)\sum_{i}^{electrons}\left(\frac{\partial^2}{\partial x_i^2}+\frac{\partial^2}{\partial y_i^2}+\frac{\partial^2}{\partial z_i^2}\right) \qquad \text{Eq 2.13}$$

and the potential is as shown in Equation 2.10.

2.2.2 Molecular orbital theory

Extension of the wavefunction to molecules requires the introduction of molecular orbital (MO) theory. The complete wavefunction is approximated by molecular orbitals. Each molecular orbital depends on spatial coordinates as well as the spin of the two electrons it contains.

The spatial orbital function $\psi(x,y,z)$ describes the Cartesian space accessible to electrons[15,76]. The electron spin quantum number introduces electron spin as a spin coordinate ξ which may be $\frac{1}{2}$ or $-\frac{1}{2}$ (that is "spin up" or "spin down"). Spin along the positive z axis is $\alpha(\xi)$ while negative spin is $\beta(\xi)$ where

$$\alpha\left(\frac{1}{2}\right)=1 \qquad \alpha\left(-\frac{1}{2}\right)=0 \qquad \beta\left(\frac{1}{2}\right)=0 \qquad \beta\left(-\frac{1}{2}\right)=1.$$

For a single electron, a spin orbital $\chi(x,y,z,\xi)$ is the product of the spatial orbital and the spin function:

$$\chi(x,y,z,\xi)=\psi(x,y,z)\alpha(\xi) \text{ or } \chi(x,y,z,\xi)=\psi(x,y,z)\beta(\xi). \qquad \text{Eq 2.14a, b}$$

For an N-electron wavefunction the product of spin orbitals is constructed:

$$\Psi = \Psi(\chi_1(X_1),\chi_2(X_2).....\chi_N(X_N)) \qquad \text{Eq 2.15}$$

where X represents both spatial and spin coordinates. This is the *hartree* product or uncorrelated wavefunction; the electrons motions represented here are completely independent of each other[77].

2.2.3 Antisymmetry and the determinantal wavefunction

A many-electron wavefunction must be antisymmetric with respect to the interchange of two electrons[76]. Thus interchanging electrons i and j must give

$$\Psi(\chi_1(X_1),....,\chi_i(X_i),....,\chi_j(X_j),....,\chi_N(X_N)) = -\Psi(\chi_1(X_1),....,\chi_i(X_j),....,\chi_j(X_i),....,\chi_N(X_N))$$
. Eq 2.16

Let us regard a two electron example with two spin orbitals χ_i, χ_j and electrons i and j. An electron i can be placed into orbital i and similarly for electron j or an electron i can be placed into orbital j and similarly for electron j. This gives

$$\Psi_{i,j}(X_i,X_j) = \chi_i(X_i)\chi_j(X_j) \text{ and } \Psi_{j,i}(X_i,X_j) = \chi_i(X_j)\chi_j(X_i). \qquad \text{Eq 2.17 a, b}$$

A wavefunction which obeys the antisymmetry principle and provides indistinguishable electrons can be formed by a linear combination of Hartree products. This wavefunction is

$$\Psi_{i,j}(X_i, X_j) = 2^{-\frac{1}{2}}(\chi_i(X_i)\chi_j(X_j) - \chi_i(X_j)\chi_j(X_i)) \qquad \text{Eq 2.18}$$

where $2^{-\frac{1}{2}}$ is a normalisation constant.

Clearly,

$$\Psi_{i,j}(X_i, X_j) = -\Psi_{j,i}(X_i, X_j) \qquad \text{Eq 2.19}$$

holds true.

Equation 2.18 can be more conveniently expressed as a determinant:

$$\Psi_{i,j}(X_i, X_j) = 2^{-\frac{1}{2}}\begin{vmatrix}\chi_i(X_i) & \chi_j(X_i)\\ \chi_i(X_j) & \chi_j(X_j)\end{vmatrix} \qquad \text{Eq 2.20}$$

, which is known as a **Slater Determinant.**

For an N-electron wavefunction the Slater determinant is :

$$\Psi_{i,j}(X_1, X_2, ..., X_N) = (N!)^{-\frac{1}{2}}\begin{vmatrix}\chi_i(X_1) & \chi_j(X_1) & \cdots & \chi_N(X_1)\\ \chi_i(X_2) & \chi_j(X_2) & \cdots & \chi_N(X_2)\\ \vdots & \vdots & & \vdots\\ \chi_i(X_N) & \chi_j(X_N) & \cdots & \chi_N(X_N)\end{vmatrix} \qquad \text{Eq 2.21}$$

Introduction of the determinantal form includes some useful determinant characteristics. Specifically, interchanging the coordinates of two electrons means interchanging two rows of the determinant. The sign of the determinant changes and Slater determinants obey the antisymmetry rule. Furthermore, if two electrons occupy the same *spin orbital* then two columns in the determinant are equal making the determinant zero. This agrees with the Pauli exclusion principle, which states that two electrons of the same spin may not occupy the same molecular orbital[76,78].

A convenient short hand notation for a normalised Slater Determinant is

$$\Psi_{i,j}(X_1, X_2, ..., X_N) = \left| \chi_i(X_1)\chi_j(X_2)\cdots\chi_N(X_N) \right\rangle \qquad \text{Eq 2.22}$$

where the normalisation constant is inclusive and only the diagonal elements of the determinant are shown.

2.2.4 Properties of the molecular orbitals

Two important properties of the MO's is that they are *orthogonal* and *normalised*[77].

For orthogonality:

$$S_{ij} = \int \psi_i^* \psi_j \, dxdydz = 0 \qquad for \quad i \neq j \qquad \text{Eq 2.23}$$

and for normalisation:

$$S_{ii} = \int \psi_i^* \psi_i \, dxdydz = 1 \qquad \text{Eq 2.24}$$

where S is called the overlap integral and * denotes the complex conjugation (especially significant if the molecular orbitals are not real functions).

A more convenient expression for the overlap integral is

$$S_{ij} = \int \psi_i^* \psi_j \, dxdydz = \delta_{ij} \qquad \text{Eq 2.25}$$

where δ_{ij} is the Kronecker delta which is 1 if i=j and 0 otherwise.

2.2.5 Linear combination of atomic orbitals (LCAO) and molecular orbital (MO) coefficients

In practice (that is computationally) each molecular orbital is written as a linear combination of one-electron functions. N one-electron *basis functions* form a basis set. We describe each orbital ψ_i as

$$\psi_i = \sum_{\mu=1}^{N} c_{\mu i}\phi_\mu \qquad \text{Eq 2.26}$$

where ϕ_μ are the N basis functions and $c_{\mu i}$ are the molecular orbital expansion coefficients.

Generally, a set of basis functions are associated with each nucleus. In other words, atomic orbitals are represented by a set of basis functions[78]. There are two types of atomic basis function, Slater-type atomic orbitals (STO's) and Gaussian-type atomic orbitals (GTO's)[57,78].

The Slater-type orbitals provide a good description of the atomic orbitals but are not computationally efficient. They must be solved numerically and are cumbersome to deal with computationally. The normalised form of the first STO is

$$\phi_{1s} = \left(\frac{\zeta_1^3}{\pi}\right)^{\frac{1}{2}} e^{-\zeta_1 r} \qquad \text{Eq 2.27}$$

where r is the radial coordinate and ζ_1 is a constant determining the orbital size[78].

The Gaussian-type orbital functions are computationally more amenable as their integrals can be evaluated explicitly. The s and p orbitals are shown below:

$$g_s = \left(\frac{2\alpha}{\pi}\right)^{\frac{3}{4}} e^{-\alpha r^2} \text{ (Represents an s-type orbital)} \qquad \text{Eq 2.28}$$

$$g_x = \left(\frac{128\alpha^5}{\pi^3}\right)^{\frac{1}{4}} xe^{-\alpha r^2}$$ (Represents a p_x-type orbital) Eq 2.29

$$g_y = \left(\frac{128\alpha^5}{\pi^3}\right)^{\frac{1}{4}} ye^{-\alpha r^2}$$ (Represents a p_y-type orbital) Eq 2.30

$$g_z = \left(\frac{128\alpha^5}{\pi^3}\right)^{\frac{1}{4}} ze^{-\alpha r^2}$$ (Represents a p_z-type orbital) Eq 2.31

where α is a constant determining the orbital size[78].

Two problems with the Gaussian-type orbital (GTO) approach exist. The cusp of zero electron probability at the origin is not represented by these functions and the Gaussian functions for d-, f- and higher orbitals do not have the angular symmetry of atomic orbitals [78]. These problems are solved by simply forming appropriate linear combinations of the Gaussians to represent the atomic orbitals[78].

For a general discourse on the postulates and other fundamental aspects of quantum mechanics and quantum chemistry several texts are available (*Atkins' Physical Chemistry*[55]; *Molecular quantum mechanics*[79]; *Modern Quantum Chemistry*[76])

Before describing in more depth how the wave equation is calculated consider Figure 2.3.

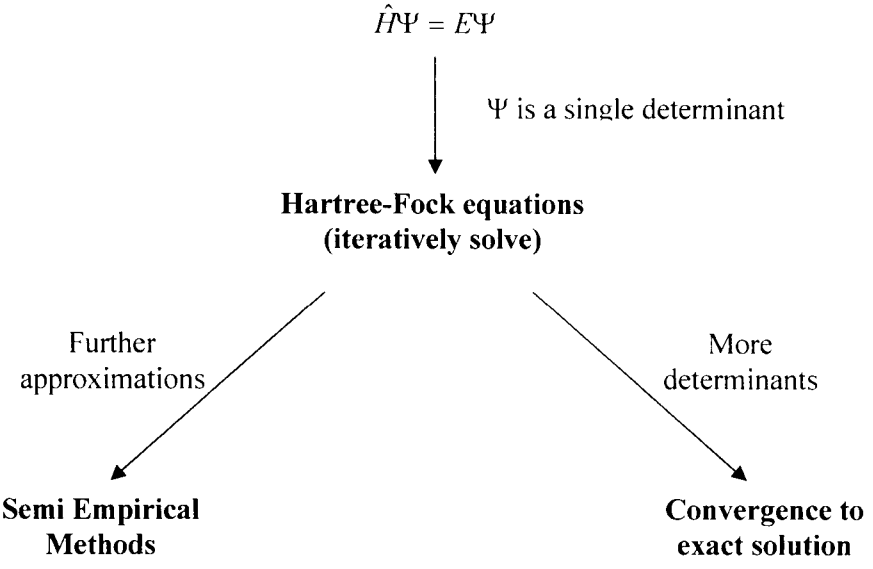

Figure 2.3 – Three ways in which the Hartree-Fock (HF) equations can be applied. The most approximate being Semi Empirical methods, intermediate is single-determinant HF and addition of determinants leads to the most accurate or exact solution (Modified from Jensen[77]).

This schematic shows how the wave equation is solved. For a single determinant we apply the Hartree-Fock model (described in the next section) to obtain a solution. If more approximations are made, a system can be modelled with the more approximate Semi Empirical treatment. Convergence to the exact solution is achieved upon addition of more determinants. The accuracy of a QM calculation will depend on the level of theory chosen (c.f. MM methods).

2.2.6 The Hartree-Fock Model

Two important ideas in this model are the variational condition and the Roothaan-Hall equation.

2.2.6.1 The variational condition

The variational condition stipulates that if a guess of the wavefunction is made, an energy that is greater or equal to the "true" energy will be calculated. More formally solving the Schrödinger equation by integrating over all space with a trial wavefunction Ξ gives an expectation value for the energy E':

$$E' = \frac{\int \Xi^* \hat{H} \Xi d\tau}{\int \Xi^* \Xi d\tau} . \qquad \text{Eq 2.32a}$$

For a normalised trial wavefunction (as is usually the case):

$$E' = \int \Xi^* \hat{H} \Xi d\tau . \qquad \text{Eq 2.32b}$$

If the guess at the wavefunction Ξ is the true wavefunction Ψ then the actual energy is calculated:

$$E = \int \Psi^* \hat{H} \Psi d\tau \qquad \text{Eq 2.33}$$

and the following applies:

$$E' = \int \Xi^* \hat{H} \Xi d\tau > E . \qquad \text{Eq 2.34}$$

Furthermore, as the trial wavefunction is improved upon, the true energy is more closely approximated (See Equation 2.34). These approximate energies are always greater than or equal to the true energy[80].

2.2.6.1.1 Minimising the energy

The best solution to a single-determinant wavefunction is found by minimising the energy with respect to the molecular orbital coefficients (previously described in Equation 2.26) these are the variational equations:

$$\frac{\partial E'}{\partial c_{\mu i}} = 0 \qquad \forall \mu, i \quad . \qquad \text{Eq 2.35}$$

2.2.6.2 Solution to the molecular orbital coefficients - The Roothaan-Hall equations

The Roothaan-Hall equations are derived from the variational equations. These equations have been derived for closed-shell systems. Closed shell systems are those with no unpaired electrons. The Hartree-Fock theory with a closed-shell single determinant is used for molecules in their ground state. This is termed Restricted Hartree-Fock (RHF) theory. The Roothaan-Hall equations are

$$\sum_{\nu=1}^{N}(F_{\mu\nu} - \epsilon_i S_{\mu\nu})c_{\nu i} = 0 \qquad \mu = 1,2,...,N \qquad \text{Eq 2.36}$$

with the normalisation condition

$$\sum_{\mu=1}^{N}\sum_{\nu=1}^{N} c_{\mu i}^{*} S_{\mu\nu} c_{\nu i} = 1 . \qquad \text{Eq 2.37}$$

The one-electron energy of a molecular orbital ψ_i is ϵ_i. The overlap matrix is an N by N matrix with elements $S_{\mu\nu}$:

$$S_{\mu\nu} = \int \phi_{\mu}^{*}(1)\phi_{\nu}(1)dx_1 dy_1 dz_1 \qquad \text{Eq 2.38}$$

where ϕ are the basis functions.

The Fock matrix is an N by N matrix with elements $F_{\mu\nu}$.

$$F_{\mu\nu} = H_{\mu\nu}^{core} + \sum_{\lambda=1}^{N}\sum_{\sigma=1}^{N} P_{\lambda\sigma}[(\mu\nu \mid \lambda\sigma) - \frac{1}{2}(\mu\lambda \mid \nu\sigma)] \qquad \text{Eq 2.39}$$

The Fock matrix represents the average field effect of all the electrons on each orbital[57].The $H_{\mu\nu}^{core}$ term of the Fock matrix represents the energy of a single electron in a field of "bare" nuclei [78].

$$H_{\mu\nu}^{core} = \int \phi_{\mu}^{*}(1)\hat{H}^{core}(1)\phi_{\nu}(1)dx_1 dy_1 dz_1,$$
$$\hat{H}^{core}(1) = -\frac{1}{2}\left(\frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial y_1^2} + \frac{\partial^2}{\partial z_1^2}\right) - \sum_{A=1}^{M}\frac{Z_A}{r_{1A}} \qquad \text{Eq 2.40 a,b}$$

The summation is carried out for all atoms where Z_A is the atomic number for an atom A.

$P_{\lambda\sigma}$ is the one-electron density matrix for the RHF solution. It is defined by a summation over the occupied orbitals as follows:

$$P_{\lambda\sigma} = 2\sum_{i=1}^{occ} c_{\lambda i}^{*} c_{\sigma i} \qquad \text{Eq 2.41}$$

The two-electron repulsion integrals are the quantities (μν|λσ),(μλ|νσ) from Equation 2.39 where

$$(\mu\nu \mid \lambda\sigma) = \int\int \phi_{\mu}^{*}(1)\phi_{\nu}(1)\left(\frac{1}{r_{12}}\right)\phi_{\lambda}^{*}(2)\phi_{\sigma}(2)dx_1 dy_1 dz_1 dx_2 dy_2 dz_2 \qquad \text{Eq 2.42}$$

and similarly for (μλ|νσ).

The Roothaan-Hall equations can be represented in matrix form as

$$FC = SC\varepsilon$$ Eq 2.43

where F is the Fock matrix, S contains the overlap elements of the basis functions, C contains the molecular coefficients and ε is a diagonal matrix containing the one-electon orbital energies ε_i of each molecular orbital χ_i.

Both the Fock matrix (because of the density matrix) and the orbitals depend upon the molecular orbital coefficients which means that the matrix equation is not linear. The matrix equation must be solved iteratively[57].

2.2.6.3 Self-Consistent Field

The iterative procedure used to solve the Roothaan-Hall equations is termed the Self-Consistent Field (SCF) method. Upon iteratively converging to the minimum energy the orbitals generate a field which will produce the same orbitals – that is they are self-consistent[57].

The SCF process is represented in Figure 2.4. A more rigorous illustration (showing the computational resources required) is available in *Ab initio molecular orbital theory* on page 53[78].

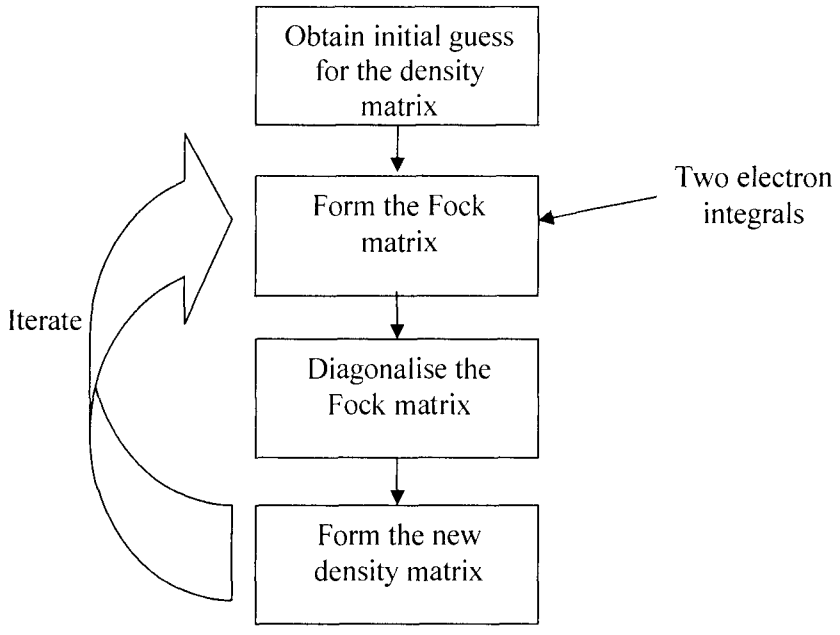

Figure 2.4 – Illustration of the SCF process (modified from Jensen[77])

2.2.6.4 Open Shell systems

Radicals or other systems with unpaired electrons are called open shell systems. These can be solved using the Pople-Nesbet instead of the Roothaan–Hall equations (Refer to *Modern Quantum Chemistry*[76]). Spin-restricted open-shell Hartee-Fock (ROHF) and Spin-unrestricted Hartree-Fock (UHF) are the two approaches available using the Roothaan–Hall equations and the Pople-Nesbet equations respectively. These approaches are not applicable to simple neutral carbohydrates and further information about the description of such systems is available in several texts[15,54,77,78].

2.2.7 Choice of basis set

To solve the total electronic wavefunction the approximate wavefunction has been expressed as a combination of orbitals (Refer to Section 2.2.5). A mathematical description for each orbital can be formulated and included into a set, the basis set[57].

The probability of finding an electron is unity if an integration is carried out over all space. In other words, the electron has a finite probability of existing anywhere in space. A basis set describes the orbitals available electrons. The less restrictive these orbitals, the more accurately the electrons are described[57].

Mathematically, almost any function can be a basis function. A very simple case is choosing a basis set for 3-Dimensional Cartesian space. Three orthonormal vectors are the x-, y- and z-unit vectors and these form a basis set. With the appropriate constants to form linear combinations of the x-, y- and z-unit vectors, the entirety of Cartesian space can be traversed – this is a mathematically complete basis set.

In a chemical sense, it is impossible to form a complete set of functions for the electrons! Both mathematical and physical constraints exist and a "physically sensible" group of functions must be chosen[81].

Four types of basis sets will be discussed: minimal, split-valence, polarised and diffuse basis sets[57].

2.2.7.1 Minimal basis sets

These have the minimum number of basis functions needed to describe all the electrons in a system. The set is "stripped down" in the interest of performance. STO-nG is such a set. STO refers to Slater-type orbitals (discussed earlier) while the nG refers to the number (n) of Gaussian functions used to define each STO (As mentioned earlier STO's are not easy to use numerically - a linear combination of Gaussians is used to approximate each STO).

2.2.7.2 Split-valence basis sets

To better represent the vector space available to electrons, the number of basis functions is increased for each valence orbital of an atom. Two or more sizes of basis function can exist

for each valence orbital. 6-31G and 6-311G are split-valence sets with two and three sizes of valence basis functions respectively (See Figure 2.5)[57]. The most important feature of this basis set is that the core and valence orbitals are not treated in the same way. That is core orbitals have one size of basis function whereas valence orbitals have more than one size of basis function, e.g. two valence basis functions for 6-31G.

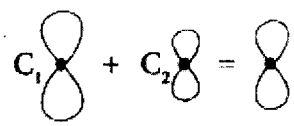

Figure 2.5– Shows the combinations of split-valence orbitals (Extracted from Frisch[57])

2.2.7.3 Polarised basis sets

Polarisation of orbitals gives them an angular momentum beyond that of the ground state. This allows orbitals of a different shape. A polarised basis commonly used for medium sized molecules is the 6-31G(d) or 6-31G* basis[57]. For a carbon atom this introduces d-functions. Further polarisation to 6-31G(d,p) also named 6-31G** adds d-functions for heavy atoms and p-functions to hydrogen atoms (See Figure 2.6)[57].

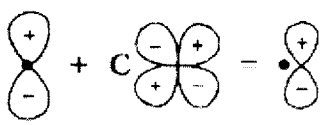

Figure 2.6 – Shows the effect of polarisable orbitals (Extracted from Frisch[57])

2.2.7.4 Diffuse basis sets

In systems with lone pairs or anions a better representation is achieved if a larger region of space is accessible to the electrons. Diffuse basis sets contain diffuse functions which are larger versions of s and p type functions. 6-31+G and 6-31++G contain diffuse functions

for heavy atoms and heavy and light atoms respectively, where the symbol + indicates a diffuse function. Diffuse functions on light atoms such as hydrogen seldom make a significant difference in accuracy (See Figure 2.7)[57].

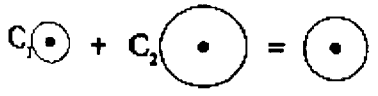

Figure 2.7 – Shows the effect of addition of diffuse functions (Extracted from Frisch[57])

2.2.7.5 Basis Set Superposition Error (BSSE)

The classic example used to describe BSSE is by looking at the hydrogen-bonded water dimer[54,58].Two water molecules react to form the product which is the hydrogen-bonded dimer. The energy of dimer formation could logically be calculated as $E_{products}$-$E_{reactants}$ = E_{dimer}-$E_{2\ infinitely\ separated\ water\ molecules}$. In reality the energy calculated tends to be incorrect. Increasing the size the basis set used in a calculation usually negates this problem (reduces this error to a minimum). This type of error is only expected to occur with minimal basis sets[54] such as (STO-nG) and although BSSE has been applied in certain studies[82], the 6-31G* basis set was found to be adequate at describing monosaccharides[83]. BSSE is mentioned for the sake of completeness.

2.2.8 Electron correlation

Previously Hartree-Fock (HF) theory and the self-consistent field (SCF) method were discussed. Hartree-Fock theory describes electrons as moving in an average field (potential) of other electrons[54]. The interaction between electrons is actually more 'correlated' than this theory describes. This lack of description for electron-correlation is a major flaw of Hartree-Fock theory.

Popular methods which include electron correlation are Configuration Interaction (CI), Many-body perturbation methods and Density Functional Theory (DFT). Many-body

perturbation and CI are traditional post SCF methods[57] in the sense that they are applied after the Hartree-Fock implementation. DFT is considered in the most detail although it is not truly a post Hartree-Fock method. Instead electron correlation is included implicitly in the SCF procedure. Further information on all of these methods is available in the *Encyclopaedia of Computational Chemistry*[15] and *Modern Quantum Chemistry*[76].

2.2.8.1 Many-body Perturbation Theory

Møller and Plesset introduced this method in which the Hartree-Fock approximation is corrected by adding small perturbations to the wavefunction[54]. The 'true' Hamiltonian can be written as the sum of a zeroth-order Hamiltonian and a perturbation V. This is parameterised with λ which varies between 0 and 1 such that it is possible to make gradual improvements to the true Hamiltonian (the calculated one).

$$H = H_0 + \lambda V \qquad \text{Eq 2.44}$$

Expressing the eigenfunctions and eigenvalues of the Hamiltonian as functions of λ gives:

$$\Psi_i = \Psi_i^{(0)} + \lambda\Psi_i^{(1)} + \lambda^2\Psi_i^{(2)} \cdots = \sum_{n=0} \lambda^n \Psi_i^{(n)} \qquad \text{Eq 2.45}$$

$$E_i = E_i^{(0)} + \lambda E_i^{(1)} + \lambda^2 E_i^{(2)} \cdots = \sum_{n=0} \lambda^n E_i^{(n)} \qquad \text{Eq 2.46}$$

where $E_i^{(1)}$ is the first order correction to the energy and so on. The term MPn is used to describe the use of Møller-Plesset theory with an n-th order correction. Although popular, this method is not variational leaving the possibility that energies lower than the true energy may be calculated[54,76].

2.2.8.2 Configuration Interaction (CI)

CI is based upon the variational method. Instead of solving the wave equation with a single ground state Slater determinant a number of excited Slater determinants are included[77]. Using a set of K basis function implies 2K spin orbitals. If N electrons are used to fill the 2K orbitals (N<2K) then 2K-N virtual unoccupied orbitals exist[54]. The excited states represent the replacement of occupied spin orbitals by virtual spin orbitals where the total number of permutations is (2K!)/N!(2K-N)! It is only possible to consider all of these permutations (full-configuration interaction) for small molecules[54]. The number of permutations can be truncated to only include single orbital permutations – that is configuration interaction singles (CIS). Further information on CI is available in the following texts: *Molecular Modelling*[54] and *Modern Quantum Chemistry*[76].

Full CI is the most complete description possible within the restrictions of a specific basis set. This point is illustrated in Figure 2.8.

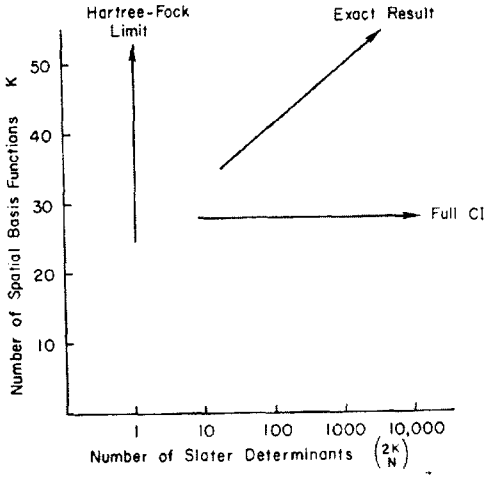

Figure 2.8 – Illustrates the dependence of a true solution on the number of basis functions and the number of Slater determinants used. (Extracted from Ostlund[76])

2.2.8.3 Density Functional Theory

DFT methods are very similar to *ab initio* Hartree-Fock methods. Instead of solving for the N-electron wavefunction they instead aim to determine the energy as a function of the overall electron density. The equivalent to the HF Roothaan-Hall equations are the Kohn-Sham equations.

This can be considered a more "realistic" approach as the electron density is a measurable quantity, unlike the wavefunction. The efforts of Hohenberg and Kohn gave this theory the rigour which has led to its more common usage[84]. The ground-breaking theorems proven by them are called the existence and variational theorems[85].

