Developing Methods to Construct Ring Pucker Free Energy Hypersurfaces Applied to the Analysis of Glycosidase Enzyme Catalytic Mechanisms

Christopher Bevan Barnett

Supervisor: Associate Professor Kevin J. Naidoo

A thesis submitted in accordance with the requirements for the degree of

Doctor of Philosophy

in the

Scientific Computing Research Unit

the Department of Chemistry

University of Cape Town

October 2010

ABSTRACT

DEVELOPING METHODS TO CONSTRUCT RING PUCKER FREE ENERGY HYPERSURFACES APPLIED TO THE ANALYSIS OF GLYCOSIDASE ENZYME CATALYTIC MECHANISMS

CHRISTOPHER BEVAN BARNETT | OCTOBER 2010

Carbohydrates consist of one or more sub-units usually various 5- and 6-membered cycles (furanoses and pyranoses) which can twist, bend or flip into a variety of conformers that differ in strain - this is ring puckering. These puckers notably the strained puckering conformers are observed during enzymatically assisted bond formation or cleavage of the glycosidic bonds of carbohydrate substrates.

In this thesis, the free energy of ring puckering is calculated by implementing the Hill-Reilly reduced coordinate pucker description into the sampling enhancing Free Energies from Adaptive Reaction Coordinate Forces (FEARCF) method. FEARCF non-Boltzmann simulations of prototypical sugars β-D-ribose and β-D-glucose converged to yield free energy pucker surfaces and volumes when using several semi-empirical QM methods - AM1, PM3, PM3CARB-1 and SCC-DFTB. From this, the accessible puckering conformations and minimum free energy paths of puckering were reasoned

An analysis of the furanose and pyranose free energy pucker surfaces and volumes compared with both Density Functional Theory RB3LYP/6-311++G** optimised structures and a Hartree-Fock free energy surface revealed that SCC-DFTB provides the best semi-empirical description of 5- and 6-membered carbohydrate ring deformation. This illustrates that necessary high energy ring conformations observed in enzymatic binding sites requires the enzyme to induce and preserve high energy conformations required for successful hydrolyses and synthesis of the glycosidic bond.

To further test this hypothesis, a 5- and 6-membered cycle were studied within enzymatic environments. The polysaccharide cellulose contains β 1-4 linked glucose subunit and is degraded by cellulase, a glycosidase. Specifically, the retaining cellobiohydrolase I (CBHI) of *Trichoderma Reesei* which cleaves cellobiose units from crystalline cellulose. The free energy volumes of puckering for the glucose sub-unit (in the catalytic position of an 8 unit cellulosic fragment - cellooctaose) were calculated and explored in vacuum, water and in the active site of CBHI. It was observed that the binding pocket of enzymes limits the ring pucker and that the active site amino acids preferentially stabilise certain puckering conformations. For CBHI, the first part of the glycosidase reaction is the glycosylation step.

This was driven to completion during SCC-DFTB QM/MD FEARCF calculations where GLU212, ASP214 and GLU217 and part of the substrate were treated quantum mechanically. The general hybrid orbital method was used to connect the QM and MM regions. The free energy barriers of glycosylation were computed and the puckering statistics during the conversion of cellooctaose to products were correlated with this. Guanosine, a 5-membered ribose derivative is phosphorylated by Purine Nucleoside Phosphorylase (PNP) in order to salvage the guanine base. The effect of the PNP protein environment on ring pucker was studied by using FEARCF SCC-DFTB QM/MD non Boltzmann free energy calculations to quantify the pucker change induced in guanosine when changing environment from vacuum, to water and to the protein. *In vacuo*, the E_4 and E_1 pucker conformers were observed as minima. Upon solvation, the puckering phase space became less restricted with the ${}^{3}T_{4}$ and ${}^{2}T_{3}$ pucker conformers as minima. In the PNP active site pucker became restricted with only the ${}^{4}E$ conformer observed.

DECLARATION

I declare that the work in this thesis, "Developing Methods to Construct Ring Pucker Free Energy Hypersurfaces Applied to the Analysis of Glycosidase Enzyme Catalytic Mechanisms" is based on research performed at the Scientific Computing Research Unit, the Department of Chemistry, University of Cape Town, South Africa. No part of this thesis has been submitted elsewhere for any other degree or qualification. I have received aid from fellow postgraduates and my supervisor where warranted. Unless referenced to the contrary, all work in this text is my own.

Chris Barnett

October 2010

ACKNOWLEDGEMENTS

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.

Prof. Naidoo thank you for the opportunity to work for and with you. Thank you for your encouragement, your pleasant nature, your genuine concern, your positive drive and ambition and all the productive time you were able to contribute to my growth, learning and understanding of this complex research field.

Much thanks to all the support from friends and family. For cakes, cookies, clothes and cash; for liquor, liquorice and love; and for those of you who understood a word of the thesis and how frustrating and time-demanding it could be. I am really grateful for everything and **Love all of you very much.**

Thanks to friends and "colleagues" (sic.) from the Scientific Computing Research Unit a.k.a. "The lab". Thank you to all of you who took a quick peek at this thesis and offered advice. I'd like to extend further thanks to Karl Wilkinson for aid in the electronic analysis for the cellulase paper, Riedaa Gamieldien for taking over my responsibilities, Ranga Jayakody for interesting discussions, Richard Matthews for being there, Gerhard Venter for "interesting" discussions and Kyle Fernandes for the Galaxy.

"Thank you!" to all, as yet unmentioned, friends and colleagues from the Chemistry department and from the University. A debt of gratitude is owed to the administrative and support staff of the Department of Chemistry, who are always efficient, helpful and cheerful.

ABBREVIATIONS

Å: Angstrom

ABF: Adaptive biasing force

AO: Atomic orbital

CBH: cellobiose hydrolase

DFT: Density functional theory

FEARCF: Free energies from adaptive reaction coordinate forces

GH: Glycoside hydrolase

GPU: Graphics processing unit

IR: infrared spectroscopy

IUBMB: International union of biochemistry and molecular biology

MD: Molecular dynamics

MM: Molecular mechanics

NMR: Nuclear magnetic resonance spectroscopy

NPT: Isothermal-isobaric ensemble

ns: nanosecond

NVT: Canonical ensemble

PMF: potnential of mean force

PNP: Purine nucleoside phosphorylase

ps: picosend

QM: Quantum mechanics

QM/MM: Quantum mechanic/molecular mechanics

SCF: Self consistent field

TS: transition state

WHAM: Weighted histogram analysis method

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 Introduction

1.1 Biochemical Importance of Carbohydrates

Sugars, also called carbohydrates or glycosides, are a class of biomolecules that are a fundamental part of any organism's biochemical lifecycle. Green plants (*Viridiplantae*) photosynthetically create simple sugars which are stored as polymeric starches (amylose) or are converted to molecules that provide structural support (hemicelluloses, xylan and cellulose). *Animalia* such as herbivores and omnivores eat plants, degrading starches to simple sugars for energy and as an entry point for biosynthesis (Figure 1-1). The molecules which transfer genetic information, deoxyribonucleic acids (DNA), have sugars as a core component of their building blocks (nucleotides). Simple sugars or monosaccharides are also a core component of nucleosides, the building blocks of ribonucleic acids (RNA), which transcribes the DNA and acts as a template for protein biosynthesis. Adenosine triphosphate (ATP) is a sugar-containing molecule that is an immediate source of free energy in biological systems, while proteoglycans (proteins with sugars attached) function as lubricants and structural components in connective tissue. The recognitive ability of cells, particularly that of the immune cells, is mediated by sugar molecules. Immuno-recognition is necessary for the correct immune response, although immune systems can be deceived (for example, the human immune system by the human immune-deficiency virus, HIV[1]).

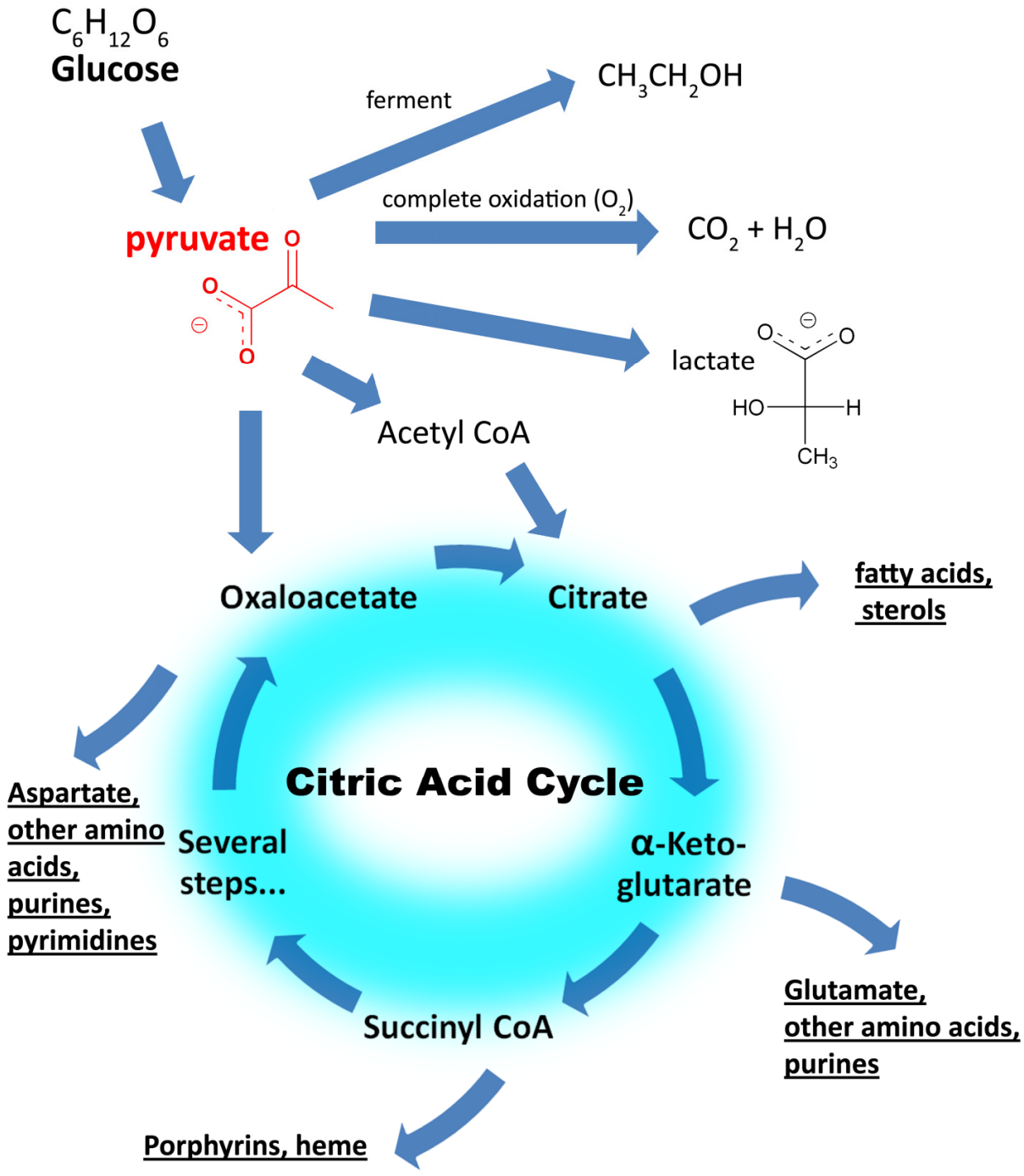

Figure 1-1 The various fates for glucose. On conversion to pyruvate, it is an entry point into the citric acid cycle and necessary for biosynthesis (adapted [2, 3])

1.2 Importance of Hydrolysis

Hydrolysis is the addition of water to cleave chemical bonds and is the underpinning reaction for digesting food molecules. For example, in the human mouth and gut, amylases digest starches, while trypsin and pepsin digest proteins in the gut. The hydrolysis reaction is also useful in the carefully designed destruction of damaged proteins and enzymes in cells, controlled by proteases. Obsolete or damaged cells are degraded by caspases.[4]

In sugars, which are composed of monosaccharide subunits joined by glycosidic linkages, the hydrolysis reaction cleaves the glycosidic bond. This is necessary for recycling and re-use of polysaccharides. Yet, the half-life of the hydrolysis reaction for the linkage between two glucoses in glucopolymers is in the range of 5 million years.[5] Cleavage of this bond for use in reactions on the biological time scale is catalysed by a class of enzymes known as glycoside hydrolases (GH) or glycosidases. The uncatalysed hydrolysis reaction for a disaccharide proceeds as in Figure 1-2. On addition of water, the glycosidic linkage is broken to yield two glucose monosaccharides. Hydrolysis may result in inversion or retention of the anomeric configuration.

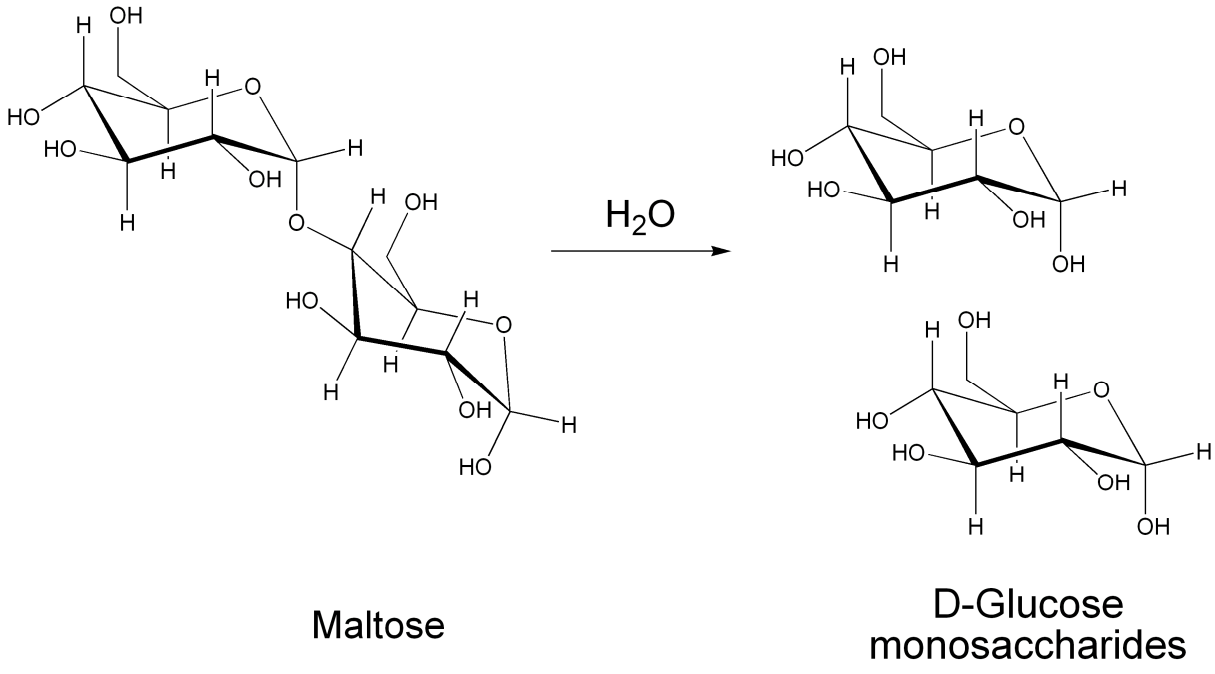

Figure 1-2 The hydrolysis of a disaccharide, maltose, by addition of water yields two glucose monosaccharide units (adapted[2, 3])

1.3 Link between Conformation and Degrees of Freedom

Carbohydrates are diversely functional because of the complex features of the monomeric unit. The monosaccharide unit has many degrees of freedom, which implies that there are a variety of conformations accessible to this moiety. Thus it has complex behavior and diverse functionality.

1.3.1 Carbohydrate monomers

The monomeric units of carbohydrates are diverse but fall into two categories; hexoses and pentoses. Hexoses contain 6 carbon atoms while pentoses contain 5 carbon atoms. These molecules are not linear and prefer to cyclise and form an intramolecular acetal. The six membered cycle is termed a pyranose, while the five membered cycle is a furanose. The cyclisation reaction for glucose is shown in Figure 1-3. An alcohol of the molecule (coloured red in the open chain D-glucose of Figure 1-3) intramolecularly reacts with the aldehyde (coloured green in the open chain D-glucose of Figure 1-3) to form a hemi-acetal.

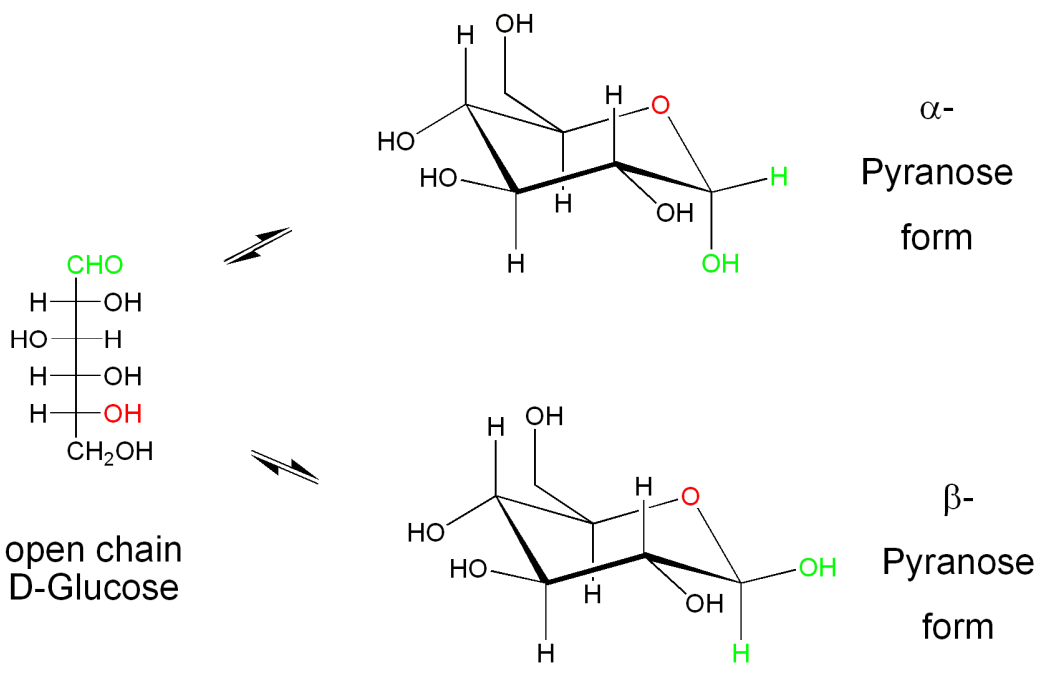

Figure 1-3 The monomeric units of sugars in the linear (open) form has aldehyde and alcohol functionality. Intramolecular acetal formation occurs to yield a closed form, cyclic hemi-acetal (adapted [2, 3])

On cyclisation, a new chiral centre is formed at C1, also called the anomeric carbon. When the hydroxyl group points downwards (below the plane of the ring) it is in the α position and when pointing

upwards, it is in the β position. These forms are called α- and β-anomers (or axial and equatorial) and either are possible after cyclisation. Although, depending on the electronegative substituents at the anomeric position of pyranoses, there is a tendency for the axial configuration. This is termed the anomeric effect [6, 7] and provides stabilisation of the axial substituent such that the inherent steric bias of the substituent is overcome. Dipole-dipole interactions may be the root cause for this effect. The conformation in which dipoles partially cancel out (Figure 1-4) is described as favoured over the one in which dipoles strengthen each other.[8]

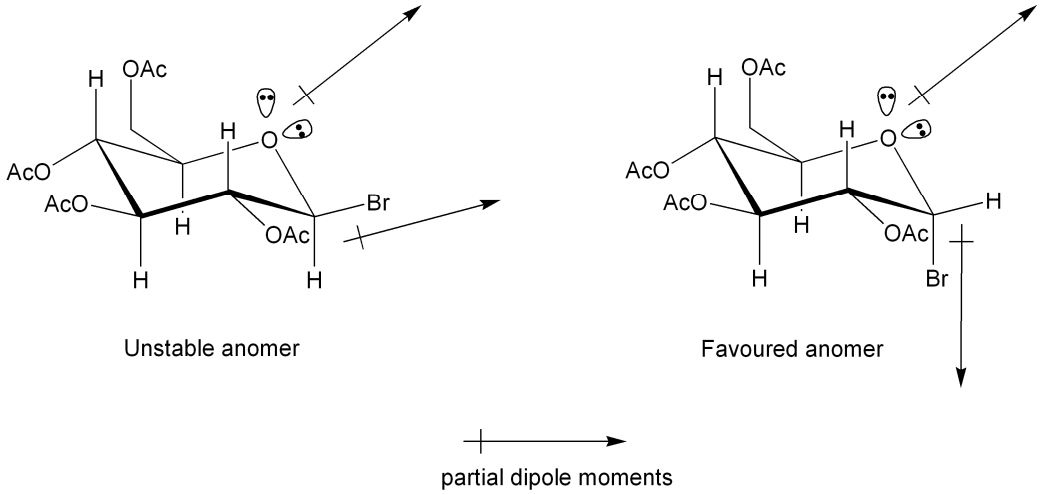

Figure 1-4 Dipole moments for an acetylated and brominated derivative of glucose. The equatorial conformer is on the left and the axial conformer on the right (adapted[8]).

An alternative explanation is that of a negative hyperconjugation effect due to the interaction of a C-X antibonding lowest unoccupied molecular orbital (LUMO) and a C-Y non-bonding highest occupied molecular orbital (HOMO). In Figure 1-5, the axial conformer has a good n→σ*orbital overlap that is not possible in the equatorial conformer. This causes energy lowering.[7, 8]

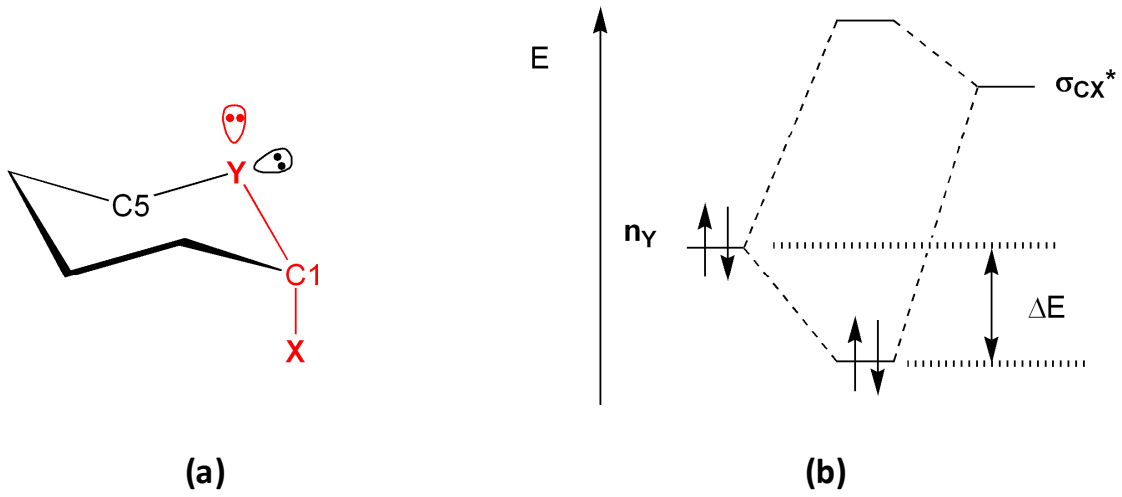

Figure 1-5 An axial anomer of a substituted pyranose derivative with the non-bonding orbitals of Y (red) and the antibonding orbital of the C-X bond anti-periplanar. The filled nonbonding (n_Y) and empty antibonding orbitals (σ_{CX}^*)can mix to form a filled lower energy orbital and an empty higher energy orbital (adapted[8]).

1.3.2 Multiple linkages

The monosaccharide units have several alcohol groups. These alcohol groups can undergo a condensation reaction to form a linkage, the glycosidic linkage, between two monosaccharides. Two monosaccharide monomers glycosidically linked are termed a disaccharide. Several glycosidically linked monosaccharides are oligosaccharides, while many units linked together are termed polysaccharides. Due to the large number of hydroxyl groups on the periphery of the monomeric units (Figure 1-3), many types of linkages are possible. Several commonly occurring molecules: cellobiose, adenosine, sucrose, lactose and maltose, shown in Figure 1-6, illustrate this variety. Cellobiose is a disaccharide of glucose, which can be further linked to form cellulose a structural polymer found in plants. In cellobiose, the two glucoses are O-glycosidically linked between C1 of one glucose and the hydroxyl of C4 on the other sugar in such a way that the glycosidic linkage lies β to the C1 atom. This is termed a β (1→4) glycosidic linkage. The nucleoside adenosine, comprises a base and a furanose sugar (ribose). A nitrogen atom (N9) of the base is N-glycosidically linked to the C1 atom of ribose. Sucrose (table sugar), lactose (milk sugar) and maltose (which polymerises to forms starches) are abundant disaccharides which have different linkages. Sucrose is α (1→2) linked, lactose is β (1→4) linked and maltose is α 1(→4) glycosidically linked.

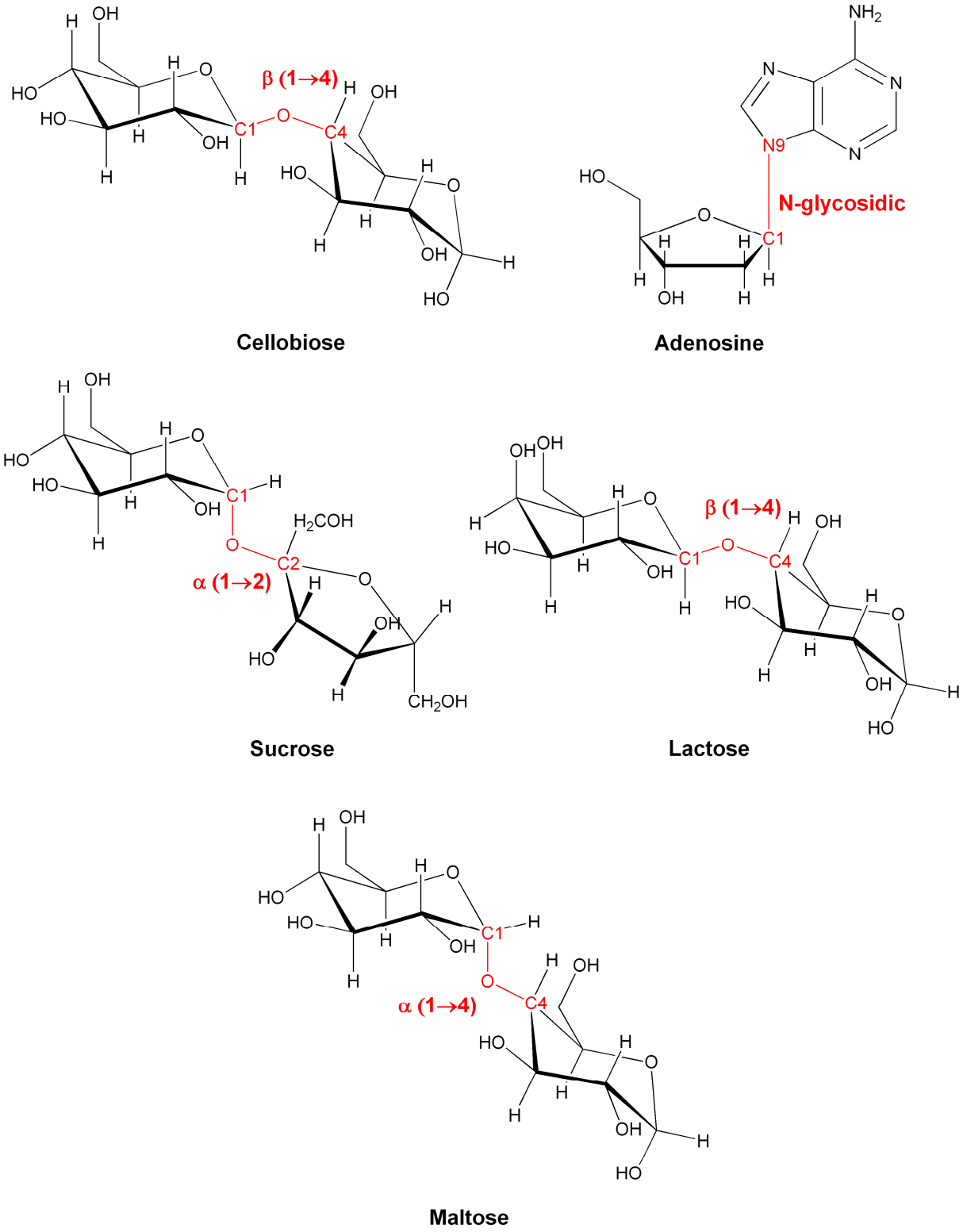

Figure 1-6 Some of the possible glycosidic linkages (highlighted in red) in commonly occurring molecules

1.3.3 Rotational degrees of freedom

Part of the conformational complexity of the monosaccharide unit is due to the hydroxyl groups on the periphery. These can be used to form linkages between other monomers and they exhibit rotational degrees of freedom. The rotational conformation of the secondary alcohol groups is such that they engage in hydrogen bonding interactions, whether these are intra- or inter-molecular. The primary alcohol group (-CH_2OH) has more degrees of freedom than the secondary alcohol groups, and is also able to hydrogen bond. However, there is a preference for this group to form a torsional angle with the ring oxygen (O-C-C-O) such that the oxygens' adopt a synclinal (gauche) conformation. The tendency for the torsional angle about C-X-Y-C or X-C-C-Y molecular fragments to prefer the synclinal (gauche) conformations (where X and Y are electronegative atoms) is termed the gauche effect. The gauche effect is a stereoelectronic preference for conformations in which the best donor lone pair or bond is anti-periplanar to the best acceptor bond.[6, 9, 10] The effect is exhibited in simple substituted alkanes such as 1, 2-difluoroethane[11] and 1, 2-ethane-diol[12, 13] but is especially relevant in determining the conformational freedom about the -O-C-C-O- fragment (the hydroxymethyl or primary alcohol) of monosaccharides[14] and 1-6 linkages between subunits of polysaccharides.