The existence theorem states that there is a unique connection between the ground state energy and the ground state electron density function[57,84]. This connection is by means of a functional – a function whose definition is itself a function[57]. This is illustrated below

$$E[\rho(r)] = \int V_{ext}(r)\rho(r)dr + F[\rho(r)] \qquad \text{Eq 2.47}$$

where the energy E depends on a function of the electron density which a function of r (denoted [ρ(r)]). The external potential $V_{ext}(r)$ is the Coulomb interaction with the nuclei and $F[\rho(r)]$ is the sum of kinetic energy of the electrons and inter-electronic contributions.

The existence and variational theorem are existence rather than constructive proofs[81]. All that is known is that a functional exists that connects the energy to the electron density. A description for $F[\rho(r)]$ is also not known. Kohn and Sham[85] suggested that the total electronic energy must consist of the following terms:

$$E = E^{T} + E^{V} + E^{J} + E^{XC} \qquad \text{Eq 2.48}$$

where E is the total energy previously denoted $E[\rho(r)]$, E^T is the kinetic energy due to electron motion, E^V is the potential energy of nuclei-electron attraction and nuclei-nuclei

repulsion, E^J is the electron-electron repulsion term and E^{XC} is the exchange correlation term describing the effects neglected by the HF equations[57].

E^V is already known and was previously described as $\int V_{ext}(r)\rho(r)dr$. In fact, what Kohn and Sham's work gives is a value for $F[\rho(r)]$ which consists of the electronic kinetic energy, electron-electron Coulombic energy and both exchange and correlation electron contributions[54]. Expressing $F[\rho(r)]$ and its constituent terms in terms of the electron density gives

$$F[\rho(r)] = E_{KE}[\rho(r)] + E_H[\rho(r)] + E_{XC}[\rho(r)] \qquad \text{Eq 2.49}$$

where $E_{KE}[\rho(r)]$ is the kinetic energy, $E_H[\rho(r)]$ is the electron-electron repulsion energy and $E_{XC}[\rho(r)]$ is the exchange and correlation contribution to the energy[54].

The $E_{KE}[\rho(r)]$ and $E_H[\rho(r)]$ functionals are chosen to be the following

$$E_{KE}[\rho(r)] = \sum_{i=1}^{N} \int \psi_i(r)(-\frac{\nabla^2}{2})\psi_i(r)dr \qquad \text{Eq 2.50}$$

$$E_H[\rho(r)] = \frac{1}{2}\int\int \frac{\rho(r_1)\rho(r_2)}{|r_1 - r_2|} dr_1 dr_2 \qquad \text{Eq 2.51}$$

where $E_{KE}[\rho(r)]$ describes a system of non-interacting electons and $E_H[\rho(r)]$ is the Hartree electrostatic energy – the sum of all pairwise electrostatic interactions.

The functional that describes the exchange and correlation must be approximated and is usually divided into exchange and correlation parts. The exchange interactions are those due to same-spin interactions while correlational are mixed-spin interactions (See Equation 2.52)[57].

$$E_{XC}[\rho(r)] = E_X[\rho(r)] + E_C[\rho(r)] \qquad \text{Eq 2.52}$$

2.2.8.3.1 Functionals for exchange and correlation

The functionals used are approximate and at present there is not a rigorous way to improve upon them[15,54,77]. Two types of functionals exist: Traditional functionals and Hybrid Functionals.

2.2.8.3.1.1 Traditional functionals

Traditional functionals differ in the way they treat exchange and correlation. Local functionals are based on the electron spin densities (ρ) while Gradient-corrected functionals depend on the electron spin densities (ρ) as well as their gradient ($\nabla\rho$). Local density approximations (LDA) treat the density as a uniform electron gas while the more advanced gradient methods treat the density as a non-uniform electron gas. Popular gradient-corrected functionals include the Becke exchange; Lee, Yang, Parr (LYP) correlation; Perdew-Wang (PW) and Vosko, Wilk, Nusair (VWN) functionals[57,86].

2.2.8.3.1.2 Hybrid functionals

Hybrid functionals linearly combine (or mix) the Hartree-Fock exchange with linear and gradient-corrected exchange terms and correlation terms. Becke's three parameter functional including LYP correlation (B3LYP)[87,88] is the most well known and the Becke-style hybrid functional is considered the best to date.

B3LYP is defined as

$$E_{XC}^{B3LYP} = (1-a_0)E_X^{LSDA} + a_0 E_X^{HF} + a_X \Delta E_X^{B88} + a_c E_C^{LYP} + (1-a_c)E_C^{VWN}. \qquad \text{Eq 2.53}$$

The exchange part is composed of local spin density approximated exchange (LSDA), the Hartree-Fock exchange and Becke's original exchange function (Becke-88 or B88). The correlational part comprises the LYP and the VWN correlational functional. The empirically derived coefficients a_0, a_x and a_c are 0.20, 0.72 and 0.81 respectively[57,77].

2.2.8.3.1.3 Specifying the functional and method

DFT calculations are specified by the underlying method e.g. restricted, and then by the two functionals used. A restricted calculation with the Becke-3-parameter exchange functional and the Lee, Yang and Parr correlational functional is described as RB3LYP. If the 6-31G* basis set is being used then the following description applies: RB3LYP/6-31G*. If multiple levels of theory have been applied consecutively the "//" notation is applied with the most recently applied theory to the left, i.e. (most recently applied)//(initially applied theory). If RB3LYP/6-31G* was used to optimise a molecular structure and RB3LYP/6-31+G** was used to calculate the single-point energy then this would be denoted RB3LYP/6-31+G**//RB3LYP/6-31G*.

2.2.8.3.2 The SCF and Kohn-Sham equations

The variational theorem applied here is equivalent to that used in HF methods[81]. If an approximate density function $\rho_{0,\,approximation}$ is chosen, the energy $E_{0,\,approximation}$ calculated is greater or equal to the exact energy E_0[81]. The trial density is based upon the superimposition of atomic densities after which a procedure[77] similar to the one described in Figure 2.4 is followed.

The Kohn-Sham equation is solved in an iterative method analogous to the HF SCF method[57] (that is we improve upon a trial electron density function) where in matrix form this is

$$H^{KS}C = SC\varepsilon \qquad \text{Eq 2.53}$$

and H^{KS} is the Kohn-Sham matrix.

Electron correlation effects are included in each DFT calculation and this will require roughly the same amount of time as a HF calculation.

2.2.9 Relative performance of *ab initio* techniques

The single-point energy of a molecule can be calculated using a variety of techniques and the calculation times compared. Methylamine single point calculations were carried out with the 6-31G* basis set by Radom et al[78].

Radom et al[78] compared the time taken for a variety of methods HF, CI, MP2, MP3 (for more information Refer to *Ab initio molecular orbital theory*[78]). The time for a Hartree-Fock single-point calculation on methylamine was 9 minutes 50 seconds in the 1980's! The ratio of HF:CI:MP2:MP3 was 1:15.0:1.5:3.6. At present these calculations are quicker. However, these ratios conveniently show that a hefty time price is paid when implementing electron correlation techniques.

2.3 Electron Population Analysis

After calculating the energy and wavefunction of a system we can further analyse the wavefunction. Population methods partition the electron density among the nuclei such that a certain number of electrons are associated with each one[54]. Cioslowski[15] emphasised the need for rigorous analysis of electronic wavefunctions. Many of the available analyses such as Mulliken and Löwdin population analysis[54,57] are not applicable to a wide range of molecules or usable in all types of electronic structure methods. The definitions in such analyses are usually arbitrary and both qualitatively and quantitatively of limited usefulness[15,54].

Two types of analysis methods which can be used to rigorously interpret electronic wavefunctions are Bader's theory of Atoms in Molecules[50,51], which gives an electron

density distribution description of a system, and Weinhold's Natural Bond Orbital Theory[53] which allows for orbital analysis of a system.

2.3.1 Quantum-Mechanical Theory of Atoms in Molecules

Bader introduced the concept of Atoms in Molecules[51,89] and provided a rigorous method of dividing the molecular volume into atomic subspace. This method is particularly meaningful and rigorous because the allocation of atomic subspace is not arbitrary but dependent on a topological analysis of the electron density.

The electron density is simply the square of the wavefunction[54,77]

$$\rho(r) = |\Psi(r)|^2 . \qquad \text{Eq 2.54}$$

Each atom consists of a nucleus and a disjoint portion of Cartesian space, the atomic basin(Ω)[15]. The electron density surface is divided into these non-overlapping atomic basins by surfaces (or borders) on which the zero flux condition is observed[90]. That is

$$\nabla\rho(r) \bullet n = 0 \qquad \text{Eq 2.55}$$

where ρ is the molecular electron density and n a vector normal to the surface. The collection of all such points defines the atomic basin.

Atomic populations can be accounted for with:

$$N_A = \int_{\Omega_A} \rho(r) dr \qquad \text{Eq 2.56}$$

where Ω_A is the basin of atom A.

The nuclei act as attractors of the electron density[77]. Similarly, on the inter-atomic basin surfaces are points where the total derivative is zero – these are termed basin attractors.

Basin attractors are minima of the electron density surface in one dimension and are termed (3,-1) or bond critical points. The (3, -1) notation signifies that in 3 dimensions (3,) there is a saddle point of the electron density with a maximum of electron density in 2 directions of space and a minimum in the third direction (,-1). It is possible to trace out the maximum electron density between two nuclear attractors – this is termed a bond path and passes through the basin attractors (bond critical points)[77].

Other critical points that exist include the ring and cage critical points which are found at the centre of a ring and cage respectively. Examples include the centre of a benzene ring and centre of cubane[77]. This topological analysis can be applied in a relatively simple and visualisable way using the AIM2000 program[89,91,92]. This method has been used to study the anomeric effect in the O-C-O unit by Mosquera[93].

2.3.2 Natural Bond Orbital Theory

Chemists often use Molecular Orbital Theory to discuss the shape and interactions of orbitals of a molecule[94]. It is often found that these are not the most "natural" orbitals that describe the molecular wavefunction. Natural orbitals (introduced by Löwdin[53]) describe the *unique* set of orthonormal 1-electron functions $\theta_i(\vec{r})$ that are intrinsic to the N-electron wavefunction $\Psi(1,2,...,N)$. These orbitals are the "best possible" for describing the electron density $\rho(\vec{r})$ of $\Psi(1,2,...,N)$, in terms of highly localised one, two (or three) centre orbitals.

Like the canonical molecular orbitals these natural orbitals may include non-physical (unrealistic) but significant delocalisation effects. Pauli's exclusion principle coupled with Weinhold's formulation of the localised criterion[15] for orbitals removes the non-physical resonance. The resulting orbitals are referred to as natural bond orbitals (NBO's).

A collection of transformation algorithms exist which can be used to convert between the different types of orbitals. The natural localised orbitals[95,96] sets fit neatly in between the standard atomic orbital view and the molecular orbital view as shown below.

$$\text{AOs} \xrightarrow{T_{NAO}} \text{NAOs} \xrightarrow{T_{NHO}} \text{NHOs} \xrightarrow{T_{NBO}} \text{NBOs} \xrightarrow{T_{NLMO}} \text{NLMOs} \xrightarrow{T_{CMO}} \text{CMOs} \qquad \text{Eq 2.57}$$

These natural orbitals are also in some sense independent of the basis set chosen, since they are "chosen" by the wavefunction. All of the transformations T are unitary except for the non-unitary transformation T_{NAO}.

The natural bond orbital description is in excellent agreement with the theory proposed by Lewis[53]. Lewis structures are often used to interpret chemical structure and bonding[94]. In most cases the "Lewis-type" bonding orbitals in NBOs are more than 99% full.

All the natural orbital sets described are complete and orthonormal. The orthogonality of these orbitals means that they cannot overlap. Fortunately, each natural orbital is uniquely associated with a "pre-orthogonal" natural orbital (PNAO, PNBO etc.[53]). Pre-orthogonal orbitals on the same atom are still orthogonal to each other but are allowed an overlap with other atoms orbitals allowing for bond interactions to occur[53].

2.3.2.1 The NBO program

NBO theory has been implemented by Weinhold in the NBO program[52,53,96,97] to qualitatively describe the wavefunction in terms of natural orbitals. Multiple analyses are included when incorporating the default functionality of the NBO program into an electronic structure system (ESS – programs such as GAMESS-UK[49]). These are Natural population analysis, Natural Bond Orbital (NBO) analysis, Natural Hybrid Orbital (NHO) directionality, bond bending analysis and 2nd-order perturbation theory analysis[52,53].

With 2nd-order perturbation theory analysis one can look at conjugative effects. Occupied (Lewis) (donor, σ) to unoccupied (non-Lewis) (acceptor, σ^*) orbital interactions are

tabulated along with the estimated stabilisation energy of conjugation. The stabilisation energy ΔE for $\sigma \longrightarrow \sigma^*$ is:

$$\Delta E^{(2)}_{i \to j^*} = q_i \frac{\left\langle \sigma_i \left| \hat{F} \right| \sigma_j^* \right\rangle^2}{\varepsilon_{j^*} - \varepsilon_i} \qquad \text{Eq 2.58}$$

$$\varepsilon_i = \left\langle \sigma_i \left| \hat{F} \right| \sigma_i \right\rangle, \quad \varepsilon_{j^*} = \left\langle \sigma_j^* \left| \hat{F} \right| \sigma_j^* \right\rangle \qquad \text{Eq 2.59}$$

where q_i is the donor orbital occupancy, $\varepsilon_i, \varepsilon_j$ are the orbital energies of donor and acceptor orbitals and $\hat{F}$ is the orbital Hamiltonian (the Fock or Kohn-Sham operator)[52,53].

2.4 Semi Empirical models – AM1, PM3, PM3C

The *ab initio* quantum mechanical methods we have discussed so far only use basic physical constants to solve the Schrödinger equation. Semi Empirical Quantum Mechanics methods use sets of experimental data and other approximations to simplify the problem. The Roothaan-Hall equations are still used but certain integrals are neglected or approximated[54].

2.4.1 Why use Semi Empirical Methods?

The cost of *ab initio* methods scales formally as the fourth power of the number of basis functions[77]. Most of the calculation involves the manipulation of integrals, specifically the large number of two-electron integrals required in constructing the Fock-matrix[77]. The cost associated with these integrals can be reduced by neglecting or approximating them[54]. This is exactly what Semi Empirical methods do.

The first approximation is to only explicitly consider the valence electrons as these are the ones involved in chemical bonding. Secondly, the majority of Semi Empirical methods only use basis sets with Slater-type s and p orbitals. The final approximation which is applied is Zero Differential Overlap. The overlap matrix S is set to the identity matrix I, one-electron integral integrals involving three centres are set to zero and three- and four-centre two-electron integrals are neglected[77].

Theses approximations reduce the "effectiveness" of Semi Empirical methods. To counteract this, parameterisation methods are used. The integrals that are still left are fitted to experimental data or data from *ab initio* calculations.

2.4.2 How do *ab initio* and Semi Empirical methods compare?

The advantages and disadvantages of these methods are listed in the next four paragraphs.

Advantages of Semi Empirical: It is relatively inexpensive (when compared to high level *ab initio*). These methods are qualitatively good at describing large systems and quantitatively good where decent parameter sets exist[57].

Limitations of Semi Empirical: Systems which are unparameterised, transition structures and in some cases hydrogen bonding[57].

Advantages of *ab initio*: It can handle any atom (as no parameters are required) and investigate excited states[57].

Limitations of *ab initio*: Calculations are computationally expensive. Memory and hard drive requirements are generally quite large. Small basis sets will generally give a poor description of a system[57,77].

For most systems a Semi Empirical method with reasonable parameters is a good choice.

2.4.3 NDDO models

The most commonly available and often used Semi Empirical models are Austin Model 1 (AM1) and Parameterised Model 3 (PM3). A further model called PM3-Carbohydrate (PM3CARB-1) will also be discussed. All of these are based upon the Neglect of Diatomic Differential Overlap (NDDO) approximation[54,77,78] although core-core repulsions are treated in a different manner for each method[54].

NDDO models reduce the number of integrals discarded by the ZDO approximations (Refer to Section 2.4.1). In NDDO all two electron two-centre integrals on the same atom are kept, only differential overlap between atomic orbitals on different atoms is neglected[54].

2.4.3.1 AM1

Austin Model 1 (AM1) was developed by Dewar in the 1980's[54,98]. Most of it's deficiencies are well known and it can be applied to a wide range of compounds. Carbohydrate studies have been carried out with this model[99,100].

2.4.3.2 PM3

The first automated parameterisation procedure was implemented by J.J.P. Stewart to develop Parameter Model 3 (PM3)[54,77]. Like AM1 it has been parameterised for a variety of molecules and its fallibilities are well known[77]. This model has been successfully implemented in disaccharide simulations[42].

2.4.3.3 PM3CARB-1

In an effort to more accurately predict conformations of carbohydrates McNamara et al. have recently proposed the PM3CARB-1 model[5]. The PM3CARB-1 model is a reparameterised version of the PM3 model but makes several improvements. A shortcoming of the PM3 model is the favouring of the $^{1}C_4$ chair over the $^{4}C_1$ chair; this disagrees with experiment and is corrected in the PM3CARB-1 model. The two types of chair for β-glucopyranose are shown in Figure 2.9. The $^{1}C_4$ chair with its numerous unfavourable axial interactions should not be favoured.

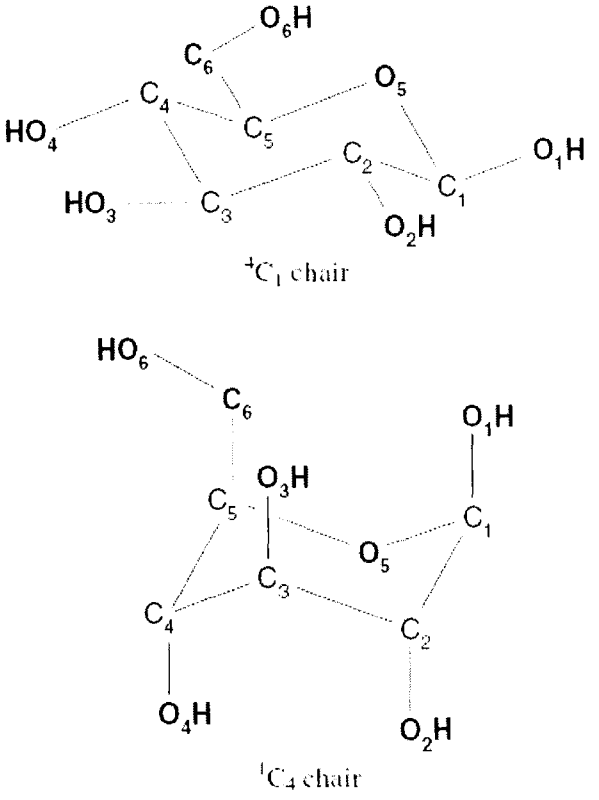

Figure 2.9 – Illustration of the $^{4}C_1$ chair and the $^{1}C_4$ chair where subscript labels refer to atom numbering about the ring (From McNamara[5]).

The PM3CARB-1 model was parameterised using MP2 methods on a small set of carbohydrate analogues and an optimum parameter fitting was achieved using statistical analysis. Electron correlation effects are parameterised into the Semi Empirical parameter set because of the use of electron correlation methods (MP2). Further details of the reparameterisation procedure are available[5].

"PM3CARB-1, appears in general more able to accurately predict the structure and energetics of a set of small carbohydrate analogues than does PM3" *[sic.]*[5]. This is illustrated by Figure 2.10 (extracted from the paper by McNamara[5]) The $^1C_4$ chair with its unfavourable axial interactions has a strange "hydrogen bond interaction" in PM3 (See Figure 2.10 (a)) which is corrected in PMCARB-1 (See Figure 2.10 (b)). The PM3CARB-1 model better represents the *ab initio* model in Figure 2.10 (c).

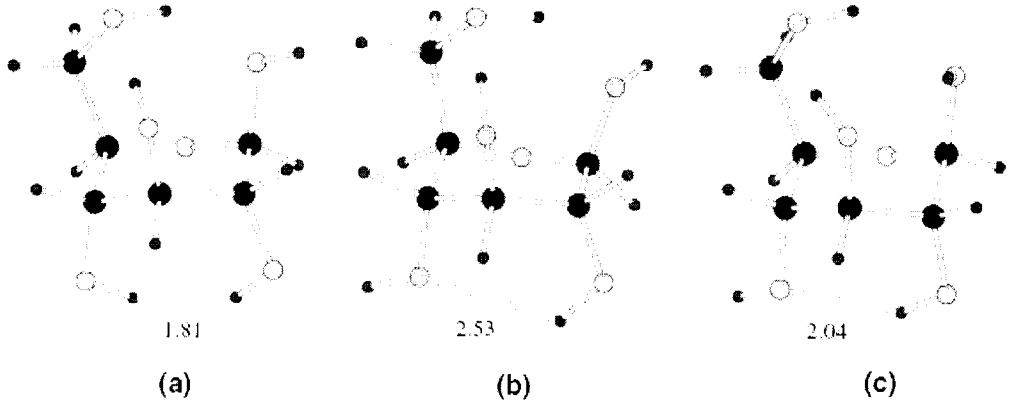

Figure 2.10 – The minimum energy structures of the β-D-glucose conformer with the $^1C_4$ chair conformation computed by (a) PM3, (b) PM3CARB-1 and (c) HF/6-31G* methods, where the distances shown are in Angstroms (Å).

Table 2.2 compares PM3 and the adjusted PM3CARB-1 parameters. This was extracted from the paper by McNamara[5] and implemented in CHARMM[48] at runtime as input to the Semi Empirical quantum module (Refer to Appendix A.1).

Table 2.2: PM3 and PM3CARB-1 parameters extracted from McNamara[5].

PM3 and optimised PM3CARB-1 parameters for O and H

Parameter	PM3	PM3CARB-1
U_{ss}(O) (eV)	86.993 002	90.938 073
U_{pp}(O) (eV)	71.879 580	76.932 200
β_s(O) (eV)	45.202 651	44.449 581
β_p(O) (eV)	24.752 515	35.343 869
α(O) (Å^{-1})	3.217 102	3.031 867
U_{ss}(H) (eV)	13.073 321	13.514 849
β_s(H) (eV)	5.626 512	4.011 786
α(H) (Å^{-1})	3.356 386	2.753 199

2.5 The hybrid approach – combining QM and MM methodologies

QM calculations are not feasible for large systems while MM techniques cannot always provide the rigour required. A combined QM/MM approach is ideal, especially for systems where electronic effects are likely to exist but where a pure QM approach would be too expensive.

A good example is a large enzyme where electronic processes such as bond-breaking and/or forming are of interest. The system can be partitioned into a QM region, an MM region and a QM/MM interaction region.

This technique can be implemented with the following Hamiltonian[101]:

$$H_{effective} = H^{\circ}_{QM} + H_{MM} + H^{Elec}_{QM/MM} + H^{vdW}_{QM/MM} + H_{boundary} \qquad \text{Eq 2.60}$$

where the energy is simply

$$E_{total} = E_{QM} + E_{MM} + E_{QM/MM} + E_{boundary}. \qquad \text{Eq 2.61}$$

H°_{QM} is the Hamiltonian describing the quantum potential energy for the quantum part of the system while H_{MM} describes the MM potential for the MM part of the system. The interactions between the QM and MM parts of the system are described by electrostatic

and van der Waals interactions using the $H^{Elec}_{QM/MM}$ and $H^{vdW}_{QM/MM}$ terms. The boundary term $H_{boundary}$ is used to mimic bulk effects and is the periodic boundary term as applied to Molecular Dynamics (MD) simulations (discussed in the next chapter).

In the case where bonds exist between QM and MM atoms special consideration must be made. The ghost atom method (or ghost hybrid orbital method, GHO) is commonly used but other approaches are available and the reader is directed to the *Encyclopaedia of Computational Chemistry*[15] and *A combined quantum mechanical and molecular mechanical potential for Molecular Dynamics simulations*[101] for further details.

The system considered in this dissertation was neatly partitioned such that whole molecules were either quantum mechanically or molecular mechanically treated. Specifically the solute (carbohydrate) was treated with QM methods and the solvent (water) was treated with MM methods. A schematic of a QM/MM system is shown below in Figure 2.11.

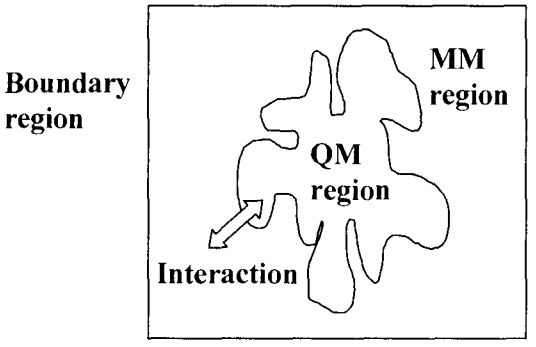

Figure 2.11 – Illustration of a generic QM/MM system (From Field[101])

Ab initio or Semi Empirical methods can be used to model the quantum mechanical region of a QM/MM simulation. The calculations conducted in this study used Semi Empirical methods, specifically the PM3CARB-1 model. Both the Semi Empirical QM region and MM region are treated within CHARMM[48]. The hybrid QM/MM[101] approach as implemented in CHARMM[48] can be used in the same way as the usual Force Field approach. Specifically geometry optimisation, MD simulations and free energy

perturbations can be applied to QM/MM systems[54,77,101]. It is important to define the Semi Empirical QM region correctly within CHARMM[48], it must not contain any MM charges. (Refer to Appendix A.2 for the topology definition used).

2.6 Optimisation Techniques

Optimisation or minimisation techniques attempt to minimise the overall energy of a molecular system and attempt to find minima on the energy surface of a molecular system. Even a simple unimolecular system contains a large number of degrees of freedom. For a non-linear molecule there are 3N-6 molecular vibrational modes[55]. Intramolecular modes such as bond vibration, angle and torsional angle rotation make this optimisation process more complex. Through-space interactions like electrostatic interactions also affect molecular energy. Note that a multidimensional energy surface conceptually describes the degrees of freedom available to a molecule.

This multidimensional hypersurface is denoted a potential energy surface or PES and is almost always impossible to visualise, yet low energy areas on the PES are of interest. A simple, descriptive and visualisable example is the PES for bond vibration in a diatomic molecule. As the bond length is modified, the energy changes and a minimum or most preferred bond length is found to exist, i.e. the equilibrium bond length. This can be expressed by the Morse potential function[102]

$$E = D_e(1 - e^{-ax})^2 \qquad \text{Eq 2.62}$$

where D_e is the bond dissociation energy at infinite separation, x is $r\text{-}r_e$ (the extension of the bond from its equilibrium distance), E is the energy and a is a constant.

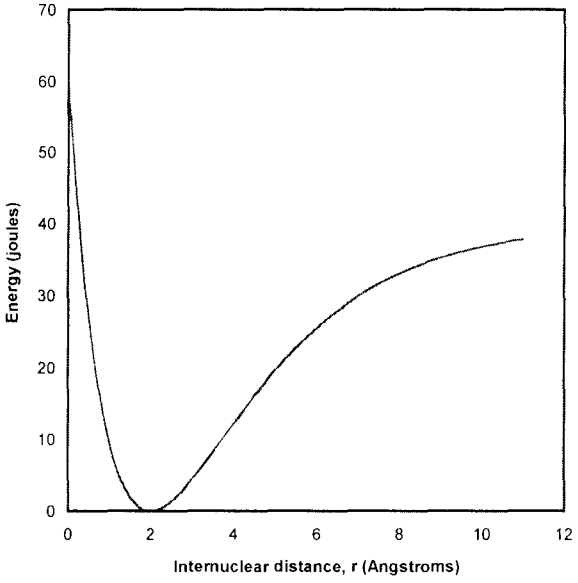

Figure 2.12 – Plot of the Morse potential function with D_e=40, a=0.4 and r_e=2.