When sugar monomers are glycosidically linked there is a torsional degree of freedom about the linkage. NMR studies and molecular modelling studies of methyl-α-cellobioside have indicated that a large flexibility is present in the linkage.[15] There is a preference for synclinal (gauche) conformers about the O-C-O-R motifs and this is termed the exo-anomeric effect. Alkoxy substituents at the anomeric position as shown in Figure 1-7, tend to prefer the conformation that minimises steric clashing and maximises any hyperconjugative effects.[6, 16]

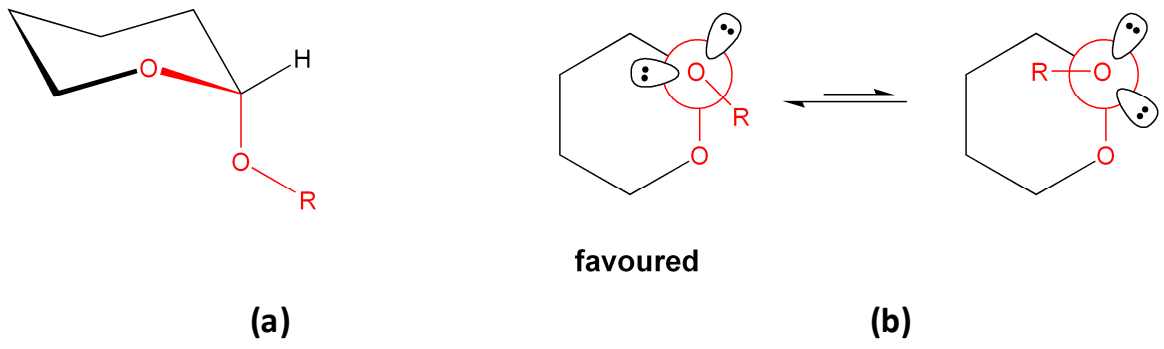

Figure 1-7 (a) The -O-C-O-R motif of a pyranose ring is indicated in red. (b) The two gauche C-O bond rotamers are shown where the dynamic equilibrium between them lies to the left. The favoured rotamer has favourable stereoelectronic interactions and minimises unfavourable steric clashes (Adapted[7]).

1.3.4 Pucker

The monomers of sugars, being cyclic molecules not dissimilar to cyclohexane, can also librate to adopt a range of ring conformers or puckers. The pucker or conformational behaviour of cyclic molecules has been a subject of investigation for more than a century.[17, 18] The existence of various ring puckering conformers was put forward by Haworth.[19] The preference for the $^{4}C_{1}$ chair conformer was established by both Hassel and Ottar[20] and Reeves.[21-23] The various conformers for a 6-membered ring are shown in Figure 1-8. There are 5 types of canonical conformer: the chair, half-chair, boat, twist- or skew-boat and the envelope.

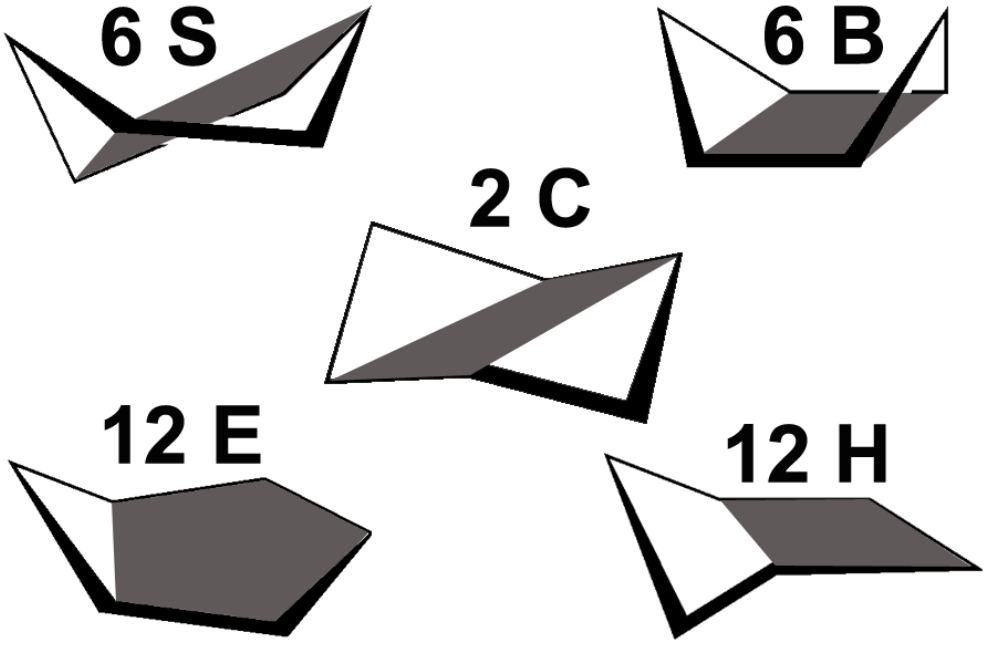

Figure 1-8 The 38 canonical conformers available to a six membered ring: Two chairs (C), six boats (B), six twists (S), twelve half-chairs (H) and twelve envelopes (E). The shaded area connects atoms in the same plane.

Studies have shown that there is dynamical inversion between these puckering conformers[24, 25] and that for cyclohexane the chair form is preferred.[26, 27] Reports of other ring forms include the boat form observed in cuprammonium galactose solution[28] and the skew forms seen in disaccharides.[29] The pathways of interconversion between these differing ring conformers are often considered using Stoddart's diagram (Figure 1-9). This diagram indicates the possible routes between different pucker conformers but gives no indication of the energy cost to travel between conformers.

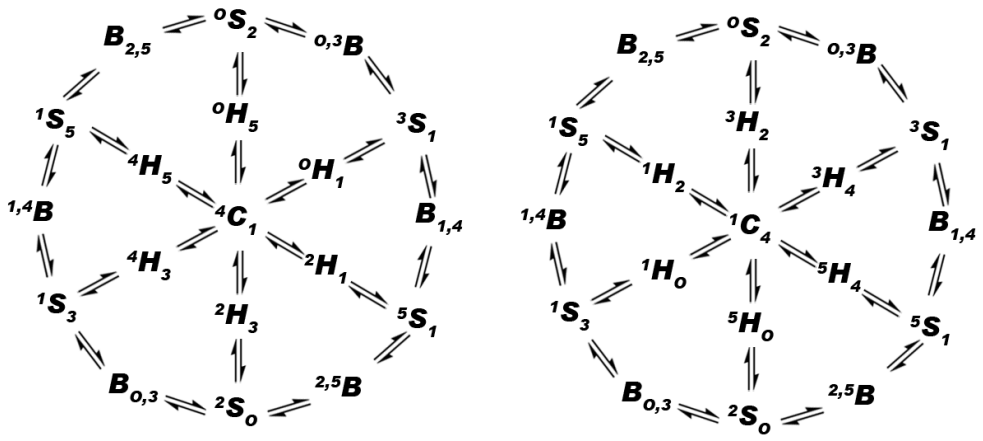

Figure 1-9 Stoddart's diagram of possible pyranoside ring inversions. (adapted [30])

1.4 Connection between Pucker and Hydrolysis

A major mode by which enzymes catalyse reactions is by stabilisation of the transition state (TS). The TS is high in energy due to the electronic and structural changes taking place in the reactants as they progress to products. Stabilisation of this high energy state is one way in which enzymes provide catalytic reaction rate enhancements of up to the order 10^{10}-10^{17}.[31] In the hydrolysis reaction, briefly discussed in §1.2, catalytic enhancement is required to cleave glycosidic linkages rapidly.

Possible intermediates depend on the mechanism of the reaction. In the reactions of sugars, differing puckers may facilitate the attack of the nucleophile by opening up the face of the electrophilic anomeric centre (C1), in a concerted reaction. Then again, in a stepwise reaction, a carbocation is formed. In sugars, there is evidence for the carbocation to be stabilized by the donation of electrons from the ring oxygen to the carbon centre. This oxocarbenium intermediate is planar at C5-O5-C1-C2 and necessitates a change in ring pucker conformation. Due to the finely tuned active site environment provided by enzymes, there is little likelihood that the pucker conformer of sugars in the TS is left to chance. Rather, enzymes must carefully induce ring puckering to enhance alignment of reacting molecules such that the specificity and reaction rate are optimal. The ability of a sugar to pucker in enzymatic active sites is necessarily constrained. Firstly, by the size of the active site cavity and secondly, by the specific interactions available with the active site amino acids.

There are only a select few conformers, shown in Figure 1-10 and Figure 1-11, that satisfy the requirements for an oxocarbenium intermediate. For furanoses, from the 20 non-planar conformers,

just two of these are able to satisfy the condition for the oxocarbenium intermediate. These are the ^{3}E and E_3 conformers illustrated in Figure 1-10. For pyranoses, 12 of the 38 non-planar conformers can satisfy this condition (Figure 1-11). The pucker conformers that allow planarity at C5-O5-C1-C2 are specific half-chair, envelope and boat conformers - $^{4}H_3$, $^{3}H_4$, $^{4}H_5$, $^{5}H_4$, $^{2}H_3$, $^{3}H_2$, ^{4}E, E_4, ^{3}E, E_3, $^{2,5}B$, $B_{2,5}$.

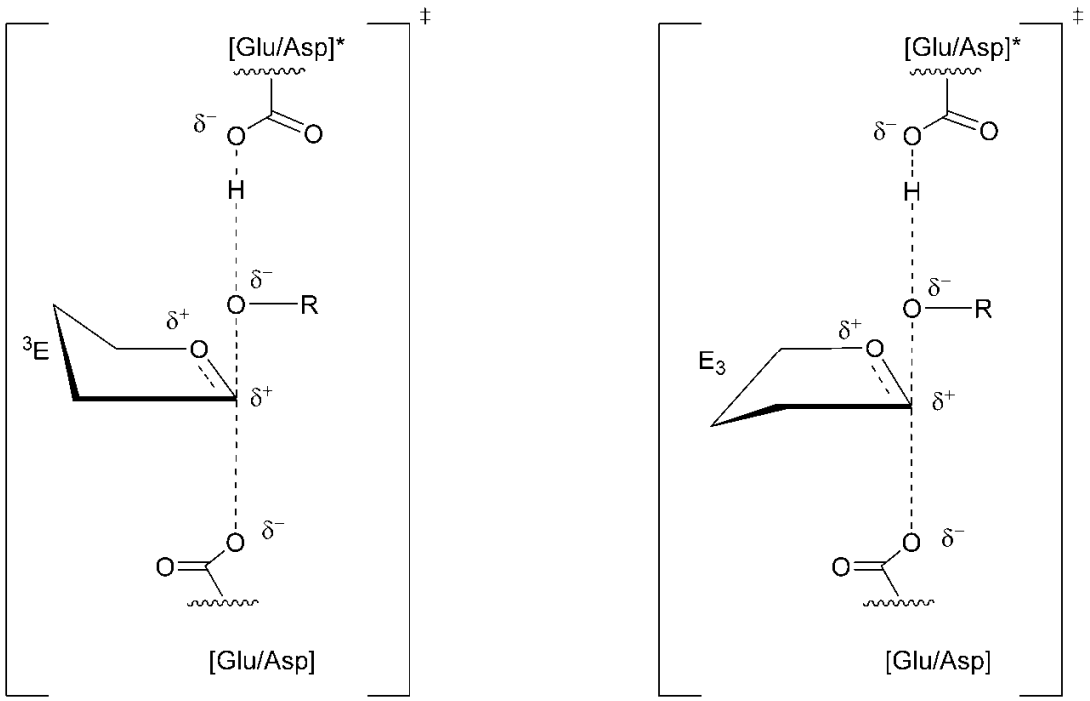

Figure 1-10 The possible oxocarbenium furanose transition states that may occur in enzyme catalysed sugar-linkage hydrolysis. The anomeric carbon, which has planar character may only exist as the ^{3}E (also known as 3-endo) or E_3 (3-exo) conformers. * indicates that the amino acid is protonated.

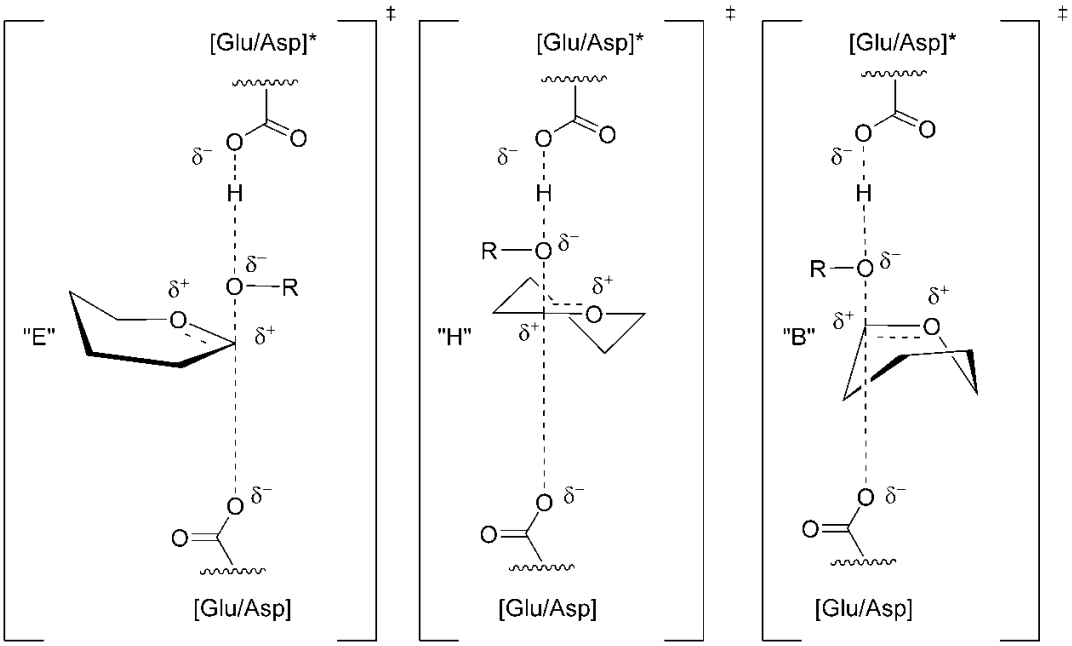

Figure 1-11 The possible oxocarbenium pyranose transition states that may occur in enzyme catalysed sugar-linkage hydrolysis.. The anomeric carbon, which has planar character may only exist in envelope ("E"), half-chair ("H") or boat ("B") conformers as represented by figures on the left, centre and right respectively. The specific conformers that can be accessed are $^{4}H_{3}$, $^{3}H_{4}$, $^{4}H_{5}$, $^{5}H_{4}$, $^{2}H_{3}$, $^{3}H_{2}$, ^{4}E, E_{4}, ^{3}E, E_{3}, $^{2,5}B$, $B_{2,5}$. * indicates that the amino acid is protonated.

There has been evidence for the existence of these conformers from X-ray crystallographic studies of proteins with bound substrate and transition state analogues. Davies et al. mapped the conformational possibilities for a variety of GH's and found that the conformational pathways varied (Figure 1-12).[31] In Figure 1-12, which indicates possible pucker conformational pathways on the $^{4}C_{1}$ side of Stoddart's diagram, there is no clear puckering trend based on the reaction catalysed. For example , GH26 β retaining mannanases (retain the stereochemistry at the anomeric position) and GH5 retaining cellulases belong to the same clan[32] of GH's and have a similar fold (tertiary structure) yet these enzymes have different pucker conformational reaction pathways (Figure 1-12).

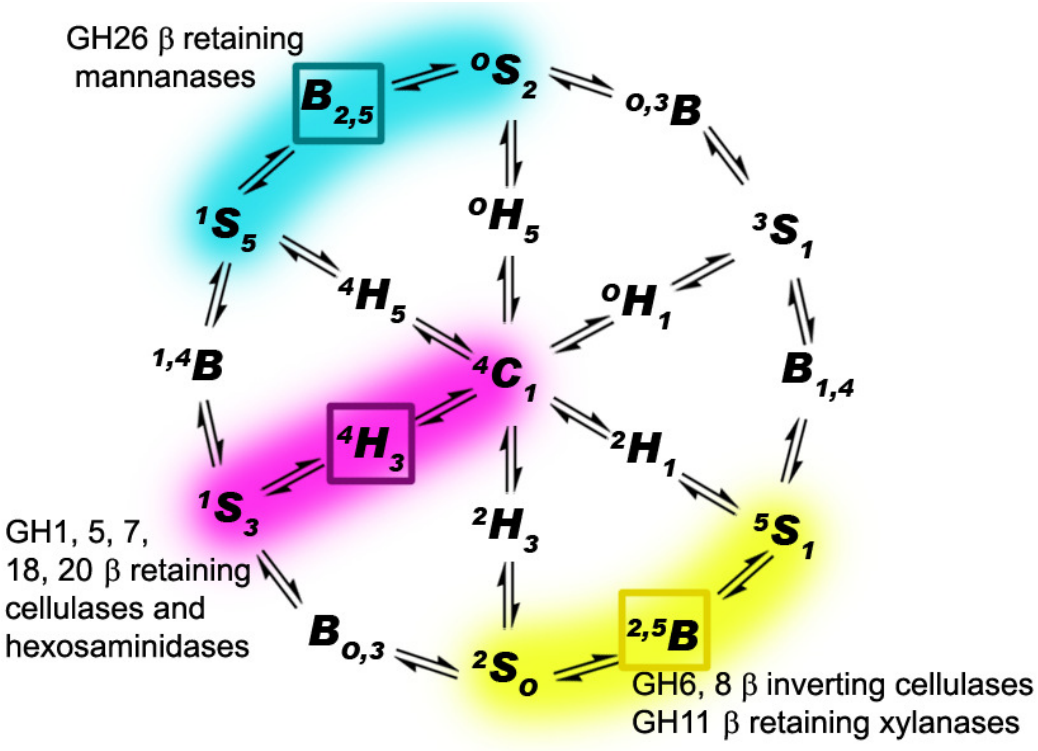

Figure 1-12 Proposed pyranoside conformations for the TS of the glycosylation reaction placed onto Stoddart's diagram.[31, 33] Highlighted regions indicate possible pucker transitions in the GH's indicated. Boxed conformers are the puckers proposed to occur at the TS. Retaining enzymes retain the stereochemistry at the anomeric position and inverting enzymes invert the stereochemistry at the anomeric position.

1.5 Important Enzymes

The International Union of Biochemistry and Molecular Biology (IUBMB) nomenclature scheme classes protein enzymes based on the reaction they catalyse into Enzyme Classifications (EC). For example, EC2 is the specification for transferases while EC3 specifies a hydrolase. Sub classes further specify enzymatic character based on the nature of the substrate, for example EC3.2 catalyse hydrolysis of saccharide linkages and EC3.2.2 catalyse the hydrolysis of O- and S-glycosyl linkages.

Several of the many important enzymes that cleave polymers of the pyranoses include amylases, cellulases and lysozyme. Amylases play a role in carbohydrate digestion by cleaving the bonds in starches.[34] Cellulases aid in the breakdown of plant biomass and may also play a role in degrading plant cell walls to allow for plant cell growth.[35] Lysozyme, found in tears and nasal sputum, has a specificity for lysing peptidoglycans (a sugar-based polymer) of gram positive bacteria cell walls and has bactericidal properties.[36] The glycosidases are structurally diverse and are classed into families that show similar

primary (the protein amino acid sequence) and tertiary structures (the folded protein). For example, human amylase exhibits a $(\beta/\alpha)_8$ barrel fold with eight αhelices and eight parallel β strands that alternate along the peptide backbone (Figure 1-13). While, a cellulase called cellobiohydrolase I (CBHI) folds into a β jelly roll where four pairs of anti-parallel β sheets are folded-over to form a barrel(Figure 1-13). These enzymes are classified into different families and clans.[37-39]

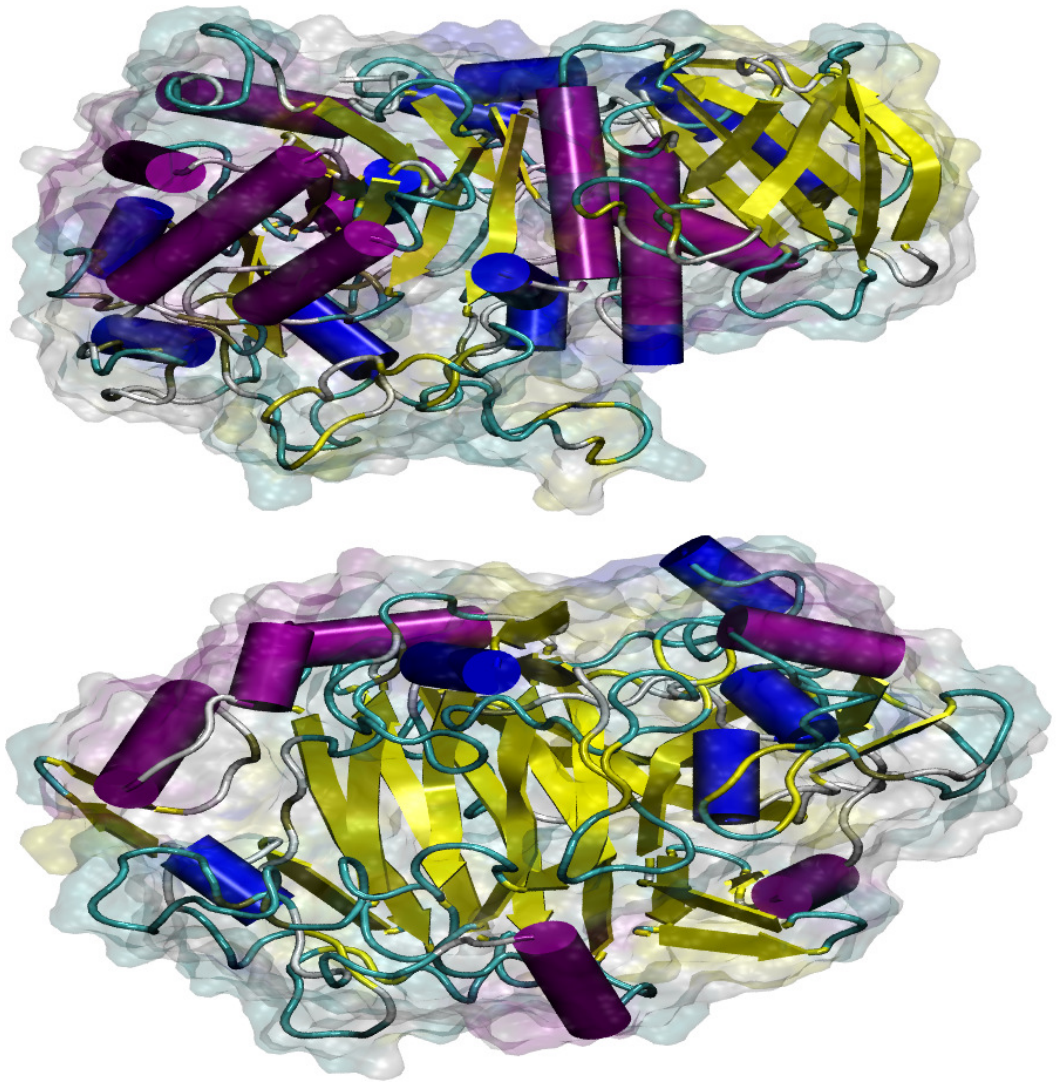

Figure 1-13 The differing tertiary structures of human amylase (PDB: 1SMD[40]) and cellobiohydrolase I (PDB: 7CEL[41]). Amylase (top) folds into a $(\beta/\alpha)_8$ barrel-fold. Cellulase (bottom) folds into a β jelly roll. The secondary structure elements (helices, sheets) are represented in colour. α-helices are purple tubes, 3/10 helices are blue tubes, β-sheets are yellow arrows. The thin tubes represent random coil conformations.

5-membered sugars, or furanoses, are the key sugar moiety of the building blocks of DNA and RNA. Both the maintenance of the genetic code and the synthesis or salvaging of new building blocks are important processes. Base excision repair enzymes catalyse the removal of damaged DNA bases by cleavage of the sugar-base C-N glycosidic linkage.[42] Transferases such as purine nucleoside phosphorylase (PNP) can catalyse the creation of nucleosides from ribose-phosphate and the requisite purine base. Similarly, the cleavage of the sugar-base C-N glycosidic and transfer of phosphate is possible.[43]

1.6 Pucker – How Is It Studied?

X-ray crystallography[44-48] and neutron diffraction[49] studies of pyranoses such as α-glucose[45, 49], β-glucose,[46, 47] cellobiose[44] and maltose[48] show the $^4C_1$ conformer in the solid state. For furanoses, in the solid state, there is little tendency for the planar conformer, as free sugars and nucleotides prefer envelopes.[50-53]

NMR is a valuable experimental technique for studying the conformation of sugars in solution.[54] This is despite the fact that low temperature studies are necessary because of the rapid conformational inversion of sugars rings at room temperature, causing broad unresolved signals in the spectra. Durette and Horton have extensively applied this technique to pyranoid and furanoid derivatives.[24, 55-59] The furanoid sugars are considered to have C2 or C3 "meta" to the ring oxygen[60], that is to be in the E_2, ^2E, E_3 or ^3E envelope conformers. The combination of molecular mechanics and NMR has been used to further the understanding of the puckering conformations.[61, 62]

The role of sugar ring pucker is intrinsically tied to the hydrolysis reaction mechanism and the TS. It is for this reason that the role of pucker in enzymatic substrates has been studied. X-ray crystallography and studies of the kinetic isotope effect (KIE) tend to be used in the majority of experimental work. The application of these methods to hexoses of β retaining glycosidases and pentoses of N-glycosyl hydrolases and transferases will be discussed in this section.

1.6.1 Hexoses of β retaining glycosidases

1.6.1.1 Crystal structure

The retaining β-glycosidases catalyse the formation (glycosylation) and consequent hydrolysis (deglycosylation) of a covalent intermediate via oxocarbenium-like transition states. Substrate distortion

(into puckers other than $^4C_1$) has been observed in the crystal structures of enzyme-substrate (Michaelis) complexes for glycoside hydrolases.[63] For instance, in studies of *Fusarium oxysporum* endoglucanase I[64], ring distortion was observed with a non-hydrolysable substrate analogue in the active site cavity. The distortion gave rise to a quasi-axial-orientation of the leaving glycosidic bond. Similar conformations, have been observed in an unhydrolysed substrate complex of chitobiase[65] and another retaining endoglucanase.[66] Unhydrolysed substrates have been observed to adopt $^1S_3$ skew-boat pyranoside conformations or a closely related ^4E envelope.[63] The covalent intermediate has been observed as an undistorted $^4C_1$ chair.[31, 63, 67]

A further understanding of the complex puckering route undertaken by pyranoses was provided by Davies et al. in their study of a retaining β-glycosidase.[66, 68] Structures of the native, substrate-bound, covalent intermediate, and product complexes of Cel5A from *Bacillus agaradhaerens* were obtained from X-ray crystallographic experiments and the bound-enzyme-substrate complex (Michaelis complex) was distorted into a $^1S_3$ skew-boat conformation. A distorted $^{2,5}B$ (boat) conformation was observed in the covalent substrate-enzyme intermediate in xylanase.[69, 70] These xylanases appear to undergo hydrolysis through a boat transition state rather than the half chair that appears to be used by glucosidases and cellulases.[68]

Hydrogen bonding interactions observed for the sugar 2-hydroxyl have been shown to be important for catalysis, contributing to transition state stabilisation.[71, 72] Direct observation of such interactions within the glycosyl-enzyme intermediate has been achieved.[68] The interaction between the 2-hydroxyl and incoming nucleophile (in glycosylation, this is a glutamate residue) is suggested to be responsible for most of the stabilisation at the 2-position.[73, 74]

1.6.1.2 Kinetic isotope effects

Isotope effects are the possible change in rate and equilibrium constant of a reaction that occur upon isotopic substitution.[42] The KIE is defined as the ratio of light to heavy isotope rates; KIE=$^{light}k/^{heavy}k$. When isotope substitution occurs, the primarily result is modification of the vibrational bond, angle and torsional frequencies. The KIEs are interpreted computationally by treating molecules as balls, and bonds as springs, with bonds acting as ideal harmonic oscillators. The vibrational environments of atoms between the reactant and TS are gleaned from the KIE[75] and so this experimental technique probes the TS directly.[42] Often, multiple positions in the substrate are labelled and the KIEs are then determined. This is followed by a computational analysis. Both synthesis of isotopically labeled molecules and the KIE

measurement are practically challenging to accomplish.[75] In 1995, Singleton et al.[76] established a method of measuring multiple, high precision ^{2}H and ^{13}C KIEs by NMR on unlabelled materials (in natural isotopic abundance). This has been successfully employed to measure KIEs in enzymes.[75]

KIEs on the atoms undergoing bond distortion are termed primary (α), while those at positions not directly involved in the chemical bond distortion are termed secondary (β). The mechanism of glycoside hydrolysis reactions, which may be bimolecular or stepwise, has been studied with KIEs as has the importance of substrate ring conformation.[75] The pucker of the TS is known to have considerable oxocarbenium character and may adapt the conformers illustrated in Figure 1-11.[63] Large α-^{2}H KIEs have been used as evidence for TS's with oxocarbenium ion character.[75] Hyper conjugation between the C2-H2 σ bond into the empty p-orbital of the anomeric carbon can help stabilise the TS, although this interaction is strongly angle dependent.[77-79] Hyperconjugation causes a weakened C2-H2 bond which is observed as large β-secondary ^{2}H or ^{3}H KIEs.[42] Based on the match of calculated with experimental KIEs, certain β-glucosides are known to proceed through the TS with E_3 or $^{4}H_3$ conformations.[75]

1.6.2 Pentoses of transferases and N-glycosyl transferases

1.6.2.1 Crystal structure

The crystal of a DNA substrate analogue with bound uracil (a furanose derivative) shows a flattening of sugar pucker.[80] Nucleosides and their derivatives exhibit ^{4}E conformers in PNP (PDBS: 1JE1[81], 1RFG[82], 3IEX[83], 1PK7[84], 1A9S[85], 1PR0[84], 1V45[82]) and in another nucleosidase(1JYS[86]). E_4 conformers have also been observed(PDB: 1RCT[87]). Sundaralingam[51, 88] analysed furanose sugar derivatives in the solid state and observed ^{2}E, E_2, ^{3}E and E_3 conformers. Ribose-1-phosphate in PNP exhibits E^3 conformers and E_2 conformers (3FB1[89]), while, transition state analogue inhibitors may adopt a planar conformer in (1RT9[90]).

1.6.2.2 Kinetic isotope effects

Investigation of the transition state via the measurement of multiple kinetic isotope effects has been valuable in the design of enzyme inhibitors for nucleoside hydrolases and ribosyltransferases (many of these developed by Schramm et al.[91-94]).[68] Large magnitude β-deuterium KIEs have been observed which are indicative of strong hyperconjugation between the elongated glycosidic bond and the C2−H2 bond. Maximal hyperconjugation corresponds to an E_3 conformation and on interpretation of the β-deuterium KIEs, the E_3 conformer is expected.[42, 77, 80]

1.7 Computational Advances and Advantages

With the advent of the desktop computer, supercomputing and now GPU computing[95], more complex computational experiments can be used to delve into the intricate depths of molecular systems. This is evident in the drive towards study of larger molecular systems[96-98] such as proteins, complex fluids and membranes and the use of more stringent quantum chemical theories[12, 99] which explicitly treat electrons. Time-based simulations for study of molecular motion, such as molecular dynamics (MD) have also been extended to longer time scales.[100-102] The use of hybrid simulations, where the active region is treated quantum mechanically (QM) and the rest of the system less stringently, is popular for the study of reactions in enzymes.[103-108]

Computational modelling is an essential complement to experimental techniques, as certain lab experiments are not feasible or cannot provide the atomistic detail sought after. Where sufficiently high-quality crystals can be obtained, X-ray crystallography experiments can provide a "snapshot" of a molecular state. This "photograph" can provide essential insight into the workings of molecules but does not provide information about the dynamic solution structure. NMR experiments of solutions usually provide time averaged spectra which can sometimes be resolved at low temperatures.

The advantage of computational simulations is that they can provide insight into the finer aspects of electronic structure with quantum mechanical methods. The time dependent properties of a system can also be calculated and understood with MD. On consideration of molecules that are flexible, the types of motions that occur are translation, rotation, torsion and vibration. Such motions are quick, especially for light atoms and in strained systems such as molecular transition states. For this reason, it is prudent to consider simulations where the time interval between successive molecular dynamics steps is on the femtosecond timescale.

A key aspect of molecular simulations is the validation of the model to the target applications. In the context of molecular modelling the validation of the model is not simple and common validation procedures involve comparison of simulation to experiment, or where possible, to very high level quantum mechanics simulations.[99]A comment by Cramer[109], on the validity of the validation process, is to be cautious; not only because of the magnitudes of experimental and computational error, but also because the experimental model and computational model may be measuring different properties of molecules in different environments. On a positive note, "Validation is often defined as the process of

determining the degree to which a model is an accurate representation of the real world from the perspective of its intended uses.”.[110]

1.8 Exploring Conformational (Phase) Space

Statistical thermodynamics / mechanics theory relates the microscopic and macroscopic properties of a system. As large molecular systems have many degrees of freedom it is necessary to explore many possible configurations and conformations for a sufficiently good sampling of these. Molecular dynamics methods drive the system dynamically using Newton's equations. The questions " when is a simulation long enough?” and “how long should a simulation be?" are often posed in computational chemistry. Within the constraints of technology available, a simulation that can be carried out in a "short" time frame (week to months) and can be analysed quickly is considered acceptable. In the 90's, MD simulations of at most a nanosecond were commonplace, while classical MD simulations are now at least tens of nanoseconds in length. These questions can be rephrased as “what is a long enough simulation, for the periodicity of the motion in a system being studied, so that I sample this motion sufficiently enough and so that the chemical interpretation is realistic?” Calculation of properties such as diffusion, depend on the average movement of all the molecules over time. To observe realistic evolution of the molecular system and to calculate such properties, adequate sampling is required.

A dilemma presents itself for very lengthy sampling. Firstly, most simulations are run at constant NVE, NVT or NPT (constant number of particles (N), volume (V), pressure (P), temperature (T)) and so have a finite energy available to them. Systems may reside trapped in certain regions of configurational and conformational space and not have the energy to escape these regions. In short, "lengthy simulations" may not yield much more information than a shorter simulation. Secondly, there is a trade-off between quantum mechanical and classical mechanical accuracy. Tens of nanoseconds of classical MD simulations are routine but this is not the case for quantum mechanical methods. Depending on the nature of the QM calculation, minimisation of systems with as few as 40 atoms may require more than 24 hours. QM dynamics is possible and the prominent algorithm in this field is Car and Parrinello’s method. This method has been implemented in the Car-Parrinello molecular dynamics (CPMD) code.[111] QM/MM dynamics are a compromise between an entirely QM or entirely MM system. The electronic structure of certain part of a molecular system is treated as QM and for the rest, where the explicit electronic character is not as important, is treated with MM. [106, 112]

It is not often possible to fully traverse the phase space of a molecular system using standard MD. Furthermore, a particular degree of freedom may be important for the evolution of the molecular system; for example, the attack of a nucleophile on an initially sp^3 carbon. In this situation, methods that bias the molecular system to improve sampling are needed. These methods are called free energy methods and improve sampling so as to obtain a complete depiction of the phase space available to a specific coordinate. Free energies depend on sampling of not just easily accessible states, but also high energy states, and many approaches apply biasing forces to encourage sampling towards as yet unsampled regions of phase space.

1.9 Objectives

The primary goal of this thesis is to implement a computational method for calculating the free energy of ring puckering. This method is then applied to the glucopyranose and furanose sugar substrates, of a glycosidase and a phosphorylase. Suitable methods to calculate free energies with respect to a chosen coordinate of interest are those based on the adaptive biasing potential schemes such as metadynamics, adaptive biasing forces (ABF) and Free Energies from Adaptive Reaction Coordinate Forces (FEARCF). In these methods, which are discussed further in Chapter 3, the free energy estimate in the chosen coordinate phase space is iteratively improved upon as phase space is sampled by applying forces to a system. The FEARCF method which applies biasing force to the chosen reaction coordinate and iteratively improves upon the free energy estimate, was chosen to calculate the free energies in this thesis.

A description of the puckering phase space available to N-membered sugar rings such as furanose and pyranose can be defined in terms of a reduced coordinate set. The IUPAC conformers are often described in terms of the Cremer and Pople reduced coordinate set, but here the Hill-Reilly triangular decomposition scheme was chosen. This scheme, detailed in Chapter 3, reduces the number of variables needed to describe the ring pucker from N to N-3 and the dimensionality of the free energy problem. The calculation and analysis of lesser dimensionality free energy (hyper)surfaces is simpler. A free energy code for calculating ring puckering of 5- and 6-membered in terms of the triangular decomposition scheme needed to be composed. In the context of the FEARCF code, it was decided to include a new pucker parameter where the pucker variables were those calculated from the triangular decomposition scheme and the puckering forces were derived in terms of the FEARCF approach.

Due to the stereoelectronic preferences shown by saccharides, a method which explicitly includes the electronic character is necessary for the study of their free energy surface. This method should also be appropriate for further studies where saccharides are placed in the active site of enzymes. Purely empirical treatments are not appropriate while high level *ab initio* methods are not yet accessible, especially because the phase space should be well sampled. Semi empirical methods seem appropriate, although these have been used in carbohydrate studies, it is not clear which is best at describing ring pucker and which is most appropriate for use in more complex systems. Using the free energy of puckering as a metric a number of semi empirical methods are compared for a prototypical furanose (ribose) and pyranose (glucose) in Chapter 4. With a method that is capable of describing the puckering behavior of furanose and pyranose, the effect of the enzyme environment on the puckering phase space can be considered.

1.10 Overview

This thesis is laid out such that the earlier chapters discuss theory and code development while the latter discuss the chemical applications thereof. A brief introduction to computational chemistry is presented in Chapter 2. Free energy methods and innovations in this thesis are presented in Chapter 3. The goal of Chapter 3 is to introduce the reader to the FEARCF method, from which puckering free energies are calculated, and to explain how the pucker free energy is calculated. The question of which semi-empirical method is appropriate for modelling 5- and 6-membered carbohydrates, especially puckering conformations, is the focus of Chapter 4. This is done by considering the ring puckering free energies of ribose and glucose, and an attempt is made to validate the AM1, PM3, PM3CARB-1 and SCC-DFTB semi-empirical methods.

The thesis outlook, Chapter 5, discusses key elements of the applications chapters that ensue. The application of the QM/MM FEARCF simulations to a 6-membered ring in a glycosidase is explained in Chapter 6. The glycosidase and its pyranose substrate are discussed, but specifically the role the enzyme plays in stabilizing the ring pucker of the react according the puckering free energy calculations. The free energy surface for the glycosylation reaction is calculated. Chapter 7 deals with the puckering of the 5-membered ring nucleoside substrate of PNP, guanosine in PNP. The impact of PNP on the free energy of puckering conformational space of guanosine is compared to that *in vacuo* and water.

1.11 References

1. L. Krishnamoorthy, J. W. Bess, A. B. Preston, K. Nagashima and L. K. Mahal, *Nature Chemical Biology*, 2009, **5**, 244-250.
2. J. M. Berg, J. L. Tymoczko and L. Stryer, *Biochemistry*, W. H. Freeman and Company, New York, 2002.
3. L. Stryer, *Biochemistry*, W.H. Freeman and Company, New York, 1988.
4. N. N. Danial and S. J. Korsmeyer, *Cell*, 2004, **116**, 205-219.
5. R. Wolfenden, X. Lu and G. Young, *Journal of the American Chemical Society*, 1998, **120**, 6814-6815.
6. A. J. Kirby, *The anomeric effect and related stereoelectronic effects at oxygen*, Springer-Verlag, Heidelberg, 1983.
7. D. A. Evans, Harvard - CEM206 notes, 2002.
8. T. K. Lindhorst, *Essentials of Carbohydrate Chemistry and Biochemistry*, Wiley, 2003.
9. G. R. J. Thatcher, *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.
10. S. Wolfe, *Accounts of Chemical Research*, 1972, **5**, 102-111.
11. N. C. Craig, A. Chen, K. H. Suh, S. Klee, G. C. Mellau, B. P. Winnewisser and M. Winnewisser, *Journal of the American Chemical Society*, 1997, **119**, 4789-4790.
12. O. Guvench and A. D. MacKerell, Jr., *Journal of Physical Chemistry A*, 2006, **110**, 9934-9939.
13. P. I. Nagy, W. J. I. Dunn, G. Alagona and C. Ghio, *Journal of the American Chemical Society*, 1991, **113**, 6719-6729.
14. C. B. Barnett and K. J. Naidoo, *Journal of Physical Chemistry B*, 2008, **112**, 15450-15459.
15. E. A. Larsson, M. Staaf, P. Saderman, C. Haag and G. Widmalm, *The Journal of Physical Chemistry A*, 2004, **108**, 3932-3937.
16. A. J. Kirby, *Stereoelectronic Effects*, Oxford University Press, 1996.
17. E. Juaristi, ed., *Conformational Behaviour of Six-Membered Rings : Analysis, Dynamics and Stereoelectronic effects*, VCH Publishers Inc., 1995.
18. R. S. Tipson, ed., *Advances in Carbohydrate Chemistry and Biochemistry*, Academic Press Inc., New York, 1971.
19. W. N. Haworth, *The constitution of sugars*, Arnold, London, 1929.
20. O. Hassel and B. Ottar, *Acta Chemica Scandinavica*, 1947, **1**, 929.
21. R. E. Reeves, *Journal of the American Chemical Society*, 1949, **71**, 215-217.
22. R. E. Reeves, *Journal of the American Chemical Society*, 1951, **73**, 957-959.
23. R. E. Reeves and F. A. Blouin, *Journal of the American Chemical Society*, 1957, **79**, 2261-2264.
24. P. L. Durette, D. Horton and N. S. Bhacca, *Carbohydrate Research*, 1969, **10**, 565-577.
25. D. Horton and N. S. Bhacca, *Journal of the American Chemical Society*, 1967, **89**, 5993.
26. K. W. F. Kohlrausch, A. W. Reitz and W. Z. Z. Stockmair, *Journal of Physical Chemistry*, 1936, **B32**, 229.
27. K. W. F. Kohlrausch and A. Z. Wittek, *Journal of Physical Chemistry*, 1941, **B48**, 177.
28. R. E. Reeves, *Journal of the American Chemical Society*, 1949, **71**, 1737-1739.
29. R. Bentley, *Journal of the American Chemical Society*, 1959, **81**, 1952-1956.
30. J. F. Stoddardt, *Stereochemistry of Carbohydrates*, Wiley-Interscience, Toronto, 1971.
31. G. J. Davies, V. M. A. Ducros, A. Varrot and D. L. Zechel, *Biochemical Society Transactions*, 2003, **31**, 523-527.

32. B. L. Cantarel, P. M. Coutinho, C. Rancurel, T. Bernard, V. Lombard and B. Henrissat, *Nucleic Acids Research*, 2009, **37**, D233-238.
33. S. Fushinobu, B. Mertz, A. D. Hill, M. Hidaka, M. Kitaoka and P. J. Reilly, *Carbohydrate Research*, 2008, **343**, 1023-1033.
34. F. A. Scannapieco, G. Torres and M. J. Levine, *Critical Reviews in Oral Biology & Medicine*, 1993, **4**, 301-307.
35. G. J. McDougall and S. C. Fry, *Plant physiology*, 1990, **93**, 1042-1048.
36. D. S. Feingold, J. N. Goldman and H. M. Kuritz, *The Journal of Bacteriology*, 1968, **96**, 2118-2126.
37. B. Henrissat, *Biochemistry Journal*, 1991, **280 (Pt 2)**, 309-316.
38. B. Henrissat and A. Romeu, *Biochemistry Journal*, 1995, **311 (Pt 1)**, 350-351.
39. B. Henrissat, T. T. Teeri and R. A. J. Warren, *FEBS letters*, 1998, **425**, 352-354.
40. N. Ramasubbu, V. Paloth, Y. Luo, G. D. Brayer and M. J. Levine, *Acta Crystallographica D Biological Crystallography*, 1996, **52**, 435-446.
41. C. Divne, J. Ståhlberg, T. T. Teeri and T. A. Jones, *Journal of Molecular Biology*, 1998, **275**, 309-325.
42. P. J. Berti and J. A. B. McCann, *Chemical Reviews*, 2006, **106**, 506-555.
43. M. D. Erion, K. Takabayashi, H. B. Smith, J. Kessi, J. Wagner, S. Hönger, S. L. Shames and S. E. Ealick, *Biochemistry*, 1997, **36**, 11725-11734.
44. S. S. C. Chu and G. A. Jeffrey, *Acta Crystallographica*, 1968, **B24**, 830.
45. T. R. R. McDonald and C. A. Beevers, *Acta Crystallographica*, 1952, **5**, 654.
46. W. G. Ferrier, *Acta Crystallographica*, 1960, **13**, 678.
47. W. G. Ferrier, *Acta Crystallographica*, 1963, **16**, 1023.
48. G. J. Quigley, A. Sarko and R. H. Marchessault, *Journal of the American Chemical Society*, 1970, **5**, 834.
49. G. M. Brown and H. A. Lev, *Science*, 1965, **147**, 1038.
50. M. Sundaralingam, *Biopolymers*, 1969, **7**, 821.
51. M. Sundaralingam, *Journal of the American Chemical Society*, 1965, **587**, 599.
52. M. Spencer, *Acta Crystallographica*, 1959, **12**, 59.
53. D. J. Patel and C. Shen, *Proceedings of the National Academy of Science*, 1978, **75**, 2553-2557.
54. J. Boisbouvier, B. Brutscher, A. Pardi, D. Marion and J.-P. Simorre, *Journal of the American Chemical Society*, 2000, **122**, 6779-6780.
55. P. L. Durette and D. Horton, *Carbohydrate Research*, 1971, **18**, 403-418.
56. P. L. Durette and D. Horton, *Carbohydrate Research*, 1971, **18**, 289-301.
57. P. L. Durette and D. Horton, *Carbohydrate Research*, 1971, **18**, 57-80.
58. P. L. Durette and D. Horton, *Carbohydrate Research*, 1971, **18**, 389-401.
59. P. L. Durette and D. Horton, *Carbohydrate Research*, 1971, **18**, 419-425.
60. R. U. Lemieux and R. Nagarajan, *Canadian Journal of Chemistry*, 1964, **42**, 1270.
61. Q. Xu and C. A. Bush, *Biochemistry*, 1996, **35**, 14521-14529.
62. J. Raap, J. H. Van Boom, H. C. Van Lieshout and C. A. G. Haasnoot, *Journal of the American Chemical Society*, 1988, **110**, 2736-2743.
63. V. M. A. Ducros, D. L. Zechel, G. N. Murshudov, H. J. Gilbert, L. Szabó, D. Stoll, S. G. Withers and G. J. Davies, *Angewandte Chemie*, 2002, **114**, 2948-2951.
64. G. Sulzenbacher, H. Driguez, B. Henrissat, M. Schulein and G. J. Davies, *Biochemistry*, 1996, **35**, 15280-15287.
65. I. Tews, A. Perrakis, A. Oppenheim, Z. Dauter, K. S. Wilson and C. E. Vorgias, *Nature Structural Biology*, 1996, **3**, 638-648.

66. G. J. Davies, L. Mackenzie, A. Varrot, M. Dauter, A. M. Brzozowski, M. Schulein and S. G. Withers, *Biochemistry*, 1998, **37**, 11707-11713.
67. D. J. Vocadlo, G. J. Davies, R. Laine and S. G. Withers, *Nature*, 2001, **412**, 835-838.
68. C. S. Rye and S. G. Withers, *Current Opinion in Chemical Biology*, 2000, **4**, 573-580.
69. G. Sidhu, S. G. Withers, N. T. Nguyen, L. P. McIntosh, L. Ziser and G. D. Brayer, *Biochemistry*, 1999, **38**, 5346-5354.
70. E. Sabini, G. Sulzenbacher, M. Dauter, Z. Dauter, P. L. Jorgensen, M. Schulein, C. Dupont, G. J. Davies and K. S. Wilson, *Chemical Biology*, 1999, **6**, 483-492.
71. D. L. Zechel and S. G. Withers, *Accounts of Chemical Research*, 2000, **33**, 11-18.
72. A. White, D. Tull, K. Johns, S. G. Withers and D. R. Rose, *Nature Structural Biology*, 1996, **3**, 149-154.
73. S. J. Williams, V. Notenboom, J. Wicki, D. R. Rose and S. G. Withers, *Journal of the American Chemical Society*, 2000, **122**, 4229-4230.
74. M. Roberge, C. Dupont, R. Morosoli, F. Shareck and D. Kluepfel, *Protein Engineering*, 1997, **10**, 399-403.
75. J. K. Lee, A. D. Bain and P. J. Berti, *Journal of the American Chemical Society*, 2004, **126**, 3769-3776.
76. D. A. Singleton and A. A. Thomas, *Journal of the American Chemical Society*, 1995, **117**, 9357.
77. D. E. Sunko, I. Szele and W. J. Hehre, *Journal of the American Chemical Society*, 1977, **99**, 5000-5005.
78. M. Ashwell, X. Guo and M. L. Sinnott, *Journal of the American Chemical Society*, 1992, **114**, 10158-10166.
79. B. A. Horenstein, D. W. Parkin, B. Estupinan and V. L. Schramm, *Biochemistry*, 1991, **30**, 10788-10795.
80. R. M. Werner and J. T. Stivers, *Biochemistry*, 2000, **39**, 14054-14064.
81. T. C. Appleby, I. I. Mathews, M. Porcelli, G. Cacciapuoti and S. E. Ealick, *Journal of Biological Chemistry*, 2001, **276**, 39232-39242.
82. F. Canduri, R. G. Silva, D. M. dos Santos, M. S. Palma, L. A. Basso, D. S. Santos and W. F. de Azevedo, Jr., *Acta Crystallographica D Biological Crystallography*, 2005, **61**, 856-862.
83. M. S. Castilho, M. P. Postigo, H. M. Pereira, G. Oliva and A. D. Andricopulo, *Bioorganic Medicinal Chemistry*, **18**, 1421-1427.
84. E. M. Bennett, C. Li, P. W. Allan, W. B. Parker and S. E. Ealick, *Journal of Biological Chemistry*, 2003, **278**, 47110-47118.
85. C. Mao, W. J. Cook, M. Zhou, A. A. Federov, S. C. Almo and S. E. Ealick, *Biochemistry*, 1998, **37**, 7135-7146.
86. J. E. Lee, K. A. Cornell, M. K. Riscoe and P. L. Howell, *Structure*, 2001, **9**, 941-953.
87. F. Canduri, D. M. dos Santos, R. G. Silva, M. A. Mendes, L. A. Basso, M. S. Palma, W. F. de Azevedo and D. S. Santos, *Biochemical and Biophysical Research Communincations*, 2004, **313**, 907-914.
88. M. Sundaralingam, T. P. Haromy and J. Raleigh, *Biochemistry*, 1980, **19**, 1718-1722.
89. H. M. Pereira, R. C. Garratt and G. Oliva, *To be published 10.2210/pdb3fb1/pdb*.
90. W. Shi, A. Lewandowicz, P. C. Tyler, R. H. Furneaux, S. C. Almo and V. L. Schramm, *To be Published (10.2210/pdb1rt9/pdb)*.
91. W. Shi, V. L. Schramm and S. C. Almo, *Journal of Biological Chemistry*, 1999, **274**, 21114-21120.
92. A. Fedorov, W. Shi, G. Kicska, E. Fedorov, P. C. Tyler, R. H. Furneaux, J. C. Hanson, G. J. Gainsford, J. Z. Larese, V. L. Schramm and S. C. Almo, *Biochemistry*, 2001, **40**, 853-860.

93. W. Shi, C. M. Li, P. C. Tyler, R. H. Furneaux, S. M. Cahill, M. E. Girvin, C. Grubmeyer, V. L. Schramm and S. C. Almo, *Biochemistry*, 1999, **38**, 9872-9880.
94. E. A. Taylor, K. Clinch, P. M. Kelly, L. Li, G. B. Evans, P. C. Tyler and V. L. Schramm, *Journal of the American Chemical Society*, 2007, **129**, 6984-6985.
95. M. Harris, in *ACM SIGGRAPH 2005 Courses*, ACM, Los Angeles, California, 2005.
96. M. L. Klein and W. Shinoda, *Science*, 2008, **321**, 798-800.
97. X. Li, S. Hassan and E. Mehler, *Proteins*, 2005, **60**, 464-484.
98. P. D. Blood and G. A. Voth, *Proceedings of the National Academy of Sciences*, 2006, **103**, 15068-15072.
99. L. Hemmingsen, D. E. Madsen, A. L. Esbensen, L. Olsen and S. B. Engelsen, *Carbohydrate Research*, 2004, **339**, 937-948.
100. R. O. Dror, M. A. Jensen, D. W. Borhani and D. E. Shaw, *The Journal of General Physiology*, **135**, 555-562.
101. Y. Duan and P. A. Kollman, *Science*, 1998, **282**, 740-744.
102. U. Mayor, C. M. Johnson, V. Daggett and A. R. Fersht, *Proceedings of the National Academy of Sciences of the United States of America*, 2000, **97**, 13518-13522.
103. K. Byun, Y. Mo and J. Gao, *Journal of the American Chemical Society*, 2001, **123**, 3974-3979.
104. L. S. Devi-Kesavan and J. Gao, *Journal of the American Chemical Society*, 2003, **125**, 1532-1540.
105. J. E. Gready, I. Rostov and P. L. Cummins, *Modelling Molecular Structure and Reactivity in Biological Systems*, 2006, 101-118.
106. L. Petersen, A. Ardevol, C. Rovira and P. J. Reilly, *Journal of the American Chemical Society*, 2010, **132**, 8291-8300.
107. T. Rungrotmongkol, A. J. Mulholland and S. Hannongbua, *Journal of Molecular Graphics & Modelling*, 2007, **26**, 1-13.
108. H. L. Woodcock, M. Hodošcek and B. R. Brooks, *Journal of Physical Chemistry A*, 2007, **111**.
109. C. J. Cramer, *Essentials of Computational Chemistry : Theories and Models* John Wiley and Sons, Chichester, 2004.
110. D. Sornette, A. B. Davis, K. Ide, K. R. Vixie, V. Pisarenko and J. R. Kamm, *Proceedings of the National Academy of Sciences*, 2007, **104**, 6562-6567.
111. R. Car and M. Parrinello, *Physical Review Letters*, 1985, **55**, 2471.
112. L. Petersen, A. Ardevol, C. Rovira and P. J. Reilly, *Journal of Physical Chemistry B*, 2009, **113**, 7331-7339.

2 Computer Simulations

In chemistry, a computer simulation broadly refers to the use of a combination of theoretical and mathematic models which have been written (coded) in a computer language and executed on a computer. This is done in order to explore the energetics, conformations, configurations, dynamics and reactivity of molecular systems. Such an approach complements the experimental viewpoint of molecular systems, as a greater molecular understanding of current systems can be gained. Intricacies and anomalies of molecular systems can be contemplated, while as yet undiscovered molecules can be virtually created and screened for use as, for example, drug molecules. Visualisation is another important part of this field. For chemists the ability to interact with molecules on screen, tumble them into interesting viewpoints, change from atomistic detail to views of molecular surfaces or secondary structure and even produce three dimensional images, greatly enhances the ability to understand complex molecules.

Computational chemistry is no longer a new field. It has several journals dedicated to it and computational tools are valuable in day-to-day X-ray crystal structure, drug molecule and materials modelling. X-ray crystallographers and chemists using structures from the Protein Data Bank (PDB) use modelling programs to place hydrogen atoms in molecular structures (hydrogen atoms do not contain enough electron density to scatter X-rays). Quantum chemists routinely probe the electronic character of molecules using high-level quantum theories based on the Schrödinger equation. The dynamic properties / observables of molecular systems can be monitored / calculated using molecular dynamics or Monte Carlo methods which are rooted in statistical mechanics theory.