The Morse potential (See Figure 2.12) simply describes the energy of system relative to its geometry i.e. the internuclear distance, r. A minimum exists at the equilibrium bond distance (r_e) and dissociation occurs as the bond is over-stretched.

By analogy this picture can be expanded to multiple dimensions for a more complex molecule where the additional axes may describe the change for any degree of freedom. This multidimensional potential will likely contain many stationary points. Some of these may be minima, others transition states (saddle-points) and finally some may be maxima (higher-order saddle-points or hilltops)[58].

Chemists are primarily interested in minima and transition states. Much effort goes into finding all the minima of a PES under exploration, especially to find the overall lowest energy denoted the global minimum. Generally, the PES is the hypersurface calculated from the energy as a function of nuclear coordinates (Only the potential energy is considered). The potential function used may be from electronic structure methods or MM methods[58].

2.6.1 How to find the minima?

The simplest approach is to vary a coordinate in increments while comparing the energy of the structures obtained. For example, a coordinate x_i could be changed to generate two new structures with $x_i+\partial x_i$ and $x_i+2\partial x_i$ (A commonly used value for the size of the change in coordinate is 0.2 Bohr or 0.2 Radians[49]). The energies of the old and new structures can be compared. A parabola can be fitted through the old coordinate and the two new coordinates. The minimum of the parabolic function is determined and the coordinate x_i is adjusted to the position of the minimum[54] (in practice alternative and more complex approaches can be used). This procedure continues until the "lowest energy" is achieved. That is the change in coordinate and energy is so small (within a certain tolerance e.g. 0.0001 au) that one assumes a minimum has been reached. Better methods include the use of the first and second derivatives of the energy to direct the search, but these methods are more time-consuming.

The methods available include[76]:

(1) - methods that do not use gradients (univariant methods such as the simplex method)

(2) - methods that use numerical gradients and second derivatives

(3) - methods with analytical gradients and numerical second derivatives

(4) - methods with both analytical gradients and second derivatives[76].

Type 4 methods seem the most appropriate as they contain the best information and are necessary for computing vibrational spectra. Unfortunately, analytical derivatives are cumbersome to calculate and Type 3 methods with numerical second derivatives are most popular[54,76]. The algorithms applied in methods 2, 3 and 4 are the same (i.e. the same algorithm can be applied) however the accuracy and time taken when on implementation is dependant on the method chosen to calculate the derivatives (numerical or analytical).

2.6.2 How does optimisation affect the calculation?

The optimisation process for an *ab initio* calculation can be represented by Figure 2.13

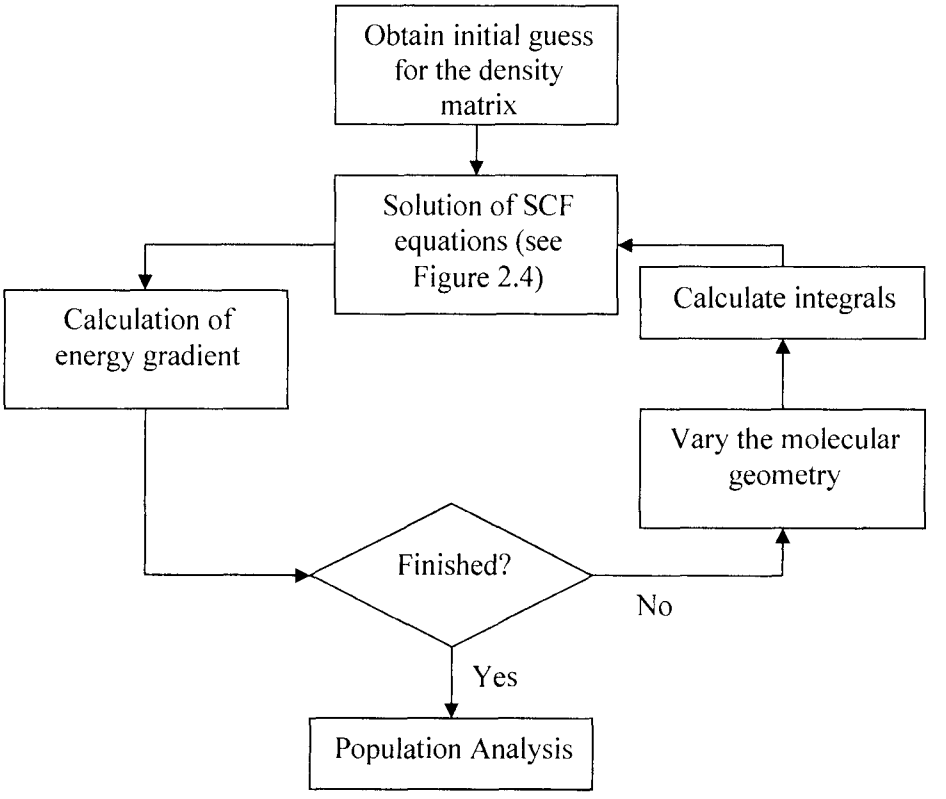

Figure 2.13 – Schematic representing the optimisation process from the initial guess through to population analysis (modified from Hehre[78]).

It is clear that a lot more calculation is required than for the simple single-point energy case represented in Figure 2.4.

Chapter 3

BACKGROUND TO COMPUTATIONAL METHODS – STATISTICAL MECHANICS AND FREE ENERGY

The previous chapter introduced methods of determining the energy of static configurations of molecular systems. *What about multiple system configurations?* In other words, *can dynamic properties be modelled and average properties be calculated?*

Experiments (such as NMR, IR and ORD) generally give information about bulk (macroscopic) properties of systems. With computational chemistry techniques it is only possible to evaluate the atomic detail of microscopic systems (a few thousand to tens of thousands of molecules) because of computational limitations (However, mesoscale simulations where a molecule can be modelled as one entity (instead of discrete atoms and bonds) can be conducted. Other computational scientists (e.g. Physics, Climatologists) can model large macroscopic systems very effectively (although often ignoring atomic detail)). *How can we extrapolate information obtained from this microscopic viewpoint to obtain macroscopic properties?*

3.1 Statistical Mechanics

Statistical mechanics allows one to consider average properties of a system as well as extrapolate the microscopic system observables to estimate macroscopic system properties. Statistical mechanics theory is the application of the mathematical tools provided by probability theory to the motion of particles[103]. Two important concepts in statistical mechanics are the ensemble average and the partition function. The idea of an ensemble was first introduced by J.W. Gibbs[102]. An ensemble is simply a large collection of imaginary replicas of a system.

The simplest example of an ensemble is to consider a closed system with adiabatic walls which contains a gas where the following are fixed: the number of gas molecules, N; the volume, V and the energy, E. If this system is replicated M times, a collection or ensemble of M systems is obtained each with N, V and E constant. This is called a microcanonical ensemble and is depicted in Figure 3.1.

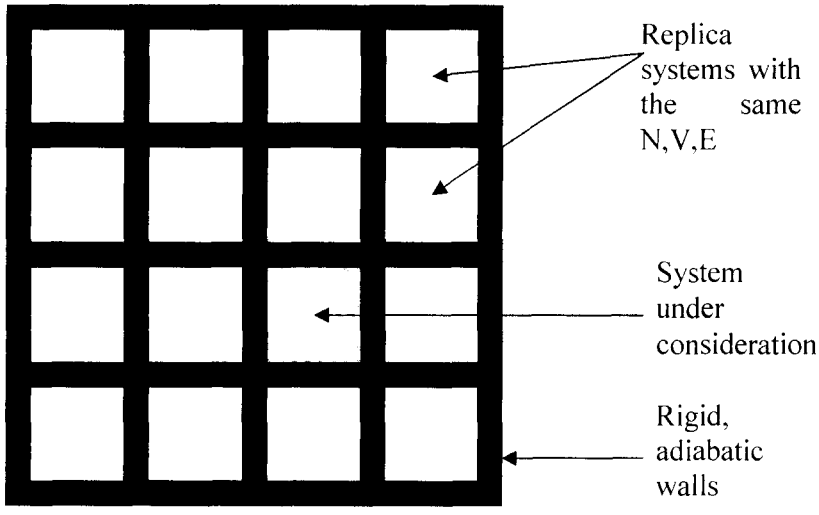

Figure 3.1 – Schematic of an NVE ensemble (modified from Laidler and Meiser[102]).

There are different energy levels available to each molecule in each system of the NVE ensemble. The number of ways or complexions, W, in which we can distribute the molecules is

$$W = \frac{N!}{\prod_i n_i!} \qquad \text{Eq 3.1}$$

where N is the total number of molecules and n_i is the number of molecules in energy level i.

In an NVE ensemble both the overall number of molecules and the overall energy must be constant. Therefore the following constraints apply:

$$\sum n_i = N \qquad \text{Eq 3.2}$$

$$\text{and } \sum_i n_i \varepsilon_i = E \qquad \text{Eq 3.3}$$

where εi is the energy of energy level i and E the total energy.

This implies that each replica system may contain particles with different instantaneous positions and momenta, as long as the total N, V and E remain constant. A specific property A, related to the momenta and positions of the particles can be measured for each replica. An average of A for all the instantaneous configurations gives the ensemble average, $\langle A \rangle$ and describes the experimental average of A[54,55,103].

The fraction of molecules in each energy state contributes to the overall energy and to the average value of A. To find the fraction of molecules in a specific state we can apply the Boltzmann Distribution, thus

$$\frac{n_i}{N} = \frac{e^{-\varepsilon_i / k_B T}}{\sum_i e^{-\varepsilon_i / k_B T}} \qquad \text{Eq 3.4}$$

where k_B is the Boltzmann constant and T, the temperature in Kelvin.

The Boltzmann distribution is more generally expressed as the partition function. The molecular partition function is

$$q \equiv \sum_i e^{-\varepsilon_i / k_B T} \qquad \text{Eq 3.5}$$

and is the sum of all states available to a molecule[55].

The system partition function, Q, is

$$Q \equiv \sum_i e^{-E_i / k_B T} \ . \qquad \text{Eq 3.6}$$

This is the summation of all energy levels available to all molecules in the system[103].

Most experiments tend not to have the number of molecules, volume and energy constant. A more pertinent ensemble is the constant N, V and T ensemble, the canonical ensemble.

Other commonly used ensembles include the grand canonical (constant μ (chemical potential), V, T) and the isothermal-isobaric (constant N, P (pressure), T) ensemble.

For a canonical ensemble we can calculate the ensemble average $\langle A \rangle$ of a property A using

$$\langle A \rangle_{NVT} = \frac{\int A(v_N) e^{-\beta U(v_N)} dr_N}{\int e^{-\beta U(v_N)} dr_N} = \frac{\sum_{r=1}^{T} A(v_N) e^{-\beta U(v_N)}}{\sum_{r=1}^{T} e^{-\beta U(v_N)}} \qquad \text{Eq 3.7}$$

where β is $(k_B T)^{-1}$ and v_N is a potential function that depends on the positions r_N and momenta p_N. It would be very difficult, and impractical, to estimate the numerator and denominator separately by using the uniform sampling method, i.e., performing T independent measurements and hoping to observe all states of A. However, non-uniform sampling methods can be designed and used to carry out the simulation.

The equilibrium states of an ensemble can be described thermodynamically. The equilibrium state for the microcanonical ensemble (NVE) would give the maximum entropy S, while the canonical ensemble (NVT) gives the minimum Helmholtz free energy A[54].

3.1.1 How do we generate ensembles computationally?

In general equilibrium methods such as Molecular Dynamics or Monte Carlo computer simulations sample configurations (microscopic states) in such a way that the regions of configurational space that make the largest contributions to the integral $\langle A \rangle_{NVT}$ are also regions that are sampled most frequently. If ρ(v) is the probability of choosing a configuration $v(r_N, p_N)$ then

$$\langle A \rangle_{NVT} = \frac{\sum_{v=1}^{T} \frac{A(v_N) e^{-\beta U(v_N)}}{\rho(v)}}{\sum_{v=1}^{T} \frac{e^{-\beta U(v_N)}}{\rho(v)}} \qquad \text{Eq 3.8}$$

where $\rho(v) = e^{-\beta\varepsilon_i}$ and so samples the Boltzmann distribution (See Equation 3.4). This produces a canonical ensemble

$$\langle A \rangle_{NVT} = \frac{1}{T}\sum_{v=1}^{T} A(v_N) \qquad \text{Eq 3.9}$$

as an unweighted average over the configurations sampled.

3.2 Monte Carlo

A Monte Carlo (MC) ensemble is generated by random changes to molecular coordinates. Only the potential energy contributes to the total energy of the system. If the energy is lowered this new state is kept. Higher energy structures can be accepted if their Boltzmann difference between the previous and current configuration is less than a random number between 0 and 1. In other words, if the increase in energy is small the probability of accepting this higher in energy state is large. Further details of this method are available in standard texts[54,77,104].

3.3 Molecular Dynamics

Molecular Dynamics (MD) generates an ensemble by following the time evolution of a system using Newton's second law of motion. The ergodic hypothesis states that the time average (to the limit of infinity) is equal to the ensemble average thus the MD generated ensemble is a valid one. Using this approach it is possible to calculate time-dependent properties such as transport coefficients. The MD approach was used exclusively in this dissertation[54].

A MD computer simulation is typically broken down into four stages: initialisation, heating, equilibration and production. Initialisation is setting up or building the molecular system. Initial positions and velocities of atoms are assigned. Heating involves increasing

the kinetic energy of a system to increase the overall temperature. This is possible as the temperature is directly related to the kinetic energy as follows[54]:

$$K = \sum_{i=1}^{N} \frac{|p_i|^2}{2m_i} = \frac{k_B T}{2}(3N - N_C) \qquad \text{Eq 3.10}$$

where K is the kinetic energy, N is the total number of particles (atoms), p and m are the momentum and mass of a particle i. The equipartition theorem states that each degree of freedom contributes $(k_BT)/2$ [55]. For N molecules and three spatial degrees of freedom (x, y and z) this gives $3(Nk_BT)/2$. However, if we constrain the system we reduce the degrees of freedom which contribute to the kinetic energy. The N_C term is equal to the number of constraints applied to the system[54].

Once a target temperature has been reached the system is allowed to evolve over time until it has been sufficiently equilibrated i.e. it is in low energy states. Finally, production phase dynamics are used to sample phase space. The production phase data is used for analysis.

3.3.1 Equilibrium Dynamics

Experimentally considered chemical systems are often in equilibrium. In computational studies equilibrium dynamics can be conducted which sample the low energy states of molecular phase space.

Newton's laws of motion are integrated to produce successive time-dependent configurations of the system[54,104]. The following differential equations are solved:

$$\frac{d^2x_i}{dt^2} = \frac{F_{x_i}}{m_i} = a_x \qquad \frac{d^2y_i}{dt^2} = \frac{F_{y_i}}{m_i} = a_y \qquad \frac{d^2z_i}{dt^2} = \frac{F_{z_i}}{m_i} = a_z \qquad F_i = \frac{-\partial U(r_i,...r_N)}{\partial r_i} \qquad \text{Eq 3.11-14}$$

where x, y, z represent the position change of a specific atom whose mass is m. Time is represented by t while the force and acceleration in the x, y, z direction are F and a respectively. U is the potential energy as a function of r_i[54,104] where r_i represents the x, y and z coordinates of the *i*th atom which is in the set of all N atoms of a system.

Full MD simulations are generated from a number of discrete steps in time. Analagous to movies in the cinema; which are composed of multiple static snapshots. At each time step (t) the total force on each particle is calculated. The accelerations are then calculated and combining this information with the known velocities and positions, the new velocities and positions for the next time step (t+δt) can be calculated. This process continues for the entire simulation and this method of integrating time in finite steps is termed a finite difference method.

The verlet algorithm is the most widely used for integrating the equations of motion[54]. The following is applied:

$$r(t+\delta t) = 2r(t) - r(t-\delta t) + \delta t^2 a(t) \qquad \text{Eq 3.15}$$

where r is the position of an atom and a its acceleration.

The main drawback of this method is addition of the small term $[\delta t^2 a(t)]$ to the difference of large terms $[2r(t) - r(t-\delta t)]$ which leads to numerical imprecision (floating point errors)[54].

The leap-frog verlet method is much more accurate and avoids the inaccuracies associated with the standard verlet method[54,104]. The following equations are employed:

$$r(t+\delta t) = r(t) + \delta t v(t+\frac{1}{2}\delta t) \qquad \text{Eq 3.16}$$

$$v(t+\frac{1}{2}\delta t) = v(t-\frac{1}{2}\delta t) + \delta t a(t) \qquad \text{Eq 3.17}$$

The velocity equation, Equation 3.16, is applied first and is always calculated halfway between time steps. Applying Equation 3.17, the new positions can then be calculated from the previous positions using the velocities calculated from Equation 3.16. The velocities and positions (coordinates) appear to "leap" over each other as suggested by the name of the algorithm[104].

3.3.2 Time step

The length of the time step is a vital component of any MD simulation. To traverse as much of conformational space as possible one might consider making the time step as large as possible[105]. In practice, this is not the case. Using the cinema analogy, if the time step is too large the introductory sequence would be followed immediately by the closing credits and if the time step was too small the show (played in slow motion) would take weeks instead of hours to watch. In computational terms, a large time step creates instabilities as molecular collisions would not be modelled accurately. A minute time step results in phase space being covered extremely slowly.

Simulations involving flexible molecules (e.g. polymers) should have a time step which is 10% of the time of the shortest period of motion[54]. The suggested time step for flexible molecules with flexible bonds is 10^{-15} or $5*10^{-16}$ seconds (1-0.5 femtosecond (fs)). Carbohydrates are flexible because of the ability of the pyranose ring to pucker and the rotations of the primary and secondary alcohol groups. A 1 fs time step is adequate for carbohydrates[54]. However, certain molecular motions are very quick and do not contribute significantly to our understanding of a molecular system. Bonds formed between hydrogen and heavy atoms (H-X) tend to vibrate at very high frequencies. These H-X bonds can be constrained to their equilibrium bond length values[54]. The SHAKE[104] algorithm is commonly used. In carbohydrate simulations C-H bonds are constrained and the carbohydrate is treated with Molecular Mechanics (MM). The water O-H bond is also treated with SHAKE[104].

3.3.3 Simulation boxes and periodic boundaries

Systems tend to be modelled in simulation boxes with a shape that allows infinite repetition by 3-D tessellation. Popularly used shapes include a simulation cube, truncated octahedron and rhombic dodecahedron[54,104]. The cubic simulation box is used because it is geometrically simple to deal with[104].

The concept of infinite periodic repetition of a system (usually called periodic boundary conditions) is related to the discussion about ensembles. This approximation stems from the need to eliminate edge effects and thereby better emulate macroscopic or bulk solution. An example which illustrates the importance of removing edge effects is from *Computer Simulations of Liquids*[104]: Consider a simulation cube of size $10 \times 10 \times 10$ units containing 1000 molecules. 488 of these molecules can be seen from the cube faces and experience "edge" effects, that is they do not experience the same forces as the rest of the molecules from the bulk solution. Replication of the simulation box as an infinite lattice eliminates these surface or edge effects. A 2-Dimensional example of this is shown in Figure 3.2.

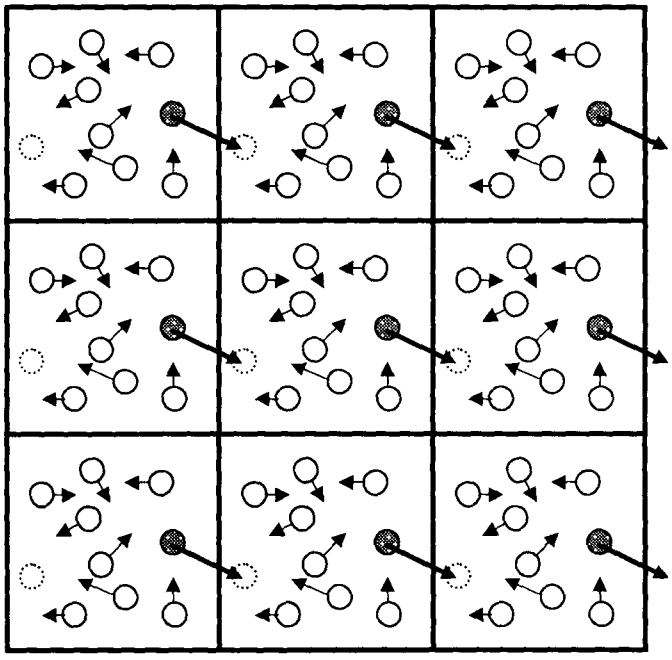

Figure 3.2 – Schematic of periodic boundary conditions for a two-dimensional system. The shaded particle is shown leaving the simulation box and re-entering from the opposite side (Adapted from Leach[54] and Lewars[58]).

When dealing with potentially infinite copies of system there are a number of potential pitfalls.

Firstly, *can the system be infinitely represented?* No. Instead, for a cubic simulation box only the first "layer" of three dimensional boxes around the original box are considered. As

for a rubic cube there are 3*3*3 cubes which means 27 cubes, but the centre cube is the original simulation box. Thus, 26 (3^3-1^3) imaginary replicas of the original simulation box are created.

Secondly, *is a molecule interacting with the image of itself or seeing the same molecule twice?* No, because both of these interactions give incorrect forces and energies. One makes sure that this doesn't occur by applying the *minimum-image convention.* This stipulates that after a certain radius (known as the cut-off) a molecule will not interact with other molecules. This image neighbourhood list cut-off is sensibly set to less or equal to half the box length (See Figure 3.3).

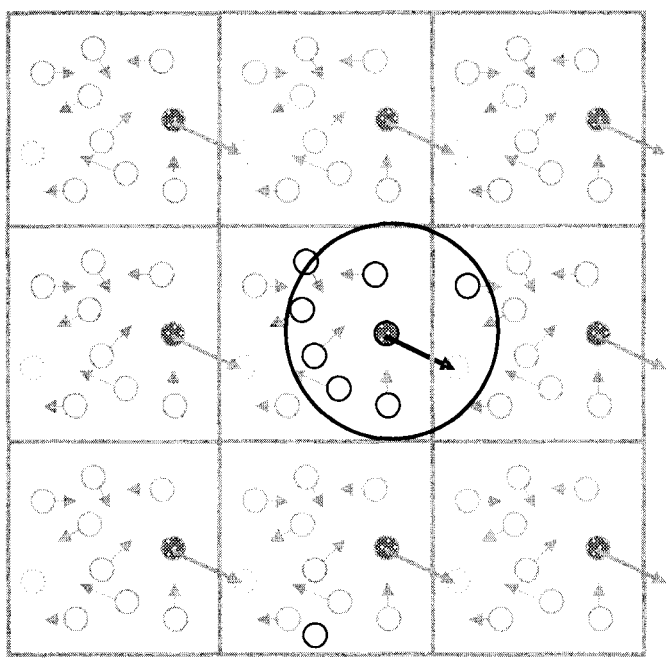

Figure 3.3 – Schematic showing a spherical cut-off around the shaded particle and the minimum image convention as applied to a periodic system (Adapted from Leach[54] and Lewars[58]).

Another type of cut-off is the non-bonded cut-off (or long-range interaction cut-off). The non-bonded cut-offs truncate the electrostatic and non-bonded interaction potential. Calculation of non-bonded interactions is a time-consuming part of any MD simulation. The Lennard-Jones potential (used to consider van der Waals interactions) usually falls off to approximately 1% of its σ value (where σ is the value at which the interparticle potential

is zero) at approximately 2.5σ[104]. Interactions past this point (2.5σ) will not significantly contribute and can be truncated. Direct truncation will lead to discontinuities in energy. Switching (Sw) or shifting (Sf) can be applied such that the potential function changes smoothly to zero at the specified cut-off point[106]. These techniques are described by the following equations[54,106]:

$$S_w(r_{ij}) = 1 \qquad r_{ij} \le r_{on} \qquad \text{Eq 3.18}$$

$$S_w(r_{ij}) = \frac{(r_{off}^2 - r_{ij}^2)(r_{off}^2 + 2r_{ij}^2 - 3r_{on}^2)}{(r_{off}^2 - r_{on}^2)^3} \qquad r_{on} < r_{ij} \le r_{off} \qquad \text{Eq 3.19}$$

$$S_w(r_{ij}) = 0 \qquad r_{ij} > r_{off} \qquad \text{Eq 3.20}$$

$$Sf(r_{ij}) = 1 - \frac{2r_{ij}^2}{r_{cut}^2} + \frac{r_{ij}^4}{r_{cut}^4} \qquad r_{ij} < r_{cut} \qquad \text{Eq 3.21}$$

$$Sf(r_{ij}) = 0 \qquad r_{ij} > r_{cut} \qquad \text{Eq 3.22}$$

For switching, a cut-on value r_{on} describes the distance at which modification to the potential starts and the cut-off value r_{off} designates the value at which the potential is truncated. For shifting, r_{cut} is simply the cut-off distance and r_{ij} is the current distance between atoms i and j.

The non-bonded neighbourhood list cutoff stores a list of molecules that fall just outside the potential truncation radius of a molecule. It is very costly to check all molecules to see if they fall into the energy truncation region. It is less costly to keep a "buffer region", a list of molecules just outside the spherical truncation cut-off, in memory. At each time step molecules that previously interacted with the current molecule will be checked as well as those in the "buffer region". This buffer region should be large enough such that a molecule will not be able to diffuse through it in one time step[104].

3.3.2 Non-equilibrium dynamics

Standard MD tends to sample low energy states and generally will not overcome barriers greater than $2k_BT$[104]. At 298.15K, barriers of greater than c.a 1-2 kcal/mol are unlikely to be traversed. To adequately sample all of phase space, not just low energy states biased or non-equilibrium dynamics are used. A simulation is biased such that the molecular system is directed to traverse the less frequently sampled, higher in energy regions of phase space.

3.3.2.1 Why bias simulations?

Statistical properties (often termed entropic properties) such as the Gibbs (G) and Helmholtz (A) free energies are directly related to the partition function of a system. Entropic properties are directly impacted by the high energy states of a system[54,77]. To correctly predict such properties biased sampling must be applied.

3.3.2.2 Langevin Dynamics / Stochastic Dynamics

Stochastic dynamics are often employed in biased simulations. The equation of motion used in Langevin dynamics is the following for each atom i in the system:

$$\frac{d^2\vec{r}_i(t)}{dt^2} = m_i^{-1}\vec{F}_i + m_i^{-1}\vec{R}_i - \gamma_i\frac{d\vec{r}_i(t)}{dt} \qquad \text{Eq 3.23}$$

where m_i and $\vec{r}_i$ are the mass and position of the ith atom, γ_i is a frictional coefficient, $\vec{R}_i$ is a stochastic term and $\vec{F}_i$ is the force arising from the Force Field.

Langevin dynamics controls the temperature by adding a stochastic term and frictional force $(\gamma_i\frac{d\vec{r}_i(t)}{dt})$ to the force arising from Force Field terms. The stochastic term introduces energy the system and can be considered a random force due to constant molecular bombardment of the solute by the solvent. The frictional force removes energy from the system and is equivalent to frictional drag on the solute by the solvent. The bombardment and drag effects are real effects that are experimentally observed as

Brownian motion[107]. Stochastic dynamics can be used to control the temperature and thus avoid velocity rescaling techniques (and the problems associated with them)[54]. Stochastic techniques are crucial for Potential of Mean Force (PMF) calculations as the random force improves the traversal of phase space.