An investigation of the dynamic structures and free energies of molecular processes (e.g. bond formation / cleavage and ring puckering) is currently inaccessible to experiment. Computer simulation of chemical systems, particularly QM/MM dynamics and free energy methods, are useful in studying these phenomena. There are a large number of tools available to the computational scientist and this chapter discusses certain of them. To supplement what is covered here, the following texts may be useful: Encyclopaedia of Computational Chemistry[1], Leach[2], Jensen[3], Lewars[4]; for quantum mechanics theory refer to Ostlund[5] and Levine[6], and for Statistical Mechanics refer to Chandler[7] and McQuarrie.[8]

2.1 Molecular Energies

The energies of molecules as a function of their structure is calculated using molecular mechanics (MM), quantum mechanics (QM) or a hybrid of these methods, namely quantum mechanics/molecular mechanics (QM/MM). MM or force field methods are based on a Newtonian ball and spring approach and the application of Hooke's law where the parameters used in this method are often derived from experimental spectra, such as infrared spectroscopy. QM methods solve the Schrödinger equation. Where *ab initio* methods attempt to do this with as little approximations as possible, semi empirical methods neglect certain electron orbital overlap terms and approximate certain terms with experimentally or *ab initio* determined parameters. QM/MM methods partition the system into QM and MM regions combining the adhoc accuracy of QM with the speed and relative accuracy of MM.

As there are many degrees of freedom in a molecular system, the potential energy surface is complex. Stationary points on the potential energy surface can be calculated by taking derivatives with respect to the potential energy function. Techniques to minimise molecular structure include steepest descent, where the gradient of the potential energy surface is followed downhill and more complex second derivative methods, such as adapted Newton Raphson, where both the gradient and Hessian the second derivative of the energy function are used to find stationary states.

2.1.1 Molecular mechanics (MM)

Empirical force fields are fitted or parameterised so as to produce acceptable behaviour for model systems. The system is treated using the laws of classical physics. MM is often referred to as the ball and spring model because the bonds between atoms are modelled with Hooke's law. The potential energy of the molecule is a function of the nuclear coordinates of each atom where the Born-Oppenheimer approximation is applied.[2] In general, electrons are not treated explicitly[9] and the MM approach cannot model electronic processes such as bond breaking[2] (although reactive force fields have been developed[10]).The total potential energy of a system is the sum of bonded and non-bonded interactions as shown in Equation 2-1.[2]

Equation 2-1

$$U = U_{bonds} + U_{angles} + U_{torsional\ angles} + U_{improper\ \ torsional\ angles} + U_{vanderWaals} + U_{electrostatic}$$

The bonded interactions are bond stretching, angle bending and torsional angle rotations. Improper torsional angle interactions (the angle used to keep four atoms in a plane) also contribute to the overall potential. Non-bonded interactions include electrostatic and van der Waals interactions and are shown in Figure 2-1.

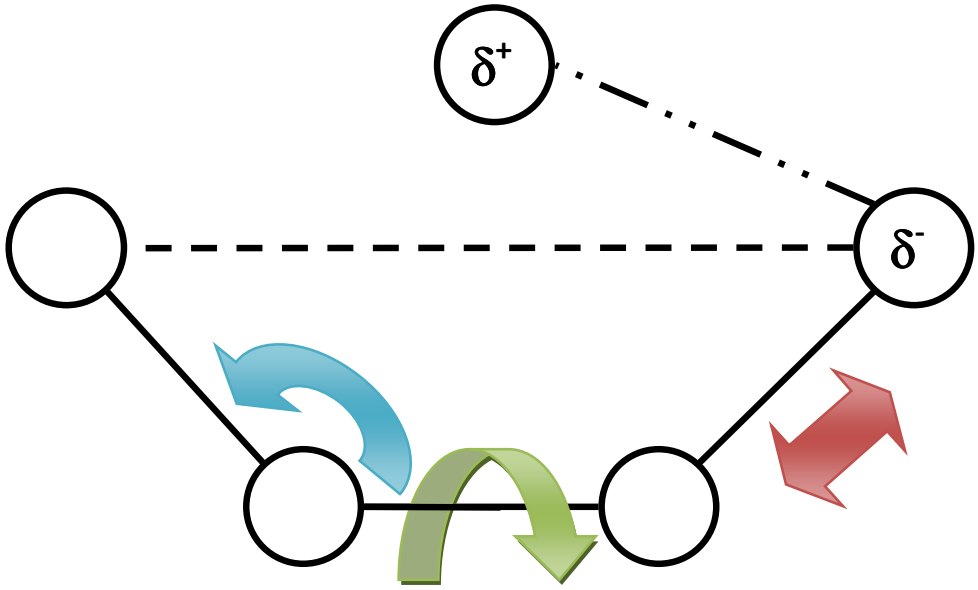

Figure 2-1 Representations of a fictional molecule where atoms are represented as circles, covalent bonds by lines and non-bonding interactions - van der Waals and electrostatic interactions by a dashed and a dashed stipple line respectively. The red arrow indicates bond stretching interaction, the blue arrow angle bending interactions and the green arrow torsional angle rotation (adapted [2,4]).

The energy function implemented in the CHARMM[11] program describes bonded terms as Hooke's law type functions (a quadratic). The non-bonded van der Waals term is described by a Lennard-Jones 6-12 potential[2] while the non-bonded electrostatic term is described by the standard coulomb charge expression. Pairwise interactions are taken into account and higher-order (many-body) interactions are parameterised into the pair potential, thus the term effective pair potential is used. More sophisticated force fields exist, such as the polarisable force fields, and the reader is directed to Leach[2] for further information on these.

Equation 2-2

$$U(R^N) = \sum_{bonds} K_b(b-b_0)^2 + \sum_{angles} K_\theta(\theta-\theta_0)^2 + \sum_{dihedrals} K_\chi[1+\cos(n\chi-\gamma)] +$$
$$\sum_{impropers} K_{imp}(\varphi-\varphi_0)^2 + \sum_{\substack{non-bonded \\ pairs(i\neq j)}} \left\{ \varepsilon_{ij}^{\min} \left[\left(\frac{R_{ij}^{\min}}{r_{ij}} \right)^{12} - 2\left(\frac{R_{ij}^{\min}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon r_{ij}} \right\}$$

Equation 2-2 defines the potential, U, for the set of all molecular coordinates R^N. The first two terms model the deformations from the equilibrium bond length b_0 and angle θ_0, and similarly for the improper dihedral angle φ_0, where the K's are the respective force constants. 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 $K\chi$ the dihedral force constant.[11] The final term, the non-bonded term, includes van der Waals interactions modelled by the Lennard-Jones potential and electrostatic interactions modelled by the Coulomb potential.[2] $R_{ij}^{\min}$ 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 ε is the relative dielectric constant and ε_0 is the permittivity of vacuum.

Certain of the many MM packages available, such as AMBER[12], CHARMM[11, 13] and GROMOS[14], have their own generally applicable force field. Force fields are often parameterised for a certain subset of molecules. In this case, force fields appropriate for carbohydrates, proteins and water were chosen.

2.1.1.1 Force fields for carbohydrates, proteins and water

Carbohydrate, protein and generalised force fields include the AMBER[15-18], CHARMM[19-21] and GROMOS[22-24] force fields. These force fields reproduce molecular behavior well enough, although they can be poor for molecules which deviate from the "training set". CHARMM is packaged with a generalised protein force field which was utilised here. The carbohydrate solution force field (CSFF) recently developed[25] to improve upon the CHARMM carbohydrate force field[19, 20] was implemented in this thesis for MM sugar residues.

Force field models for water range from simple models such as TIP3P[26, 27] and SPC/E[28], to polarisable and *ab initio* water models.[2] Water models which explicitly model atomistic detail are termed explicit

water models. Common features in these models include rigid water geometries and 3 interaction sites (Figure 2-2). The TIP3P[26, 27] water (available in CHARMM[13]) used in this study has 3 electrostatic interaction sites. A partial negative charge on the oxygen (q (O)) is balanced by the two positive partial 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.[2]Although these models have varying parameters, for example, charge (q), bond length (r) and HOH angle (as shown in Table 2-2), they are known to represent molecular properties sufficiently accurately for most simulations.[2] The TIP3P model is acceptable for these simulations and the CHARMM protein force field has been parameterised for use with this water model.

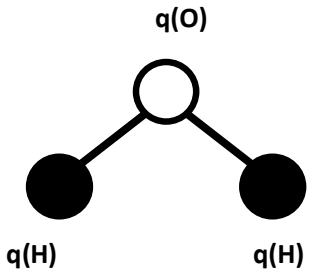

Figure 2-2 The 3 site water model used for SPC, SPC/E and TIP3P water models. (adapted[2, 27])

Table 2-1 A comparison of the parameters for certain water models (adapted[2, 27])

Coordinates	SPC	SPC/E	TIP3P
r(OH) [Å]	1.0	1.0	0.9572
HOH [°]	109.47	109.47	104.52
q(O) [*e*]	-0.82	-0.8472	-0.834
q(H) [*e*]	0.41	0.4238	0.417

2.1.2 Quantum mechanics (QM)

Quantum mechanics but particularly *ab initio* methods[29] are considered the most accurate for describing chemical systems. Electrons are included explicitly when solving the Schrödinger equation, which can be done using several commonly available electronic structure programs such as Gaussian[30] and GAMESS-UK.[31] Solutions to the Schrödinger equation depends on several postulates.[9] The time independent Schrödinger equation $\hat{H}\Psi = E\Psi$ is

Equation 2-3

$$\left\{\frac{-\hbar^2}{2m}\nabla^2+\frac{1}{4\pi\varepsilon_0}\left(-\sum_{i}^{electrons}\sum_{I}^{nuclei}\left(\frac{Z_I e^2}{|R_I-r_i|}\right)+\sum_{i}^{electrons}\sum_{j<i}^{electrons}\left(\frac{e^2}{|r_i-r_j|}\right)+\sum_{I}^{nuclei}\sum_{J<I}^{nuclei}\left(\frac{Z_I Z_J e^2}{|R_I-R_j|}\right)\right)\right\}\Psi(\vec{r})=E\Psi(\vec{r})$$

where *m* is the mass, *e* is the electronic charge and Z the atomic number. The summation terms represent the Coulomb interactions of electron-nuclei attraction, electron-electron repulsion and nuclear-nuclear repulsion between nuclei *I, J* and electrons *i, j*.

2.1.2.1 Hartree-Fock theory

Solutions of the Schrödinger equation for ground-state molecules with unpaired electrons (closed-shell or restricted) can be obtained iteratively using Hartree-Fock theory with a single Slater determinant. This model neglects nuclear kinetics (Born-Oppenheimer approximation[5]) and incorporates a linear combinations of atomic orbitals (LCAO), where K one-electron orbitals ϕ_ν are linearly combined to produce spin orbitals ψ_I with coefficients $c_{\nu i}$ as follows:

Equation 2-4

$$\psi_i=\sum_{\nu=1}^{K}c_{\nu i}\phi_\nu$$

The Slater determinant includes the Hartree product and Pauli's exclusion principle. The solutions to this depend on the Roothaan-Hall equations.[2, 5] The Roothaan-Hall equations are

Equation 2-5

$$\sum_{\nu=1}^{N}(F_{\mu\nu}-\varepsilon_i S_{\mu\nu})c_{\nu i}=0 \qquad \mu=1,2,\ldots,K$$

with the normalisation condition

Equation 2-6

$$\sum_{\mu=1}^{N}\sum_{v=1}^{N} c_{\mu i}^{*} S_{\mu v} c_{v i} = 1 .$$

The one-electron energy of a molecular orbital ψ_i is ε_i and N is the number of basis functions. The overlap matrix, $S_{\mu v}$, is an N by N matrix with elements

Equation 2-7

$$S_{\mu v} = \int \phi_{\mu}^{*}(1)\phi_{v}(1) dx_1 dy_1 dz_1 .$$

The Fock matrix is an N by N matrix with elements $F_{\mu v}$ such that

Equation 2-8

$$F_{\mu v} = H_{\mu v}^{core} + \sum_{\lambda=1}^{N}\sum_{\sigma=1}^{N} P_{\lambda\sigma}[(\mu v \mid \lambda\sigma) - \frac{1}{2}(\mu\lambda \mid v\sigma)].$$

The Fock matrix represents the average field effect of all the electrons on each orbital.[32] The $H_{\mu v}^{core}$ term of the Fock matrix represents the energy of a single electron in a field of "bare" nuclei[33] , where the summation is carried out for all atoms (M, the number of atomic nuclei) and Z_A is the atomic number for an atom A.

Equation 2-9

$$H_{\mu v}^{core} = \int \phi_{\mu}^{*}(1)\hat{H}^{core}(1)\phi_{v}(1) dx_1 dy_1 dz_1 ,$$
$$\hat{H}^{core}(1) = -\frac{1}{2}\nabla^2 - \sum_{A=1}^{M}\frac{Z_A}{r_{1A}}$$

P is the charge density matrix, for the restricted Hartree-Fock solution defined as a summation over the occupied orbitals, whose elements are defined as

Equation 2-10

$$P_{\lambda\sigma} = 2\sum_{i=1}^{N/2} c_{\lambda i} c_{\sigma i} ; P_{\mu\nu} = 2\sum_{i=1}^{N/2} c_{\mu i} c_{\nu i}$$

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

Equation 2-11

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

and similarly for (μλ|νσ).

The matrix form for a ground state or closed-shell system (no unpaired electrons) is derived from the Roothaan-Hall equations

Equation 2-12

$$FC = SC\varepsilon$$

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-electron 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 and the matrix equation must be solved iteratively. This procedure is known as self-consistent field (SCF) [32] and it is termed self-consistent because upon iteratively converging to the minimum energy the orbitals generate a field which will produce the same orbitals. The solutions are also variational. For a trial wavefunction, Ξ, the expectation value of an energy, E', is always greater than or equal to the "true" energy, E

Equation 2-13

$$E' = \frac{\int \Xi^* \hat{H} \Xi d\tau}{\int \Xi^* \Xi d\tau} \geq E .$$

2.1.2.2 Density functional theory

Hartree-Fock theory describes electrons as moving in an average field (potential) of other electrons[2], but electronic motions are more correlated with each other than this model suggests. Although post Hartree-Fock methods[32] are able to include correlation, a Density Functional Method (DFT) is applied in this thesis. DFT methods solve for the N-electron wavefunction as a function of the overall electron density. Kohn and Hohenberg proved that there exists a unique connection between the ground state energy and the ground state electron density function[32, 34] by means of a functional – a function whose definition is itself a function – and that this is variational in nature.[35] Specifically

Equation 2-14

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

where the energy E depends on a function of the electron density $F[\rho(r)]$, which is a function of the nuclear and electronic coordinates *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,

Equation 2-15

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

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.[2]

The matrix solution to these equations takes the form of the Kohn-Sham equations[2, 5],

Equation 2-16

$$H^{KS}C = SC\varepsilon$$

where the H^{KS} is the Kohn-Sham Hamiltonian. These equations are variational and self-consistent[3, 32, 36] where an approximate density function ρ_0 is chosen and iteratively improved upon.

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

Equation 2-17

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

Equation 2-18

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

where $E_{KE}[\rho(r)]$ describes a system of non-interacting electrons and $E_H[\rho(r)]$ is the Hartree electrostatic energy, that is 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 the correlation interactions are mixed-spin interactions (Equation 2-19).[32]

Equation 2-19

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

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. Examples of gradient-corrected functional are the Becke exchange; Lee, Yang, Parr (LYP) correlation; Perdew-Wang (PW) and Vosko, Wilk, Nusair (VWN) functional.[32, 37]

Becke's three parameter functional including LYP correlation (B3LYP)[38, 39] hybrid functionals linearly combine (or mix) the Hartree-Fock exchange with linear and gradient-corrected exchange terms and correlation terms.

Equation 2-20

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

The exchange part (the terms with E_X) 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 (the terms with E_C) 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.[3, 32]

2.1.2.3 Basis sets

The total electronic wavefunction is solved by expressing it as an approximate wavefunction. This is in turn defined as a linear combination of orbitals, which are defined by a set of functions - a basis set.[32] Basis sets are often classified into minimal, split-valence, polarised and diffuse basis sets.[32] Minimal basis sets 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. Minimal basis sets may not correctly describe infinitely separated molecules and this type of error, called basis set superposition error (BSSE), is only expected to occur with minimal basis sets.[2] Split-valence basis sets better represent the vector space available to electrons by increasing the number of basis functions for each valence orbital of an atom. Two or more sizes of basis function can exist for each valence orbital. Basis sets which include polarisation functions include orbitals with an angular momentum beyond that of the ground state. Basis sets which include diffuse functions include larger versions of s and p type function and are good for describing the orbital nature of molecules which have lone pairs or are anionic.

The Pople sets used in this thesis use the X-YZg specification where X is the number of primitive Gaussians that define each core atomic orbital. YZ designates the existence of two basis functions which describe the valence orbitals. The first being a linear combination of Y primitive Gaussians and the second a linear combination of Z primitive Gaussians. This particular example X-YZg is known as a split-valence double zeta basis set and triple (X-YZWg) and quadruple (X-YZWVg) also exist. The most important feature of this basis set is that the core and valence orbitals are described using a different number of Gaussians. The designation * indicates the addition of polarised orbitals. A single * or (d) includes d-orbitals on heavy atoms while ** or (d, p) augments the d-orbital on heavy atoms with p-orbitals on Hydrogen atoms. '+' represents the addition of diffuse functions for heavy atoms, while '++' adds diffuse functions to heavy and light atoms. Diffuse functions on light atoms such as hydrogen seldom make a significant difference in accuracy.[32]

There have been many studies of carbohydrate conformations [40-44] with varying basis sets and theories applied. Barnett[44] used DFT B3LYP/6-31G* and B3LYP/6-31+G** in studies of the gauche effect.

While studies by Csonka[45, 46] and Momany[43] indicate B3LYP/6-31+G** or B3LYP/6-311+G** should be sufficiently high a level for carbohydrates when using Density Functional Theory.

2.1.2.4 Semi empirical methods

The computational cost associated with *ab initio* methods scales formally as the fourth power of the number of basis functions[3] and so as an alternative, semi empirical approaches are often used. Most of the cost, associated with *ab initio* calculations, involves the manipulation of the large number of two-electron integrals required to construct the Fock-matrix[3], and these may be neglected or approximated.[2] Although other methods exist, the Neglect of Diatomic Differential Overlap (NDDO) method only neglects differential overlap between atomic orbitals on different atoms.[2, 3, 33]

Such methods are relatively inexpensive and are qualitatively good at describing large systems where parameter sets exist[32], but may be limited in modelling molecular systems for which there are not similar structures in the parameter set. TS structures tend to fall in this category.[32] Semi empirical models discussed here, as they may be applied to carbohydrate systems, include Austin Model 1 (AM1), Parameterised Model 3 (PM3) and a reparameterised carbohydrate specific PM3 model, the PM3 Carbohydrate (PM3CARB-1) model. AM1 was developed by Dewar[2, 47] and it is applicable to a wide range of compounds including carbohydrates.[48, 49] PM3 was parameterized, for a variety of molecules, in an automated fashion by Stewart[2, 3] and disaccharide simulations have been carried out with this model.[50] The fallibilities of both AM1 and PM3 are well known.[3] The PM3CARB-1 model developed by McNamara[51] , a reparameterised version (see Table 2-2) of the PM3 model, makes several improvements. Specifically, it corrects the shortcoming of the PM3 model where upon minimization the $^{1}C_{4}$ chair of glucose lies lower in energy than the $^{4}C_{1}$ chair. "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.].[51]

Table 2-2 PM3 and optimised PM3CARB-1 parameters for oxygen and hydrogen adapted from McNamara.[51]

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

The modified NDDO methods are parameterised in terms of atomic variables and there are several parameters per atom. These are orbital exponents $\zeta_{s/p}$; one electron terms (the energy of an electron experiencing the full nuclear charge), $U_{s/p}$ and (atomic "resonance parameters") $\beta_{s/p}$, two electron terms, G_{ss}, G_{sp}, G_{pp}, G_{p2}, H_{sp}; parameters for the core-core repulsion, α; and fitting constants a_k, b_k, c_k.[3]

The Fock elements for the driving equations for modified NDDO methods such as AM1 are:

Equation 2-21

$$F_{\mu\mu} = H_{\mu\mu}^{core} + \sum_{\nu \subset A}[P_{\nu\nu}(\mu\mu|\nu\nu) - 0.5P_{\nu\nu}(\mu\nu|\mu\nu)] + \sum_{B \neq A}\sum_{\lambda \in B}\sum_{\sigma \in B} P_{\lambda\sigma}(\mu\mu|\lambda\sigma)$$

$$H_{\mu\mu}^{core} = U_{\mu\mu} - \sum_{B \neq A} V_{\mu\mu B}$$

Equation 2-22

$$F_{\mu\nu} = H_{\mu\nu}^{core} + 1.5P_{\mu\nu}(\mu\nu|\mu\nu) - 0.5P_{\mu\nu}(\mu\mu|\nu\nu) + \sum_{B \neq A}\sum_{\lambda \in B}\sum_{\sigma \in B} P_{\lambda\sigma}(\mu\nu|\lambda\sigma)$$

$$H_{\mu\nu}^{core} = -\sum_{B \neq A} V_{\mu\nu B}$$

where the orbital μ, ν are both contained on atom A or are

Equation 2-23

$$F_{\mu\nu} = H_{\mu\nu}^{core} - 0.5 \sum_{\lambda \in B} \sum_{\sigma \in A} P_{\lambda\sigma} (\mu\sigma | \nu\lambda)$$
$$H_{\mu\nu}^{core} = 0.5 S_{\mu\nu} (\beta_\mu + \beta_\nu) - \sum_{B \neq A} V_{\mu\nu B}$$

when μ is contained on atom A and ν is contained on atom B.

The two-centre, one-electron attractions, $V_{\mu\mu B}$ and $V_{\mu\nu B}$, between the electronic distributions on atom A and the core of atom B are:

Equation 2-24

$$V_{\mu\mu B} = -Z_B (\mu_A \mu_A | s_B s_B)$$
$$V_{\mu\nu B} = -Z_B (\mu_A \nu_A | s_B s_B) .$$

In this approach the core-core repulsions between atoms A and B are calculated but the core-core repulsions terms for O-H and N-H are treated separately (X=O, N)[2] :

Equation 2-25

$$E_{AB} = Z_A Z_B (s_A s_A | s_B s_B) \left\{ 1 + e^{-\alpha_A R_{AB}} + e^{-\alpha_B R_{AB}} \right\} +$$
$$\frac{Z_A Z_B}{R_{AB}} \left\{ \sum_i K_{A_i} e^{\left[-L_{A_i} (R_{AB} - M_{A_i})^2 \right]} + \sum_j K_{B_j} e^{\left[-L_{B_i} (R_{AB} - M_{B_i})^2 \right]} \right\}$$
$$E_{XH} = Z_X Z_H (s_X s_X | s_H s_H) \left\{ 1 + R_{XH} e^{-\alpha_X R_{XH} / R_{AB}} + e^{-\alpha_H R_{XH}} \right\}$$

The last term of E_{AB}, the sums of Gaussians, is used to make corrections to the core-core repulsion. The spherical Gaussian functions have widths defined by L. The parameters M and K are optimised for each atom. These methods generally use only s-p orbital bases, and in such a basis there are five types of one-centre two-electron integral that are approximated, these are

Equation 2-26

$$\langle ss|ss\rangle = G_{ss}$$
$$\langle sp|sp\rangle = G_{sp}$$
$$\langle ss|pp\rangle = H_{sp}$$
$$\langle pp|pp\rangle = G_{pp}$$
$$\langle pp'|pp'\rangle = G_{p2}$$

where the G parameters are Coulomb terms, with the G_{p2} integral are over different types of p-functions and the H parameter is the exchange integral.[3]

2.1.2.5 Self-consistent charge density functional tight-binding (SCC-DFTB)

SCC-DFTB is an approximate DFT method, which is an alternative to semi empirical methods such as AM1 and PM3. It is formally not a semi empirical method itself because its parameters are based solely on DFT calculations.[52, 53] The energies are calculated from a second order expansion of the DFT total energy[54] relative to reference density ρ_0 such that the total energy is

Equation 2-27

$$E = \sum_{i}^{occ} \langle \psi_i | \hat{H}_0 | \psi_i \rangle + 0.5 \int\int \left(\frac{1}{|r-r'|} + \left. \frac{\delta^2 E_{ex}}{\delta\rho\delta\rho'} \right|_{\rho_0} \right) \times \delta\rho(r)\delta\rho'(r')drdr' +$$
$$\left\{ -0.5 \int\int \frac{\rho'_0(r')\delta\rho_0(r)}{|r-r'|} drdr' + E_{XC}[\rho_0(r)] - \int V_{XC}(\rho_0)\rho_0(r)dr \right\} + E_{core}$$

where the charge density variation is $\delta\rho$, E_{core} is the core-core repulsion energy and the effective Kohn-Sham Hamiltonian $\hat{H}_0$, depends only on the reference density, ρ_0, as follows

Equation 2-28

$$\hat{H}_0 = -0.5\nabla^2 - \sum_k \frac{Z_k}{|R_k - r|} + \int \frac{\rho_0(r')}{|r-r'|} dr' + V_{XC}[\rho_0(r)] .$$

Further approximations are made to Equation 2-27 so that the SCC-DTFB energy[53] becomes,

Equation 2-29

$$E^{SCC-DFTB} = \sum_{i}^{OCC} \langle \psi_i | \hat{H}_0 | \psi_i \rangle + E_{rep} + 0.5 \sum_{\alpha\beta} \Delta q_\alpha \Delta q_\beta \gamma_{\alpha,\beta} \,.$$

E_{rep} approximates the last four terms of Equation 2-27 as the sum of a set of atom-atom pairwise potentials $U^{A\text{-}B}$ by

Equation 2-30

$$E_{rep} = \sum_{A,B} U^{A-B}(R_{AB}) \,.$$

The Hamiltonian matrix elements, $\langle \psi_i | \hat{H}_0 | \psi_i \rangle$, are represented in a minimal basis of localised pseudo-atomic Slater orbitals[52, 53], χ_μ

Equation 2-31

$$\psi_i = \sum_{\mu} c_\mu^i \chi_\mu \,.$$

Diagonal elements of the Hamiltonian matrix, $H_{\mu\nu}$, in this minimal basis are calculated using atomic eigenvalues and the non-diagonal elements are calculated with a two-centre approximation

Equation 2-32

$$H_{\mu\nu} = \langle \chi_\mu | \hat{T} + V_{eff}\left[\rho_0^A + \rho_0^B\right] | \chi_\nu \rangle, \mu \in A, \nu \in B$$

where V_{eff} is the effective Kohn-Sham potential, A and B are two atoms and ρ_0^A is the electron density on a neutral atom A. The Hamiltonian matrix elements, $H_{\mu\nu}$, and the overlap matrix elements, $S_{\mu\nu}$, are precomputed and tabulated for the interatomic distance between A and B.[53] During an SCC-DFTB calculation the values of these matrix elements are interpolated from the tabulated data. This approximation is one of the reasons that SCC-DFTB calculations are quite rapid (when compared to DFT or *ab initio)*.

The charge density variation, $\delta\rho$, relative to the reference density, ρ_0, is approximated as a superposition of neutral atomic density contributions, $\delta\rho^A$ by

Equation 2-33

$$\delta\rho = \sum_A \delta\rho^A \ .$$

The charge fluctuations on an atom A, $\delta\rho^A$, are approximated with

Equation 2-34

$$\Delta q^A = q^A - q_0^A$$

where q^A is the Mulliken electron population and q_0^A is the valence electrons on a neutral atom A.

The second term of Equation 2-27, which includes the second derivative of the total energy with respect to the charge density fluctuations, $\left.\frac{\delta^2 E_{ex}}{\delta\rho\delta\rho'}\right|_{\rho_0}$, is approximated by a function γ_{AB}. There are several functional forms for γ_{AB}. The CHARMM[11, 53] implementation of SCC-DFTB models γ_{AB} as the Coulombic interaction between two exponentially decaying spherical charge distributions centred on two nuclei, that is

Equation 2-35

$$\gamma_{AB} = \int\int \frac{s(\tau_A, R_A)s(\tau_B, R_B)}{|r-r'|} dr dr' =$$
$$\frac{1}{R_{AB}} - \left[e^{-\tau_A R_{AB}} K(\tau_A, \tau_B, R_{AB}) + e^{-\tau_B R_{AB}} K(\tau_B, \tau_A, R_{AB})\right]$$

, where $s(\tau_A, R_A)$ is a 1s Slater orbital centred on R_A with an exponent τ_A and $K(\tau_A, \tau_B, R_{AB})$ is

Equation 2-36

$$K(\tau_A, \tau_B, R_{AB}) = \frac{\tau_B^4 \tau_A}{2(\tau_A^2 - \tau_B^2)^2} - \frac{\tau_B^6 - 3\tau_B^4 \tau_A^2}{(\tau_A^2 - \tau_B^2)^3 R_{AB}}.$$

τ_A depends on the chemical hardness U_A for an atom A as

Equation 2-37

$$\tau_A = \frac{16}{5} U_A$$

A semi empirical dispersion term[52] can be added to improve pi-pi stacking interactions.[55] This term is simply a damped van der Waals-like function. The SCC-DFTB energy with dispersion (SCC-DFTB/d) is

Equation 2-38

$$E^{SCC-DFTB/d} = E^{SCC-DFTB} - \sum_{\alpha,\beta} f(R_{\alpha,\beta}) \frac{C_6^{\alpha,\beta}}{R_{\alpha,\beta}^6}$$

where $f(R_{\alpha,\beta})$ is a damping function that rapidly approaches zero and $C_6^{\alpha,\beta}$ are molecular polarisabilities derived from experimentally determined atomic polarisabilities. The function $f(R_{\alpha,\beta})$ is

Equation 2-39

$$f(R_{\alpha,\beta}) = \left[1 - e^{-d\left(R_{\alpha\beta} / R^0_{\alpha\beta} \right)^N} \right]^M$$

with d=3.0, N=7 and M=4 for all atoms. $R^0_{\alpha\beta}$ values are calculated from atomic R^0_α using a combination rule

Equation 2-40

$$R^0_{\alpha\beta} = \frac{\left(R^0_\alpha\right)^3 + \left(R^0_\beta\right)^3}{\left(R^0_\alpha\right)^2 + \left(R^0_\beta\right)^2}.$$

The R^0_α parameters 3.5 Å for H and 3.8 Å for C, N, O although these parameters are adjustable.

Hydrogen bonding energies are slightly underestimated in SCC-DFTB. These can be improved by the addition of a damped hydrogen bond term.[56] For further details of the SCC-DFTB implementation into CHARMM see Elstner.[53]

2.1.3 Hybrid computations - quantum mechanics / molecular mechanics (QM/MM)

When studying large complex systems, such as proteins, it is not computationally feasible to treat the entire system with QM methods. With the advent of GPU computing, Car-Parrinello dynamics[57, 58] and access to supercomputing facilities, studies of complex systems using pure QM methods is more feasible than before, however, QM/MM methods are still widely used. QM/MM is particularly useful for the study of a system where a reactive region undergoes electronic structural changes, while the rest of the system may require less rigour. Such a system can be partitioned into a QM region and an MM region which interact and the Hamiltonian is modified to reflect this[53, 59]:

Equation 2-41

$$\hat{H}_{effective} = \hat{H}^{\circ}_{QM} + \hat{H}_{MM} + \hat{H}^{Elec}_{QM/MM} + \hat{H}^{vdW}_{QM/MM} + \hat{H}^{bonded}_{QM/MM} + \hat{H}_{boundary}$$

$\hat{H}^{\circ}_{QM}$ is the Hamiltonian describing the quantum potential energy for the quantum part of the system while $\hat{H}_{MM}$ describes the MM potential. The interactions between the QM and MM parts of the system are separated into electrostatic $\hat{H}^{Elec}_{QM/MM}$ and van der Waals $\hat{H}^{vdW}_{QM/MM}$ terms, where the electrostatic terms arise from the electric field of the MM atoms and the vdW terms account for dispersion interactions and Pauli repulsions between the QM and MM regions. $\hat{H}_{boundary}$ includes boundary condition terms such as periodic boundary or stochastic boundary terms.

The QM region can be embedded into the MM system using mechanical, electrostatic or polarised embedding schemes; such that electrostatic interactions, $\hat{H}^{Elec}_{QM/MM}$ differ. In mechanical embedding, the electrostatic interactions are treated using the MM charge model (often atomic point charges). In this scheme, the *in vacuo* QM region couples classically to the MM region with point charges placed at the nuclear sites of the QM region. The major shortcoming of this scheme is that the QM region is not polarised by the charges in the MM region. This can be avoided by electrostatic embedding, where the

MM point charges are included as one-electron terms in the Hamiltonian and so the QM region is polarised by the MM region. The electrostatic QM/MM operator gains additional terms such that [60]

Equation 2-42

$$\hat{H}_{QM/MM}^{Elec} = -\sum_{i}^{N}\sum_{B\in MM}^{L}\frac{q_B}{|r_i - R_B|} + \sum_{A\in QM}^{M}\sum_{B\in MM}^{L}\frac{q_J Z_A}{|R_A - R_B|}$$

where there are N electrons, L point charges and M QM nuclei indexed by A and B. The MM point charges, q_B, are at positions defined by the position vectors R_B, while Z_A are the nuclear charges of the QM atoms located at R_A and r_i are the electron positions.[53, 60] The final type of embedding, polarised embedding, requires a polarisable MM force field such that there can be mutual polarisation between the QM and MM regions.

Where there are no covalent bonds between the QM and MM regions, the $\hat{H}_{QM/MM}^{bonded}$ term is not necessary but in proteins the active site amino acids should be modelled quantum mechanically, as they may polarise and induce conformational changes in the substrate, and so the $\hat{H}_{QM/MM}^{bonded}$ term accounts for the partitioning of the QM and MM regions across a chemical bond. Although several methods[11] exist, only the General Hybrid Orbital (GHO) and Single Link atom (SLA) methods are discussed here.

In the link atom approach, the valency of the QM region is usually satiated by a hydrogen atom. This atom is made to exist co-linearly with the covalent bond between the QM and MM regions. It has been found that calculations may not converge unless the partial charges on the MM atoms covelently closest to the link-atom are removed.[11] A particular criticism of this approach is that on inclusion of the link atom there are additional degrees of freedom to a system.[61] The GHO method[62], unlike the link atom approach, does not add extra degrees of freedom to the system. It is an extension of the frozen localised orbital approach, partitioning the system at an sp^3 atom. The boundary atom is included in both the QM and MM parts of the calculation. It retains its classical character and also includes a hybrid orbital (which is fully optimised by the self-consistent QM calculation) and three auxiliary orbitals (as expected for an sp^3 hybridised atom, which do not partake in the QM optimization but provide an effective electric field for interactions).[61] The GHO method was used to join the QM and MM regions (Figure 2-3). This method originally developed by Field[62] for semi empirical QM/MM has been extended for use in *ab initio*[63] QM/MM and SCC-DFTB/MM calculations.[64]

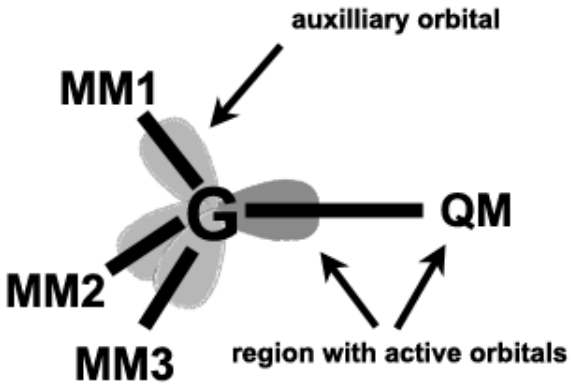

Figure 2-3 A generic QM/MM system with G a GHO boundary atom which has 3 auxilliary and 1 active orbital. The active orbital is optimised with the QM region. (adapted[59, 64])

In the GHO method, the single hybrid orbital is included in the LCAO such that

Equation 2-43

$$\phi_i^H = \sum_{\mu}^{N} c_{\mu i}\chi_\mu + c_{Bi}\eta_B$$

where the superscript H emphasises that this is a hybrid orbital basis and c_{Bi} is the coefficient for the hybrid orbital η_B from the GHO atom.

Elements of the effective one-electron matrix, $H_{\mu\nu}^H$, are defined by

Equation 2-44

$$H_{\mu\nu}^{eff} = H_{\mu\nu}^H + I_{\mu\nu}^H + 0.5\sum_{b}^{3} P_{bb}^H[(\mu\nu, bb) - 0.5(\mu b, b\nu)]$$

where $H_{\mu\nu}^H$ is the "standard" one-electron matrix element for the QM region, $I_{\mu\nu}^H$ is the QM/MM one-electron integral resulting form the partial charges of the classical MM region. The boundary auxilliary orbitals are specified by *b* and the charge density of the auxilliary hybrid orbitals is P_{bb}^H. The

electron-nucleus attraction and Coulomb-exchange repulsion interaction due to the auxilliary orbitals is defined as

Equation 2-45

$$E_{aux} = \sum_{b}^{3} P_{bb}^{H} \left(H_{bb}^{H} + I_{bb}^{H} \right) + 0.5 \sum_{b,c}^{3} P_{bb}^{H} P_{cc}^{H} [(bb|cc) - 0.5(bc|cb)]$$

where the electron-nucleus attraction integral of the auxilliary orbitals for the QM nucleus charges is H_{bb}^{H} and for the MM nucleus charges is I_{bb}^{H}.

2.2 Statistical Mechanics

A single optimised structure is of no use when considering properties that depend on the dynamical nature of a system. To this end, a probabilistic theory[7] (statistical mechanics theory) is applied where the average properties of a system can be "measured". This is possible because the observables of a microscopic system are related to macroscopic quantities.

An isolated system where the internal energy E, volume V and number of particles N are fixed is the starting point for this theory. If it is assumed that the number of quantum-mechanical states, Ω(N, V, E), characterised by the N, V, E variables is finite then the Boltzmann distribution relates the entropy S of the system and the number of states Ω(N, V, E)

Equation 2-46

$$S(N,V,E) = k_B \ln \Omega(N,V,E)$$

where k_B is Boltzmann's constant. Clearly the macroscopic thermodynamic property, S, is related to the microscopic property, Ω.

If it is supposed that a large collections of such systems exist which may vary in terms of microscopic state but retain identical N, V, E, then such a collection is termed an ensemble.[65] In the NVE or microcanonical ensemble, all the microscopic states available are equally likely at thermodynamic equilibrium, so the probability of a specific state is inversely dependent on the number of available states Ω

Equation 2-47

$$P_i = \frac{1}{\Omega(N,V,E)}$$

An average of a property, Y, for all the instantaneous configurations gives the ensemble average or observable, $\langle Y \rangle$, and describes the experimental average of Y.[2, 7, 9] The ergodic hypothesis states that the ensemble average of a property is equivalent to the time average (as the time limit approaches infinity) and so it is possible to generate ensembles with molecular dynamics (MD) simulations.

For comparison to experimental observables, the constant number of particles N, volume V, temperature T ensemble, otherwise called the canonical ensemble, is more useful than the isolated NVE ensemble. The internal energies in the NVT system are not fixed and so the probability of the system Pr(E) being found at a specific energy E is

Equation 2-48

$$\Pr(E) = \frac{\Omega(N,V,E)e^{(-E/k_BT)}}{Q(N,V,T)}$$

Taking into account the normalization condition, that the sum of the likelihood of the *i* energy states being occupied must be unity

Equation 2-49

$$\sum_i \Pr(E) = 1$$

where the canonical partition function Q(N, V, T) in Equation 2-50, is the summation of all *i* energy levels available to all molecules in the system.[7]

Equation 2-50

$$Q(N,V,T) = \sum_i \Omega(N,V,E)e^{(-E_i/k_BT)} = \sum_i e^{-E_i/k_BT}$$

For the NVT ensemble, the canonical partition function is related to the Helmholtz free energy by

Equation 2-51

$$A(N,V,T) = -k_B T \ln Q(N,V,T) .$$

The ensemble used in this thesis is the constant NVT ensemble, the canonical ensemble. For a canonical ensemble the ensemble average $\langle Y \rangle$ of a property Y is calculated using

Equation 2-52

$$\langle Y \rangle_{NVT} = \frac{\int Y(v_N) e^{-\beta U(v_N)} dr_N}{\int e^{-\beta U(v_N)} dr_N} = \frac{\sum_{v=1}^{T} Y(v_N) e^{-\beta U(v_N)}}{\sum_{v=1}^{T} e^{-\beta U(v_N)}}$$

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 of the N atoms in the molecular system. The minimum Helmholtz free energy A can be measured from the equilibrium state for the canonical ensemble.[2] In general, equilibrium methods such as MD simulations sample configurations (microscopic states) in such a way that if $\rho(v_N)$ is the probability of choosing a configuration $v(r_N, p_N)$ then

Equation 2-53

$$\langle Y \rangle_{NVT} = \frac{\sum_{v=1}^{T} \frac{Y(v_N) e^{-\beta U(v_N)}}{\rho(v_N)}}{\sum_{v=1}^{T} \frac{e^{-\beta U(v_N)}}{\rho(v_N)}}$$

where $\rho(v_N) = e^{-\beta \varepsilon_i}$ and so samples the Boltzmann distribution of the canonical ensemble as an unweighted average over the configurations sampled (Equation 2-54).

Equation 2-54

$$\langle Y \rangle_{NVT} = \frac{1}{T} \sum_{v=1}^{T} Y(v_N)$$

2.3 Molecular Dynamics

In this thesis, ensembles were generated in a time-dependent manner using molecular dynamics. Velocities and accelerations are generated from the integration of Newton's Equations of motion using the leap-frog Verlet method. The leap-frog methods avoids the inaccuracies associated with the standard Verlet method[2, 66] and employs the following equations

Equation 2-55

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

Equation 2-56

$$r(t+\delta t)=r(t)+\delta t v(t+\frac{1}{2}\delta t)$$

The velocity equation, Equation 2-55 , is applied first and is always calculated halfway between time steps. Applying Equation 2-56, the new positions can then be calculated from the previous positions using the velocities calculated from Equation 2-55. The velocities and positions (coordinates) appear to "leap" over each other, as suggested by the name of the algorithm.[66]

Certain simulations employed the velocity-Verlet integrator (named VV2 in CHARMM) which is similar to leap-frog but calculates the velocity and position simultaneously at each time step

Equation 2-57

$$v(t+\delta t)=v(t)+\frac{a(t)+a(t+\delta t)}{2}\delta t$$

Equation 2-58

$$r(t+\delta t)=r(t)+\delta t v+\frac{1}{2}a(t)(\delta t)^2$$

When employing the velocity-Verlet integrator, the Nose-Hoover thermostat is used. This method couples a heat bath to a system to regulate the temperature.[67, 68]

2.3.1 Non-bonded interactions

There are many through space interactions and as the distance between particles increases these taper off until they are negligible. With this in mind, the computational efficiency of simulations can be improved by applying a non-bonded cut-off scheme where through space interactions are cut off at a certain distance. A switching or shifting scheme is then applied as direct truncation will lead to discontinuities in potential energy. Switching (Sw) or shifting (Sf) can be applied such that the potential energy function smoothly approaches zero at the specified cut-off distance.[69] These techniques are described by the following equations, where the potential is multiplied by the a function of the forms described below,[2, 69]

Equation 2-59

$$\left.\begin{array}{ll} S_w(r_{ij}) = 1 & r_{ij} \le r_{on} \\ S_w(r_{ij}) = \dfrac{(r_{off}^2 - r_{ij}^2)(r_{off}^2 + 2r_{ij}^2 - 3r_{on}^2)}{(r_{off}^2 - r_{on}^2)^3} & r_{on} < r_{ij} \le r_{off} \\ S_w(r_{ij}) = 0 & r_{ij} > r_{off} \end{array}\right\} \quad switching$$

Equation 2-60

$$\left.\begin{array}{ll} Sf(r_{ij}) = 1 - \dfrac{2r_{ij}^2}{r_{cut}^2} + \dfrac{r_{ij}^4}{r_{cut}^4} & r_{ij} < r_{cut} \\ Sf(r_{ij}) = 0 & r_{ij} > r_{cut} \end{array}\right\} \quad shifting$$

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.

2.3.2 Boundary conditions

When studying molecular systems of finite size, the boundary regions play an effect on the integrity of the molecular calculation.

The periodic boundary scheme emulates bulk properties and avoids edge effects in systems contained in regular shaped simulation volumes such as cubes. The simulation cell is replicated, or imaged, using translation symmetry elements to produce identical images of a system around a central real

system(Figure 2-4). An atom must not interact with its own image, to avoid interaction between "real" molecules and image molecule in the periodic boxes, the minimum image convention is applied. An image cutoff is applied after which molecular interactions cease (Figure 2-4).

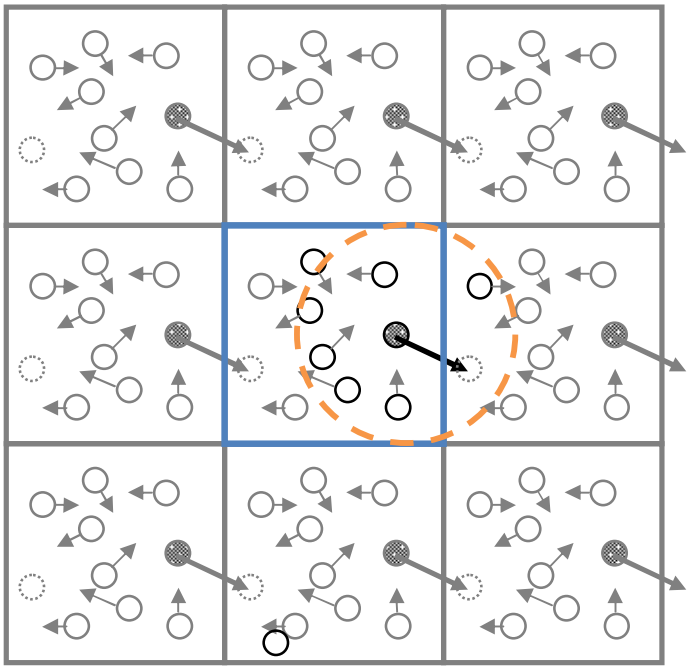

Figure 2-4 2D representation of a system showing the real (central) and periodically imaged (surrounded) boxes. The dashed circle represents the image cut off that is applied (adapted[2, 4]).

For certain systems such as large enzymes the periodic boundary scheme is not appropriate as very large simulation cells are required. In such a situation, the stochastic boundary conditions[70-72] are used, where a system is divided into three spherical shells centred at the reactive site origin as shown in Figure 2-5. The outer shell or the reservoir region between radii R_2 and R_3 contains all other particles and the atoms in this region *may* be held fixed to maintain the shape of a system. The second shell or buffer region, between R_1 and R_2, is necessary for stabilizing the temperature of the system and atom are treated with Langevin dynamics. The inner or reaction region, within radius R_1, is treated using leap-frog Verlet dynamics.

Figure 2-5 A protein in a stochastic boundary system sphere. The reaction region is that within R_1, between R_1 and R_2 is the buffer region and between R_2 and R_3 is the reservoir region.

In these systems the central region undergoes standard Verlet dynamics, while Langevin dynamics are employed in the second shell (the buffer region) and the atoms in the reservoir region are kept fixed. The buffer region is used to create a smooth transition from the dynamic reaction region to the fixed reservoir region, by applying a screening potential, S_{SB},

Equation 2-61

$$S_{SB} = \begin{cases} 0 & r \le R_B \\ \dfrac{(r-R_B)^2(3R_L - R_B - 2r)}{2(R_L - R_B)^3} & R_B < r \le R_L \\ 0.5 & r > R_L \end{cases}$$

where r is the distance from the centre of the system, R_B is the Buffer Boundary and R_L is the Langevin Boundary.

The equation of motion used in Langevin dynamics is the following for each atom i in the system:

Equation 2-62

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

where m_i and $\vec{r}_i$ are the mass and position of the *i*th 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 $\vec{\gamma}_i \frac{d\vec{r}_i(t)}{dt}$ to the force arising from force field terms. The stochastic term introduces energy into the system and can be considered as a random force due to constant molecular bombardment of the solute by the solvent. The frictional force removes energy from the system and can be considered of as frictional drag. The bombardment and drag effects are real effects that are experimentally observed as Brownian motion.[8] Stochastic dynamics can be used to control the temperature and thus avoid velocity rescaling techniques and their associated problems.[2]

Simulation artifacts may result in stochastic simulations where large parts of the system are fixed , especially near the reaction region. For systems, such as that in Figure 2-5, where the radius of the inner reaction region is larger than the non-bonded cutoff distance and the radius of the water sphere is large, the reaction region is negligibly affected by the stochastic boundary. Previous studies have shown that stochastic and periodic boundary calculations compare favourably.[73]

2.4 General Molecular System Setup Schemes

For the simulations of small molecules a standard setup procedure was applied as shown in Figure 2-6. Molecules were built in CHARMM. The molecular energies were minimised using either steepest-descent or adapted Newton Raphson optimisation techniques. Where warranted, molecules were placed in a solvent cube of TIP3P water. Heating dynamics were followed by equilibration dynamics at 298.15K.

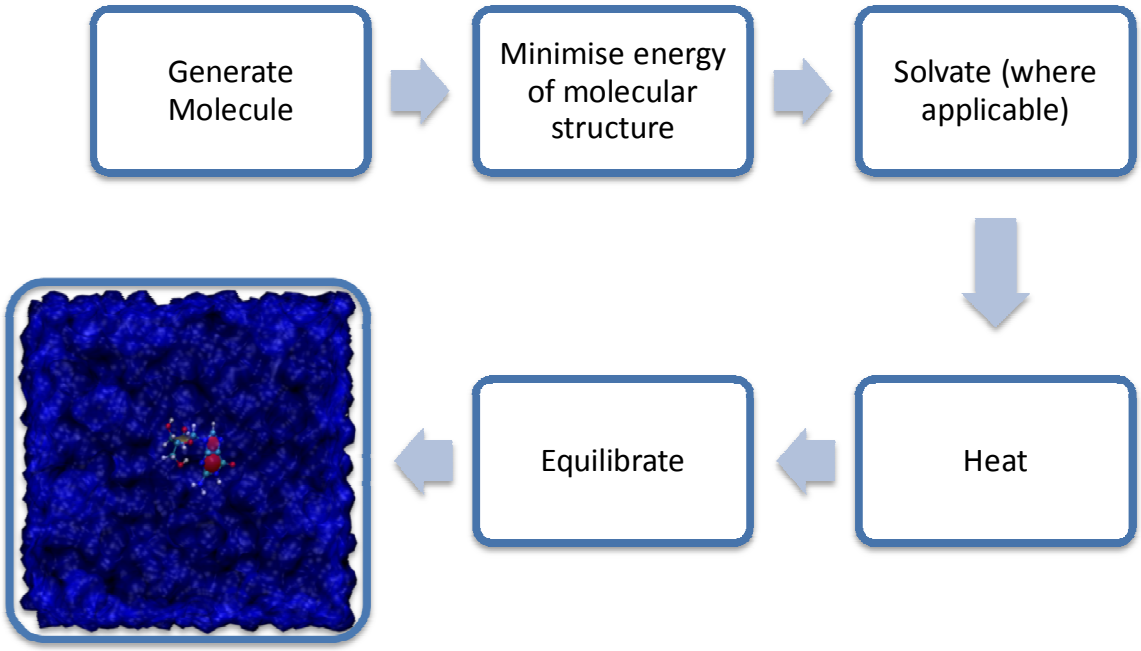

Figure 2-6 A general overview of the process of readying a molecule for free energy simulations

The setup for a protein system is outlined in Figure 2-7. The Protein Data Bank (PDB) file for a particular protein containing its atomic coordinates was downloaded from the PDB. Superfluous molecules were removed and the PDB file cleaned. The protonation states of amino acid residues are determined and hydrogens are added using the HBUILD routine in CHARMM. The QM region and GHO atoms are chosen and the structure is minimised. The system is solvated using a TIP3P water sphere of appropriate size, after which the MM and then the QM regions are each minimised separately followed by minimisation of the entire system. Heating dynamics were followed by equilibration dynamics at 298.15K. The system was resolvated if necessary and re-equilibrated. The system was then set up and ready for free energy simulations.

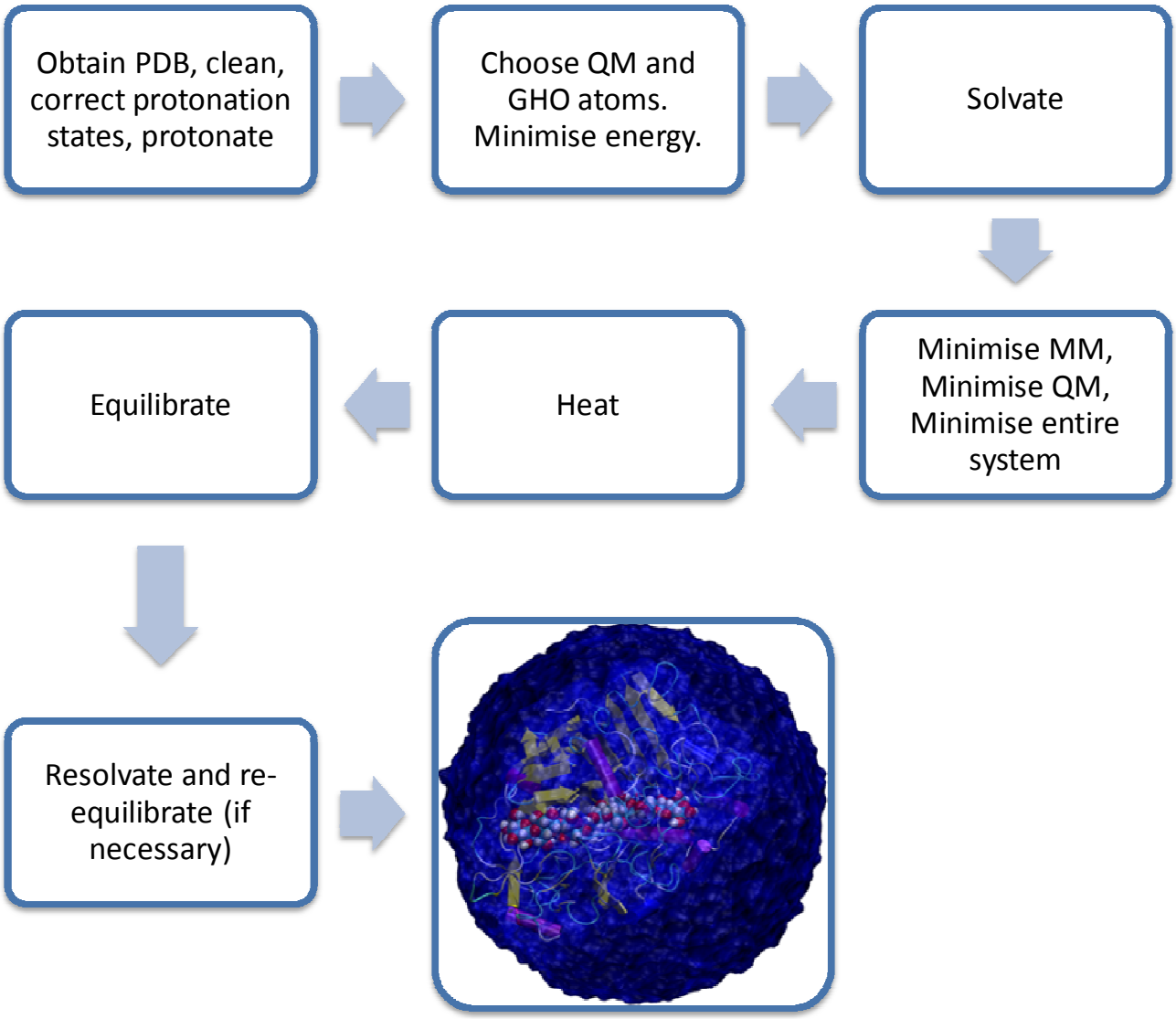

Figure 2-7 A general overview of the process of readying a protein for free energy simulations

2.5 References

1. *Encyclopaedia of Computational Chemistry*, Wiley, Chichester, 1998.
2. A. R. Leach, *Molecular Modelling: Principles and Applications*, Longman, 2001.
3. F. Jensen, *Introduction to Computational Chemistry*, J. Wiley and sons, England, 1999.
4. E. Lewars, *Computational Chemistry: Introduction to the Theory and Applications of molecular and Quantum Mechanics*, Kluwer Academic Publishers, Boston, 2004.
5. N. S. Ostlund and A. Szabo, *Modern Quantum Chemistry: Introduction to Advanced Electonic Structure Theory*, Dover Publications, New York, 1996.
6. I. N. Levine, *Quantum Chemistry*, Prentice Hall, New Jersey, 2000.
7. D. Chandler, *Introduction to Modern Statistical Mechanics*, Oxford University Press, Oxford, 1987.
8. D. A. McQuarrie, *Statistical Mechanics*, University Science Books, California, 2000.
9. P. Atkins and J. de Paula, *Atkins' Physical Chemistry*, Oxford University Press, 2002.
10. A. C. van Duin, S. Dasgupta, F. Lorant and W. A. Goddard, III, *Journal of Physical Chemistry A*, 2001, **105**, 9396-9409.
11. B. R. Brooks, C. L. Brooks, 3rd, A. D. Mackerell, Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York and M. Karplus, *Journal of Computational Chemistry*, 2009, **30**, 1545-1614.
12. P. K. Weiner and A. Kollman Peter, *Journal of Computational Chemistry*, 1981, **2**, 287-303.
13. B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus, *Journal of Computational Chemistry*, 1983, **4**, 187-217.
14. W. F. Van Gunsteren and H. J. C. Berendsen, University of Groningen, 1988.
15. T. M. Glennon, Y. Zheng, S. M. le Grand, B. A. Shutzberh and K. M. Merz, Jr., *Journal of Computational Chemistry*, 1994, **15**, 1019-1040.
16. F. A. Momany and J. L. Willet, *Carbohydrate Research*, 2000, **326**, 194-209.
17. H. Senderowitz, W. C. Still and C. Parish, *Journal of the American Chemical Society*, 1996, **118**, 2078-2086.
18. H. Senderowitz and W. C. Still, *Journal of Organic Chemistry*, 1997, **62**, 1427-1438.
19. J. W. Brady, M. J. Field, M. Giammona and S. N. Ha, *Carbohydrate Research*, 1988, **180**, 207-221.
20. J. W. Brady, R. Palma, M. E. Himmel and G. Liang, *Molecular mechanics studies of cellulases*, ACS, Washington, 2001.
21. S. Reiling, M. Schlenkrich and J. Brickman, *Journal of Computational Chemistry*, 1996, **17**, 450-468.
22. B. Meyer and K. Ott, *Journal of Computational Chemistry*, 1996, **17**, 1068-1084.
23. W. Damm, A. Frontera, J. Tirado-Rives and W. L. Jorgensen, *Journal of Computational Chemistry*, 1997, **18**, 1955-1970.
24. D. Kony, W. Damm, S. Stoll and W. F. Van Gunsteren, *Journal of Computational Chemistry*, 2002, **23**, 1416-1429.
25. M. Kuttel, J. W. Brady and K. J. Naidoo, *Journal of Computational Chemistry*, 2002, **23**, 1236-1243.
26. W. L. Jorgensen and C. Jenson, *Journal of Computational Chemistry*, 1998, **19**, 1179-1186.
27. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *Journal of Chemical Physics*, 1983, **79**, 926-935.
28. H. J. C. Berendsen, J. R. Grigera and T. P. J. Straatsma, *Journal of Physical Chemistry*, 1987, **91**, 6269-6271.