In gaseous phase simulations the frictional force is applied to the "solute" while in condensed phase (liquid) simulations the frictional force is applied to the solvent.

3.4 Free Energy Calculations and Perturbation Methods

Chemists are interested in molecular properties and behaviors but perhaps the most fascinating are binding and conformational preferences because they are difficult to determine. These preferences can be explored by looking at the free energy, which is the total amount of energy in a physical system which can be converted to do work[108]. An accurate method for predicting the free energy gives us the ability to understand and predict molecular properties[109].

The free energy whether it is the Gibbs (G) or Helmholtz (A) energy is an entropic or thermal property[54]. As mentioned earlier, thermal properties are directly related to the partition function. High energy states make important contributions to the partition function and thus to the free energy. The Helmholtz free energy, A depends directly on the system partition function, Q as shown below:

$$A = -k_B T \ln Q . \qquad \text{Eq 3.24}$$

Consider two conformations characterised by potentials $E_1(\mathbf{r})$ and $E_0(\mathbf{r})$, the free energy difference between the two conformations can be determined by (following the formulation of R. W. Zwanzig[110] for two fluids):

$$A_1 - A_0 = -k_B T \ln \frac{Q_1}{Q_0} = -k_B T \ln \left\langle e^{-\beta \Delta U} \right\rangle_0 \qquad \text{Eq 3.25}$$

where $\Delta E = E_1(\mathbf{r}) - E_0(\mathbf{r})$, and the ensemble $\langle A \rangle_0$ is taken in the reference system $E_0(\mathbf{r})$. This equation is used in computer simulations to compute free energy differences in a procedure known as the free energy perturbation (FEP) method[111-113].

3.4.1 Umbrella Sampling and the Potential of Mean Force (PMF)

A general solution to sample regions of large negative ΔE, which cannot be efficiently sampled using equation 3.25, is on the basis of a general non-Boltzmann distribution[114]

$$\rho_w(r) = \frac{W(r)e^{-\beta E_0(r)}}{\int W(r)e^{-\beta E_0(r)}dr} \qquad \text{Eq 3.26}$$

where $W(\mathbf{r})$ is a positive-valued weighting function. The average of any property in the original ensemble, $\langle A \rangle_0$, can be related to averages taken over MC trials or MD trajectories in the weighted ensemble:

$$\langle A \rangle_0 = \frac{\langle A/W \rangle_W}{\langle 1/W \rangle_W}. \qquad \text{Eq 3.27}$$

Both averages $\langle A/W \rangle_W$ and $\langle 1/W \rangle_W$ are needed to get the final result. This technique is referred to as umbrella sampling[115-117].

The Potential of Mean Force (PMF) method can be used to describe how n degrees of freedom are related to the change in energy of a molecular system. Thus any sensibly chosen degree of freedom can be related to the free energy change in a system.

The PMF is the average potential a coordinate of interest feels due to all interactions of and with the solvent[103]. We can apply a reversible force to the coordinate such that it traverses phase space. Integration of this reversible force yields the reversible work (the Potential of the Mean Force). If the PMF is calculated at constant N, V and T then the reversible work is the change in the Helmholtz free energy.

These methods have been previously applied to carbohydrate systems by Kuttel[40,118,119] and Bryce[41,42].

3.4.2 Implementing PMF routines in CHARMM

The PMF code used in this dissertation to record the probability of the system and to bias dynamics was written and implemented into CHARMM 27b1[120].

The 1-Dimensional Potential of Mean Force code required for hydroxymethyl (primary alcohol) rotation analysis was incorporated by modification of the USERE routine in the latest version of CHARMM (c33b2)[48]. The code was re-compiled to produce a serial dynamics program. The code has been previously used in parallel for standard MD simulations. However, the Semi Empirical procedures in CHARMM currently do not parallelise and a serial version of the code was used.

3.4.3 Theoretical background – Umbrella Sampling and the PMF

The PMF represented by $W(\bar{\xi})$ can be calculated for a generalised multidimensional degree of freedom $\bar{\xi}$ using the following equation

$$W(\bar{\xi}) = -k_B T \ln P(\bar{\xi}). \qquad \text{Eq 3.28}$$

The full derivation for the multidimensional degree of freedom[40] is practically equivalent to the 1-Dimensional case although the implementation is somewhat different. The 1-Dimensional case for a torsional degree of freedom was used in this project and it will be described further.

The PMF, $W(\omega)$ for the free energy of rotation of the torsional angle ω is related to the probability of states of a system by

$$W(\omega) = -k_B T \ln P(\omega) \qquad \text{Eq 3.29}$$

where $P(\omega)$ is the probability distribution function for ω, k_B is the Boltzmann constant and T is the temperature. When applied to canonical dynamics simulations (constant NVT) the Helmholtz free energy rotation curve is obtained. High energy states make a large contribution to the PMF and a reasonable estimate requires proper sampling of phase

space. Biased sampling methods are applied to achieve improved sampling in the form of an umbrella potential[115-117].

The modified system potential (the Hamiltonian, H' which includes the biasing) is a sum of the system Hamiltonian (H°) and the applied umbrella potential biasing $U(\omega)$ as shown below:

$$H' = H^{\circ} + U(\omega) \text{ or} \qquad \text{Eq 3.30}$$

$$H' = H^{\circ}_{QM} + H_{MM} + H^{Elec}_{QM/MM} + H^{vdW}_{QM/MM} + H_{boundary} + U(\omega). \qquad \text{Eq 3.31}$$

Application of the biased potential, H' to a simulation yields a biased probability density distribution $P'(\omega)$. To obtain the unbiased probability density distribution the following is applied:

$$P(\omega) = CP'(\omega)e^{\beta U(\omega)} \qquad \text{Eq 3.32}$$

where C is an arbitrary integration constant and $\beta = -\frac{1}{k_B T}$.

With the purpose being to sample phase space as evenly as possible, the ideal choice for $U(\omega)$ is the inverse of the PMF ($-W(\omega)$) thus;

$$U(\omega) = kT \ln P(\omega). \qquad \text{Eq 3.33}$$

At the beginning of these calculations the $W(\omega)$ function is unknown and the initial simulation is an unbiased one (apply $U(\omega) = 0 \forall \omega$). The probability distribution for this simulation allows one to calculate a first guess for the PMF, $W(\omega)$ from which an improved biasing potential, $U(\omega)$ is calculated for use in the next simulation. This is an iterative process and is repeated until all phase space is adequately sampled.

3.4.3.1 Potential of Mean Force convergence

"Adequate sampling" is a very generic description. An adequate description of phase space implies that enough of phase space is sampled such that quantities calculated from the

statistical distribution are meaningful. More specifically the ratio of most to least sampled should in theory be equivalent. This is not actually feasible and instead the ratios of highest to lowest populations for each glucose and galactose PMF simulation were required to be less than 5:1. In studies by Kumar et al., a less strict convergence criterion was enforced[121,122]. This ratio can be calculated by summing and binning (binning by dividing phase space into equal portions of a finite size and filling these bins) all the populations sampled throughout the iterative PMF procedure.

3.4.4 Theoretical background for Weighted Histogram Analysis Method

The Weighted Histogram Analysis Method (WHAM[115,121-123]) is an iterative procedure used to improve the sampling of conformational space and therefore make free energy calculations converge more rapidly. The WHAM procedure optimises links between simulations and produces the best possible estimation of free energies. It was used in this work and it was first implemented in CHARMM simulations of carbohydrates[71] to combine biased population distributions from multiple simulations using computed weighting factors to yield the best unbiased probability distribution, p_k. After i simulations for a torsional degree of freedom ω which has been divided into k bins, the unbiased probability histogram p_k is given by the following equations:

$$p_k = \frac{\sum_i n_{i,k}}{\sum_i N_i f_i c_{i,k}} \qquad \text{Eq 3.34}$$

$$f_i = \frac{1}{\sum_k c_{i,k} p_k} \qquad \text{Eq 3.35}$$

$$c_{i,k} = e^{-\beta U_i(\omega_k)} \;\; ; \;\; N_i = \sum_k n_{i,k} \; . \qquad \text{Eq 3.36}$$

Here, n_i are the total number of configurations stored in the *i*th simulation, f_i are the free energy weighting factors, each ω_k is a value of ω that falls in the *k*th bin, $n_{i,k}$ is the histogram for ω from the *k*th bin in the *i*th simulation. The WHAM equations are applied

iteratively until the maximum difference between the previous and current iteration weighting coefficients ($\max\left|f_j^{i} - f_j^{i-1}\right|$) is less than 0.001[71].

For the PMF calculations in this dissertation, a histogram that allowed full rotation of a torsional angle through 360 degrees was constructed. This rotational range was labelled from -180 degrees to +180 degrees and was divided into 72 bins of width 2.5 degrees.

3.5 Pucker Analysis

The purpose of pucker analysis is to evaluate the puckering of the pyranose ring during a MD simulation. The Cremer and Pople definition[124] is used and Figure 3.4 illustrates the coordinate system for determining ring puckering. For six-membered rings, e.g. pyranoses, three Cremer and Pople parameters define ring conformation. These are Q the magnitude as well as θ and ϕ the angles that represent ring puckering. The most important of these is the angle θ which represents the degree of puckering away from the chair conformer to the boat or twist boat conformer. The chair structure is designated by an angle of 0 or 180 degrees for θ.

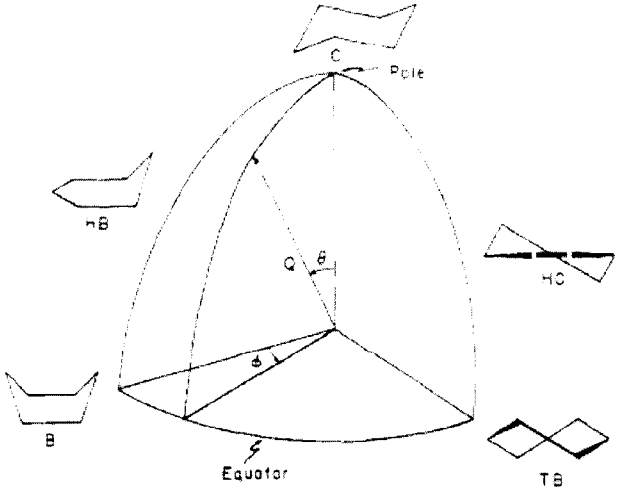

Figure 3.4 – An illustration of how Q, θ and ϕ (the magnitude and the angles that represent ring puckering) are related to the pyranose conformers (chair, half Boat, boat, half chair and twist boat)(from Cremer[124]).

This analysis is required to ensure that the preferred $^4C_1$ chair has not flipped to a boat or to the $^1C_4$ chair during a simulation.

Chapter 4

THE HYDROXYMETHYL GROUP CONFORMATIONAL PREFERENCE

The results for both glucose and galactose simulations are presented in this chapter. The methodology used in the gaseous and aqueous phase calculations will be described for each monosaccharide followed by a discussion of the simulation results.

4.1 Gaseous phase methodology

4.1.1 Potential of Mean Force Dynamics simulations

Gas phase dynamics for the β-anomers of glucose and galactose were done using CHARMM 33b2[48]. Both conformers were simulated although only the β-anomers are shown here because of their dominance while the α-anomer data is supplied as supplementary data (Refer to Appendix B). From here on the terms glucose and galactose refer to the β-anomer of the said sugar. The saccharides were treated with PM3CARB-1 where the initial structures were minimised for 100 steps (using steepest descent minimisation techniques[54] available in CHARMM[48]) followed by a 0.4 nanoseconds (ns) heating phase to a final temperature of 298.15K.

The PMFs were constructed from twenty constrained dynamics simulations of 0.4 nanoseconds (ns) totalling 8ns for each saccharide. A complementary set of unconstrained simulations were carried out (Refer to Appendix C for the PMF curves). Weak restraints were applied to the pyranoside ring dihedrals in the constraint dynamics to preserve the $^4C_1$ ring conformation (Refer to Appendix D for the details of the restraints applied). The ring can pucker out of the $^4C_1$ chair in the gaseous phase because of strong intramolecular interactions. Without the screening effect of solvent, intramolecular electrostatic interactions are strong and the hydroxyl groups interact quite strongly. In the gaseous phase these hydrogen bond type interactions are stronger in the $^1C_4$ chair (the axial chair).

The canonical ensemble (constant N, V and T) was used to describe the system at 298.15K. Langevin dynamics with the leapfrog-verlet integrator and an integration time step of 1 fs were employed[54,104]. The following equation of motion was applied (Refer to Chapter 3):

$$\frac{d^2\vec{r}_i(t)}{dt^2} = m_i^{-1}\vec{F}_i + m_i^{-1}\vec{R}_i - \gamma_i\frac{d\vec{r}_i(t)}{dt} \qquad \text{Eq 4.1}$$

The stochastic term $(\vec{R}_i)$ and a frictional force proportional to a frictional coefficient (γ_i) control the simulation temperature. A frictional coefficient of magnitude 62.5 ps^{-1} was applied to all atoms[119]. Other than providing a heat bath to control the temperature, Langevin dynamics also improves sampling of conformational space by virtue of the random stochastic force. The CHARMM switching functions were applied on a group-by-group basis and in these vacuum dynamics the long range interactions were effectively not truncated as they were switched to zero between 8 Å and 9 Å. Non-bonded interactions were updated heuristically.

4.1.2 Quantum Mechanics Simulations

4.1.2.1 The rotational profile

Ab initio calculations were completed for glucose and galactose with GAMESS-UK 7.0[49]. Electron correlation was included in the description of glucose and galactose using Density Functional Methods (Refer to Chapter 2). B3LYP[87,88] was employed with restricted closed shell Hartree-Fock for all calculations.

In the gaseous phase the orientations of the secondary hydroxyls of the saccharide ring must be either clockwise or reverse clockwise for the best intramolecular interactions to occur[36]. All conformers considered have the secondary hydroxyls oriented in the lower energy reverse clockwise direction[36]. The gas phase rotational curves were constructed for glucose and galactose using single-point energies from the RB3LYP/6-31G* and RB3LYP/6-31+G** levels of theory (Refer to Chapter 2). These curves were generated by

rotation of the ω-torsion angle (-175, -170, -125, -120, -115, -70, -65, -60, -55, -50, -5, 0, 5, 50, 55, 60, 65, 70, 115, 120, 125, 170, 175, 180) (See Figure 4.1).

The orientation of the hydrogens of hydroxyl groups are important for finding the lowest possible stationary points[46]. The θ-torsional angle (rotation of the hydrogen of the 6-hydroxyl group (HO6-O6-C6-C5); -60, 60,180) was rotated for glucose and galactose. The profile of galactose also included rotation of the τ-torsional angle (rotation of the hydrogen of the 4-hydroxyl group (HO4-O4-C4-C5); -60, 60,180).

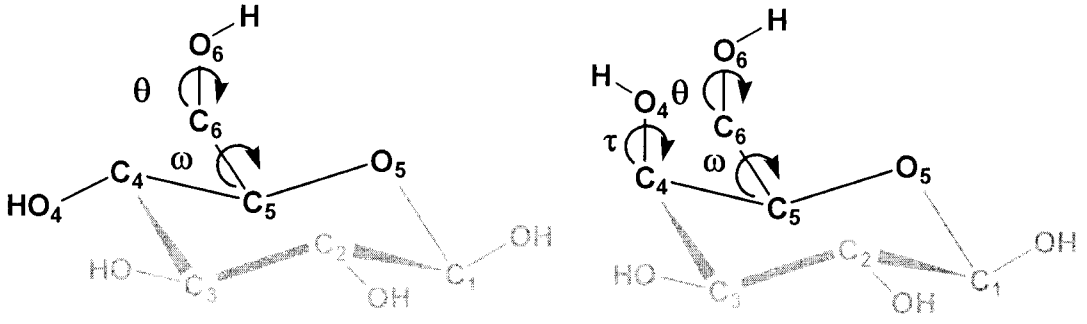

Figure 4.1 – Chemical sketches for glucose (left) and galactose (right) with the O5-C5-C6-O6 (ω) and HO6-O6-C6-C5 (θ) torsional angles indicated for both monosaccharides and the HO4-O4-C4-C5(τ) indicated for galactose

The stationary points at gg, gt and tg positions for each carbohydrate were each geometry optimised at the RB3LYP/6-31G* and RB3LYP/6-31+G** levels of theory. These basis sets are commonly used for medium to large molecules and a comparison of basis sets in a previous study showed that these are adequate for monosaccharides[125].

4.1.2.2 Atoms in Molecules Analysis (Wavefunction analysis)

The wavefunctions for the optimised structures were analysed with Atoms in Molecules (AIM) [50]. A topological analysis of the electron density is considered. Bond critical points

(or (3,-1) points) were searched for as these indicate shared electron density between atoms.

4.1.2.3 Natural Bond Orbital Analysis

Natural Bond Orbital (NBO)[53] analysis was also carried out. The wavefunction is partitioned into its "most natural" orbitals. The NBO approach has been successfully used to explain other chemical phenomenon such as the unusual acidity of Meldrum's acid[99]. The second order perturbation energies about the hydroxymethyl group were studied using NBO 3.0[52,53,96,97] as implemented in GAMESS-UK 7.0[49]

4.1.2.4 Single point energies of extracted PMF coordinates

From the gaseous phase PMF QM/MM dynamics calculations a random collection (ensemble) of structures at each of the stationary points were extracted. The single-point energies were calculated for each extracted structure at the RB3LYP/6-31G*//PM3CARB-1 and RB3LYP/6-31−G**//PM3CARB-1 level. The single-point energies were also calculated with the Semi Empirical PM3CARB-1[5] model using CHARMM version 33b2[48]. The RB3LYP/6-31−G** lowest energy structure for each ensemble was chosen and geometry optimised at the RB3LYP/6-31G*, RB3LYP/6-31+G** and the PM3CARB-1 level of theory.

This data will be presented first as when included in the discussion of the PMF, the discussion becomes disjoint and cumbersome. Its purpose is to investigate the validity of PM3CARB-1 as a generally applicable Semi Empirical method.

Figures 4.2 to 4.5 compare single-point energy profiles for RB3LYP/6-31G*//PM3CARB-1 and RB3LYP/6-31+G**//PM3CARB-1 and PM3CARB-1//PM3CARB-1.

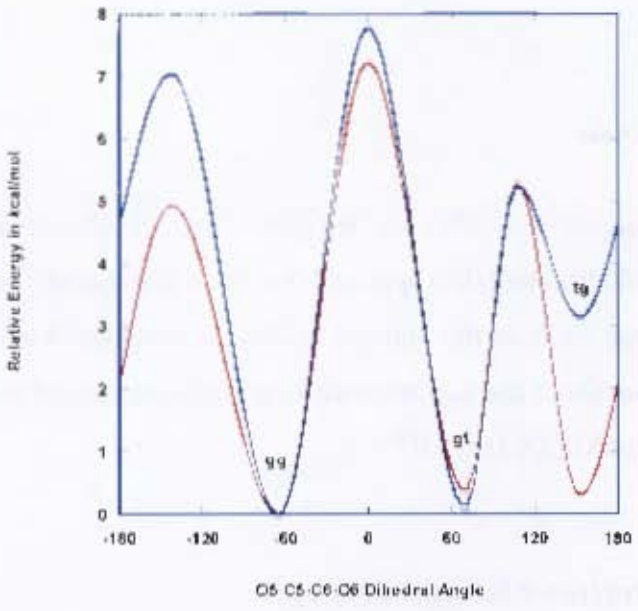

Figure 4.2 – Comparison of an ensemble of static *ab initio* calculations for rotation of the hydroxymethyl group of glucose using RB3LYP/6-31G* (red line) and RB3LYP/6-311G** (blue line).

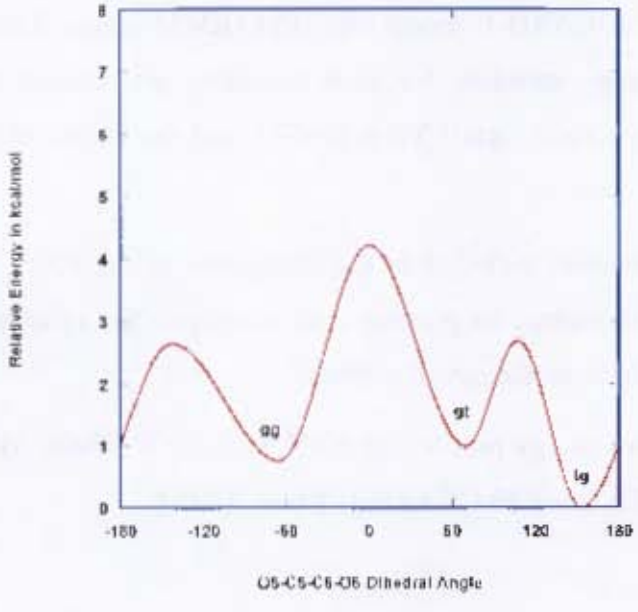

Figure 4.3 – Comparison of an ensemble of static PM3CARB-1 calculations for rotation of the hydroxymethyl group of glucose.

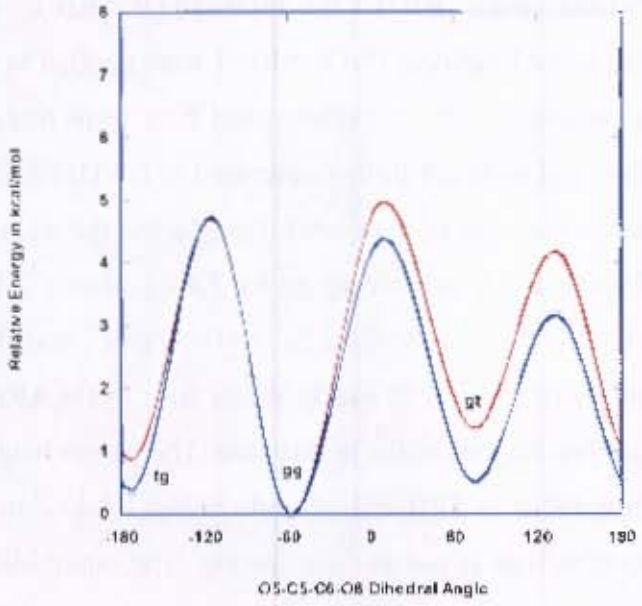

Figure 4.4 – Comparison of an ensemble of static *ab initio* calculations for rotation of the hydroxymethyl group of galactose using RB3LYP/6-31G* (red line) and RB3LYP/6-31+G** (blue line).

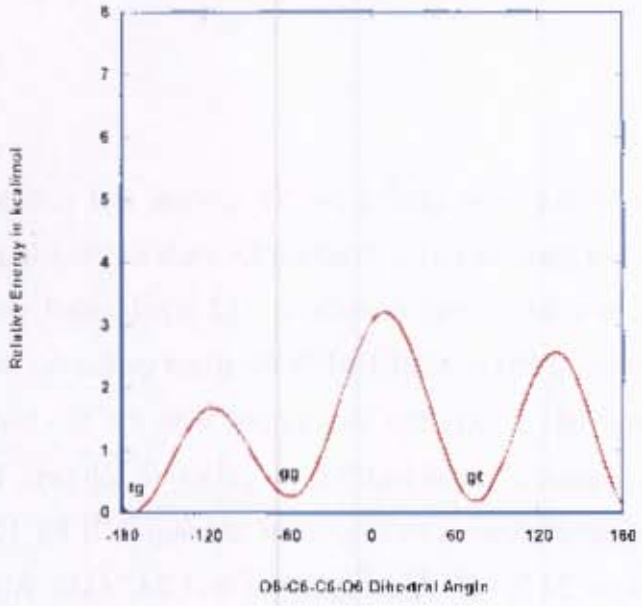

Figure 4.5 - Comparison of an ensemble of static PM3CARB-1 calculations for rotation of the hydroxymethyl group of galactose.

The conformational energies calculated from RB3LYP/6-31G*//PM3CARB-1 and RB3LYP/6-31+G**//PM3CARB-1 and Semi Empirical PM3CARB-1 were applied to the same structures and can be directly compared. The structures used here were directly extracted from the dynamics simulations and were not further optimised in GAMESS-UK. The effect of applying different basis sets can also be compared. For glucose the addition of diffuse functions lowers the preference for tg relative to gg by 2.83 kcal·mol^{-1}. The barrier heights between gg to gt and tg to gg also increase by 2.11kcal·mol^{-1} and 0.55 kcal·mol^{-1} respectively. The tg conformer of glucose is overly stable with PM3CARB-1 suggesting over-estimation of hydrogen bonding strengths in this case. The barrier heights with PM3CARB-1 are also lower than those in DFT calculations being 3-4 kcal·mol^{-1} lower in energy. For galactose the comparison is not as illuminating. The ensemble of conformers for gg is predicted to be the lowest energy ω-angle for both DFT methods. PM3CARB-1 predicts the minima to be tg. The gt conformer is predicted to be much more stable in PM3CARB-1 than with DFT. There is no obvious explanation for this discrepancy. Once again the barrier heights in PM3CARB-1 are much lower than when using DFT methods by approximately 1-3 kcal·mol^{-1}.

4.2 Aqueous phase methodology

Dynamics simulations in the aqueous phase were carried out for glucose and galactose with explicit water molecules. The initial gaseous phase structure for each saccharide was immersed into a 24.86260 Å cubic simulation box containing 512 equilibrated water molecules. The saccharide was treated using PM3CARB-1 while the water molecules were treated using the TIP3P[72,73] MM potential. Waters that overlapped with the saccharide were removed leaving 498 waters in glucose solution and 500 for galactose solution. The system was then minimised (using steepest descent minimisation techniques[54]) for 1000 steps. The box length was adjusted to 24.72020 Å for glucose and 24.74580 Å for galactose such that the correct experimental density (1.006 g·cm-3) for the mass

percentage of these solutions was reflected[7]. The systems were then heated to 298.15K and equilibrated for 2 ns.

Twenty constrained dynamics simulations of 0.4 nanoseconds (ns) in length were performed to give a total of 8ns for each saccharide solution. Weak restraints were applied to the pyranoside ring dihedrals to prevent the transition from a chair to a boat structure, although this was not expected to occur in solution (Refer Appendix D for details of the restraints applied). Unconstrained dynamics simulations were also performed (Refer to Appendix C for the PMF curves).

Solution dynamics were carried out using the canonical ensemble (NVT) at 298.15K and the Langevin dynamics method. The leapfrog-verlet integrator was used with an integration time step of 1 fs. The frictional coefficient of $62.5ps^{-1}$ was applied to all MM atoms (the water molecules)[119]. The SHAKE algorithm[126] was applied to all MM hydrogens. Langevin dynamics provides temperature control and is also vital for improving sampling of phase space. This is particularly important in the aqueous phase. The CHARMM switching functions were applied on a group-by-group basis. The long range interactions were switched to zero between 8 Å and 10.36010 Å for glucose and between 8 Å and 10.37628 Å for galactose. Non-bonded interactions were updated heuristically with the non-bonded list cut-off being set to half the simulation cube side length (12.36010 Å for glucose and 12.37628 Å for galactose).

Periodic boundary conditions were implemented using the crystal keyword in a cubic simulation box. Images were updated heuristically and the minimum energy convention was applied. The image cut-off was set to half the simulation cube side length to eliminate edge effects.

4.3 Results - Vacuum electronic effects and conformational preferences

4.3.1 The Potential of Mean Force

The relative Free Energy curves for the rotation about the ω-torsional angle for glucose and galactose are Figures 4.6 and 4.7. The minimum energy conformations for glucose (with conformer name and ω in brackets) are 0.000 kcal·mol^{-1} (gg, -65.0°), 0.104 kcal·mol^{-1} (tg, 152.5°), 0.311 kcal·mol^{-1} (gt, 70°). Similarly for galactose, minimum energy conformers are 0.000 kcal·mol^{-1} (gt, 75.0°), 0.208 kcal·mol^{-1} (gg, -57.5°), 0.552 kcal·mol^{-1} (tg, -172.5°).