29. C. J. Cramer and D. G. Truhlar, *Journal of the American Chemical Society*, 1993, **115**, 5745-5753.
30. M. J. T. Frisch, 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.
31. M. F. Guest, I. J. Bush, H. J. J. v. Dam, P. Sherwood, J. M. H. Thomas, J. H. v. Lenthe, R. W. A. Havenith and J. Kendrick, *Molecular Physics*, 2005, **103**, 719-747.
32. A. Frisch and J. B. Foresman, *Exploring Chemistry with Electronic Structure Methods*, Gaussian Inc., Pittsburgh, 1996.
33. W. J. Hehre, L. Radom, P. v. R. Schleyer and J. A. Pople, *Ab initio molecular orbital theory*, John Wiley and Sons, New York, U.S.A, 1986.
34. J. P. Lowe and K. A. Peterson, *Quantum Chemistry*, Elsevier Academic Press, San Diego, 2006.
35. W. Kohn and L. J. Sham, *Physical Reviews*, 1965, **140**, A1133-A1138.
36. D. B. Cook, *Handbook of Computational Quantum Chemistry*, Dover Publications, New York, 2005.
37. S. H. Vosko, L. Wilk and M. Nusair, *Canadian Journal of Physics*, 1980, **58**, 1200-1211.
38. A. D. Becke, *Journal of Chemical Physics*, 1993, **98**, 5648-5652.
39. C. Lee, W. Yang and R. G. Parr, *Physical Reviews B37*, 1988, 785-789.
40. C. J. Cramer and D. G. Truhlar, *Journal of the American Chemical Society*, 1993, **115**, 5745-5753.
41. S. E. Barrows, J. W. Storer, C. J. Cramer, A. D. French and D. G. Truhlar, *Journal of Computational Chemistry*, 1998, **19**, 1111-1129.
42. P. L. Polavarapu and C. S. Ewig, *Journal of Computational Chemistry*, 1992, **13**, 1255-1261.
43. M. Appell, G. Strati, J. L. Willett and F. A. Momany, *Carbohydrate Research*, 2004, **339**, 537-551.
44. C. B. Barnett and K. J. Naidoo, *Journal of Physical Chemistry B*, 2008, **112**, 15450-15459.
45. G. I. Csonka, *Journal of Molecular Structure: THEOCHEM*, 2002, **584**, 1-4.
46. G. I. Csonka, A. D. French, G. P. Johnson and C. A. Stortz, *Journal of Chemical Theory and Computation*, 2009, **5**, 679-692.
47. M. J. S. Dewar, E. G. Zoebisch, E. F. Healy and J. J. P. Stewart, *Journal of the American Chemical Society*, 1985, **107**, 3902-3909.
48. K. Byun, Y. Mo and J. Gao, *Journal of the American Chemical Society*, 2001, **123**, 3974-3979.
49. M. E. Brewster, M. Huang, P. Emil, J. Pitha, M. J. S. Dewar, J. J. Kaminski and N. Bodor, *Carbohydrate Research*, 1993, **242**, 53-67.
50. A.-M. Muslim and R. A. Bryce, *Chemical Physics Letters*, 2004, **388**, 473-478.
51. J. P. McNamara, A.-M. Muslim, H. Abdel-Aal, H. Wang, M. Mohr, I. H. Hillier and R. A. Bryce, *Chemical Physics Letters*, 2004, **394**, 429-436.
52. M. Elstner, *Theoretical Chemistry Accounts*, 2006, **116**, 316-325.
53. Q. Cui, M. Elstner, E. Kaxiras, T. Frauenheim and M. Karplus, *Journal of Physical Chemistry B*, 2001, **105**, 569-585.

54. T. Frauenheim, G. Seifert, M. Elstner, T. Niehaus, C. Köhler, M. Amkreutz, M. Sternberg, Z. Hajnal, A. Di Carlo and S. Suhai, *Journal of Physics: Condensed Matter*, 2002, **14**, 3015-3047.
55. H. Zhou, E. Tajkhorshid, T. Frauenheim, S. Suhai and M. Elstner, *Chemical Physics*, 2002, **277**, 91-103.
56. D. Riccardi, P. Schaefer, Y. Yang, H. Yu, N. Ghosh, X. Prat-Resina, P. König, G. Li, D. Xu, H. Guo, M. Elstner and Q. Cui, *Journal of Physical Chemistry B*, 2006, **110**, 6458-6469.
57. M. Iannuzzi, A. Laio and M. Parrinello, *Physical Review Letters*, 2003, **90**, 238302.
58. R. Car and M. Parrinello, *Physical Review Letters*, 1985, **55**, 2471-2474.
59. M. J. Field, P. A. Bash and M. Karplus, *Journal of Computational Chemistry*, 1990, **11**, 700-733.
60. H. M. Senn and W. Thiel, *Angewandte Chemie International Edition*, 2009, **48**, 1198-1229.
61. A. J. Mulholland, in *Chemical Modelling: Applications and Theory*, ed. A. Hinchliffe, RSC Publishing, Cambridge, 2006, pp. 23-60.
62. J. Gao, P. Amara, C. Alhambra and M. J. Field, *The Journal of Physical Chemistry A*, 1998, **102**, 4714-4721.
63. J. Pu, J. Gao and D. G. Truhlar, *The Journal of Physical Chemistry A*, 2004, **108**, 632-650.
64. J. Pu, J. Gao and D. G. Truhlar, *The Journal of Physical Chemistry A*, 2004, **108**, 5454-5463.
65. K. J. Laidler, J. H. Meiser and B. C. Sanctuary, *Physical Chemistry*, Houghton-Mifflin, 2002.
66. M. P. Allen and D. J. Tildesley, *Computer Simulations of Liquids*, Clarendon Press, Oxford, 1989.
67. S. Nose, *Molecular Physics*, 1984, **52**, 255-268.
68. W. G. Hoover, *Physical Reviews A*, 1985, **31**, 1695-1697.
69. B. R. Brooks and P. J. Steinbach, *Journal of Computational Chemistry*, 1994, **15**, 667-683.
70. R. W. Pastor and M. Karplus, *The Journal of Physical Chemistry*, 1988, **92**, 2636-2641.
71. C. L. Brooks, 3rd, A. Brunger and M. Karplus, *Biopolymers*, 1985, **24**, 843-865.
72. A. T. Brunger, C. L. Brooks, 3rd and M. Karplus, *Proceedings of the National Academy of Sciences U S A*, 1985, **82**, 8458-8462.
73. A. T. Brunger, C. L. Brooks, 3rd and M. Karplus, *Chemical Physics Letters*, 1984, **105**, 495-500.

3 Free Energy Methods and Code Development[1]

Innovative method development, in modern theoretical and computational chemistry, is vital for studying and understanding the origins of fundamental chemistry with regards to conformation, configuration and reaction mechanisms. In light of this, the core goal of this thesis is to develop code applicable to novel simulation of enzymes and carbohydrates. The major concern is the evaluation of molecular ring deformations and their critical role in key molecular reactions and processes[1], especially the puckering conformations of pyranose and furanose enzymatic pathways.[2] As discussed in Chapter 1, on consideration of several classes of enzymes, the puckering of a carbohydrate rings is noted on degradation of the glycosidic bond through hydrolysis (e.g. glycoside hydrolases) or N-glycosidic linkage degradation via phosphorolysis (e.g. nucleoside phosphorylases).

Calculation of the free energies associated with molecular processes is not trivial, but from such a quantity molecular properties such as binding and conformational preferences can be predicted and understood. [3] In this thesis, an adaptive potential of mean force approach is used to calculate free energies associated with ring pucker of selected glycosides. The bond creation and degradation of the glycosidic linkage is also considered. This chapter discusses the concepts of free energies from reactive coordinate forces (FEARCF), a general method for calculating the free energy of a reaction coordinate space, which improves phase space sampling of high energy configurations / conformations. Other adaptive free energy methods are also discussed. In order to calculate the free energy of ring puckering, a method for calculating ring puckering is included in FEARCF.

[1] Parts of this chapter have been adapted from published work:

Barnett, C. B. & Naidoo, K. J. Calculating Ring Pucker Free Energy Surfaces From Reaction Coordinate Forces. Journal Name: AIP Conference Proceedings; Journal Volume: 1102; Journal Issue: 1; Conference: TACC 2008: International conference on theory and application of computational chemistry - 2008, Shanghai (China), 23-27 Sep 2008; Other Information: DOI: 10.1063/1.3108377; (c) 2009 American Institute of Physics; Country of input: International Atomic Energy Agency (IAEA), Medium: X; Size: page(s) 214-218 (2009).

Barnett, C. B. & Naidoo, K. J. Free Energies from Adaptive Reaction Coordinate Forces (FEARCF): an application to ring puckering. Molecular Physics: An International Journal at the Interface Between Chemistry and Physics 107, 1243 - 1250 (2009).

3.1 Biasing Free Energy Methods

The free energy of a system (the total amount of energy in a physical system which can be converted to do work[4]) is a state function which can be used to explore binding and conformational preferences. In the previous chapter, the relevance of statistical thermodynamics theory was discussed. In this theory, the relationship between the Helmholtz free energy, A, and the canonical partition function is as follows

Equation 3-1

$$A(N,V,T) = -k_B T \ln Q(N,V,T)$$

Equation 3-1 shows that the free energy depends on the canonical partition function and is not of the form of a canonical average over phase space. Unlike for mechanical quantities, such as Pressure and Volume, which can be expressed as ensemble averages, thermal properties[5] (such as entropy and the free energy) cannot be expressed as ensemble averages and are not measurable from conventional Boltzmann sampling simulations.

Following the formulation of Zwanzig[6] for two fluids, it is possible to calculate the free energy difference of two conformations characterised by differing potentials

Equation 3-2

$$A_1 - A_0 = -k_B T \ln \frac{Q_1}{Q_0} = -k_B T \ln \left\langle e^{-\beta \Delta E} \right\rangle_0$$

where $\Delta E = E_1(r) - E_0(r)$, and the ensemble average $\langle A \rangle_0$ is taken in the reference system $E_0(r)$. The energy E is a function of the state *r* of the system. This equation can be used to compute free energy differences in a procedure known as the free energy perturbation (FEP) method. [7-9] This method is applicable in systems, to estimate free energy difference, where the two systems chosen are not too dissimilar. The statistical error in such an evaluation of the free energy is large, as the greatest contribution to the ensemble average $\langle e^{-\beta \Delta E} \rangle_0$ stems from the regions of configurational space where the probability distribution of state 0 is small but $e^{-\beta \Delta E}$ is large. A standard Boltzmann sampling is found not to be an efficient way to explore the relevant parts of phase space needed to calculate the ensemble average $\langle e^{-\beta \Delta E} \rangle_0$. [10, 11]

Nonphysical sampling distributions for estimating free energies were developed by Torrie and Valleau.[10, 11] Such an approach is a general solution to sample regions of large negative ΔE, which cannot be efficiently sampled using Equation 3-2. A non-negative weighting function, *W(ξ)*, is chosen to favour configurations in the chosen phase space, *ξ*, that are important for the ensemble average $\langle e^{-\beta\Delta E}\rangle_0$ to provide a sampling distribution, $\rho_w(\xi)$. $\rho_w(\xi)$ is not a Boltzmann distribution[10] but a Markov chain as follows

Equation 3-3

$$\rho_w(\xi) = \frac{W(\xi)e^{-\beta E_0(\xi)}}{\int W(\xi)e^{-\beta E_0(\xi)}d\xi}.$$

The weighting function, *W(ξ)*, can be adjusted to simultaneously sample regions of configurational space accessible to two or more physical systems and so this technique is referred to as umbrella sampling. [10-14] In such a calculation, it is surprisingly not advantageous to refine the weighting function so that all relevant configurations are sampled in a single simulation. Instead, it is better to perform several umbrella sampling simulations with different weighting functions to localise exploration of the phase space into overlapping regions (several overlapping windows of configurational space are chosen and several calculations are done). The average of any property in the original ensemble, $\langle Y\rangle_0$, an unbiased ensemble average can be recovered by taking ensemble averages of the $\rho_w(\xi)$, distribution. Both averages $\langle Y/W\rangle_W$ and $\langle 1/W\rangle_W$ are required to achieve the final result, thus

Equation 3-4

$$\langle Y\rangle_0 = \frac{\langle Y/W\rangle_w}{\langle 1/W\rangle_w}.$$

A good estimate of the probability density of $\left\langle e^{-\beta\Delta E}\right\rangle_0$ where it takes on its largest values can be obtained from

Equation 3-5

$$\left\langle e^{-\beta\Delta E} \right\rangle_0 = \frac{\left\langle e^{-\beta\Delta E} / W \right\rangle_W}{\left\langle 1/W \right\rangle_W} .$$

Adaptive Umbrella sampling[14, 15] extends the method of Torrie and Valleau such that a uniform sampling of a potential energy surface can be achieved by carrying out several simulations, *i,* with a modified Hamiltonian, H_i.

Equation 3-6

$$H_i = H^{\circ} + U_i$$

An umbrella potential, U_i, is applied to bias phase space sampling. After each simulation, U_i, the umbrella potential can be updated based on the sampling of the previous *j* simulations where *j=1, i-1*.

3.1.1 Distributions functions and reversible work

A classical interpretation of the microscopic state of a system is typified by a point in phase space which specifies the coordinates and momenta of a system.

Equation 3-7

$$(r_1, r_2, \ldots, r_N; p_1, p_2, \ldots, p_N) = (r^N, p^N)$$

The positions are represented by r_i and p_i are the momenta of a particle *i,* while r^N an p^N represent all the points in configuration and momentum space. In addition to the overall configurational distribution, $P(r^N)$, reduced distributions such as the radial distribution function can be defined. The radial distribution function g(r) is a reduced distribution function which gives the probability for finding any two particles a distance r apart.

Following from Kirkwood's equation where assuming that $r_1 \ldots r_N$ are the position of N particles at constant V, T ; the correlation function g of a subset of particles [16, 17] is related to the Helmholtz energy (defined here as $w^{(n)}$) by

Equation 3-8

$$g^{(n)}(r_1,...,r_n) \equiv e^{-\beta w^{(n)}}{}_{r_1,...,r_n}$$

This is the reversible work theorem and it states that the free energy (or reversible work) can be defined in terms of a distribution function for a particular process, for example, the movement of two molecules from coordination to infinite separation. Furthering this derivation, [16, 17] yields the force acting on a specific particle *j*, $-\nabla_j U$, for any fixed configuration as

Equation 3-9

$$-\nabla_j w^{(n)} = \frac{\int \cdots \int e^{-\beta U}(-\nabla_j U) dr_{n+1} \cdots dr_N}{\int \cdots \int e^{-\beta U} dr_{n+1} \cdots dr_N} \qquad j = 1,2,.....,n$$

where the right hand side of this expression is the mean force $f_j^{(n)}$ acting on a particle *j* averaged over all possible configurations. The potential that yields the mean force on a particle *j* is $w^{(n)}$ the potential of mean force or free energy for the process defined by the reduced distribution.

The free energies, in terms of chosen coordinates, can be calculated based on the adaptive umbrella sampling method. [15] Free energy methods that have incorporated this concept include the metadynamics of Laio and Parrinello[18], the adaptive biasing force (ABF) method of Darve and Pohorille [19, 20] and the Free energies from adaptive reaction coordinate forces (FEARCF) method. [21, 22]

3.2 Free Energies from Reaction Coordinate Forces

FEARCF is a force biasing method, based on probability distributions and histograms, that can be applied to multi-dimensional conformation or reaction free energy problems to calculate multi-dimensional potentials of mean force (PMF). The PMF $W(\xi)$ (not the weighting function of Torrie and Valleau) is calculated as a function of a multi-dimensional set, ξ, and is related to the probability density, $P(\xi)$ by

Equation 3-10

$$W(\xi) = -k_B T \ln P(\xi)$$

where k_B is the Boltzmann constant and T is the temperature in Kelvin. The members of the set ξ are carefully chosen reaction coordinates that represent a process applicable to the system under consideration. A reaction coordinate may be a bond distance dihedral angle or similar. The available reaction coordinate space is defined as a discretised n-dimensional grid. During a simulation a running tally (history) of the visited reaction coordinates is recorded. The probability density of the reaction coordinate is derived from this history by creating a histogram using the discretised n-dimensional reaction coordinate space. The population of this grid represents a running tally of the reaction coordinate probability density, derived from the history of simulations to that point. The aim is to bias away from previously sampled regions of coordinate space. To do this, a biasing potential, *U(ξ),* is applied that adapts from iteration to iteration by incorporating the information about previously sampled reaction coordinate space i.e. the probability density. *U(ξ)* is defined as

Equation 3-11

$$U(\xi) = k_B T \ln P(\xi)$$

and is the inverse of the PMF.

A multidimensional cubic-spline interpolation routine is used to calculate the reaction coordinate biasing forces, *F(ξ),* from the biasing potential. These forces are applied to all atoms used in the reaction coordinate definition to bias the next simulation's reaction coordinate trajectory away from previously sampled areas. The entire reaction coordinate space is equally sampled when the biasing forces are derived from the true PMF. The method requires no intervention from the user other than a physical understanding of the complexity of the computer experiment, which is used to make a judicious choice of reaction coordinate and simulation length at each update of the biasing force.

Free energy hyper-surfaces approach convergence when close to uniform sampling of phase space is achieved, although ratios of 1:50 are considered acceptable. [13] To improve sampling of reaction coordinate space , especially when there are large energy barriers, a biasing force[12, 13, 15] is applied to bias the future trajectories away from the sampled regions of reaction coordinate space to unsampled regions as just discussed. This force also aids the system in traversing barrier heights larger than those thermally accessible at a chosen temperature.

At the start of the FEARCF simulations W(ξ) and likewise P(ξ) are not known. To initialise the simulations, a zero biasing potential is applied (U(ξ) = 0). The resulting probability distribution is used as a first guess for W(ξ) from which an improved biasing potential U(ξ) for the next simulation is obtained. As illustrated in Figure 3-1, the calculation of W(ξ) is an iterative process and several simultaneous simulations are batched at once to improve convergence.

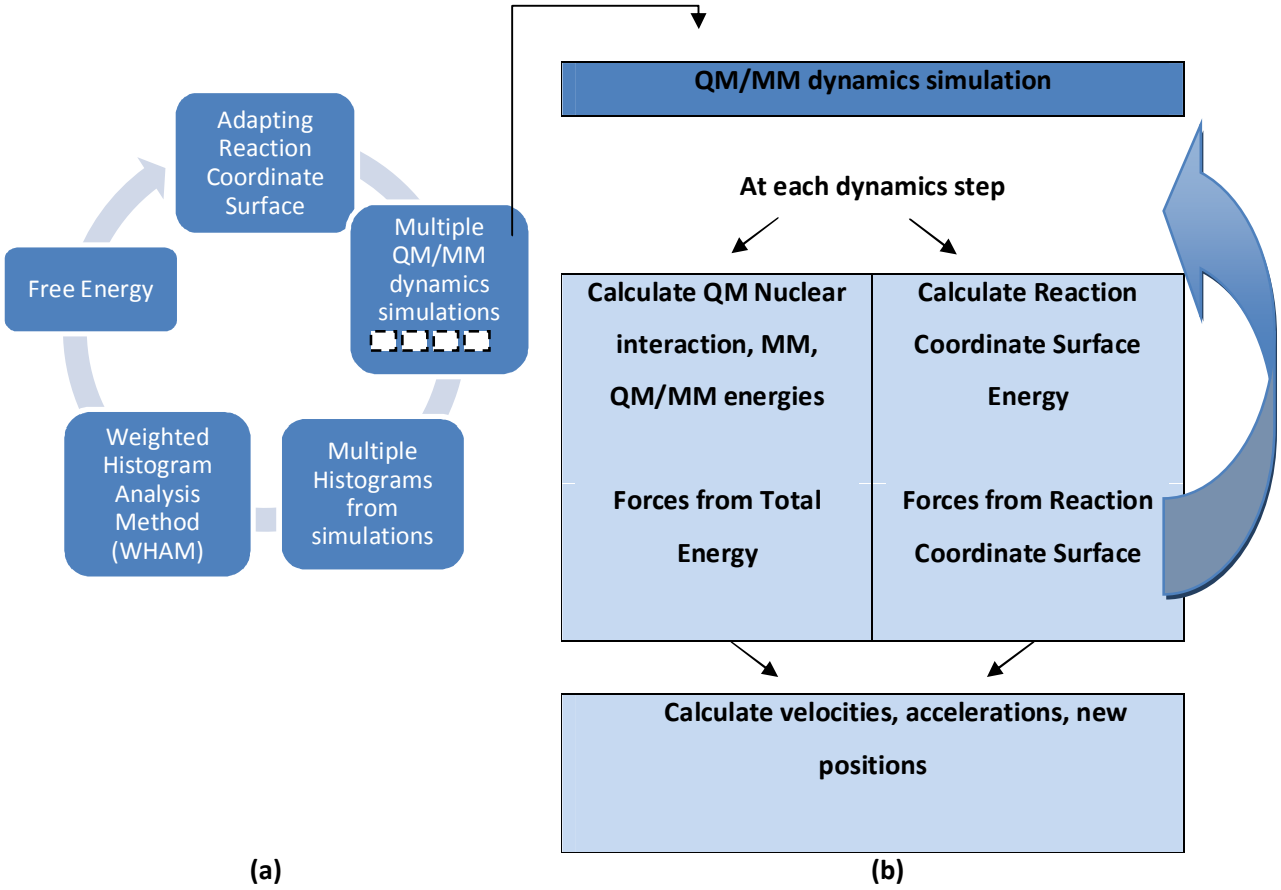

Figure 3-1 (a)The iterative FEARCF process (b) Details of the force calculated during dynamics

Also the sampling from previous simulations is not wasted. In order to overlap the histogram data from multiple simulation the weighted histogram analysis method (WHAM)[9, 23, 24] an extension of the histogram procedure developed by Ferrenberg and Swendsen[25, 26] is applied. This method ensures that the probability distribution histrograms for all simulations overlap best to provide the most accurate unbiased probability distribution p_k. The solution to the equations is self-consistent. The WHAM equations are

Equation 3-12

$$p_k = \frac{\sum_i n_{i,k}}{\sum_i N_i f_i e^{-\beta U_i(\xi_k)}} \quad ; N_i = \sum_k n_{i,k} \quad ; f_i = \frac{1}{\sum_k e^{-\beta U_i(\xi_k)} p_k}$$

where f_i are the free energy weighting factors, which depend on the solution p_k (this is an iterative procedure), the total number of configurations is N_i and each simulation, *i*, has n_i stored configurations per *kth* bin of the discretised histogram.

3.2.1 Metadynamics and adaptive biasing force

Other methods have also employed the adaptive umbrella method of Mezei. [15] A popular adaptive umbrella potential method, metadynamics, developed by Laio and Parrinello uses a combination of Gaussians as a biasing potential from which coarse-grained forces are calculated to direct the system to regions that have not been previously sampled.[18] Darve and Pohorille's ABF routine for thermodynamic integration continuously updates the biasing forces at every step during the simulation. [19, 20]

The simplest form of metadynamics[27] is 'direct metadynamics'. A collective variable (CV, essentially a reaction coordinate) is chosen and MD commences. Based on the sampling history of the CV, S(x), a small repulsive Gaussian potential is added every τ_G steps of simulation according to Equation 3-13. This external Gaussian potential, V_G depends on the Gaussian height, *w,* Gaussian width δs and frequency at which the Gaussians are added as follows

Equation 3-13

$$V_G(S(x),t) = w \sum_{\substack{t'=\tau_G,2\tau_G,\ldots \\ t'<t}} e^{\left(\frac{(S(x)-s(t'))^2}{2\delta s^2}\right)}$$

where $s(t) = V_G(S(x),t)$ is the value of the CV at a time t.

The choice of these parameters influences the accuracy of these simulations and the rate at which the system moves into new regions of phase space. Testing the efficacy of these parameters to a system

under study is most likely required. This is unlike the FEARCF approach where the accuracy is dependent on the bin size and the previous sampling of phase space.

In a similar fashion to FEARCF, the main algorithmic process involves calculating a CV, *s*, which is stored every T_G steps. Concurrently the derivative of V_G with respect to the CV, $(\partial / \partial x)V_G(S(x),t)$, is calculated and added to the forces.

A file containing the position of all the Gaussians is kept and this can be used to visualise the $V_G(s)$, which is the inverse of the free energy. All of these Gaussians are of the same height and width. As the number of Gaussians increases so does this file. To prevent the generation of large files, a gridding approach is used to store the potential, V_G, which depends on all the Gaussians.

An extension to this method, Lagrangian metadynamics, has been utilised to study reactions and ring puckering. [18, 27-29] Often a large amount of energy is applied to the chosen CV's, this may result in dynamic instabilities during dynamics or an inhomogenous temperature distribution. This is addressed by a Lagrangian coupling scheme where the CV's are treated as new dynamical variables, $s_\alpha(x)$, that are coupled to the system under consideration with harmonic restraining potentials of the form $\sum_\alpha 0.5k_\alpha[s_\alpha(x)-s_\alpha]^2$ where s_α is the 'equilibrium position' for the CV and each k_α is a force constant. A fictitious energy is also assigned $\sum_\alpha 0.5M_\alpha s_\alpha^2$ such that the Lagrangian of the system, *L*, is expressed as the original Lagrangian L_0, the Gaussian potential, V_G, the harmonic restraining potentials and the fictitious kinetic energy, as follows

Equation 3-14

$$L = L_0 + \sum_\alpha 0.5M_\alpha s_\alpha^2 - \sum_\alpha 0.5k_\alpha[s_\alpha(x)-s_\alpha]^2 + V(t,s)$$

where the CV's s_α each have a kinetic force constant k_α and a fictitious mass M_α and V takes on a different form (details in [29]). The fictitious mass can be chosen such that the evolution of the CV's is not erratic while the force constant k is a parameter chosen to be smaller than the frequency at which the collective variables oscillate. Examples of CV's that can be chosen in this approach include geometry relate variables such as distances, angles and dihedrals; coordination numbers; the potential energy and normal modes.

The ABF method is similar to the FEARCF method as external forces derived from the derivative of the free energy force the reaction coordinate to achieve improved sampling. In the ABF method these adaptive forces are calculated on the fly. The biasing force is estimated locally from sampled conformations of the system gained from previous steps in the simulation and it applied as soon as a bin has enough samples to estimate the mean force. The equation governing the PMF is of the form

Equation 3-15

$$A(\xi^*) = -k_B T \log P(\xi^*)$$

for an order parameter ξ^*, and the mean force, F_ξ, is

Equation 3-16

$$\left\langle F_\xi \mid \xi^* \right\rangle = -\frac{dA(\xi^*)}{d\xi}.$$

The sample size is usually small (~100) and the overall sampling is improved by applying an external biasing force of the form $-\left\langle F_\xi \mid \xi^* \right\rangle \nabla\xi$ or $\lambda\nabla\xi$, where λ is a Lagrange multiplier that can be related to the derivative of the free energy $\frac{dA(\xi^*)}{d\xi}$. These forces are then applied along generalised system coordinates. The result is that the instantaneous force acting on the coordinates of interest when subtracted from the equations of motion results in zero acceleration along the reaction coordinate. To reduce errors due to incomplete sampling of bins, a ramp function is applied. Due to this, it is plausible that the ABF method could converge faster in cases where the statistical errors arising from sampling are not optimal for all bins but this has not been verified. In comparison, the FEARCF method calculates new biasing forces over all bins from the histogram, $P(\xi)$, which is constructed after each iteration together with all previous iterations using the WHAM procedure. This is done numerically using a multidimensional cubic-spline routine. The major advantage of the FEARCF method is therefore that it is easily extended to multiple dimensions, while in the ABF method it is not straightforward to reconstruct $A(\xi)$ from its derivatives.