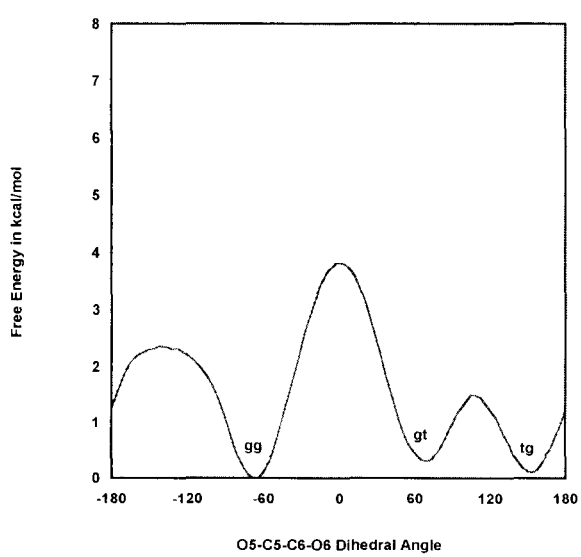

Figure 4.6 – Vacuum PMF (for the Helmholtz Free Energy) for rotation of the hydroxymethyl group of glucose in the gaseous phase using PM3CARB-1 constrained dynamics.

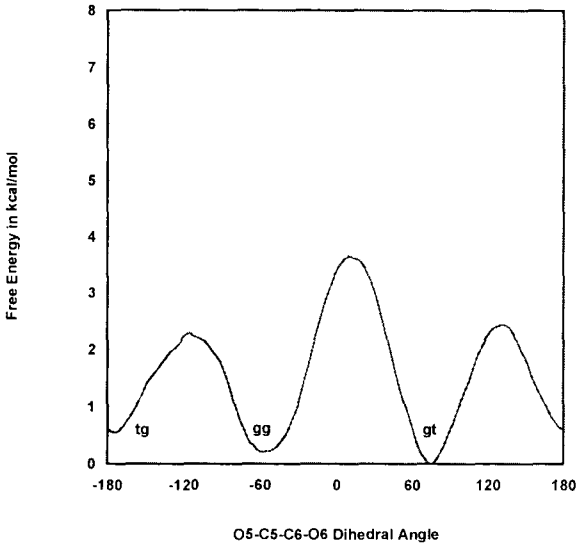

Figure 4.7 – Vacuum PMF (for the Helmholtz Free Energy) for rotation of the hydroxymethyl group of galactose in the gaseous phase using PM3CARB-1 constrained dynamics.

The gaseous phase for glucose show a similar relative ordering to methyl α- and β-D-glucopyranoside results published by Tvaroška[127].

Both saccharides exhibit some gauche preference *in vacuo* but these preferences do not exactly agree with experimental solution NMR[43,44 16] results suggesting that solvent plays a role in determining the conformation of the hydroxymethyl group. Even *in vacuo* the configuration of the 4-hydroxyl alters the order of minimum energy conformers, note the difference in rotational preference for glucose and galactose (Specifically, the gt conformer in galactose).

4.3.1.1 Potential of Mean Force convergence

The procedure described in Chapter 3 was applied to both *in vacuo* glucose and galactose PMF data.

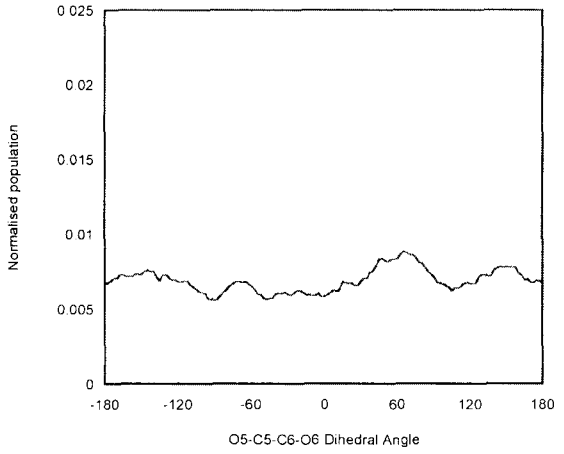

Figure 4.8 – The gaseous phase glucose population distribution from the PMF. This shows a relatively even sampling of phase space. The normalised maximum is 0.00884 and the normalised minimum is 0.00558, where the total population is normalized to 1.

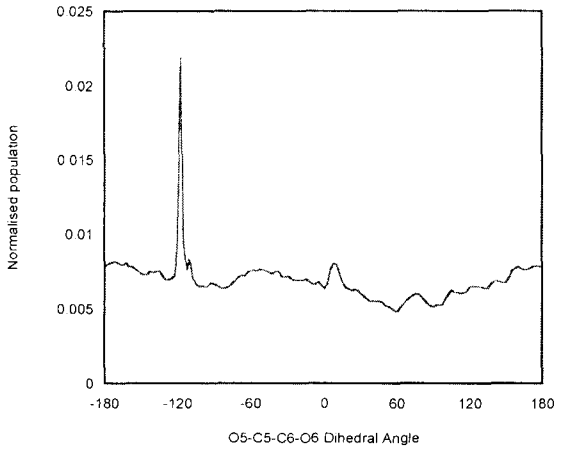

Figure 4.9 – The gaseous phase galactose population distribution from the PMF. This shows a relatively even sampling of phase space. The normalised maximum is 0.02192 and the normalised minimum is 0.00484, where the total population is normalised to 1.

Both the glucose and galactose PMF calculations show thorough sampling of phase space (See Figures 4.8 and 4.9). The ratios of the most to least sampled conformers for glucose

and galactose are 1.58:1.00 and 4.53:1.00 respectively. Figure 4.9 shows a single irregularity at approximately -120° for galactose, however the ratio of most to least sampled conformers is still within the boundaries discussed in Chapter 3.

4.3.1.2 Pucker Analysis

Although weak constraints have been applied to the endocyclic torsion angles (the internal ring torsional angles) of the glucose and galactose ring, a pucker analysis is still necessary (refer to Chapter 3). It is necessary to make sure that the ring hasn't puckered to a boat and to make sure that the ring is not completely rigid (still has some conformational freedom). Figures 4.10 and 4.11 illustrate the puckering for glucose and galactose throughout the PMF simulations.

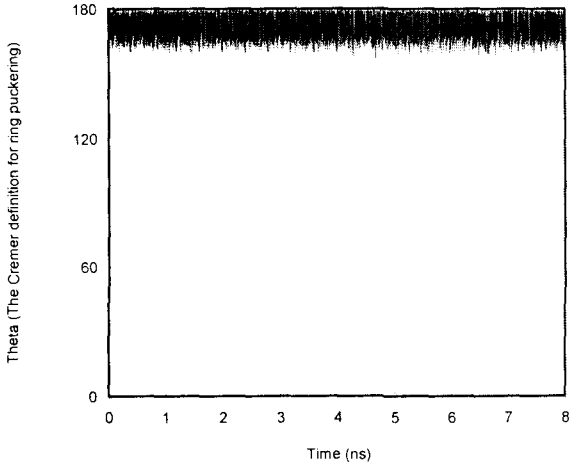

Figure 4.10 – Plot showing the ring puckering of glucose in the gaseous phase.

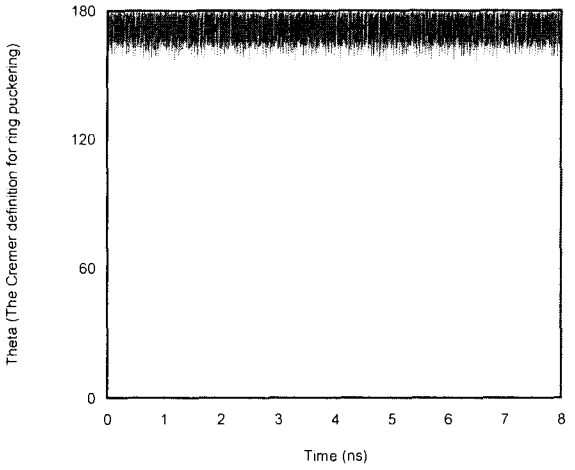

Figure 4.11 – Plot showing the ring puckering of galactose in the gaseous phase.

Even though ring constraints have been applied, a slight puckering from a perfect chair structure is observed (θ (theta) fluctuates near to 180 degrees). The inherent flexibility of the ring still exists.

4.3.2 Population Distributions

The population distribution for each well of the free energy curve was calculated. The border between a well population and a transition population is at the point which the derivative of the free energy curve is at a maximum or at a minimum. The small amount of border population is summed into the nearest well population.

The gaseous phase population distribution for glucose yields a distribution of gg:gt:tg = 39.92:21.51:32.34 and the population distribution for galactose yields gg:gt:tg = 37.32:43.76:16.53 as shown in Table 4.1.

It is valid to record this data to two decimal places as these calculations are very precise. There are a few reasons for this: the nature of the algorithm, the high precision to which numbers can be stored and finally the exceptional sampling ratios (see Section 4.3.1.1).

With ratios below 5:1 for the most to least sampled populations of a simulation the populations fractions calculated have a very low statistical error and are precise.

Table 4.1: β-D-Glucose and β-D-Galactose populations in the gaseous phase

conformer	β-D-Glucose Normalised Population of well	β-D-Galactose Normalised Population of well
gg	39.92	37.32
gt	21.51	43.76
tg	32.34	16.53

Note that the values in Table 4.1 do not sum to the total population (100%) as the PMF algorithm explicitly includes transition populations. The summed population for the all the transition states is 6.23 for glucose and 2.39 for galactose.

4.3.3 Natural Bond Orbital Analysis

Natural Bond Orbital analysis was applied to the *in vacuo* quantum mechanically optimised minima. The most interesting interactions are highlighted in bold in Table 4.2.

Table 4.2: Natural Bond Orbital Analysis for β-D-glucose and β-D-galactose in vacuum; optimised using RB3LYP/6-31+G showing σ to σ* donations**

Rotamer	Glucose Natural Bond orbital interactions (in kcal/mol)		Galactose Natural Bond orbital interactions (in kcal/mol)	
gg	$\sigma_{C5\text{-}O5} \rightarrow \sigma^*_{C6\text{-}H}$	1.27 kcal·mol^{-1}	$\sigma_{C5\text{-}O5} \rightarrow \sigma^*_{C6\text{-}H}$	1.12 kcal·mol^{-1}
	$\sigma_{C6\text{-}H} \rightarrow \sigma^*_{C5\text{-}O5}$	**5.41 kcal·mol^{-1}**	**$\sigma_{C6\text{-}H} \rightarrow \sigma^*_{C5\text{-}O5}$**	**6.23 kcal·mol^{-1}**
	$\sigma_{C6\text{-}H} \rightarrow \sigma^*_{C5\text{-}O5}$	0.69 kcal·mol^{-1}	$\sigma_{C6\text{-}H} \rightarrow \sigma^*_{C5\text{-}O5}$	0.58 kcal·mol^{-1}
	$\sigma_{C5\text{-}H} \rightarrow \sigma^*_{C6\text{-}O6}$	**4.91 kcal·mol^{-1}**	**$\sigma_{C5\text{-}H} \rightarrow \sigma^*_{C6\text{-}O6}$**	**5.40 kcal·mol^{-1}**
	$\sigma_{C6\text{-}O6} \rightarrow \sigma^*_{C5\text{-}H}$	1.33 kcal·mol^{-1}	$\sigma_{C6\text{-}O6} \rightarrow \sigma^*_{C5\text{-}H}$	1.19 kcal·mol^{-1}
gt	$\sigma_{C5\text{-}O5} \rightarrow \sigma^*_{C6\text{-}H}$	1.22 kcal·mol^{-1}	$\sigma_{C5\text{-}O5} \rightarrow \sigma^*_{C6\text{-}H}$	1.30 kcal·mol^{-1}
	$\sigma_{C6\text{-}H} \rightarrow \sigma^*_{C5\text{-}O5}$	**5.49 kcal·mol^{-1}**	**$\sigma_{C6\text{-}H} \rightarrow \sigma^*_{C5\text{-}O5}$**	**5.53 kcal·mol^{-1}**
	$\sigma_{C6\text{-}H} \rightarrow \sigma^*_{C5\text{-}O5}$	0.86 kcal·mol^{-1}	$\sigma_{C6\text{-}H} \rightarrow \sigma^*_{C5\text{-}O5}$	0.94 kcal·mol^{-1}
	$\sigma_{C6\text{-}H} \rightarrow \sigma^*_{C5\text{-}H}$	3.66 kcal·mol^{-1}	$\sigma_{C6\text{-}H} \rightarrow \sigma^*_{C5\text{-}H}$	3.71 kcal·mol^{-1}
	$\sigma_{C5\text{-}H} \rightarrow \sigma^*_{C6\text{-}O6}$	**0.58 kcal·mol^{-1}**	**$\sigma_{C5\text{-}H} \rightarrow \sigma^*_{C6\text{-}O6}$**	**0.68 kcal·mol^{-1}**
	$\sigma_{C5\text{-}H} \rightarrow \sigma^*_{C6\text{-}H}$	3.03 kcal·mol^{-1}	$\sigma_{C5\text{-}H} \rightarrow \sigma^*_{C6\text{-}H}$	3.06 kcal·mol^{-1}
tg	$\sigma_{C5\text{-}O5} \rightarrow \sigma^*_{C6\text{-}O6}$	1.88 kcal·mol^{-1}	$\sigma_{C5\text{-}O5} \rightarrow \sigma^*_{C6\text{-}O6}$	1.87 kcal·mol^{-1}
	$\sigma_{C6\text{-}O6} \rightarrow \sigma^*_{C5\text{-}O5}$	2.07 kcal·mol^{-1}	$\sigma_{C6\text{-}O6} \rightarrow \sigma^*_{C5\text{-}O5}$	2.49 kcal·mol^{-1}
	$\sigma_{C6\text{-}H} \rightarrow \sigma^*_{C5\text{-}H}$	3.36 kcal·mol^{-1}	$\sigma_{C6\text{-}H} \rightarrow \sigma^*_{C5\text{-}H}$	3.43 kcal·mol^{-1}
	$\sigma_{C6\text{-}H} \rightarrow \sigma^*_{C5\text{-}O5}$	**1.15 kcal·mol^{-1}**	**$\sigma_{C6\text{-}H} \rightarrow \sigma^*_{C5\text{-}O5}$**	**0.97 kcal·mol^{-1}**
	$\sigma_{C5\text{-}H} \rightarrow \sigma^*_{C6\text{-}O6}$	**0.58 kcal·mol^{-1}**	**$\sigma_{C5\text{-}H} \rightarrow \sigma^*_{C6\text{-}O6}$**	**0.67 kcal·mol^{-1}**
	$\sigma_{C5\text{-}H} \rightarrow \sigma^*_{C6\text{-}H}$	2.84 kcal·mol^{-1}	$\sigma_{C5\text{-}H} \rightarrow \sigma^*_{C6\text{-}H}$	2.99 kcal·mol^{-1}

These interactions are bonding orbital to antibonding orbital interactions between C-H and C-O orbitals, specifically $\sigma_{C6\text{-}H} \rightarrow \sigma^*_{C5\text{-}O5}$ and $\sigma_{C5\text{-}H} \rightarrow \sigma^*_{C6\text{-}O6}$.

For a gg conformer the $\sigma_{C6\text{-}H} \rightarrow \sigma^*_{C5\text{-}O5}$ and $\sigma_{C5\text{-}H} \rightarrow \sigma^*_{C6\text{-}O6}$ interactions are represented in Figure 4.12.

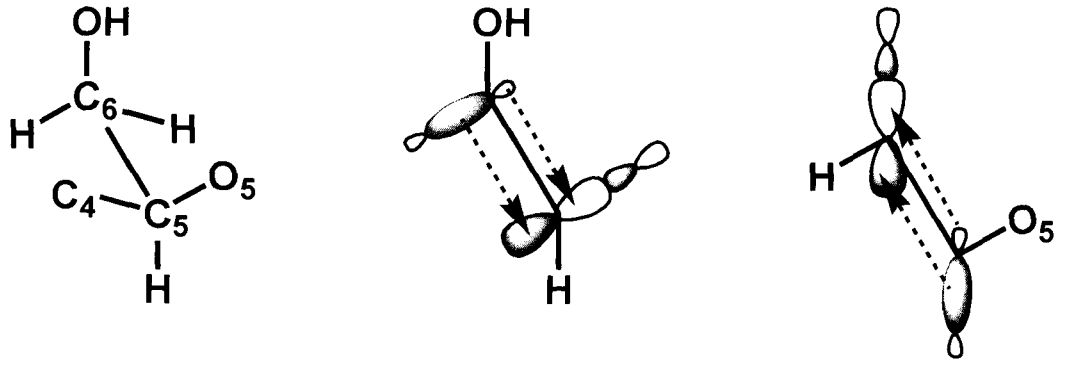

Figure 4.12 – Left : A sawhorse-type projection of the atoms connected to the C5-C6 bond Centre: Illustration of donation of electron density from $\sigma_{C6\text{-}H} \rightarrow \sigma^*_{C5\text{-}O5}$ in the gg rotamer. Right: Illustration of donation of electron density from $\sigma_{C5\text{-}H} \rightarrow \sigma^*_{C6\text{-}O6}$ in the gg rotamer.

These orbital interactions are most stabilising in the order gg > gt >tg for both glucose and galactose. In the gg conformer the best possible donor and acceptor bonds are anti to one another, that is C6-H (a bonding donor) is opposite to C5-O5 (an anti-bonding acceptor) and at the same time C5-H (a bonding donor) is opposite to C6-O6 (an anti-bonding acceptor) (Figure 4.12). The NBO model intrinsic stabilisation is the same as that predicted by Kirby's definition of the gauche effect[11] (Refer to Chapter 1).

The stabilisation effect of the $\sigma_{C6\text{-}H} \rightarrow \sigma^*_{C5\text{-}O5}$ and $\sigma_{C5\text{-}H} \rightarrow \sigma^*_{C6\text{-}O6}$ is likely to change as we deviate from the low energy stationary points of the PMF. NBO analysis was applied to conformers that show maximum ω-torsional angle deviation ($\Delta\omega$) but still belong in a specific well. These values are shown in Table 4.3 and 4.4 and the general trend is that low energy stationary points have a high NBO orbital overlap.

Table 4.3: β-D-Glucose overlap *in vacuo*

conformer	Total NBO overlap for $\sigma_{C6\text{-}H} \rightarrow \sigma^*_{C5\text{-}O5}$ and $\sigma_{C5\text{-}H} \rightarrow \sigma^*_{C6\text{-}O6}$ at 'ω -Δω' (kcal·mol^{-1})	Total NBO overlap for $\sigma_{C6\text{-}H} \rightarrow \sigma^*_{C5\text{-}O5}$ and $\sigma_{C5\text{-}H} \rightarrow \sigma^*_{C6\text{-}O6}$ at 'ω' (kcal·mol^{-1})	Total NBO overlap for $\sigma_{C6\text{-}H} \rightarrow \sigma^*_{C5\text{-}O5}$ and $\sigma_{C5\text{-}H} \rightarrow \sigma^*_{C6\text{-}O6}$ at 'ω +Δω' (kcal·mol^{-1})
gg	8.73	10.31	7.65
gt	6.27	7.04	7.45
tg	4.59	2.94	0.56

Table 4.4: β-D-Galactose NBO overlap *in vacuo*

conformer	Total NBO overlap for $\sigma_{C6\text{-}H} \rightarrow \sigma^*_{C5\text{-}O5}$ and $\sigma_{C5\text{-}H} \rightarrow \sigma^*_{C6\text{-}O6}$ at 'ω -Δω' (kcal·mol^{-1})	Total NBO overlap for $\sigma_{C6\text{-}H} \rightarrow \sigma^*_{C5\text{-}O5}$ and $\sigma_{C5\text{-}H} \rightarrow \sigma^*_{C6\text{-}O6}$ at 'ω' (kcal·mol^{-1})	Total NBO overlap for $\sigma_{C6\text{-}H} \rightarrow \sigma^*_{C5\text{-}O5}$ and $\sigma_{C5\text{-}H} \rightarrow \sigma^*_{C6\text{-}O6}$ at 'ω +Δω' (kcal·mol^{-1})
gg	10.56	12.1	6.67
gt	4.97	7.17	5.93
tg	2.84	0.84	2.28

4.3.4 Hydrogen bonding

The NBO analysis doesn't adequately describe the gaseous phase preferences for either glucose or galactose. The existence and strength of intramolecular hydrogen bonds must still be accounted for.

Tvaroška et al. considered the energy profiles of the hydroxymethyl moiety of multiple glycero-hexopyranoside rings[46] (dehydroxylated glucose rings). In these studies the hydrogen of the 6-hydroxyl group was explicitly turned away such that no intramolecular hydrogen bonds could be formed. These conformers were then compared with those that

could form intramolecular hydrogen bonds from O6-H to O5 leading to the conclusion that hydrogen bonding caused gg and gt to be preferred. Furthermore it was postulated that these intramolecular hydrogen bonds were the sole cause of the gauche effect. In some of their more recent work Tvaroška et al.[34] have considered more realistic monosaccharide derivatives and have used a similar method to postulate the existence of 6-hydroxyl to 4-hydroxyl hydrogen bonds. Brown[83] also considered hydrogen bonds using the somewhat arbitrary distance and angle criterion for defining these bonds.

The sharing of electrons (i.e. a covalent interaction) between the donor and acceptor atoms is what distinguishes between hydrogen bonds and strong electrostatic interactions. A more rigorous analysis such as the AIM analysis should be applied[15]. The Atoms in Molecules analysis of the QM optimised structures for glucose and galactose was used to determine if hydrogen bonds exist. According to articles by Popelier[128,129], there are eight criteria that should be met to characterise a hydrogen bond. However, "by observation one of these conditions has been proven to be sufficient as well"[128]. Two important and logical criteria are that a bond critical point exists (in a topologically reasonable position, see Figures 4.15 – 4.18) and that the density (in au) be between 0.002 and 0.035 au.

A (3,-1) bond critical point was found between O4 and O6-H for both the tg conformer of glucose and the gg conformer of galactose (See Figures 4.15 and 4.16). No other hydrogen bonds were found for the hydroxymethyl group. However, in galactose a hydrogen bond was found to exist in all conformers studied between the axial 4-hydroxyl and equatorial 3-hydroxyl group. The values for the electron density (ρ) were 0.0244 au for glucose and 0.0251 au for galactose (with the RB3LYP/6-31G* basis) where O6 is the donor and O4 the acceptor as is shown in Figure 4.15 and 4.16. The galactose 4-hydroxyl to 3-hydroxyl hydrogen bond (Figure 4.16-18) has similar electron densities for all minima; gg = 0.0190 au, gt = 0.0196 au, tg = 0.0204 au. All of these densities are within the range indicated by Popelier[128,129]. For comparison, the average density values for covalent C-H and O-H bonds are 0.28 au and 0.34 au respectively.

Figure 4.13 – Glucose, the gg conformer. The atomic centres are indicated as large coloured balls (Carbon is black, Oxygen is red, Hydrogen is white). Bond paths are drawn between the atomic centres. On each bond path is a bond critical point indicated by a red microsphere. The yellow microsphere indicates a ring critical point.

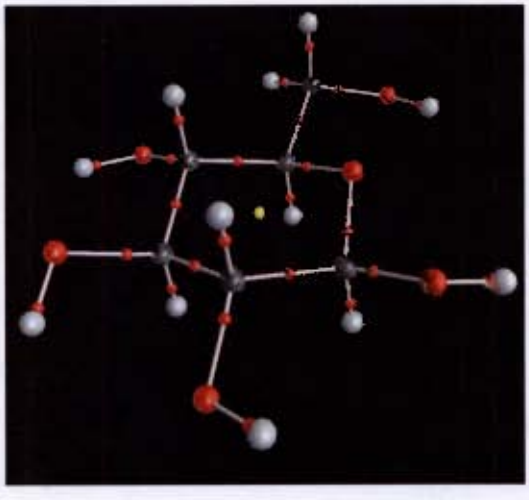

Figure 4.14 – Glucose, the gt conformer. The atomic centres are indicated as large coloured balls (Carbon is black, Oxygen is red, Hydrogen is white). Bond paths are drawn between the atomic centres. On each bond path is a bond critical point indicated by a red microsphere. The yellow microsphere indicates a ring critical point.

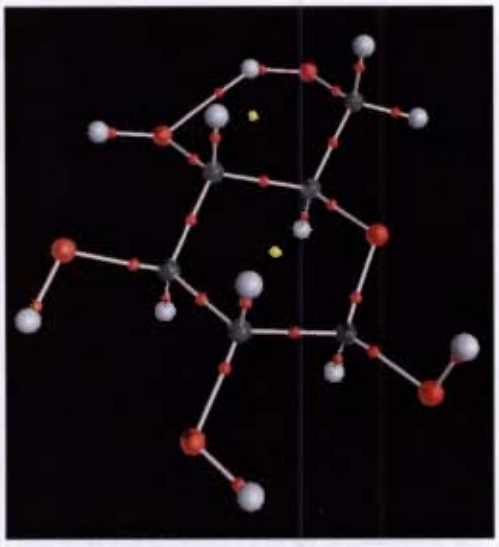

Figure 4.15 Glucose, the tg conformer. The atomic centres are indicated as large coloured balls (Carbon is black, Oxygen is red, Hydrogen is white). Bond paths are drawn between the atomic centres. On each bond path is a bond critical point indicated by a red microsphere. The yellow microsphere indicates a ring critical point. In this conformer, a bond critical point is found between the O4 and HO6 indicative of hydrogen bonding (compare to gg and gt conformers and note the new bond path and bond critical point).

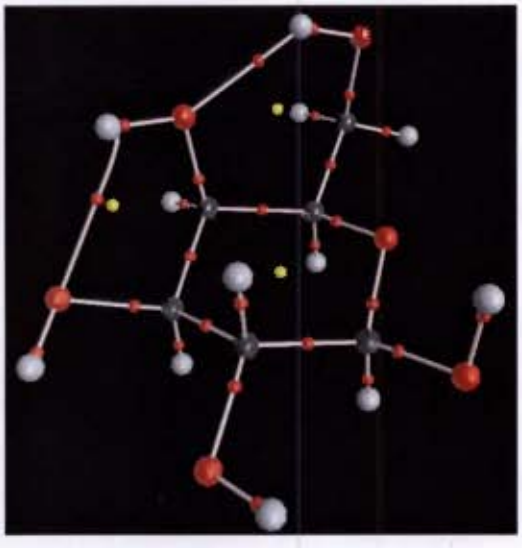

Figure 4.16 – Galactose, the gg conformer. The atomic centres are indicated as large coloured balls (Carbon is black, Oxygen is red, Hydrogen is white). Bond paths are drawn between the atomic centres. On each bond path is a bond critical point indicated by a red microsphere. The yellow microspheres indicate ring critical points. In this conformer a bond critical point is found between the O4 and HO6 indicative of hydrogen bonding. In this galactose conformer, a HO4-O3 hydrogen bond is also noted.