3.3 Reaction Coordinate Requirements for FEARCF

In order to calculate sampling for some specified reaction coordinate(s), ξ, this reaction coordinate and its derivative(s) must be calculable. A simple example coordinate is a bond distance, R, between two atoms A and B. During a molecular dynamics calculation the system evolves and so does R. At each dynamics step R must be calculated and recorded. This is calculated using

Equation 3-17

$$R = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2 + (z_A - z_B)^2} \,.$$

For subsequent molecular dynamics calculations, where a guess of the pmf, W(ξ), is known, R must be calculated, recorded and the forces must be calculated and applied. The correct force is calculated from the splined derivative dW(ξ)/dR which is found by a cubic splining procedure using the value of R on *W(ξ)*.

The reaction coordinate is calculated for each step of a dynamics simulation. The correct force is calculated using the reaction coordinates as "look-up values" for a splining routine and the force is calculated and applied. The force $F(x_A)$ is derived from $\frac{dW}{dR}$ and the chain rule yields

Equation 3-18

$$F(x_A) = \frac{dx_A}{dR} = \frac{dW}{dR}\frac{dx_A}{dW} \,.$$

3.4 Describing The Pucker of N-Membered Cycles

Carbohydrates, in addition to the conformational freedom about their secondary and primary alcohols, have a special degree of conformational freedom – ring puckering. This is the twisting, bending or flipping of the cycle as discussed in the introductory chapter. 5-membered cycles, such as cyclopentane and ribose, have 20 non-planar conformers - 10 envelope and 10 twist conformers and a planar conformer. These are depicted in Figure 3-2. For a 6-membered cycle such as glucose the conformers available are shown in Figure 1-8. There are 38 canonical conformers (as defined by IUPAC[30])

that a 6-membered cycle can conform to (the unfavourable planar conformer is rarely considered) these include chair, boat, skew-boat, half-chair and envelope.

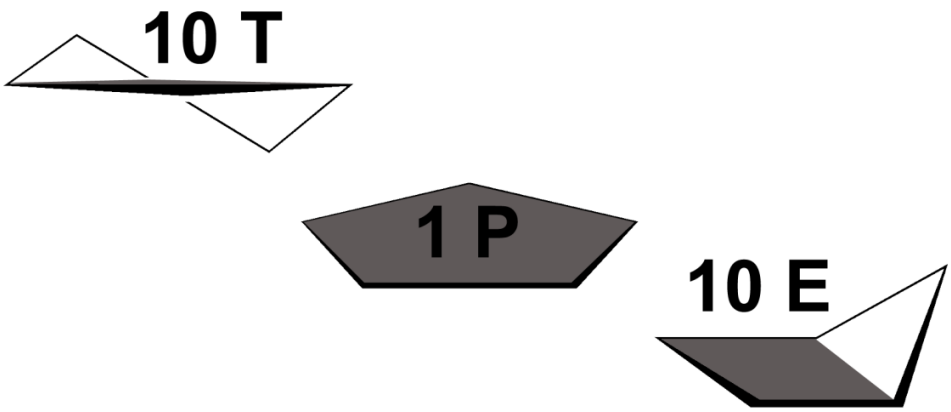

Figure 3-2 The 21 canonical conformers available to a five membered ring: Ten envelopes (E) and Ten twists (T). The shaded area connects atoms in the same plane. A planar conformer also exists.

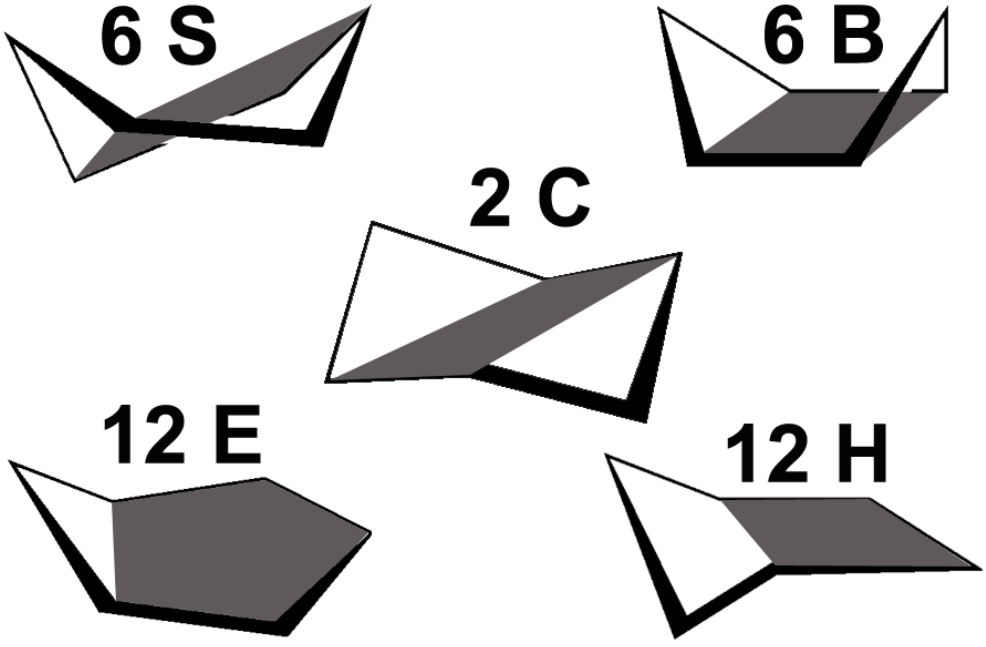

Figure 3-3 The 38 canonical conformers available to a six membered ring: Two chairs (C), six boats (B), six twists (S), twelve half-chairs (H) and twelve envelopes (E). The shaded area connects atoms in the same plane.

Ring puckering depends on all N atoms (in this context N is the number of atoms in the cycle and not the total number of particles) that define the cycle which means that there are N-degrees of freedom. It

seems that even for small cycles, not having very many degrees of freedom, that a description, calculation or analysis of ring pucker should be immensely complex. This is not the case as a reduced coordinate sets can be employed to describe ring puckering in terms of lower dimensional space. The first of these sets was introduced by Kilpatrick, Pitzer, and Spitzer[31] in their 1947 paper where they considered out-of-plane distortions of a pentagon. Other key researchers included Joshi[32], Boeyens[33, 34], Geise[35], Sundaralingam[36], Altona[35, 36] and Dunitz.[37, 38] However, the analysis and categorisation of the puckering of molecular rings is marked by the influential work of Cremer and Pople[39], who in 1975, published a paper describing a modification to the Kilpatrick, Pitzer, and Spitzer method applicable to general monocyclic puckered ring. Since then several researchers[28, 40-43] have shown interest in the calculation and designation of puckering for monosaccharides.

The basic concept involved is that the movement of the N-cycle is decomposed from N-dimensional original coordinates to a N-3 dimensional reduced coordinate set. Cremer and Pople, for instance, used the concept of normal out of plane modes and with this any N-cycle can be described by N-3 coordinates. More recently, Bérces[43] wished to describe the ring motion in terms of 3 dihedral angles, while Hill and Reilly[44] used a triangular decomposition scheme decomposing the an N-cycle into N-3 triangular flaps.

The Cremer-Pople method is undoubtedly the most familiar method and is widely applied to analyse ring conformations determined from X-ray structures[45, 46] or to determine pucker conformations from molecular dynamics simulations. One of the more recent applications of this method is as the source of molecular pucker coordinates, to create a colourful visual aid[47] for better recognition of pucker conformations when utilising the program Visual Molecular Dynamics (VMD).[48-50] While appreciating the impact of other methods, the Cremer-Pople and Hill-Reilly sets of reduced coordinates are the focus here; the Cremer-Pople method as it is widely known and the Hill-Reilly because it is conceptually simple to implement.

3.4.1 The Cremer-Pople reduced coordinate set

The Cremer-Pople set is a popular tool for describing the pucker of 6-membered rings and can describe 5-membered pucker. Cremer and Pople describe the conformers of a N-membered cycle using a normal out-of-plane modes analysis. Assuming a mean plane, the deviations of the N-cycle's atoms from this plane are calculated using the sum of sinusoidal functions.

The algorithm[39] is briefly described here for an N atom cycle. For an N atom cycle having *j* position vectors R_j =(x_j, y_j, z_j), where *j*=1, .. N , the geometrical centre is placed at the origin, that is

Equation 3-19

$$\sum_{j=1}^{N} R_j = 0$$

The deviations of the atoms in the N-cycle are relative to an as yet undefined mean plane. This mean plane is chosen such that it passes through the newly chosen origin. Its orientation is fixed perpendicular to the molecular z axis where

Equation 3-20

$$\sum_{j=1}^{N} z_j = 0$$

and the orientation of the mean plane is fixed uniquely by applying

Equation 3-21

$$\sum_{j=1}^{N} z_j \sin[2\pi(j-1)/N] = 0; \quad \sum_{j=1}^{N} z_j \cos[2\pi(j-1)/N] = 0.$$

It is possible to determine the mean plane from the position vectors, R_j, by applying

Equation 3-22

$$R' = \sum_{j=1}^{N} R_j \sin[2\pi(j-1)/N]$$

$$R'' = \sum_{j=1}^{N} R_j \cos[2\pi(j-1)/N]$$

where *R'* and *R''* give the y and x components of the position vectors respectively. The normal vector, *n*, defines the positive direction of the molecular z-axis perpendicular to the mean plane (the "topside" of the ring) by

Equation 3-23

$$n = R' \times R'' / \left| R' \times R'' \right|$$

The full set of displacements for each atom, z_j, from the mean plane are

Equation 3-24

$$z_j = R_j \bullet n .$$

From these equations, generalised puckering coordinates, q_m and ϕ_m can be defined such that if the number of ring atoms, N, is odd and N>3 then

Equation 3-25

$$q_m \sin\phi_m = -\sqrt{2/N} \sum_{j=1}^{N} z_j \sin\left[2\pi m(j-1)/N\right]; \quad q_m \cos\phi_m = \sqrt{2/N} \sum_{j=1}^{N} z_j \cos\left[2\pi m(j-1)/N\right].$$

These equations apply for m= (2, 3, … , (N-1)/2), where q_m are puckering amplitudes $\left(q_m \geq 0\right)$ and ϕ_m are puckering phase angles $\left(0 \leq \phi_m < 2\pi\right)$. If the number of ring atoms N is even, then Equation 3-25 is applicable up till m=((N/2)-1) and there is a single puckering amplitude coordinate $q_{N/2}$ described as

Equation 3-26

$$q_{N/2} = N^{-1/2} \sum_{j=1}^{N} z_j \cos\left[\pi(j-1)\right] = N^{-1/2} \sum_{j=1}^{N} (-1)^{j-1} z_j$$

which may take on either sign. The total number of puckering coordinates is then N-3 for an N atom cycle. The normalisation factors in Equation 3-25 and Equation 3-26 are such that

Equation 3-27

$$\sum_{1}^{N} z_j^2 = \sum_{m} q_m^2 = Q^2$$

where the quantity Q $\left(Q \geq 0\right)$ is termed the total puckering amplitude.

For 6-membered cycles such as hexoses, there are three puckering coordinates, an amplitude-phase pair *(q_2, ϕ_2)* and a single puckering coordinate q_3. These coordinates are popularly expressed as a spherical polar set (Q, θ, ϕ) in terms of the total pucker amplitude (Q), a phase angle (ϕ) which in combination with another angle θ ($0 \leq \theta \leq \pi$), where

Equation 3-28

$$q_2 = Q\sin\theta; \quad q_3 = Q\cos\theta$$

describe the chair, half-chair, envelope, twist-boat and boat conformations (Figure 3-4).

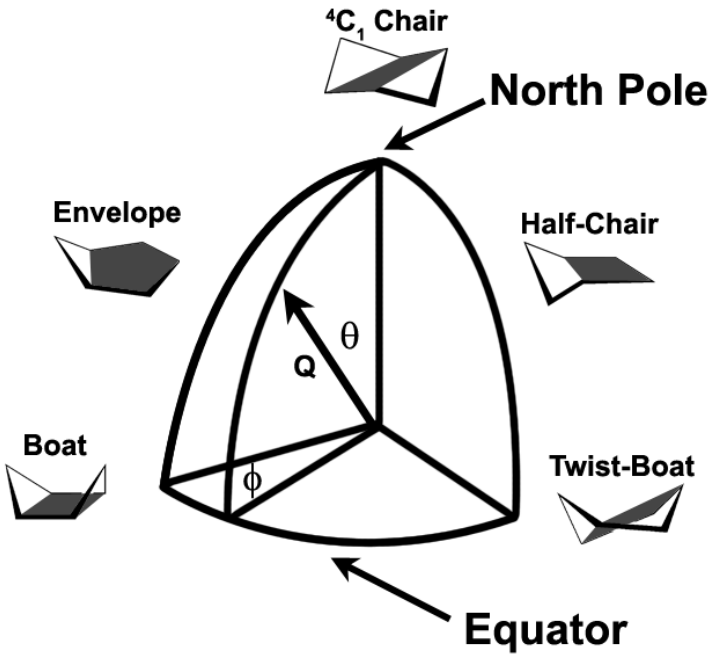

Figure 3-4 A northern hemisphere octant of a sphere (at constant Q) on which the 6-membered ring conformers can be mapped using the Cremer-Pople set.

Although forces can be applied to calculate free energies using this method, it does not lend itself well to the force-based approach. The relation between the applied reaction coordinate force and the geometrical change it induces on the geometry / topology of the furanose / pyranose ring is not simple.

3.4.2 The Hill-Reilly reduced coordinate set

Hill and Reilly proposed an intuitive description of ring puckering[44], where for an N-membered ring they reduced the Cartesian coordinates to N-3 ring pucker coordinates using a triangular decomposition scheme. Similar to the Cremer-Pople definition, a puckering conformation is described as deviations from a plane. A monocyclic ring is decomposed into rotatable planes and a reference plane. The pucker coordinates are specified by the angle, θ_i, that the rotatable planes make with respect to the reference plane. This method is suitable for a force based approach because it is easy to understand how a force applied to the rotatable plane leads to geometrical change of the ring. The Hill-Reilly coordinates are the angles of rotation (θ_i) about the puckering axes a_i as illustrated in Figure 3-5. The planes are constructed using the atomic positions x_i, where the i atoms are labelled i=(0, 1, …, N), the rotations axes a_i between planes, the bond vectors r_i, the normal n to the central plane.

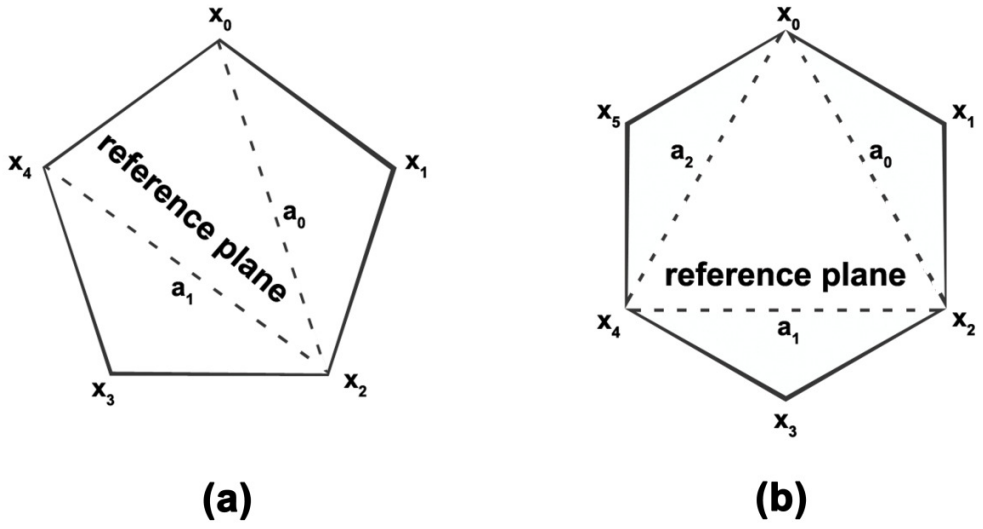

Figure 3-5 (a) A five-membered ring viewed from above with the axes of rotation, a_i, using the Hill-Reilly triangular decomposition implementation of ring puckering labelled. The rotating planes (r.p.) are measured as rotating with respect to the reference plane (formed from atoms x_0, x_2, x_4). The puckering axes are designated as a_i. (b) as for (a) but for a six-membered ring

The algorithm[44] is briefly described here for an N cycle. The axes of rotation, a_i, are calculated based on the atomic positions, x_i, using

Equation 3-29

$$a_i = x_{2(i+1)} - x_{2i}$$

where $i = i \bmod 6$ due to the cyclic nature of numbering.

The vector normal to the reference plane, n, (this is not the same as the Cremer Pople derivation of a mean plane) is calculated as the cross product of two axes with each other

Equation 3-30

$$n = a_1 \otimes a_0 .$$

Bond vectors, r_i, representing the bonds between adjacent atoms are calculated with

Equation 3-31

$$r_i = x_{i+1} - x_i$$

An atom's orientation vector, p_i, orthogonal to a plane defined by the bond vectors is calculated as

Equation 3-32

$$p_i = r_{i-1} \otimes r_i$$

where p_i is also orthogonal to the axis vectors, a_i.

To calculate the angle of puckering with an appropriate sign, a vector, q_i, orthogonal to a specific p_i and a particular axis, a_i, is defined as

Equation 3-33

$$q_i = a_i \otimes p_{2i+1} .$$

The angle of intersection of q_i and n is used to calculate the angle of puckering, as this yields $\pi/2\text{-}\theta_i$, where for the rotatable flaps (planes) above the reference plane θ_i is positive, while for flaps below the reference plane θ_i is negative. θ_i is calculated as

Equation 3-34

$$\theta_i = \pi/2 - \cos^{-1}[(q_i \bullet n) \bullet (\|q_i\| \bullet \|n\|)^{-1}]$$

and illustrated in Figure 3-6.

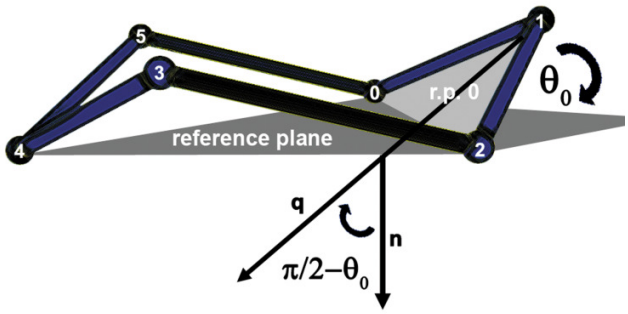

Figure 3-6 Description of pucker angle relative to the reference plane and rotatable plane 0 (r.p.0).

These equations are valid for N≤ 6 but must be modified for N>6.[44] Hill and Reilly derived the puckering coordinates for the canonical conformers of a hexose based on a cyclohexane model and compared these with the equivalent Cremer-Pople coordinates. [44]

In this thesis, 5 and 6-membered rings are considered and the triangular decomposition scheme(Hill-Reilly definitions) as illustrated in Figure 3-5. For N=5 there are two puckering coordinates θ_0, $\theta_1 \in$ [-π/2, π/2] and for N=6 there are three puckering coordinates θ_0, θ_1, $\theta_2 \in$ [-π/2, π/2] where a positive angle indicates rotation of a specific rotatable plane above the reference plane.

3.5 Incorporation of Ring Puckering into FEARCF

The Cremer-Pople method was not chosen as the default methodology for FEARCF ring puckering. Decomposition of the ring into Cremer-Pople coordinates is not intuitive – that is the when considering the pucker of a N-membered cycle it is not easy to visualise how the Cremer-Pople coordinates are directly related to the conformation. Furthermore, the forces to be applied to change between puckering conformers in Cremer-Pople space are not geometrically intuitive. Finally, the worst problem with the implementation of this method into the FEARCF methodology is that the q_i coordinates are not bound (closed). FEARCF will improve sampling in unsampled regions resulting in extensively large values of q_i (or Q) being sampled. Large q_i coordinates represent highly distorted pucker conformations. It is

possible to rectify this by applying a restraining potential to this coordinate. [28] However, this method is too laborious to use in this context.

The Hill-Reilly triangular decomposition set for N-cycles is more intuitive. The coordinates themselves can be used to construct a puckered N-cycle simply and the application of forces results in movement of the puckering flaps. The construction of these coordinates is such that all coordinates are angular and form closed sets $\theta_0, \theta_1, \theta_2 \in [-\pi/2, \pi/2]$. Hill-Reilly triangular decomposition is a sensible set to include in an adaptive force based method (FEARCF) for calculating the potential of mean force of ring pucker.

As mentioned earlier in §3.2 and §3.3, there are two parts to the FEARCF protocol within the simulation itself. Firstly the pucker coordinate must be calculated. This is based on the choice of the user who may choose N ring atoms and order the ring atoms to form the Hill-Reilly plane and flaps; this choice is important for the analysis of ring puckering.

Based on the chosen atoms the Hill-Reilly pucker coordinates are calculated as described above, with the final calculation being

Equation 3-35

$$\theta_i = \pi/2 - \cos^{-1}[(q_i \bullet n) \bullet (\|q_i\| \bullet \|n\|)^{-1}]$$

At each step of the simulation the biasing force for each θ_i must be calculated and applied to the atoms involved in the rotatable planes *i*. The FEARCF forces are applied to the ring in order to further the traversal of pucker coordinate space and this is the gradient of the umbrella potential for that puckering coordinate:

Equation 3-36

$$-\frac{\partial U(\xi)}{\partial \theta_i} = F(\theta_i)$$

where the force, $F(\theta_i)$, acting on θ_i is calculated as the partial derivative $\frac{\partial U(\xi)}{\partial \theta_i}$ of the applied biasing reaction coordinate umbrella potential. The result being that the *i*th ring plane is biased away

from the areas it has sampled with a force opposite to the accumulated sampling density as shown in Equation 3-36 and Figure 3-7.

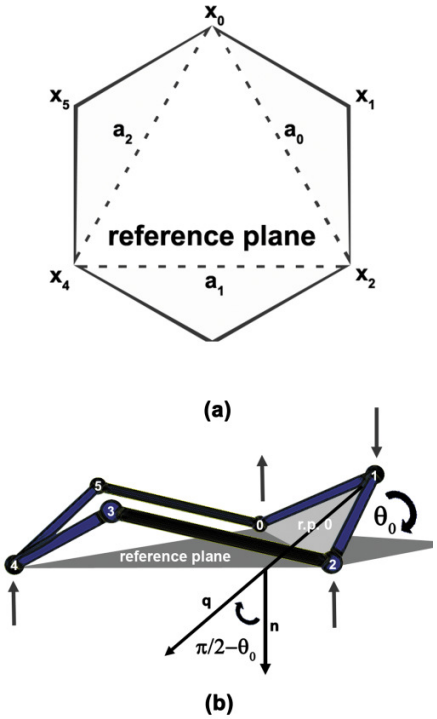

Figure 3-7 (a) A 6-membered ring viewed from above with the axes of rotation using the Hill-Reilly implementation of ring puckering. The rotating planes (r.p.) are measured as rotating with respect to the reference plane (formed from atoms x_0, x_2, x_4). The puckering axes are designated as a_i. (b) Description of pucker angle relative to the reference plane and rotatable plane 0 (r.p.0). A simple example of a negative force (to bend the flap towards the reference plane) applied to theta and perpendicular to the reference plane will apply forces to atoms 0, 1, 2 and 4 as indicated.

When implemented in a molecular dynamics routine, the biasing force for each puckering coordinate is added to the forces from the equilibrium Boltzmann dynamics forces applied to each atom. The biasing forces, F_i, on each ring atom can be recovered from the reaction coordinate forces by recasting them in terms of the PMF, $W(\xi)$ as shown in Equation 3-37.

Equation 3-37

$$\frac{\partial W(\xi)}{\partial \theta_i} = F_i$$

By application of the chain rule, the reaction coordinate force F_i can be converted to atomic coordinate forces $\frac{\partial W(\xi)}{\partial x_i}$ where *i* is the atom number.

The reaction coordinate force can be rewritten in terms of the vector q_i using the chain rule

Equation 3-38

$$\frac{\partial W(\xi)}{\partial \theta_i}\frac{\partial \theta_i}{\partial q_i} = \frac{\partial W(\xi)}{\partial q_i}.$$

Similarly the reaction coordinate forces can be written in terms of the vector *n*,

Equation 3-39

$$\frac{\partial W(\xi)}{\partial \theta_i}\frac{\partial \theta_i}{\partial n} = \frac{\partial W(\xi)}{\partial n}.$$

On combination of these (and including all other contributions), the biasing forces ($\partial W(\xi)/\partial xi$) on the ring atoms x_i, originating from the pucker axis a_i (Figure 3-7), are calculated as

Equation 3-40

$$\frac{\partial W(\xi)}{\partial a_i}\frac{\partial a_i}{\partial x_{2(i+1)}} = \frac{\partial W(\xi)}{\partial x_{2(i+1)}}; \frac{\partial W(\xi)}{\partial a_i}\frac{\partial a_i}{\partial x_{2i}} = \frac{\partial W(\xi)}{\partial x_{2i}}$$

which are due to the reaction coordinate θ_i. The total biasing force is calculated by summing the contributions from each reaction coordinate which is then applied to the ring atoms. This can be calculated for each of the ring atoms *i*, where *i* are the labels (1 to N) for the chosen ring atoms. For a 5-membered ring a two dimensional reaction coordinate (ξ) is defined as the combination of these pucker coordinates i.e., $\xi=(\theta_0, \theta_1)$ and for a 6-membered ring, a three dimensional reaction coordinate (ξ)is defined as the combination of these pucker coordinates i.e., $\xi=(\theta_0, \theta_1, \theta_2)$.

3.6 References

1. C. B. Barnett and K. J. Naidoo, *Molecular Physics*, 2009, **107**, 1243 - 1250.
2. V. L. Schramm, *Annual Reviews in Biochemistry*, 1998, **67**, 693-720.
3. W. Pearlman and R. Rao, *Encyclopaedia of Computational Chemistry*, Wiley, Chichester, 1998.
4. K. J. Laidler, J. H. Meiser and B. C. Sanctuary, *Physical Chemistry*, Houghton-Mifflin, 1999.
5. A. R. Leach, *Molecular Modelling: Principles and Applications*, Longman, 2001.
6. R. W. Zwanzig, *Journal of Chemical Physics*, 1954, **22**, 1420.
7. W. L. Jorgensen, *Journal of Chemical Physics*, 1985, **6**, 3050-3054.
8. A. Kollman Peter, *Chemical Reviews (Washington, D. C.)*, 1993, **93**, 2395-3417.
9. S. Kumar, P. W. Payne and M. Vasquez, *Journal of Computational Chemistry*, 1996, **17**, 1269-1275.
10. J. P. Valleau and G. M. Torrie, *In Modern theoretical chemistry, Statistical mechanics A, Equilibrium techniques*, Plenum Press, New York, U. S. A, 1977.
11. J. P. Valleau and G. M. Torrie, *Journal of Computational Physics*, 1977, **23**, 187-199.
12. C. Bartels and M. Karplus, *Journal of Computational Chemistry*, 1997, **18**, 1450-1462.
13. C. Bartels and M. Karplus, *Journal of Physical Chemistry B*, 1998, **102**, 865-880.
14. R. W. W. Hooft, B. P. Van Eijck and J. Kroon, *Journal of Chemical Physics*, 1992, **97**, 6690-6694.
15. M. Mezei, *Journal of Computational Physics*, 1987, **68**, 237-248.
16. D. A. McQuarrie, *Statistical Mechanics*, University Science Books, California, 2000.
17. D. Chandler, *Introduction to Modern Statistical Mechanics*, Oxford University Press, Oxford, 1987.
18. A. Laio and M. Parrinello, *Proceedings of the National Academy of Sciences U S A*, 2002, **99**, 12562-12566.
19. E. Darve and A. Pohorille, *Journal of Chemical Physics*, 2001, **115**, 9169-9183.
20. E. Darve, D. Rodriguez-Gomez and A. Pohorille, *Journal of Chemical Physics*, 2008, **128**, 144120.
21. J. Strümpfer and K. J. Naidoo, *Journal of Computational Chemistry*, 2010, **31**, 308-316.
22. R. Rajamani, K. J. Naidoo and J. Gao, *Journal of Computational Chemistry*, 2003, **24**, 1775-1781.
23. S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen and P. A. Kollman, *Journal of Computational Chemistry*, 1995, **16**, 1339-1350.
24. S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman and J. M. Rosenberg, *Journal of Computational Chemistry*, 1992, **13**, 1011-1021.
25. A. M. Ferrenberg and R. H. Swendsen, *Physical Review Letters*, 1988, **61**, 2635-2638.
26. A. M. Ferrenberg and R. H. Swendsen, *Physical Review Letters*, 1989, **63**, 1195-1198.
27. L. Alessandro and L. G. Francesco, *Reports on Progress in Physics*, 2008, **71**, 126601.
28. X. Biarnés, A. Ardévol, A. Planas, C. Rovira, A. Laio and M. Parrinello, *Journal of the American Chemical Society*, 2007, **129**, 10686-10693.
29. M. Iannuzzi, A. Laio and M. Parrinello, *Physical Review Letters*, 2003, **90**, 238302.
30. *Pure & Applied Chemistry*, 1980, **53**, 1901-1905.
31. J. E. Kilpatrick, K. S. Pitzer and R. Spitzer, *Journal of the American Chemical Society*, 1947, **69**, 2483-2488.
32. N. V. Joshi and V. S. R. Rao, *Biopolymers*, 1979, **18**, 2993-3004.