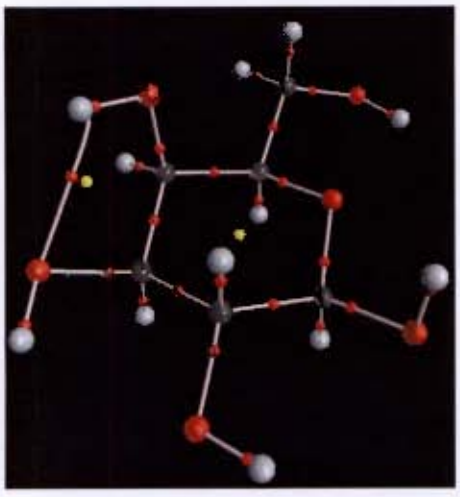

Figure 4.17 – Galactose, the gt conformer. The atomic centres are indicated as large coloured balls (Carbon is black, Oxygen is red, Hydrogen is white). Bond paths are drawn between the atomic centres. On each bond path is a bond critical point indicated by a red microsphere. The yellow microspheres indicate ring critical points. In this galactose conformer, only a HO4-O3 hydrogen bond is noted.

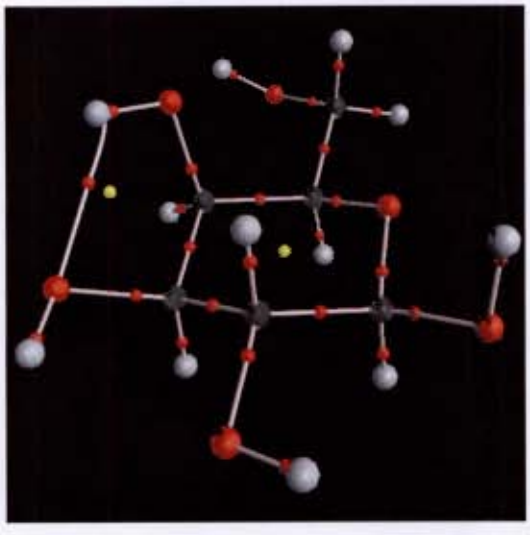

Figure 4.18 Galactose, the tg conformer. The atomic centres are indicated as large coloured balls (Carbon is black, Oxygen is red, Hydrogen is white). Bond paths are drawn between the atomic centres. On each bond path is a bond critical point indicated by a red microsphere. The yellow microspheres indicate ring critical points. In this galactose conformer, only a HO4-O3 hydrogen bond is noted.

The only hydrogen bond which can affect the rotameric distribution for the hydroxymethyl group can exist between the hydroxyl of the hydroxymethyl group (6-hydroxyl) and the

hydroxyl at C-4 (4-hydroxyl). It is clear from AIM analysis that only the tg conformer of glucose (Figure 4.15) and the gg conformer of galactose (Figure 4.16) can form a hydrogen bond between the 6-hydroxyl and 4-hydroxyl groups.

AIM analysis was only applied to static molecules and other types of hydrogen bond type interactions are possible between the 6-hydroxyl, 4-hydroxyl for the tg conformer of glucose and the gg conformer of galactose. These are shown in Figure 4.19. The darkened-wedged bonds represent bonds pointing out of the page at ca. 60° to the plane of the paper while the outlined-wedged bonds represent bonds pointing into the page at ca. 60° to the plane of the paper.

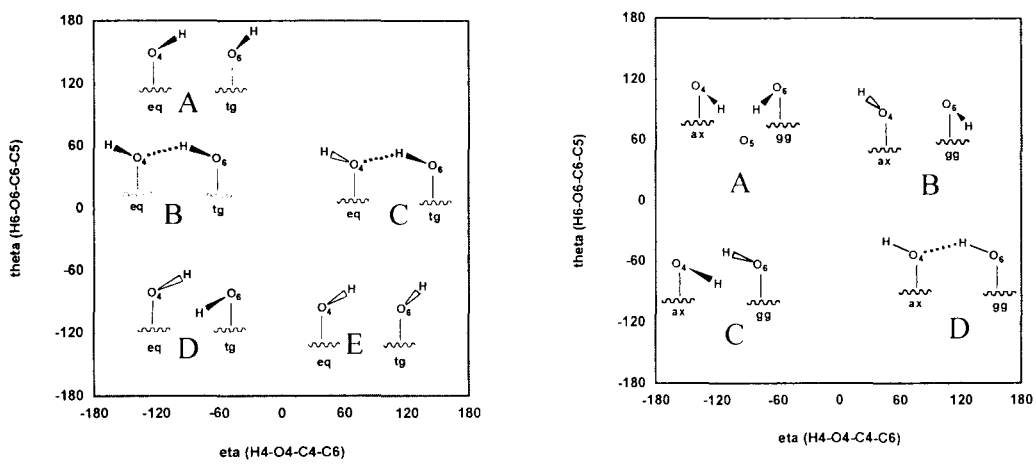

Figure 4.19 – Illustrations of the potential interactions between the 4-hydroxy and 6-hydroxy group of the tg conformer of glucose (left) and the gg conformer of galactose (right) when rotating the θ- and η-torsional angles(H6-O6-C6-O5 and H4-O4-C4-C6 respectively). Darkened wedges point out of the plane of the paper and outlined wedges point into the plane of the paper.

How important are hydrogen bonding interactions in a statistical sense? This can be gauged by analysing the direction in which the hydrogen of the 4-hydroxyl and the 6-hydroxyl are pointing for the entire gaseous phase dynamics trajectory. The torsional angles considered were H6-O6-C6-C5 (theta, θ) and H4-O4-C4-C6 (eta,η - a pseudo-

torsional angle (pseudo as not all atoms are bonded)) for the tg conformer of glucose and the gg conformer of galactose. From this analysis the probability distribution can be plotted as contours as in Figure 4.20. The tg conformer of glucose favours the B orientation as shown in Figures 4.19 and 4.20, while the gg conformer of galactose favours the D orientation as described by Figures 4.19 and 4.20 for the entire gaseous phase trajectory.

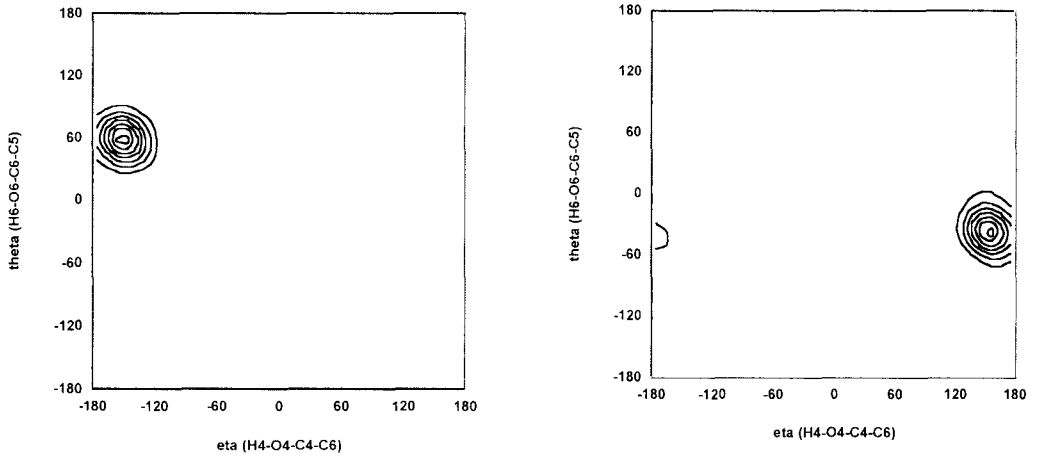

Figure 4.20 – H-bond 2D vacuum contour plot for glucose (left) and galactose (right). Every contour is a 10% increment of the maximum density. One smoothing pass has been applied to this data to reduce artifacts.

Coupling the statistical interpretation of the intramolecular hydrogen bond with AIM analysis provides a strong argument for a **real** covalent intramolecular hydrogen bond which dominates in the tg and gg conformer for glucose and galactose respectively.

4.3.5 Other molecular interactions

Steric interactions and unfavourable dipole-dipole interactions should also be considered. The most sterically hindered conformer with the most unfavourable dipole-dipole interactions between O4, O5, O6 is undoubtedly the gg conformer of galactose[34].

4.3.6 Summary

For glucose, intrinsic stereoelectronic $\sigma_{C6\text{-}H} \rightarrow \sigma^*_{C5\text{-}O5}$ and $\sigma_{C5\text{-}H} \rightarrow \sigma^*_{C6\text{-}O6}$ interactions favour the gauche conformers. However, the tg conformer is more favoured than expected because of an intramolecular hydrogen bond which is formed between the 4-hydroxyl group and 6-hydroxyl group. The overall vacuum population for glucose is gg>tg>gt.

For galactose, intrinsic stereoelectronic $\sigma_{C6\text{-}H} \rightarrow \sigma^*_{C5\text{-}O5}$ and $\sigma_{C5\text{-}H} \rightarrow \sigma^*_{C6\text{-}O6}$ interactions favour the gauche conformers. The preference for gg is modified due to 1, 3 diaxial steric interactions, dipole-dipole interactions and intramolecular hydrogen bonding. The unfavourable steric and dipole-dipole interactions seem to be moderated by the favourable intramolecular hydrogen bond between the 4-hydroxyl group and 6-hydroxyl group. A similar rationalisation was made by Tvaroška[34] although De Vries and Buck disagree[20]. The overall gaseous phase population is gt>gg>tg.

4.4 Solvent electronic effects and conformational preferences

4.4.1 The Potential of Mean Force

The relative Free Energy curves for the rotation about the ω-torsional angle for glucose and galactose in the aqueous phase are shown in Figures 4.21 and 4.22. The minimum energy conformations for glucose (with conformer name and ω in brackets) are 0.000 kcal·mol^{-1} (gg, -70°), 0.334 kcal·mol^{-1} (gt, 70°), 1.648 kcal·mol^{-1} (tg, 160°). Similarly for galactose, minimum energy conformers are 0.000 kcal·mol^{-1} (gt, 70°), 0.955 kcal·mol^{-1} (tg, 175°), 1.401 kcal·mol^{-1} (gg, -57.5°).

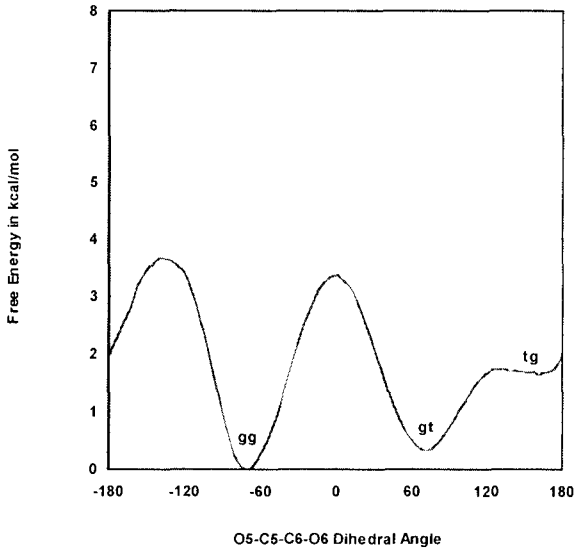

Figure 4.21 – Aqueous solution PMF (for the Helmholtz Free Energy) for glucose using PM3CARB-1 QM/MM constrained dynamics.

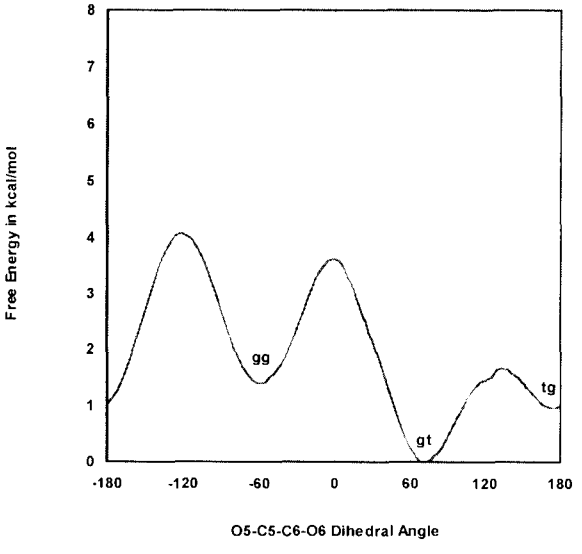

Figure 4.22 – Aqueous solution PMF (for the Helmholtz Free Energy) for galactose using PM3CARB-1 QM/MM constrained dynamics.

Clearly, inclusion of solvent impacts the rotational preference about the hydroxymethyl group. The barrier heights in the PMF curves are affected by solvation. The barrier height for a gg to gt transition is not significantly different for either glucose or galactose. The gg to gt barrier for both monosaccharides increases by ca. 2 kcal·mol^{-1}, while gt to tg transition lowers by ca. 1 kcal·mol^{-1} for galactose.

The aqueous solution phase preference now correctly reflects NMR results[16,43,44]. An important question is the following: **In what way has solvation chemically affected the distribution preference?**

4.4.1.1 Potential of Mean Force convergence

As for gaseous phase calculations the ratio of most to least sampled conformers is a means to check the convergence of the PMF procedure.

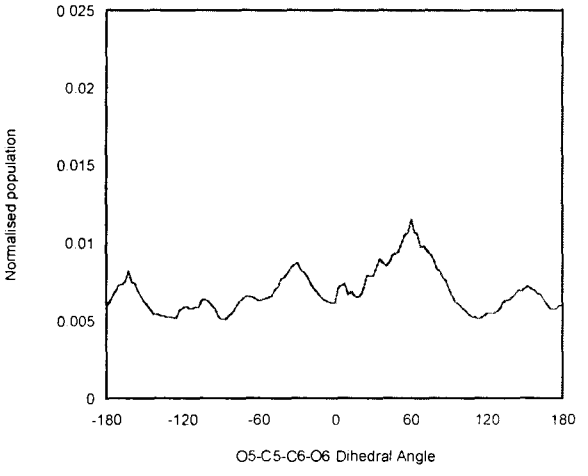

Figure 4.23 – The aqueous glucose population distribution from the PMF. This shows relatively even sampling of phase space. The normalised maximum is 0.01154 and the normalised minimum is 0.00511, where the total population is normalised to 1.

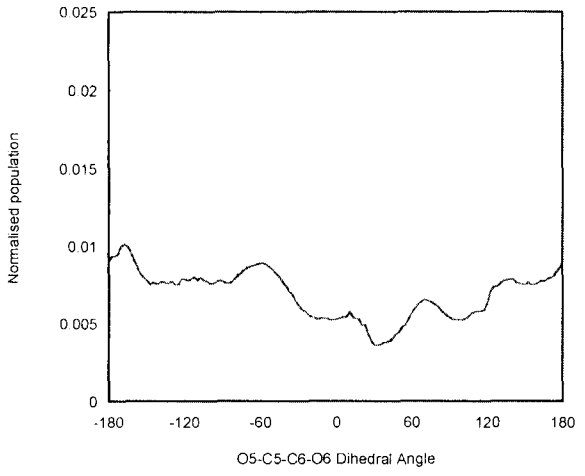

Figure 4.24 – The aqueous galactose population distribution from the PMF. This shows relatively even sampling of phase space. The normalised maximum is 0.01015 and the normalised minimum is 0.00361, where the total population is normalised to 1.

Both the glucose and galactose PMF calculations show thorough sampling of phase space (See Figures 4.23 and 4.24). The ratios of the most to least sampled conformers for glucose and galactose are 2.56:1.00 and 2.81:1.00 respectively. This is a good result as the ratios fall within the limits described in chapter 3.

4.3.1.2 Pucker Analysis

Although weak constraints have been applied to the endocyclic torsion angles of the glucose and galactose ring in solution, a pucker analysis was still carried out. As described previously, one makes sure that the ring hasn't puckered to a boat and that the ring is still conformationally flexible.

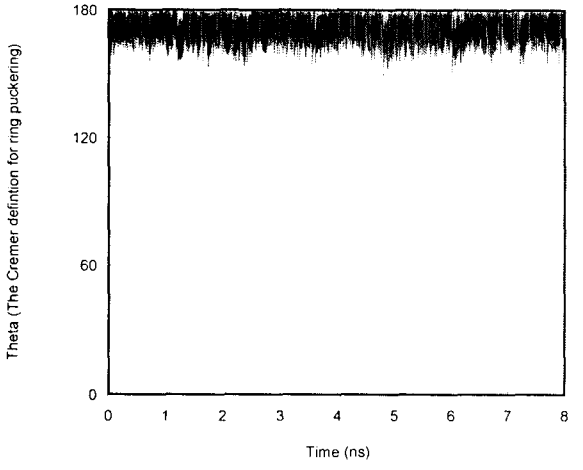

Figure 4.25 – Plot showing the ring puckering of glucose in water.

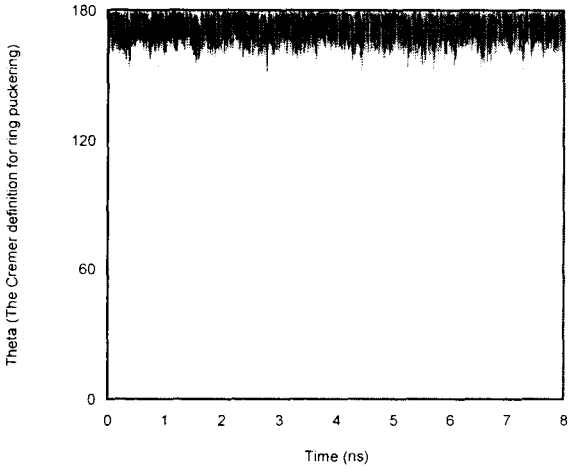

Figure 4.26 – Plot showing the ring puckering of galactose in water.

Figure 4.25 and 4.26 illustrate that the $^4C_1$ chair form is maintained throughout the simulation. However, the flexibility of the ring is still evident and this flexibility is a crucial part of monosaccharide (and higher saccharide) behaviour.

4.4.2 Natural Bond Orbital Analysis

The intrinsic stereoelectronic preference predicted by Natural Bond Analysis must still exist. Remember that the $\sigma_{C6\text{-}H} \rightarrow \sigma^*_{C5\text{-}O5}$ and $\sigma_{C5\text{-}H} \rightarrow \sigma^*_{C6\text{-}O6}$ interactions (shown in Figure 4.12) predict the gauche conformers are preferred with gg > gt > tg. After solvation the gg > gt > tg preference predicted by NBO holds true for glucose.

4.4.3 Hydrogen bonding

Once again, the effect of hydrogen bonding is important. Grindley and Rockwell[21] initially concluded that sugar-water hydrogen bonds were not important in determining hydroxymethyl preferences. In contrast, research by Kirschner and Woods on methoxylated gluco- and galacto-pyranosides in different solvents showed that intramolecular hydrogen bonding is weakened in water[130]. They further postulated that only solvation correctly reproduces the hydroxymethyl preferences observed in NMR experiments[82]. Naidoo and Chen published work[31,37] on the carbohydrate hydrogen bond strengths and on the competitive nature of sugar-water hydrogen bond interactions. De Vries and Buck showed that intramolecular hydrogen bonds can exist in solvents of relatively low polarity such as $CDCl_3$[20]. Intuitively, we do expect intramolecular hydrogen bonds to be weakened, if not completed disrupted, in a polar solvent such as water.

In the gaseous phase, the internal hydrogen bonding present in the tg conformer of glucose and the gg conformer of galactose affected the rotational freedom of the hydroxymethyl group. A statistical analysis of these conformers was carried out for the solution PMF. Although hydrogen bonding interactions were dominant in the gaseous phase, a different preference is observed in solution. Figure 4.27 shows a more expansive landscape for the hydroxyl groups. All the motifs shown in Figure 4.14 are sampled throughout the solvated dynamics simulation.

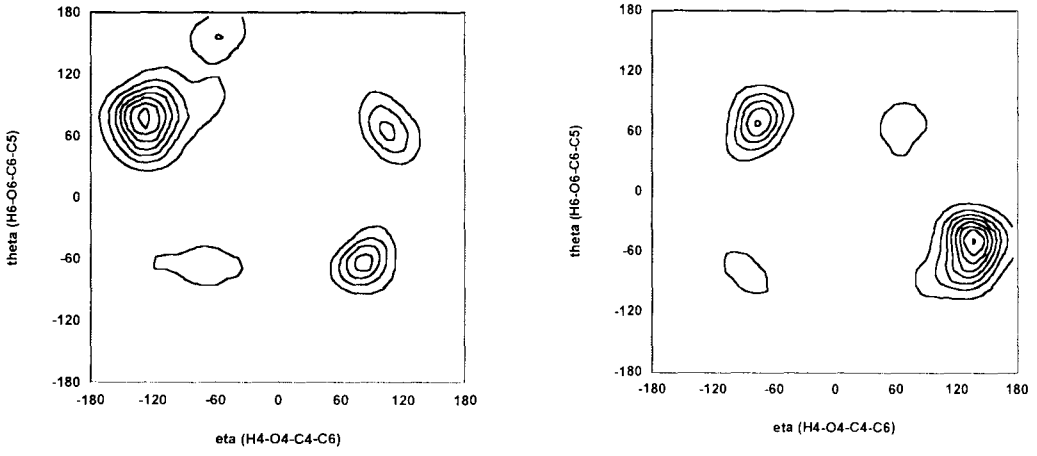

Figure 4.27 – H-bond 2D water contour plot for glucose (left) and galactose (right). Every contour is a 10% increment of the maximum density. One smoothing pass has been applied to this data to reduce artifacts.

It is quite clear that the competitive nature of a polar solvent like water which has the ability to form dynamic saccharide-water hydrogen bond interactions can disrupt or at least reduce the preference for intramolecular hydrogen bonds. Disruption of the intramolecular hydrogen bonding interaction lowers the preference for the tg conformer of glucose and the gg conformer of galactose by ca. 1.5 kcal·mol^{-1}. **This value is a good estimate of the intramolecular hydrogen bond strength**.

4.4.4 Population analysis in the aqueous phase

The population distribution for the aqueous phase can also be calculated yielding gg:gt:tg of 59.21:34.88:0.83 for glucose and gg:gt:tg of 3.32:79.60:10.15 for galactose as shown in Table 4.5. As for the gaseous case, the sum of the transition state populations is the difference between the sum of the well populations and 100%. The transition populations for glucose and galactose in solution are 5.09 and 6.93 respectively.

Table 4.5: β-D-Glucose and β-D-Galactose populations in the aqueous phase

conformer	β-D-Glucose Normalised Population of well	β-D-Galactose Normalised Population of well
gg	59.21	3.32
gt	34.88	79.60
tg	0.82	10.15

These distributions agree quite well with experimental NMR values. Bock et al.[16] reported gg:gt:tg for glucose to be 52:41:7 while Nishida et al.[43,44] reported gg:gt:tg = 53:45:2 for glucose and gg:gt:tg = 18:61:21[43] for galactose. These NMR studies used chiral deuteration at C6 to aid in the assignment of proton ^{1}H NMR spectra. A variety of glucose derivatives with protecting groups were also studied and used to validate the assignments made.

The NMR experiments conducted on monosaccharides so far cannot give any indication of transition state populations (perhaps this would be possible using low temperature NMR). Another way of comparing the computational and experimental work is to modify the computational populations by incorporating the transition state populations into the nearest well population. This gives gg:gt:tg = 59.62:38.78:1.60 for glucose and gg:gt:tg = 3.36:85.45:11.19 for galactose. These populations also compare favourably with the experimental data.

4.4.5 Summary

In summary, solvation of glucose changes the preferred population from gg>tg>gt to gg>gt>tg. The strong intramolecular hydrogen bond in the tg conformer made it lower in energy in the gaseous phase. Enthalpically this must still have approximately the same strength. In solution the water-hydroxyl hydrogen bonds compete with the intramolecular hydrogen bond between the 4-hydroxyl and 6-hydroxyl in the tg conformer. We have shown that intramolecular hydrogen bonds can be disrupted by competitive water-hydroxyl interactions and that 4-hydroxyl and 6-hydroxyl phase space is much more freely sampled in solution (See Figures 4.19 and 4.27).

Similarly, solvation of galactose changes the preferred population from gt>gg>tg to gt>tg>gg. The gg conformer had unfavourable steric and dipole-dipole interactions which were 'reduced' by a strong hydrogen bond interaction in the gaseous phase. As with the tg conformer of glucose, solvation of the gg conformer of galactose disrupts the intramolecular hydrogen bond. Figures 4.19 and 4.27 show the conformational motifs sampled during the dynamics simulation. Although the gg conformer of galactose contains the favourable hyperconjugative resonance described by NBO; the unfavourable steric and dipole-dipole interactions dominate and the gg conformer is not preferred.

CONCLUSIONS

The PM3CARB-1 model has been successful at describing the energetics and conformers of monosaccharides in the gas and solution phase. It can be considered a good substitute for time-consuming *ab initio* calculations. Although PM3CARB-1 is slower than MM methods, specific bond, angle and torsional angle parameter terms are explicitly accounted for meaning that it is more easily applicable to generic carbohydrate simulations. Furthermore, the effect of electrons is implicitly included in these simulations unlike in MM simulations. The barrier heights for rotation of the hydroxymethyl group were found to be lower in PM3CARB-1 calculations than for the DFT calculations carried out. This suggests that some improvements can be made to the PM3CARB-1 method. The comparison of the 6-31G* and 6-31+G** basis sets in the DFT calculations suggests that diffuse functions are important and should be implemented where possible.

Computationally, the PM3CARB-1 model is not particularly difficult to implement. One must make sure that the simulation program used (CHARMM[48] in this case) can correctly perform Semi Empirical calculations (by testing against benchmarks). The molecule to be simulated can be implemented as usual in CHARMM[48] with minor modifications to the topology (all atoms must have no MM charge, as shown in Appendix A). At present the lack of parallelisation of the Semi Empirical code limits its application. However, the PM3CARB-1 model is easy to implement and is a good model for describing conformations in carbohydrates.

The gauche preference in two monosaccharides, glucose and galactose, has been explained by using QM/MM dynamics. The Helmholtz free energy for the rotation of the hydroxymethyl group was calculated by implementing umbrella sampling and WHAM with PMF techniques. Electronic effects that contributed to the gauche preference in these monosaccharides were explained by applying AIM and NBO analysis. Specifically, the role of the hyperconjugative effect between $\sigma_{C\text{-}H}$ and $\sigma^*_{C\text{-}O}$ orbitals and the effect of intramolecular hydrogen bonding. Solvent (water) affected the gauche rotational preference by competitive hydrogen bond interactions with the monosaccharide hydroxyl

groups. The solution gauche preferences and populations for glucose and galactose (gg:gt:tg = 59.21:34.88:0.83 and gg:gt:tg = 3.32:79.60:10.15.) agreed well with experimental NMR[16,43,44] data.

In summary, a variety of methods have been implemented in this dissertation and these have been successful at describing the role of electronic and solvent effects in monosaccharide conformational preferences.

Appendix

A.1 – PM3CARB-1 Parameters

```
USS H -13.514849
BETAS H -4.011786
ALFA H 2.753199
USS O -90.938073
UPP O -76.932200
BETAS O -44.449581
BETAP O -35.343869
ALFA O 3.031867
END
```

A.2 – Force Field Modifications

```
RESI BGLQ   0.00000   ! 4C1 Beta-D-glucopyranose monomer
GROU                 ! Quantum Ready
ATOM C1   CBS    0.000 !
ATOM H1   HAS    0.000 !
ATOM O1   OHS    0.000 !
ATOM HO1  HOS    0.000 !
ATOM C5   CTS    0.000 !
ATOM H5   HAS    0.000 !
ATOM O5   OES    0.000 !
ATOM C2   CTS    0.000
ATOM H2   HAS    0.000
ATOM O2   OHS    0.000
ATOM HO2  HOS    0.000
ATOM C3   CTS    0.000
ATOM H3   HAS    0.000
ATOM O3   OHS    0.000
ATOM HO3  HOS    0.000
ATOM C4   CTS    0.000
ATOM H4   HAS    0.000
ATOM O4   OHS    0.000
ATOM HO4  HOS    0.000
ATOM C6   CPS    0.000
ATOM H61  HAS    0.000
```

```
ATOM H62  HAS    0.000
ATOM O6   OHS    0.000
ATOM HO6  HOS    0.000
BOND C1  O1     C1  H1     O1  HO1    C1  O5     C1  C2
BOND C2  H2     C2  O2     O2  HO2    C2  C3     C3  H3
BOND C3  O3     O3  HO3    C3  C4     C4  H4     C4  O4
BOND O4  HO4    C4  C5     C5  H5     C5  C6     C6  H61
BOND C6  H62    C6  O6     O6  HO6    C5  O5
!  I   J   K   L    R(IK)  T(IKJ)   PHI   T(JKL)  R(KL)
IC  O1  C2 *C1  H1    1.3890 105.75 114.54 108.17  1.0950
IC  O1  O5 *C1  C2    1.3890 111.55 117.06 110.06  1.5340
IC  O2  C3 *C2  H2    1.4154 112.27 -118.21 108.23  1.0919
IC  O2  C1 *C2  C3    1.4154 110.87 -125.56 111.08  1.5253
IC  O3  C4 *C3  H3    1.4157 110.61 120.65 108.81  1.1068
IC  O3  C2 *C3  C4    1.4157 108.09 120.77 109.86  1.5177
IC  O4  C5 *C4  H4    1.4252 110.90 -120.61 108.35  1.1024
IC  O4  C3 *C4  C5    1.4252 108.31 -122.08 111.17  1.5287
IC  C6  O5 *C5  H5    1.5099 108.10 118.69 109.65  1.1042
IC  C6  C4 *C5  O5    1.5099 111.57 119.10 108.69  1.4274
IC  O6  H62 *C6  H61   1.4132 110.47 -120.32 107.85  1.0945
IC  O6  C5 *C6  H62   1.4132 110.45 -121.53 108.99  1.0959
IC  O5  C1  C2  C3    1.4254 110.06  54.09 111.08  1.5253
IC  C1  C2  C3  C4    1.5340 111.08 -51.23 109.86  1.5177
IC  C2  C3  C4  C5    1.5253 109.86  53.25 111.17  1.5288
IC  C3  C4  C5  O5    1.5177 111.17 -57.46 108.69  1.4274
IC  C4  C5  O5  C1    1.5288 108.69  62.25 113.77  1.4254
IC  C5  O5  C1  C2    1.4274 113.77 -60.97 110.06  1.5340
IC  C4  C5  C6  O6    1.5287 111.57 -170.28 110.45  1.4132
IC  O5  C1  O1  HO1   1.4254 111.55  74.87 107.83  0.9684
IC  C1  C2  O2  HO2   1.5340 110.87 -100.51 112.13  0.9638
IC  C2  C3  O3  HO3   1.5253 108.09 -165.88 112.08  0.9730
IC  C3  C4  O4  HO4   1.5177 108.31 134.18 106.97  0.9713
IC  C5  C6  O6  HO6   1.5099 110.44 -143.88 107.72  0.9641
PATC FIRS NONE LAST NONE

RESI BGAQ   0.00000   ! 4C1 Beta-D-galactopyranose monomer
GROU                 ! Quantum Ready
ATOM C1   CBS    0.000 !
ATOM H1   HAS    0.000 !
ATOM O1   OHS    0.000 !
ATOM HO1  HOS    0.000 !
ATOM C5   CTS    0.000 !
```

```
ATOM H5   HAS    0.000 !
ATOM O5   OES    0.000 !
ATOM C2   CTS    0.000
ATOM H2   HAS    0.000
ATOM O2   OHS    0.000
ATOM HO2  HOS    0.000
ATOM C3   CTS    0.000
ATOM H3   HAS    0.000
ATOM O3   OHS    0.000
ATOM HO3  HOS    0.000
ATOM C4   CTS    0.000
ATOM H4   HAS    0.000
ATOM O4   OHS    0.000
ATOM HO4  HOS    0.000
ATOM C6   CPS    0.000
ATOM H61  HAS    0.000
ATOM H62  HAS    0.000
ATOM O6   OHS    0.000
ATOM HO6  HOS    0.000
BOND C1  O1     C1  H1     O1  HO1    C1  O5     C1  C2
BOND C2  H2     C2  O2     O2  HO2    C2  C3     C3  H3
BOND C3  O3     O3  HO3    C3  C4     C4  H4     C4  O4
BOND O4  HO4    C4  C5     C5  H5     C5  C6     C6  H61
BOND C6  H62    C6  O6     O6  HO6    C5  O5
!  I   J   K   L    R(IK)  T(IKJ)   PHI   T(JKL)  R(KL)
IC  O1  C2 *C1  H1    1.3890 105.75 114.54 108.17  1.0950
IC  O1  O5 *C1  C2    1.3890 111.55 117.06 110.06  1.5340
IC  O2  C3 *C2  H2    1.4154 112.27 -118.21 108.23  1.0919
IC  O2  C1 *C2  C3    1.4154 110.87 -125.56 111.08  1.5253
IC  O3  C4 *C3  H3    1.4157 110.61 120.65 108.81  1.1068
IC  O3  C2 *C3  C4    1.4157 108.09 120.77 109.86  1.5177
IC  O4  C5 *C4  H4    1.4215 107.18 119.92 110.37  1.1004
IC  O4  C3 *C4  C5    1.4215 110.06 118.59 111.17  1.5289
IC  C6  O5 *C5  H5    1.5099 108.10 118.69 109.65  1.1042
IC  C6  C4 *C5  O5    1.5099 111.57 119.10 108.69  1.4274
IC  O6  H62 *C6  H61   1.4132 110.47 -120.32 107.85  1.0945
IC  O6  C5 *C6  H62   1.4132 110.45 -121.53 108.99  1.0959
IC  O5  C1  C2  C3    1.4254 110.06  54.09 111.08  1.5253
IC  C1  C2  C3  C4    1.5340 111.08 -51.23 109.86  1.5177
IC  C2  C3  C4  C5    1.5253 109.86  53.25 111.17  1.5288
IC  C3  C4  C5  O5    1.5177 111.17 -57.46 108.69  1.4274
IC  C4  C5  O5  C1    1.5288 108.69  62.25 113.77  1.4254
```

```
IC  C5  O5  C1  C2    1.4274 113.77  -60.97 110.06  1.5340
IC  C4  C5  C6  O6    1.5287 111.57 -170.28 110.45  1.4132
IC  O5  C1  O1  HO1   1.4254 111.55   74.87 107.83  0.9684
IC  C1  C2  O2  HO2   1.5340 110.87 -100.51 112.13  0.9638
IC  C2  C3  O3  HO3   1.5253 108.09 -165.88 112.08  0.9730
IC  C3  C4  O4  HO4   1.5177 108.31  134.18 106.97  0.9713
IC  C5  C6  O6  HO6   1.5099 110.44 -143.88 107.72  0.9641
PATC FIRS NONE LAST NONE
RESI AGLQ   0.00000   ! 4C1 alpha-D-glucopyranose monomer
GROU                  ! Quantum Ready
ATOM C1   CTS    0.000 !
ATOM H1   HAS    0.000 !
ATOM O1   OHS    0.000 !
ATOM HO1  HOS    0.000 !
ATOM C5   CTS    0.000 !
ATOM H5   HAS    0.000 !
ATOM O5   OES    0.000 !
GROU
ATOM C2   CTS    0.000
ATOM H2   HAS    0.000
ATOM O2   OHS    0.000
ATOM HO2  HOS    0.000
GROU
ATOM C3   CTS    0.000
ATOM H3   HAS    0.000
ATOM O3   OHS    0.000
ATOM HO3  HOS    0.000
GROU
ATOM C4   CTS    0.000
ATOM H4   HAS    0.000
ATOM O4   OHS    0.000
ATOM HO4  HOS    0.000
GROU
ATOM C6   CPS    0.000
ATOM H61  HAS    0.000
ATOM H62  HAS    0.000
ATOM O6   OHS    0.000
ATOM HO6  HOS    0.000
BOND C1  O1     C1  H1     O1  HO1    C1  O5     C1  C2
BOND C2  H2     C2  O2     O2  HO2    C2  C3     C3  H3
BOND C3  O3     O3  HO3    C3  C4     C4  H4     C4  O4
BOND O4  HO4    C4  C5     C5  H5     C5  C6     C6  H61
```

```
BOND C6  H62    C6  O6     O6  HO6    C5  O5
!  I  J  K  L   R(IK)  T(IKJ)  PHI  T(JKL)  R(KL)
IC  O1  C2 *C1  H1   1.3889 109.35 -122.69 108.98  1.0950
IC  O1  O5 *C1  C2   1.3889 111.55 -121.57 110.06  1.5340
IC  O2  C3 *C2  H2   1.4154 112.27 -118.21 108.23  1.0919
IC  O2  C1 *C2  C3   1.4154 110.87 -125.56 111.08  1.5253
IC  O3  C4 *C3  H3   1.4157 110.61 120.65 108.81  1.1068
IC  O3  C2 *C3  C4   1.4157 108.09 120.77 109.86  1.5177
IC  O4  C5 *C4  H4   1.4252 110.90 -120.61 108.35  1.1024
IC  O4  C3 *C4  C5   1.4252 108.31 -122.08 111.17  1.5287
IC  C6  O5 *C5  H5   1.5099 108.10 118.69 109.65  1.1042
IC  C6  C4 *C5  O5   1.5099 111.57 119.10 108.69  1.4274
IC  O6  H62*C6  H61  1.4132 110.47 -120.32 107.85  1.0945
IC  O6  C5 *C6  H62  1.4132 110.45 -121.53 108.99  1.0959
IC  O5  C1  C2  C3   1.4254 110.06  54.09 111.08  1.5253
IC  C1  C2  C3  C4   1.5340 111.08 -51.23 109.86  1.5177
IC  C2  C3  C4  C5   1.5253 109.86  53.25 111.17  1.5288
IC  C3  C4  C5  O5   1.5177 111.17 -57.46 108.69  1.4274
IC  C4  C5  O5  C1   1.5288 108.69  62.25 113.77  1.4254
IC  C5  O5  C1  C2   1.4274 113.77 -60.97 110.06  1.5340
IC  C4  C5  C6  O6   1.5287 111.57 -170.28 110.45  1.4132
IC  O5  C1  O1  HO1  1.4254 111.55  74.87 107.83  0.9684
IC  C1  C2  O2  HO2  1.5340 110.87 -100.51 112.13  0.9638
IC  C2  C3  O3  HO3  1.5253 108.09 -165.88 112.08  0.9730
IC  C3  C4  O4  HO4  1.5177 108.31 134.18 106.97  0.9713
IC  C5  C6  O6  HO6  1.5099 110.44 -143.88 107.72  0.9641
PATC FIRS NONE LAST NONE

RESI AGAQ   0.00000   ! 4C1 alpha-D-glucopyranose monomer
GROU                ! Quantum Ready
ATOM C1  CTS   0.000 !
ATOM H1  HAS   0.000 !
ATOM O1  OHS   0.000 !
ATOM HO1 HOS   0.000 !
ATOM C5  CTS   0.000 !
ATOM H5  HAS   0.000 !
ATOM O5  OES   0.000 !
GROU
ATOM C2  CTS   0.000
ATOM H2  HAS   0.000
ATOM O2  OHS   0.000
ATOM HO2 HOS   0.000
```

```
GROU
ATOM C3   CTS    0.000
ATOM H3   HAS    0.000
ATOM O3   OHS    0.000
ATOM HO3  HOS    0.000
GROU
ATOM C4   CTS    0.000
ATOM H4   HAS    0.000
ATOM O4   OHS    0.000
ATOM HO4  HOS    0.000
GROU
ATOM C6   CPS    0.000
ATOM H61  HAS    0.000
ATOM H62  HAS    0.000
ATOM O6   OHS    0.000
ATOM HO6  HOS    0.000
BOND C1  O1     C1  H1     O1  HO1    C1  O5     C1  C2
BOND C2  H2     C2  O2     O2  HO2    C2  C3     C3  H3
BOND C3  O3     O3  HO3    C3  C4     C4  H4     C4  O4
BOND O4  HO4    C4  C5     C5  H5     C5  C6     C6  H61
BOND C6  H62    C6  O6     O6  HO6    C5  O5
!  I   J   K   L    R(IK)  T(IKJ)   PHI   T(JKL)  R(KL)
IC  O1  C2 *C1  H1   1.3889 109.35 -122.69 108.98  1.0950
IC  O1  O5 *C1  C2   1.3889 111.55 -121.57 110.06  1.5340
IC  O2  C3 *C2  H2   1.4154 112.27 -118.21 108.23  1.0919
IC  O2  C1 *C2  C3   1.4154 110.87 -125.56 111.08  1.5253
IC  O3  C4 *C3  H3   1.4157 110.61  120.65 108.81  1.1068
IC  O3  C2 *C3  C4   1.4157 108.09  120.77 109.86  1.5177
IC  O4  C5 *C4  H4   1.4215 107.18  119.92 110.37  1.1004
IC  O4  C3 *C4  C5   1.4215 110.06  118.59 111.17  1.5289
IC  C6  O5 *C5  H5   1.5099 108.10  118.69 109.65  1.1042
IC  C6  C4 *C5  O5   1.5099 111.57  119.10 108.69  1.4274
IC  O6  H62*C6  H61  1.4132 110.47 -120.32 107.85  1.0945
IC  O6  C5 *C6  H62  1.4132 110.45 -121.53 108.99  1.0959
IC  O5  C1  C2  C3   1.4254 110.06   54.09 111.08  1.5253
IC  C1  C2  C3  C4   1.5340 111.08  -51.23 109.86  1.5177
IC  C2  C3  C4  C5   1.5253 109.86   53.25 111.17  1.5288
IC  C3  C4  C5  O5   1.5177 111.17  -57.46 108.69  1.4274
IC  C4  C5  O5  C1   1.5288 108.69   62.25 113.77  1.4254
IC  C5  O5  C1  C2   1.4274 113.77  -60.97 110.06  1.5340
IC  C4  C5  C6  O6   1.5287 111.57 -170.28 110.45  1.4132
IC  O5  C1  O1  HO1  1.4254 111.55   74.87 107.83  0.9684
```

IC C1 C2 O2 HO2 1.5340 110.87 -100.51 112.13 0.9638

IC C2 C3 O3 HO3 1.5253 108.09 -165.88 112.08 0.9730

IC C3 C4 O4 HO4 1.5177 108.31 134.18 106.97 0.9713

IC C5 C6 O6 HO6 1.5099 110.44 -143.88 107.72 0.9641

PATC FIRS NONE LAST NONE

B – Alpha (α)-monosaccharides

B.1 – α-glucose PMFs

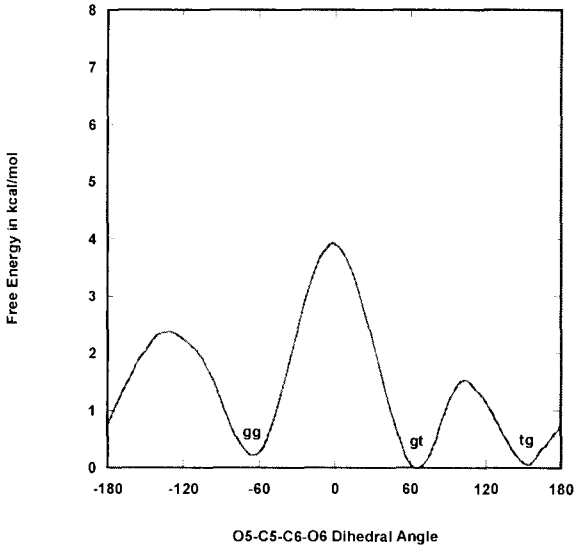

Figure B.1.1 – Vacuum PMF (for the Helmholtz Free Energy) for α-glucose in the gaseous phase using PM3CARB-1 constrained dynamics.

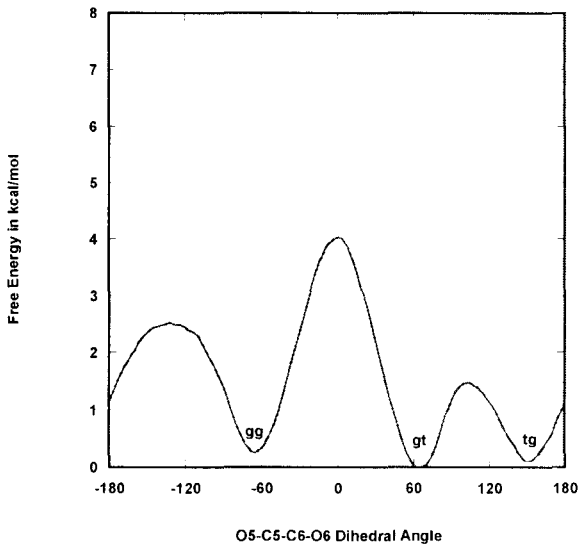

Figure B.1.2 – Vacuum PMF (for the Helmholtz Free Energy) for α-glucose in the gaseous phase using PM3CARB-1 unconstrained dynamics.

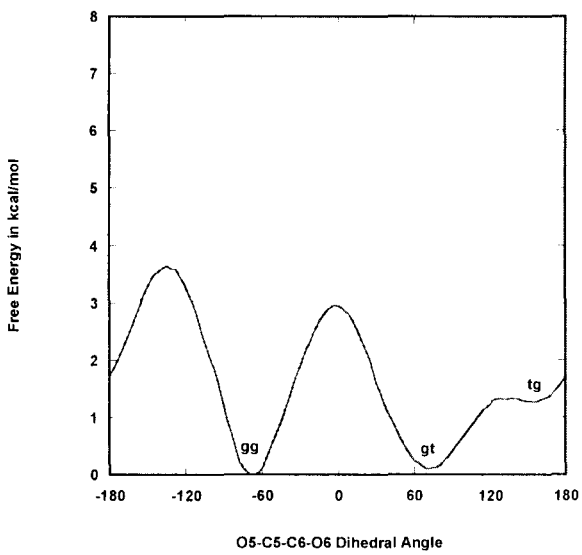

Figure B.1.3 – Aqueous solution PMF (for the Helmholtz Free Energy) for α-glucose using PM3CARB-1 QM/MM constrained dynamics.

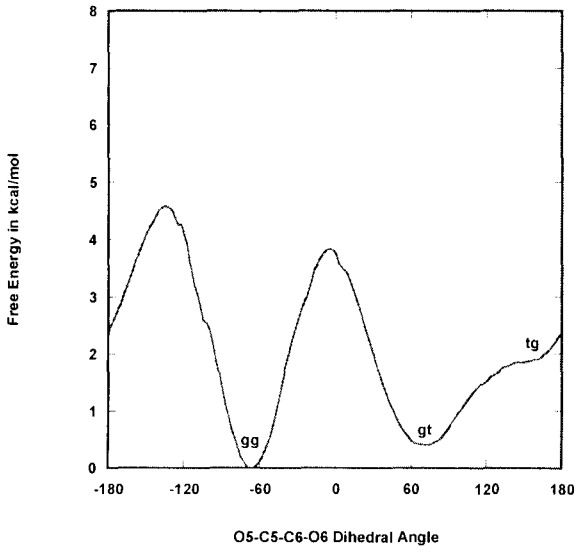

Figure B.1.4 – Aqueous solution PMF (for the Helmholtz Free Energy) for α-glucose using PM3CARB-1 QM/MM unconstrained dynamics.

B.2 – α–galactose PMFs

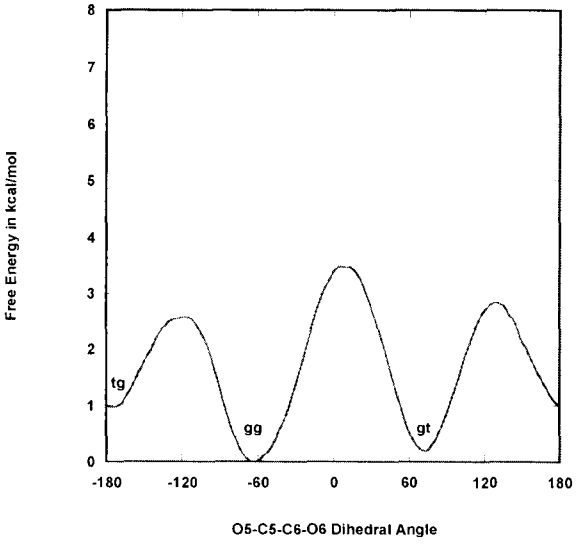

Figure B.2.1 – Vacuum PMF (for the Helmholtz Free Energy) for α-galactose in the gaseous phase using PM3CARB-1 constrained dynamics.

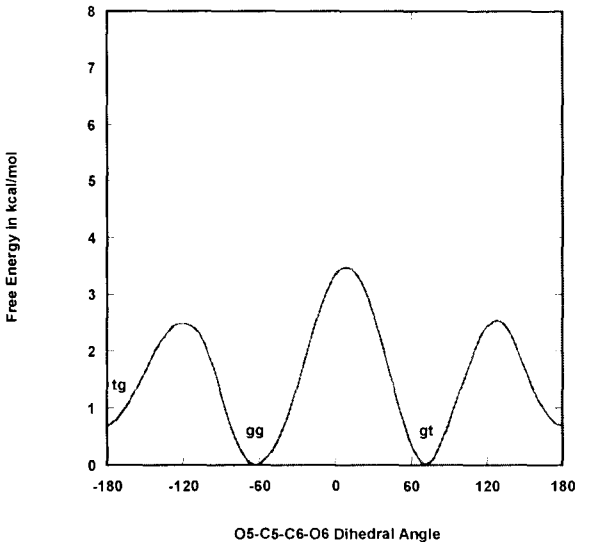

Figure B.2.2 – Vacuum PMF (for the Helmholtz Free Energy) for α-galactose in the gaseous phase using PM3CARB-1 unconstrained dynamics.

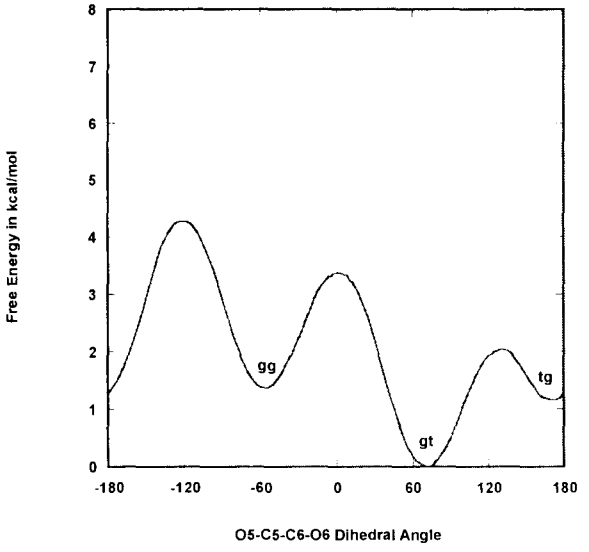

Figure B.2.3 – Aqueous solution PMF (for the Helmholtz Free Energy) for α-galactose using PM3CARB-1 QM/MM constrained dynamics.

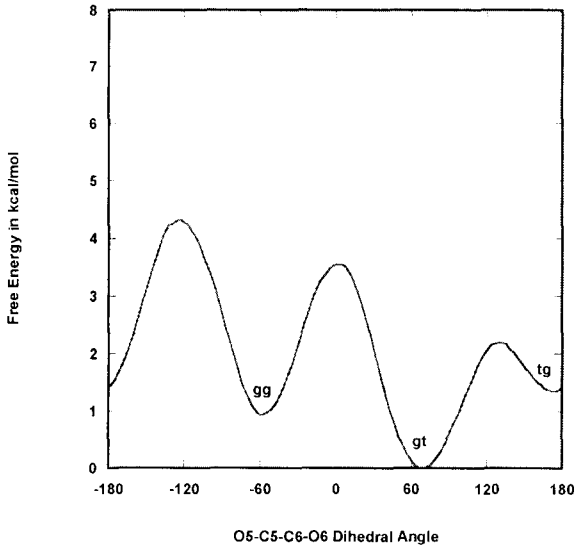

Figure B.2.4 – Aqueous solution PMF (for the Helmholtz Free Energy) for α-galactose using PM3CARB-1 QM/MM unconstrained dynamics.

C – Unconstrained dynamics PMFs for glucose and galactose

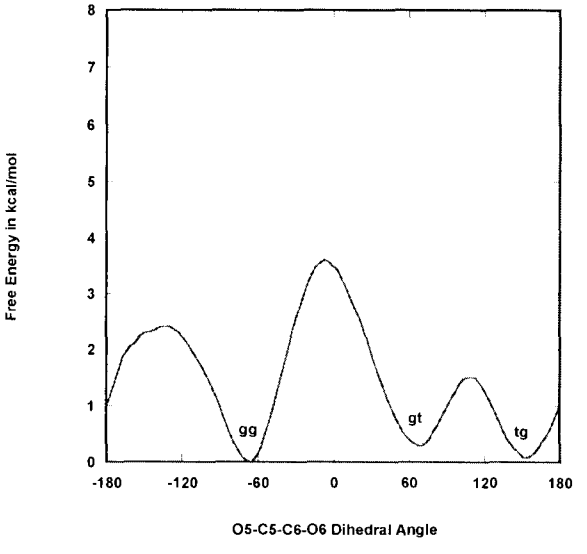

Figure C.1.1 – Vacuum PMF (for the Helmholtz Free Energy) for glucose in the gaseous phase using PM3CARB-1 unconstrained dynamics.

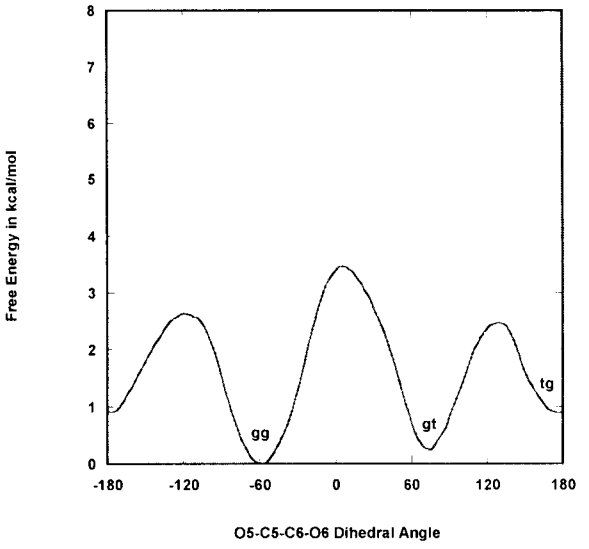

Figure C.1.2 – Vacuum PMF (for the Helmholtz Free Energy) for galactose in the gaseous phase using PM3CARB-1 unconstrained dynamics.

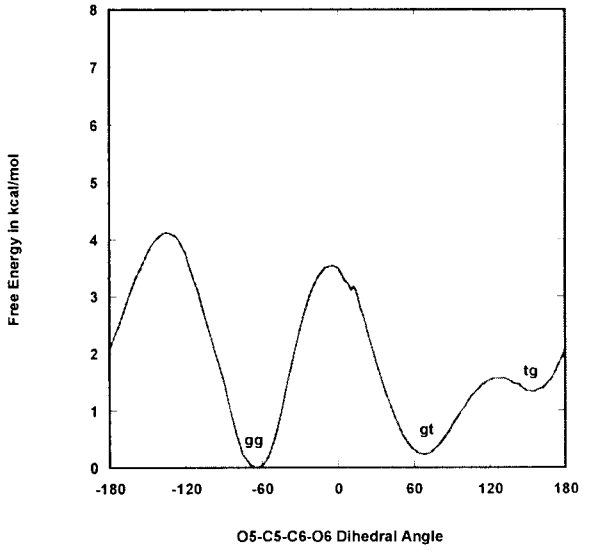

Figure C.1.3 – Aqueous solution PMF (for the Helmholtz Free Energy) for glucose using PM3CARB-1 QM/MM unconstrained dynamics.

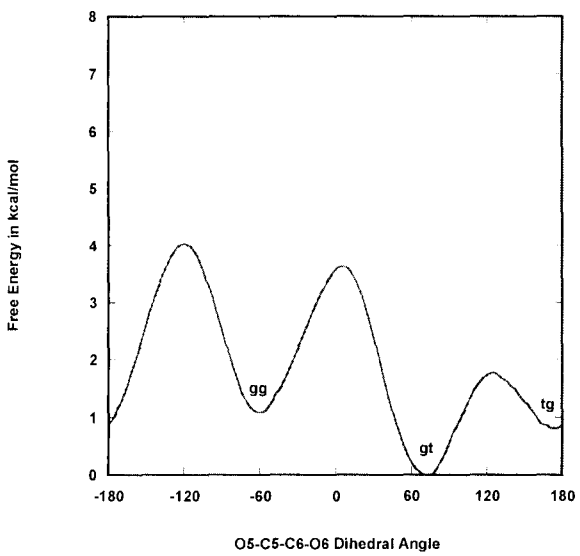

Figure C.1.4 – Aqueous solution PMF (for the Helmholtz Free Energy) for galactose using PM3CARB-1 QM/MM unconstrained dynamics.

D – Ring constraints as applied to constrained vacuum and water simulations

! CONSTRAIN CODE FOR GLUCOSE THE SAME APPLIES FOR GALACTOSE

set c 54.09

set d -51.23

set e 53.25

set f -57.46

set g 62.25

set h -60.97

! Constraining Endocyclic torsion angles

cons dihe gluc 1 O5 gluc 1 C1 gluc 1 C2 gluc 1 C3 force 5 min @c

cons dihe gluc 1 C1 gluc 1 C2 gluc 1 C3 gluc 1 C4 force 5 min @d

cons dihe gluc 1 C2 gluc 1 C3 gluc 1 C4 gluc 1 C5 force 5 min @e

cons dihe gluc 1 C3 gluc 1 C4 gluc 1 C5 gluc 1 O5 force 5 min @f

cons dihe gluc 1 C4 gluc 1 C5 gluc 1 O5 gluc 1 C1 force 5 min @g

cons dihe gluc 1 C5 gluc 1 O5 gluc 1 C1 gluc 1 C2 force 5 min @h

REFERENCES:

1. Stryer, L. *Biochemistry* (W.H. Freeman and Company, New York, 1988).
2. Berg, J. M., Tymoczko, J. L. & Stryer, L. *Biochemistry* (W.H. Freeman and Co., New York, 2002).
3. Clayden, J., Greeves, N., Warren, S. & Wothers, P. *Organic Chemistry* (Oxford University Press, Oxford, 2001).
4. Lindhorst, T. K. *Essentials of Carbohydrate Chemistry and Biochemistry* (Wiley, 2003).
5. McNamara, J. P. et al. Towards a quantum mechanical force field for carbohydrates: a reparametrized semi-empirical MO approach. *Chemical Physics Letters* **394**, 429-436 (2004).
6. Kirby, A. J. *The anomeric effect and related stereoelectronic effects at oxygen* (Springer-Verlag, Heidelberg, 1983).
7. *CRC Handbook of Chemistry and Physics* (ed. Weast, R. C.) (CRC Press, 1974).
8. Kirby, A. J. *Stereoelectronic Effects* (Oxford University Press, 1996).
9. Evans, D. A. (Harvard - CEM206 notes, 2002).
10. Wolfe, S. Gauche effect. Stereochemical consequences of adjacent electron pairs and polar bonds. *Accounts of Chemical Research* **5**, 102-111 (1972).
11. Thatcher, G. R. J. *The Anomeric effect and associated stereoelectronic effects. (Developed from a Symposium Sponsored by the Division of Carbohydrate Chemistry at the 204th National Meeting of the American Chemical Society, Washington, DC, August 23-28, 1992.) [In: ACS Symp. Ser., 1993; 539]* (1993).
12. Craig, N. C. et al. Contribution to the Study of the Gauche Effect. The Complete Structure of the Anti Rotamer of 1,2-Difluoroethane. *Journal of the American Chemical Society* **119**, 4789-4790 (1997).
13. Guvench, O. & MacKerell, A. D., Jr. Quantum Mechanical Analysis of 1,2-Ethanediol Conformational Energetics and Hydrogen Bonding. *Journal of Physical Chemistry A* **110**, 9934-9939 (2006).
14. Nagy, P. I., Dunn, W. J. I., Alagona, G. & Ghio, C. Theoretical calculations on 1,2-Ethanediol. Gauche-Trans Equilibrium in Gas-phase and Aqueous solution. *Journal of the American Chemical Society* **113**, 6719-6729 (1991).
15. *Encyclopaedia of Computational Chemistry* (eds. Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J. & Kollman, P. A.) (Wiley, Chichester, 1998).
16. Bock, K., Fernandez Bolanos Guzman, J. & Ogawa, S. A proton and carbon-13 NMR spectroscopic analysis of six pseudohexoses. *Carbohydrate Research* **174**, 354-9 (1988).
17. Best, R. B., Jackson, G. E. & Naidoo, K. J. An NMR investigation into the dynamics of panose, an a(1->4) and a(1->6)-linked trisaccharide. *Spectroscopy Letters* **35**, 625-632 (2002).

18. Best, R. B., Jackson, G. E. & Naidoo, K. J. Molecular Dynamics and NMR Study of the a(1->4) and a(1->6) Glycosidic Linkages: Maltose and Isomaltose. *Journal of Physical Chemistry B* **105**, 4742-4751 (2001).
19. Duus, J. O., Gotfredsen, C. H. & Bock, K. Carbohydrate Structural Determination by NMR Spectroscopy: Modern Methods and Limitations. *Chemical Reviews (Washington, D. C.)* **100**, 4589-4614 (2000).
20. De Vries, N. K. & Buck, H. M. Different rotamer populations around the C-5-C-6 bond for a- and b-D-galactopyranosides through the combined interaction of the gauche and anomeric effects: a 300-MHz proton NMR and MNDO study. *Carbohydrate Research* **165**, 1-16 (1987).
21. Rockwell, G. D. & Grindley, T. B. Effect of Solvation on the Rotation of Hydroxymethyl Groups in Carbohydrates. *Journal of the American Chemical Society* **120**, 10953-10963 (1998).
22. Behrends, R. et al. Ultrasonic Relaxation and Fast Chemical Kinetics of Some Carbohydrate Aqueous Solutions. *Journal of the American Chemical Society* **119**, 2182-2186 (1997).
23. Roeen, A., Padron, J. I. & Vazquez, J. T. Hydroxymethyl Rotamer Populations in Disaccharides. *Journal of Organic Chemistry* **68**, 4615-4630 (2003).
24. Vazquez, J. T., Weisler, W. T. & Nakanishi, K. Circular Dichroism Spectra of Bichromophorically derivatized methyl-D-galactopyranosides, calculable by pairwise additivity, provide a basis for novel microanalysis of oligosaccharides. *Carbohydrate Research* **176**, 175-194 (1988).
25. Vazquez, J. T., Morales, E. Q. & Padron, J. I. Alkyl Galactopyranosides: Rotational Population Dependence of the Hydroxymethyl Group and the Aglycon and Its Absolute Configuration and in the Anomeric Configuration. *Journal of Organic Chemistry* **63**, 8247-8258 (1998).
26. Padron, J. I. & Vazquez, J. T. Stereochemical study of the CD spectral differences between anomers of alkyl glucopyranosides. *Tetrahedron: Asymmetry* **9**, 613-627 (1998).
27. Mason, P. E. et al. Neutron diffraction and simulation studies of the exocyclic hydroxymethyl conformation of glucose. *Journal of Chemical Physics* **125**, 224505/1-224505/9 (2006).
28. Brady, J. W., Mason, P. E., Matthews, J. F., Naidoo, K. J. & Ueda, K. Molecular Dynamics and Neutron Diffraction Studies of the Structuring of Water by Carbohydrates and Other Solutes. *Abstracts, 36th Great Lakes Regional Meeting of the American Chemical Society, Peoria, IL, United States, October 17-20*, LAKES04-412 (2004).
29. Talon, C. et al. Dynamics of Water Molecules in Glucose Solutions. *Journal of Physical Chemistry* **108**, 5120-5126 (2004).
30. Best, R. B., Jackson, G. E. & Naidoo, K. J. Modeling the a(1->6) Branch Point of Amylopectin in Solution. *Journal of Physical Chemistry B* **106**, 5091-5098 (2002).
31. Naidoo, K. J. & Chen, J. Y.-J. The role of water in the design of glycosidic linkage flexibility. *Molecular Physics* **101**, 2687-2694 (2003).
32. Corzana, F. et al. A hydration study of 1,4 and 1,6 linked alpha-glucans by comparative 10 ns molecular dynamics simulations and 500MHz NMR. *Journal of Computational Chemistry* **25**, 573-586 (2004).

33. Liu, Q. & Brady, J. W. Model dependence of the anistropic structuring of solvent water around sugars in molecular dynamics simulations. *Journal of Physical Chemistry B* **101**, 1317-1321 (1997).
34. Tvaroška, I., R., T. F., P., U. J. & P., C. J. Quantum mechanical and NMR spectroscopy studies on the conformations of the hydroxymethyl and methoxymethyl groups in aldohexosides. *Carbohydrate Research* **337**, 353-67 (2002).
35. Tvaroška, I. & Carver, J. P. Ab initio molecular orbital calculation of carbohydrate model compounds. 2. Conformational analysis of axial and equatorial 2-methoxytetrahydropyrans. *Journal of Physical Chemistry* **98**, 9477-85 (1994).
36. Polavarapu, P. L. & Ewig, C. S. Ab initio computed molecular structures and energies of the conformers of glucose. *Journal of Computational Chemistry* **13**, 1255-61 (1992).
37. Chen, J. Y.-J. & Naidoo, K. J. Evaluating Intramolecular Hydrogen Bond Strengths in (1-4) Linked Disaccharides from Electron Density Relationships. *Journal of Physical Chemistry B* **107**, 9558-9566 (2003).
38. Naidoo, K. J. & Chen, Y.-J. Ab initio study of inter-residue hydrogen bonding in a(1-4)-linked polysaccharides. *Abstracts of Papers, 222nd ACS National Meeting, Chicago, IL, United States, August 26-30, 2001*, CARB-119 (2001).
39. Schleyer, P. v. R. & Salzner, U. Ab initio Examinatios of Anomeric Effects in Tetrahydropyrans, 1,3-Dioxanes and Glucose. *Journal of Organic Chemistry* **59**, 2138-2155 (1994).
40. Kuttel, M. M. & Naidoo, K. J. Free Energy Surfaces for the a(1->4)-Glycosidic Linkage: Implications for Polysaccharide Solution Structure and Dynamics. *Journal of Physical Chemistry B* **109**, 7468-7474 (2005).
41. Muslim, A.-M., McNamara, J. P., Abdel-aal, H., Hillier, I.-. & Bryce, R. *Hybrid QM/MM simulation of carbohydrates in aqueous solution* (eds. Vliegenhart, H. & Woods, R. J.) (Washington, American Chemical Society, Washington, DC, 2006).
42. Muslim, A.-M. & Bryce, R. A. Carbohydrate conformation in aqueous solution: calculation of a QM/MM potential of mean force. *Chemical Physics Letters* **388**, 473-478 (2004).
43. Nishida, Y., Ohrui, H. & Meguro, H. Proton NMR studies of (6R)- and (6S)-deuterated D-hexoses: assignment of the preferred rotamers about C5-C6 bond of D-glucose and D-galactose derivatives in solutions. *Tetrahedron Letters* **25**, 1575-8 (1984).
44. Nishida, Y., Hori, H., Ohrui, H. & Meguro, H. Proton NMR analysis for rotameric distribution of the C(5)-C(6) bonds of D-glucopyranoses in solution. *Journal of Carbohydrate Chemistry* 7, 239-50 (1988).
45. Cramer, C. J. & Truhlar, D. G. Quantum chemical conformational analysis of glucose in aqueous solution. *Journal of the American Chemical Society* **115**, 5745-53 (1993).
46. Tvaroška, I. & Carver, J. P. Ab Initio Molecular Orbital Calculation of Carbohydrate Model Compounds. 6. The Gauche Effect and Conformations of the Hydroxymethyl and Methoxymethyl Groups. *Journal of Physical Chemistry B* **101**, 2992-2999 (1997).

47. Lemieux, R. U. & Brewer, J. T. Conformational preferences for solvated hydroxymethyl groups in hexopyranose structures. *Advances in Chemistry Series* **No. 117**, 121-46 (1973).
48. Brooks, B. R. et al. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry* **4**, 187-217 (1983).
49. GAMESS-UK is a package of ab initio programs. See: "http://www.cfs.dl.ac.uk/gamess-uk/index.shtm", M.F. Guest, I. J. Bush, H.J.J. van Dam, P. Sherwood, J.M.H. Thomas, J.H. van Lenthe, R.W.A Havenith, J. Kendrick, "The GAMESS-UK electronic structure package: algorithms, developments and applications", Molecular Physics, Vol. 103, No. 6-8, 20 March-20 April 2005, 719-747. The initial DFT module within GAMESS-UK was developed by Dr. P. Young under the auspices of EPSRC's Collaborative Computational Project No. 1 (CCP1) (1995-1997). Subsequent developments have been undertaken by staff at the Daresbury Laboratory.
50. Bader, R. F. W. in *Handbook of Molecular Physics and Quantum Chemisty* 770-791 (John Wiley and sons, Chichester, 2003).
51. Bader, R. F. W. Atoms in Molecules. *Accounts of Chemical Research* **18**, 9-15 (1985).
52. Reed, A. E., Weinstock, R. B. & Weinhold, F. Natural population analysis. *Journal of Chemical Physics* **83** (1985).
53. Weinhold, F. & Landis, C. R. Natural bond orbitals and extensions of localized bonding concepts. *Chemistry Education: Research and Practice in Europe* **2**, 91-104 (2001).
54. Leach, A. R. *Molecular Modelling: Principles and Applications* (Longman, 2001).
55. Atkins, P. & de Paula, J. *Atkins' Physical Chemistry* (Oxford University Press, 2002).
56. van Duin, A. C., Dasgupta, S., Lorant, F. & Goddard, W. A., III. ReaxFF: A Reactive Force Field for Hydrocarbons. *Journal of Physical Chemistry A* **105**, 9396-9409 (2001).
57. Frisch, A. & Foresman, J. B. *Exploring Chemistry with Electronic Structure Methods* (Gaussian Inc., Pittsburgh, 1996).
58. Lewars, E. *Computational Chemistry: Introduction to the Theory and Applications of molecular and Quantum Mechanics* (Kluwer Academic Publishers, 2004).
59. Weiner, P. K. & Kollman Peter, A. Amber: Assisted model building program with energy refinement. A general program for modeling molecules and their interactions. *Journal of Computational Chemistry* **2**, 287-303 (1981).
60. Van Gunsteren, W. F. & Berendsen, H. J. C. (University of Groningen, 1988).
61. Glennon, T. M., Zheng, Y., le Grand, S. M., Shutzberh, B. A. & Merz, K. M., Jr. A force field for monosacharides and (1-4) linked polysaccharides. *Journal of Computational Chemistry* **15**, 1019-1040 (1994).
62. Momany, F. A. & Willet, J. L. Computational studies on carbohydrates: in vacuo studies using a revised AMBER force field AMB99C designed for alpha-1-4 linkages. *Carbohydrate Research* **326**, 194-209 (2000).

63. Senderowitz, H., Still, W. C. & Parish, C. Carbohydrates: United atom AMBER parameterisation of pyranoses and simulations yielding anomeric free energies. *Journal of the American Chemical Society* **118**, 2078-2086 (1996).
64. Senderowitz, H. & Still, W. C. A quantum mechanically derived all-atom force field for pyranose oligosaccharides, amber parameters and free energy simulations. *Journal of Organic Chemistry* **62**, 1427-1438 (1997).
65. Brady, J. W., Field, M. J., Giammona, M. & Ha, S. N. A revised potential surface for molecular studies of carbohydrates. *Carbohydrate Research* **180**, 207-221 (1988).
66. Brady, J. W., Palma, R., Himmel, M. E. & Liang, G. *Molecular mechanics studies of cellulases* (ed. Himmel, M. E.) (ACS, Washington, 2001).
67. Reiling, S., Schlenkrich, M. & Brickman, J. Force field parameters for carbohydrates. *Journal of Computational Chemistry* **17**, 450-468 (1996).
68. Meyer, B. & Ott, K. Parameterisation of GROMOS force field for oligosaccharides and assessment of efficiency of molecular dynamics simulations. *Journal of Computational Chemistry* **17**, 1068-1084 (1996).
69. Damm, W., Frontera, A., Tirado-Rives, J. & Jorgensen, W. L. OPLS all-atom force field for carbohydrates. *Journal of Computational Chemistry* **18**, 1955-1970 (1997).
70. Kony, D., Damm, W., Stoll, S. & Van Gunsteren, W. F. An improved OPLS-AA force field for carbohydrates. *Journal of Computational Chemistry* **23**, 1416-1429 (2002).
71. Kuttel, M., Brady, J. W. & Naidoo, K. J. Carbohydrate solution simulations: Producing a force field with experimentally consistent primary alcohol rotational frequencies and populations. *Journal of Computational Chemistry* **23**, 1236-1243 (2002).
72. Jorgensen, W. L. & Jenson, C. Temperature dependence of TIP3P, SPC, and TIP4P water from NPT Monte Carlo simulations: seeking temperatures of maximum density. *Journal of Computational Chemistry* **19**, 1179-1186 (1998).
73. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *Journal of Chemical Physics* **79**, 926-35 (1983).
74. Berendsen, H. J. C., Grigera, J. R. & Straatsma, T. P. J. The missing term in effective pair potentials. *Journal of Physical Chemistry* **91**, 6269-6271 (1987).
75. Frisch, M. J. T., G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, Jr., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.;

Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; and Pople, J. A. (Gaussian, Inc., Wallingford CT, 2004).

76. Ostlund, N. S. & Szabo, A. *Modern Quantum Chemistry: Introduction to Advanced Electonic Structure Theory* (Dover Publications, New York, 1996).
77. Jensen, F. *Introduction to Computational Chemistry* (J. Wiley and sons, England, 1999).
78. Hehre, W. J., Radom, L., Schleyer, P. v. R. & Pople, J. A. *Ab initio molecular orbital theory* (John Wiley and Sons, New York, U.S.A, 1986).
79. Atkins, P. *Molecular quantum mechanics : an introduction to quantum chemistry* (Clarendon Press, Oxford, 1970).
80. Hinchcliffe, A. *Chemical Modeling: from atoms to liquids* (Wiley, 1999).
81. Cook, D. B. *Handbook of Computational Quantum Chemistry* (Dover Publications, New York, 2005).
82. Kirschner, K. N. & Woods, R. J. Solvent interactions determine carbohydrate conformation. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 10541-5 (2001).
83. Wladkowski, B. D., Chenoweth, S. A., Jones, K. E. & Brown, J. W. Exocyclic Hydroxymethyl Rotational Conformers of b- and a-D-Glucopyranose in the Gas Phase and Aqueous Solution. *Journal of Physical Chemistry A* **102**, 5086-5092 (1998).
84. Lowe, J. P. & Peterson, K. A. *Quantum Chemistry* (Elsevier Academic Press, San Diego, 2006).
85. Kohn, W. & Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Physical Reviews* **140**, A1133-A1138 (1965).
86. Vosko, S. H., Wilk, L. & Nusair, M. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Canadian Journal of Physics* **58**, 1200-1211 (1980).
87. Becke, A. D. Density-functional thermochemistry III. The role of exact exchange. *Journal of Chemical Physics* **98**, 5648-5652 (1993).
88. Lee, C., Yang, W. & Parr, R. G. Development of the Colle-Salvetti correlation energy formula into a functional of the density. *Physical Reviews B37*, 785-789 (1988).
89. Bader, R. F. W. *Atoms in Molecules: A Quantum Theory* (Clarendon Press, Oxford, 1990).
90. Parr, R. G., Ayers, P. W. & Nalewajski, R. F. What is an Atom in a Molecule? *Journal of Physical Chemistry A* **109**, 3957-3959 (2005).
91. Biegler-König, F., Schönbohm, J. & Bayles, D. A Program to Analyze and Visualize Atoms in Molecules. *Journal of Computational Chemistry* **22**, 545-559 (2001).
92. Biegler-König, F. & Schönbohm, J. An Update to the AIM2000 Program for Atoms in Molecules. *Journal of Computational Chemistry* **23**, 1489-1494 (2002).
93. Mosquera, R. A. & Vila, A. Atoms in Molecules Interpretation of the Anomeric Effect in the O-C-O unit. *Journal of Computational Chemistry* **28**, 1517-1530 (2007).

94. Silberberg, M. S. *Chemistry: the molecular nature of matter and change* (McGraw-Hill, Boston, 2003).
95. Foster, J. P. & Weinhold, F. Natural Hybrid Orbitals. *Journal of the American Chemical Society* **102**, 7211-7218 (1980).
96. Reed, A. E. & Weinhold, F. Natural localized molecular orbitals. *Journal of Chemical Physics* **83** (1985).
97. Glendening, E. D., Reed, A. E., Carpenter, J. E. & Weinhold, F. in *QCPE Bulletin* 58 (1990).
98. Dewar, M. J. S., Zoebisch, E. G., Healy, E. F. & Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *Journal of the American Chemical Society* **107**, 3902-3909 (1985).
99. Byun, K., Mo, Y. & Gao, J. New Insight on the Origin of the Unusual Acidity of Meldrum's Acid from ab Initio and Combined QM/MM Simulation Study. *Journal of the American Chemical Society* **123**, 3974-3979 (2001).
100. Brewster, M. E. et al. An AM1 molecular orbital study of alpha-D-glucopyranose and Beta-maltose : Evaluation and Implications. *Carbohydrate Research* **242**, 53-67 (1993).
101. Field, M. J., Bash, P. A. & Karplus, M. A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations. *Journal of Computational Chemistry* **11**, 700-33 (1990).
102. Laidler, K. J., Meiser, J. H. & Sanctuary, B. C. *Physical Chemistry* (Houghton-Mifflin, 2002).
103. Chandler, D. *Introduction to Modern Statistical Mechanics* (Oxford University Press, Oxford, 1987).
104. Allen, M. P. & Tildesley, D. J. *Computer Simulations of Liquids* (Clarendon Press, Oxford, 1989).
105. Allen, M. P. *Introduction t to Molecular Dynamics Simulation* (eds. Attig, N., Binder, K., Grubmuller, H. & Kremer, K.) (John von Neumann Institute for Computing, Julich, 2004).
106. Brooks, B. R. & Steinbach, P. J. New spherical-cutoff methods for long-range forces in macromolecular simulation. *Journal of Computational Chemistry* **15**, 667-683 (1994).
107. McQuarrie, D. A. *Statistical Mechanics* (University Science Books, California, 2000).
108. Laidler, K. J., Meiser, J. H. & Sanctuary, B. C. *Physical Chemistry* (Houghton-Mifflin, 1999).
109. Pearlman, W. & Rao, R. *Encyclopaedia of Computational Chemistry* (eds. Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J. & Kollman, P. A.) (Wiley, Chichester, 1998).
110. Zwanzig, R. W. *Journal of Chemical Physics* **22**, 1420 (1954).
111. Jorgensen, W. L. Monte Carlo simulations of differences in free energies of hydration. *Journal of Chemical Physics* **6**, 3050-3054 (1985).
112. Kollman Peter, A. Free Energy Calculations: Applications to Chemical and Biochemical phenomena. *Chemical Reviews (Washington, D. C.)* **93**, 2395-3417 (1993).

113. Kumar, S., Payne, P. W. & Vasquez, M. Method for free-energy calculations using iterative techniques. *Journal of Computational Chemistry* **17**, 1269-1275 (1996).
114. Valleau, J. P. & Torrie, G. M. *In Modern theoretical chemistry, Statistical mechanics A, Equilibrium techniques* (ed. Berne, B. J.) (Plenum Press, New York, U. S. A, 1977).
115. Bartels, C. & Karplus, M. Multidimensional adaptive umbrella sampling: applications to main chain and side chain peptide conformations. *Journal of Computational Chemistry* **18**, 1450-1462 (1997).
116. Bartels, C. & Karplus, M. Probability Distributions for Complex Systems: Adaptive Umbrella Sampling of the Potential Energy. *Journal of Physical Chemistry B* **102**, 865-880 (1998).
117. Hooft, R. W. W., Van Eijck, B. P. & Kroon, J. An adaptive umbrella-sampling procedure in conformational analysis using molecular dynamics and its application to glycol. *Journal of Chemical Physics* **97**, 6690-4 (1992).
118. Kuttel, M. M. & Naidoo, K. J. Simulations of polysaccharide dynamics: The role of conformational transitions in determining physical properties. *Abstracts of Papers, 225th ACS National Meeting, New Orleans, LA, United States, March 23-27, 2003*, CARB-071 (2003).
119. Kuttel, M. M. & Naidoo, K. J. Ramachandran free-energy surfaces for disaccharides: trehalose, a case study. *Carbohydrate Research* **340**, 875-879 (2005).
120. Naidoo, K. J. & Brady, J. W. Calculation of the Ramachandran Potential of Mean Force for a Disaccharide in Aqueous Solution. *Journal of the American Chemical Society* **121**, 2244-2252 (1999).
121. Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H. & Kollman, P. A. Multidimensional free-energy calculations using the weighted histogram analysis method. *Journal of Computational Chemistry* **16**, 1339-50 (1995).
122. Kumar, S., Bouzida, D., Swendsen, R. H., Kollman, P. A. & Rosenberg, J. M. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry* **13**, 1011-21 (1992).
123. Boczko, E. M. & Brooks, C. L., III. Constant-temperature free energy surfaces for physical and chemical processes. *Journal of Physical Chemistry* **97**, 4509-4513 (1993).
124. Cremer, D. & Pople, J. A. General definition of ring puckering coordinates. *Journal of the American Chemical Society* **97**, 1354-1358 (1975).
125. Brown, J. W. & Wladkowski, B. D. Ab Initio Studies of the Exocyclic Hydroxymethyl Rotational Surface in a-D-Glucopyranose. *Journal of the American Chemical Society* **118**, 1190-3 (1996).
126. Van Gunsteren, W. F. & Berendsen, H. J. C. Algorithms for macromolecular dynamics and constraint dynamics. *Molecular Physics* **34**, 1311-27 (1977).
127. Tvaroška, I., Taravel, F. R., Utille, J. P. & Carver, J. P. Quantum mechanical and NMR spectroscopy studies on the conformations of the hydroxymethyl and methoxymethyl groups in aldohexosides. *Carbohydrate Research* **337**, 353-367 (2002).
128. Popelier, P. L. A. & Koch, U. Characterization of C-H-O Hydrogen Bonds on the Basis of the Charge Density. *Journal of PHysical Chemistry* **99**, 9747-9754 (1995).

129. Popelier, P. L. A. Characterization of a Dihydrogen Bond on the Basis of the Electron Density. *Journal of Physical Chemistry A* **102**, 1873-1878 (1998).
130. Gonzalez-Outeirino, J., Kirschner, K. N., Thobhani, S. & Woods, R. J. Reconciling solvent effects on rotamer populations in carbohydrates - A joint MD and NMR analysis. *Canadian Journal of Chemistry* **84**, 569-579 (2006).