33. D. G. Evans and J. C. A. Boeyens, *Acta Crystallographica, Section B: Structural Science*, 1989, **B45**, 581-590.
34. J. C. A. Boeyens and D. G. Evans, *Acta Crystallographica, Section B: Structural Science*, 1989, **B45**, 577-581.
35. H. J. Geise, C. Altona and C. Romers, *Tetrahedron Letters*, 1967, 1383-&.
36. C. Altona and M. Sundaralingam, *Journal of the American Chemical Society*, 1972, **94**, 8205-&.
37. J. D. Dunitz and J. Waser, *Journal of the American Chemical Society*, 1972, **94**, 5645-&.
38. J. D. Dunitz, *Tetrahedron*, 1972, **28**, 5459-&.
39. D. Cremer and J. A. Pople, *Journal of the American Chemical Society*, 1975, **97**, 1354-1358.
40. S. C. Harvey and M. Prabhakaran, *Journal of the American Chemical Society*, 1986, **108**, 6128-6136.
41. A. R. Ionescu, A. Bérces, M. Z. Zgierski, D. M. Whitfield and T. Nukada, *Journal of Physical Chemistry A*, 2005, **109**, 8096-8105.
42. T. Sato, *Nucleic Acids Research*, 1983, **11**, 4933-4938.
43. A. Bérces, G. Enright, T. Nukada and D. M. Whitfield, *Journal of the American Chemical Society*, 2001, **123**, 5460-5464.
44. A. D. Hill and P. J. Reilly, *Journal of Chemical Information and Modelling*, 2007, **47**, 1031-1035.
45. D. J. Patel and C. Shen, *Proceedings of the National Academy of Science*, 1978, **75**, 2553-2557.
46. K. Gohda, K. Oka, K. Tomita and T. Hakoshima, *Journal of Biological Chemistry*, 1994, **269**, 17531-17536.
47. S. Cross, M. M. Kuttel, J. E. Stone and J. E. Gain, *Journal of Molecular Graphics and Modelling*, 2009, **28**, 131-139.
48. W. Humphrey, A. Dalke and K. Schulten, *Journal of Molecular Graphics*, 1996, **14**, 33-38.
49. J. Stone, Computer Science Department, University of Missouri-Rolla, 1998.
50. M. Sanner, A. Olsen and J.-C. Spehner, in *Proceedings of the 11th ACM Symposium on Computational Geometry*, ACM, New York, 1995, pp. C6-C7.

4 Semi-empirical Methods for Carbohydrates*

4.1 Introduction

Carbohydrate monomers are conformationally very complex. The complexity stems from the possibility of multiple linkages that can be formed via the numerous hydroxyl groups decorating the periphery of the flexible monosaccharide rings and the wide range of pucker conformers available to these five membered (furanose) and six membered (pyranose) rings. This conformational diversity of the basic monomers, used to construct oligosaccharides and biopolymers, has led to the extensive presence of carbohydrates in nature and a diverse range of functions. In cellular systems these roles range from being the primary source for fuel, to acting as metabolic intermediates in cellular formative pathways and being structurally important in the genetic (RNA/DNA) framework.

The nature of furanose and pyranose ring conformational pucker is critical in the formation of transition states found along the reaction path catalysed by phosphorylases[1], ribosyl transferases, hydrolases[2] and dehalogenases.[3] Pyranose rings play a prominent role in carbohydrate processing enzymes, including glycosidases (the enzymes responsible for the breakdown of di-, oligo- and polysaccharides) and glycosyltransferases (the enzymes which transfer saccharides to other saccharide moieties). Glycoside chemistry is largely the chemistry of the oxocarbenium ion formed in the transition state (TS) as shown in Figure 1-11.

* This chapter is adapted from a paper in review : Barnett, C. B. & Naidoo, K. J. Ring puckering : a metric for evaluating the accuracy of AM1, PM3, PM3CARB-1 and SCC-DFTB carbohydrate QM/MM simulations. Journal of Physical Chemistry B In Review (2010).

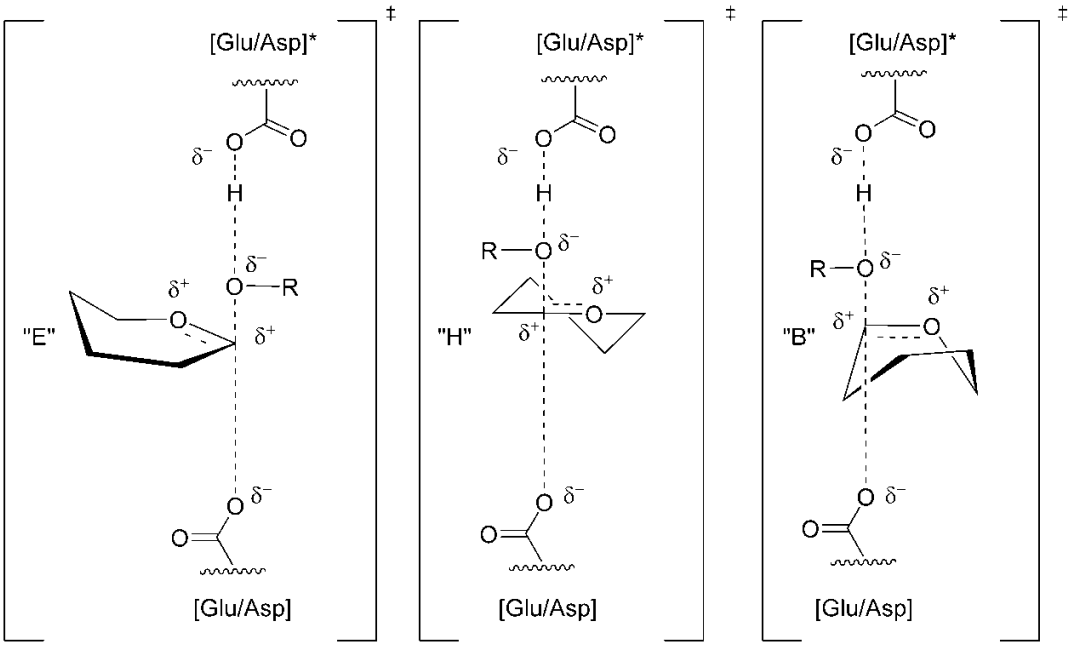

Figure 4-1 The possible oxocarbenium pyranose transition states that may occur in acid catalysed processes. The anomeric carbon, which has planar character may only exist in envelope ("E"), half-chair ("H") or boat ("B") conformers as represented by figures on the left, centre and right respectively. The specific conformers that can be accessed are $^{4}H_{3}$, $^{3}H_{4}$, $^{4}H_{5}$, $^{5}H_{4}$, $^{2}H_{3}$, $^{3}H_{2}$, ^{4}E, E_{4}, ^{3}E, E_{3}, $^{2,5}B$, $B_{2,5}$. * indicates that the amino acid is protonated.

In the case of five membered rings, an example of the importance of ring conformations is found in the mechanism of the base excision repair (BER) enzymes. These enzymes remove nucleobases from DNA in the first step of the repair of many types of DNA damage.[4] The mechanisms of N-glycoside hydrolysis and N-glycosyl transfer reactions have similarities to the O-glycoside (polysaccharide) hydrolyses. As in the pyranose case there is an accumulation of positive charge on the sugar ring at the TS. Hyperconjugation stabilises the carbocation when the dihedral angle between the C2'-H2' bond and the developing empty p-orbital on C1' is at 0° or 180°.[5] The hyperconjugation in furanose oxocarbenium ions therefore significantly depends on the ring pucker conformation. When a strong π-bond develops between C1' and the ring oxygen, O4' one of two envelope conformations, 3'-endo or 3'- exo (Figure 4-2) could form.[4]

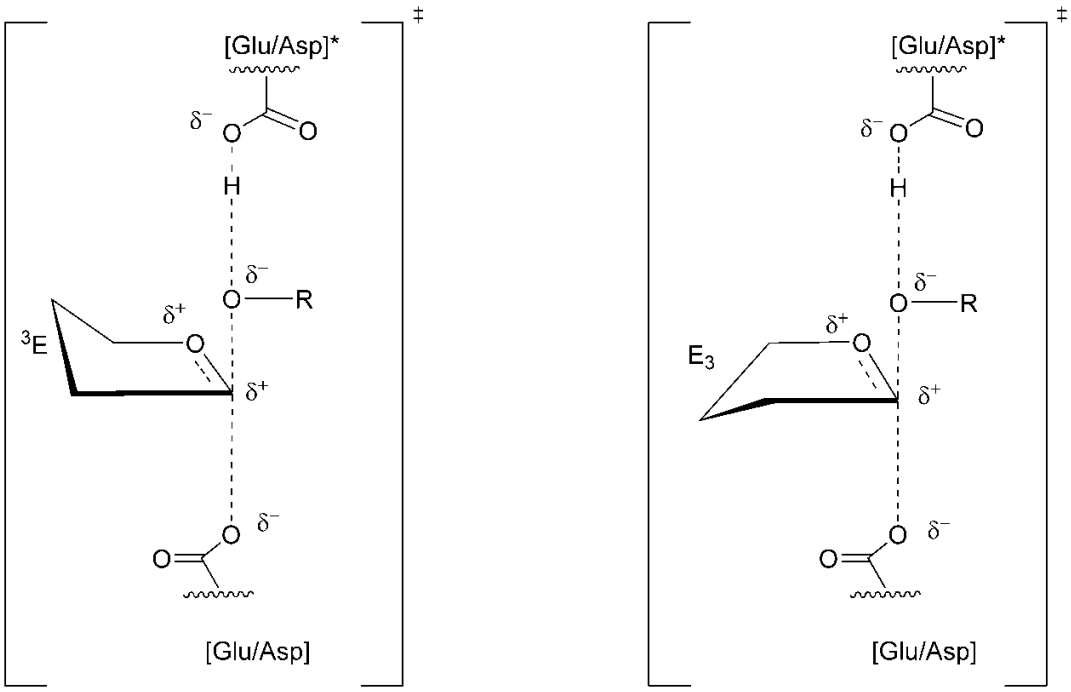

Figure 4-2 The possible oxocarbenium furanose transition states that may occur in acid catalysed processes. The anomeric carbon, which has planar character may only exist as the ^3E (also known as 3-endo) or E_3 (3-exo) conformers. * indicates that the amino acid is protonated.

The investigation of reactions in carbohydrate processing enzymes via experiments has limited possibilities. While the mechanism can be probed with i) repeated X-ray structural analysis of mutated enzymes or enzymes bound to inhibitors[6, 7] and ii) Kinetic Isotope Effect (KIE) experiments[8] the nature of the TS remains inaccessible to experimentalists. Therefore the use of computational methods, particularly hybrid quantum classical (QM/MM) ones, has become central to the investigation of enzyme reaction mechanisms and the conformational and electronic nature of the TS.

Herein lies the problem when simulating carbohydrates. Unless the semi-empirical methods used in QM/MM simulations are able to model the intricate conformational and electronic transitions of furanose and pyranose monosaccharides accurately, the mechanistic details and TS structures derived from these simulations may be meaningless. This matter is addressed here by calculating the complete free energy of pucker surfaces for ribose and the free energy of pucker volumes for glucose monosaccharides using AM1[9], PM3[10], PM3CARB-1[11] and SCC-DFTB[12] semi-empirical methods. These methods are then compared with each other, to selected DFT RB3LYP/6-311++G(d, p) stationary structures and to a free energy of ring puckering 6-31G Hartree-Fock surface.

4.2 Pucker Definitions

Five membered rings can be classified into envelopes (E) and twists (T) similarly six membered rings can be classified into chairs (C), boats (B), skew-boats (S), the form between two boats, envelopes (E) and half chairs (H).[13] The mathematical definition of these conformers was based on a spherical polar coordinate system which was introduced by Cremer and Pople in their classic 1975 publication.[14] These definitions proved very popular despite the cumbersome relationship between these coordinates and physically meaningful stresses and strains on the rings. More recently, Hill and Reilly proposed a triangular decomposition coordinate set that enables a ready description of ring conformers as a function of triangular planes deviating from a reference plane placed on the monosaccharide ring.[15]

While other descriptions of ring puckering have been proposed[16, 17] and are in popular use[14] the triangular decomposition[15] has been found to be particularly amenable to the Free Energy from Adaptive Reaction Coordinate Forces (FEARCF) method.[18] Furthermore, the decomposition of a monocyclic ring into a reference plane and rotatable (puckering) planes translates into an easily understood physical description.[15, 19] For an N-membered ring there are N-2 planes; a central reference plane and N-3 rotatable puckering planes where the puckering coordinates are the angles between each of the rotatable planes with respect to the central plane. For example a 6-membered ring has a central plane and 3 puckering planes such that there are three puckering coordinates $\theta_0, \theta_1, \theta_2 \in$ [-90°, 90°]. The angle of puckering is calculated from

Equation 4-1

$$\theta_i = \pi/2 - \cos^{-1}[(q_i \bullet n) \bullet (\|q_i\| \bullet \|n\|)^{-1}]$$

where q_i is a vector normal to the rotatable plane *i* and the axis about which the plane rotates, while ***n*** is the vector normal to the reference plane (shown in Figure 4-3a for pyranose ring).

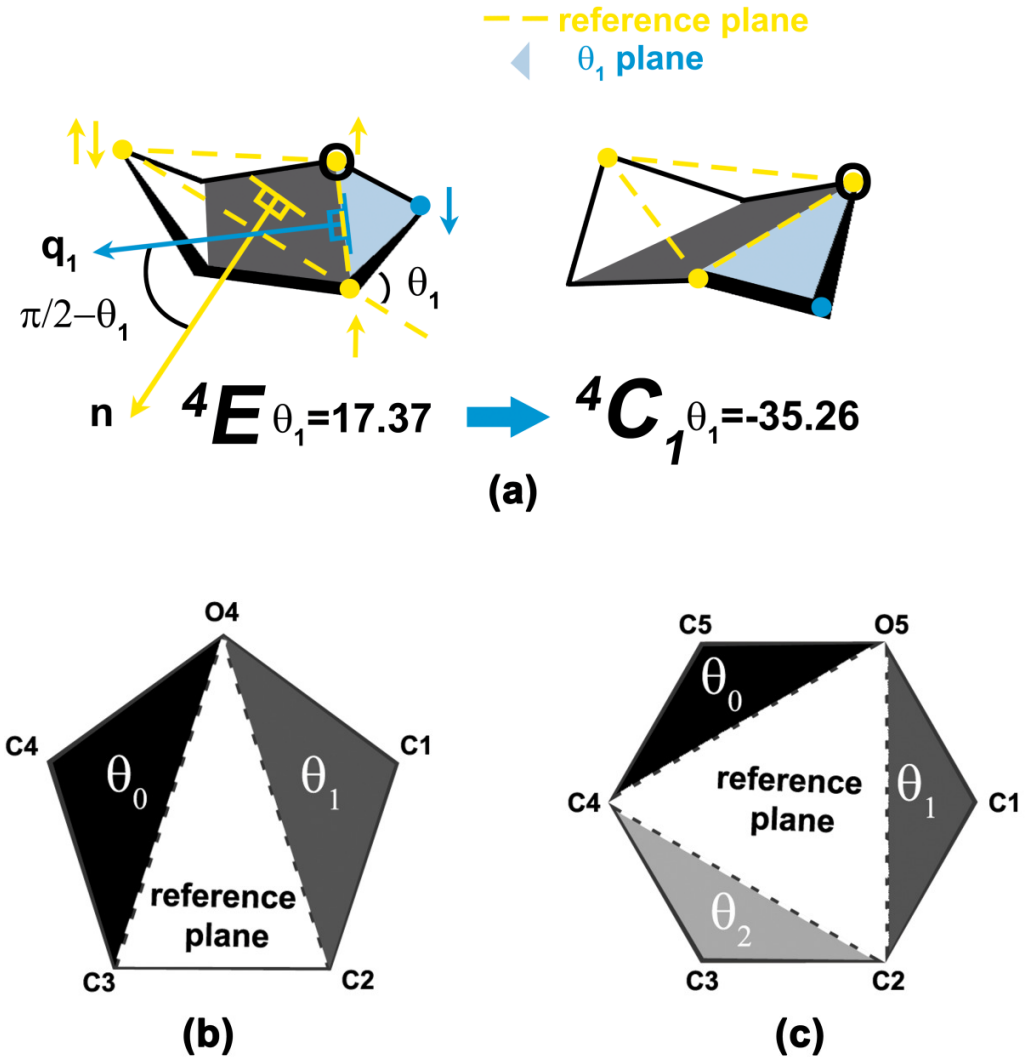

Figure 4-3 (a) FEARCF forces (shown as short arrows) originating from a rotatable plane (blue) using triangular decomposition coordinates for a pyranose ring showing a transition from ^4E conformation to the $^4C_1$ conformation. The reference plane and rotatable planes labeled for (b) β-D-ribose (c) and β-D-glucose.

These coordinates are illustrated for ribose and glucose in Figure 4-3b and c respectively. The reference plane used in the triangular decomposition[15] is not the same as the mean plane used in the Cremer-Pople[14] definition or that of the IUPAC nomenclature for carbohydrate rings. Nonetheless, as with Cremer-Pople coordinates, these puckering coordinates can be mapped to the IUPAC canonical conformers.[15] The mapping for furanose is shown in Figure 4-4 while a list of the full set of pucker coordinate values is shown in Table 4-1.

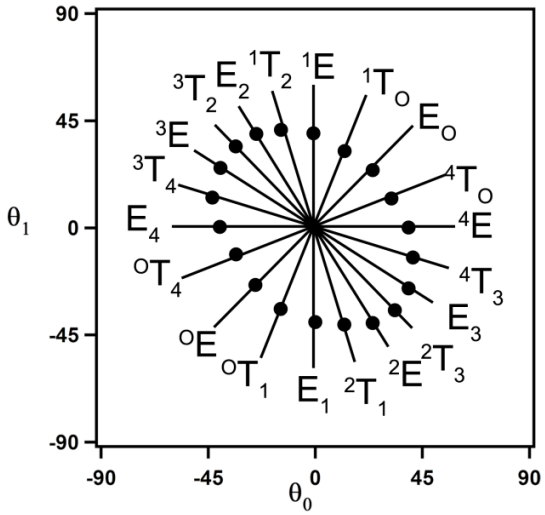

Figure 4-4 The triangular decomposition pucker space for 5-membered rings. The origin represents a planar ring. The bold nodes represent the optimised IUPAC conformations for a furanose ring (based on cyclopentane). The line segments branching from the origin represent a range of points at which a specific IUPAC conformation exists. There are 10 envelope, 10 twists and 1 planar conformer.

Table 4-1 The triangular decomposition pucker coordinates in degrees that correspond to ideal IUPAC conformers of cyclopentane.

Canonical Conformation	θ_0	θ_1
$^{O}T_1$	-13.10	-33.89
$^{3}T_4$	-42.16	13.21
$^{2}T_3$	34.50	-34.50
$^{1}T_2$	-13.21	42.16
$^{4}T_O$	33.89	13.11
$^{1}T_O$	13.10	33.89
$^{4}T_3$	42.16	-13.21
$^{3}T_2$	-34.50	34.50
$^{2}T_1$	13.21	-42.16
$^{O}T_4$	-33.89	-13.11
E_2	-24.88	40.00
E_1	0.00	-39.90
E_O	24.50	24.50
E_4	-39.50	0.00
E_3	39.50	-24.90
^{2}E	24.88	-40.00
^{1}E	0.00	39.90
^{O}E	-24.50	-24.50
^{4}E	39.50	0.00
^{3}E	-39.50	24.90
P	0.00	0.00

In Figure 4-5, the canonical conformers for pyranose are shown and a list of the full set of pucker coordinate values in are shown in Table 4-2. The exact mapping of the Hill-Reilly proposed pucker coordinates to canonical conformers does depend on the ordering of atoms used to define the reference and pucker planes. They chose C1-C3-C5 in their definition with C2 as the apex of the rotatatable puckering plane (or ring flap) defining θ_0, C3 as the apex of θ_1 of and O5 as the apex of θ_2. This nomenclature was followed in earlier work for glucose[19] and similarly for ribose.[20] However, the definitions used in this thesis are not the same as those used previously. Here, the free energy surfaces are presented such that they can be analysed as a function of the motion of the anomeric carbon and the ring carbon bearing the primary alcohol substituent. This information is necessary for understanding the relationship between the ring pucker and the oxocarbenium ion formation in the transition states of glycosidases and phosphorylases.

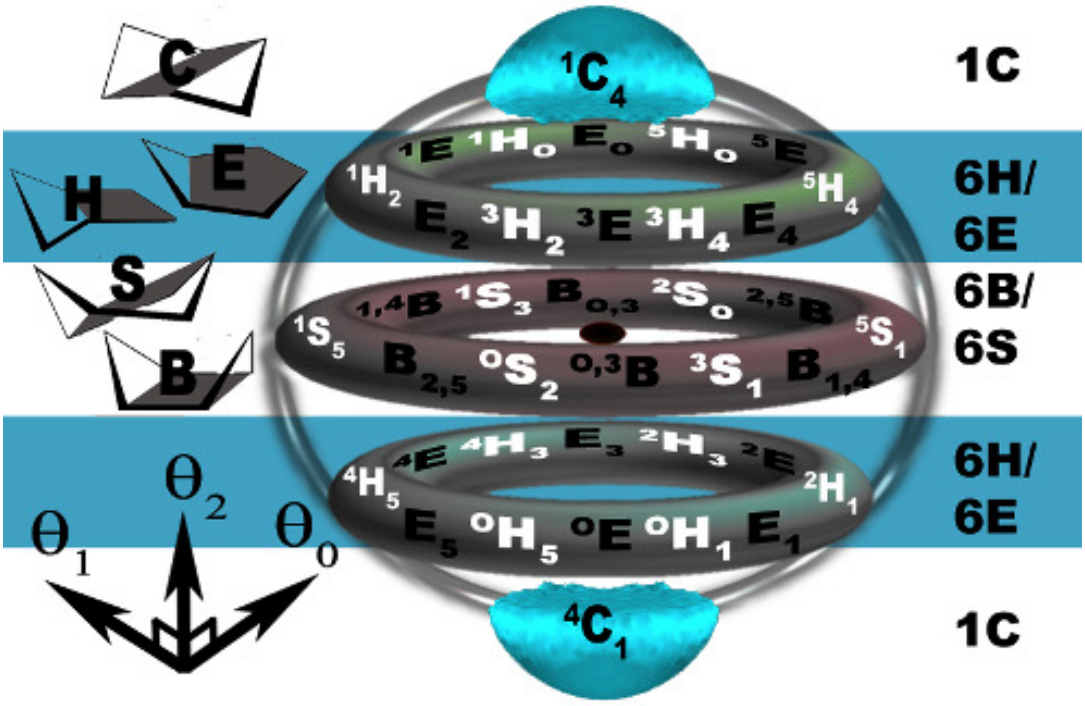

Figure 4-5 The triangular decomposition pucker volume accessible to 6-membered rings showing the positions of IUPAC labelled conformers. The origin represents a planar 6-membered ring, the axes drawn are with respect to the origin. The "poles" are chair conformers, the "tropics" are half-chair and envelopes, the "equator" has boat (sofa) and twist-boat (skew-boat) conformers.

Table 4-2 The triangular decomposition pucker coordinates in degrees for the C4-O5-C2 reference plane for a pyranose ring that corresponds to IUPAC canonical conformers.

Canonical Conformation	θ_0	θ_1	θ_2
$B_{2,5}$	-74.20	35.26	35.26
$^{1}S_{5}$	-50.84	50.84	0.00
$^{O}S_{2}$	-50.84	0.00	50.84
E_{5}	-46.86	0.00	0.00
$^{4}H_{5}$	-42.16	9.07	-17.83
$^{O}H_{5}$	-42.16	-17.83	9.06
$^{1,4}B$	-35.26	74.2	-35.26
$^{O,3}B$	-35.26	-35.26	74.2
$^{4}C_{1}$	-35.26	-35.26	-35.26
^{4}E	-35.26	17.37	-35.26
^{O}E	-35.26	-35.26	17.37
$^{4}H_{3}$	-17.83	9.07	-42.16
$^{O}H_{1}$	-17.83	-42.16	9.07
E_{2}	-17.37	35.26	35.26
$^{1}H_{2}$	-9.07	42.16	17.83
$^{3}H_{2}$	-9.07	17.83	42.16
^{1}E	0.00	46.86	0.00
$^{1}S_{3}$	0.00	50.84	-50.84
^{3}E	0.00	0.00	46.86
$^{3}S_{1}$	0.00	-50.84	50.84
E_{1}	0.00	-46.86	0.00
E_{3}	0.00	0.00	-46.86
$^{2}H_{1}$	9.07	-42.16	-17.83
$^{2}H_{3}$	9.07	-17.83	-42.16
^{2}E	17.37	-35.26	-35.26
$^{1}H_{O}$	17.83	42.16	-9.07
$^{3}H_{4}$	17.83	-9.07	42.16
$^{1}C_{4}$	35.26	35.26	35.26
$B_{1,4}$	35.26	-74.20	35.26
$B_{O,3}$	35.26	35.26	-74.2
E_{4}	35.26	-17.37	35.26
E_{O}	35.26	35.26	-17.37
$^{5}H_{4}$	42.16	-9.07	17.83
$^{5}H_{O}$	42.16	17.83	-9.06
^{5}E	46.86	0.00	0.00
$^{2}S_{O}$	50.84	0.00	-50.84
$^{5}S_{1}$	50.84	-50.84	0.00
$^{2,5}B$	74.20	-35.26	-35.26

The reference plane for the furanose ring was chosen as C3-O4-C2 and the two ring flaps were defined as C3-C4-O4 (θ_0) and O4-C1-C2 (θ_1) as shown in Figure 4-3b. Here, θ_0 describes the movement of the ring carbon bearing the primary alcohol, while θ_1 describes the movement of the anomeric carbon into and out of the reference plane. For the pyranose ring the atoms C4-O5-C2 were chosen as the reference plane (Figure 4-3c) and the three flaps were defined as C4-C5-O5 (θ_0), O5-C1-C2 (θ_1) and C2-C3-C4 (θ_2), θ_0 describes the movement of the carbon to which the primary alcohol moiety (C5) is attached, θ_1 describes the movement of the anomeric carbon (C1), and θ_2 describes the movement of a carbon (C3) to which a secondary alcohol is attached.

4.3 Computational Methods

Using the FEARCF method and the Hill-Reilly[15] definition, the free energy of puckering surface for β-D-ribose and volume for β-D-glucose rings *in vacuo* were calculated.

4.3.1 Calculation of the free energy of ring pucker

The reaction coordinate (ξ) for ribose is $\xi_{ribose}=(\theta_0, \theta_1)$ while for glucose it is $\xi_{glucose}=(\theta_0, \theta_1, \theta_2)$. The potential of mean force (PMF), $W(\xi)$ is then calculate as discussed in the previous chapter as a function of the (N-3)-dimensional coordinate set ξ (where N is the number of atoms in the ring) and is related to the probability density, $P(\xi)$ by

Equation 4-2

$$W(\xi) = -k_B T \ln P(\xi)$$

where k_B is the Boltzmann constant and T is the temperature in Kelvin. The PMF is derived in a canonical ensemble (constant NVT) giving the Helmholtz free energy. In the FEARCF method barrier heights separating conformational minima are traversed by applying an adaptive biasing force $F(\xi)$. $F(\xi)$ is derived from the reaction coordinate potential energy $U(\xi)$ defined from the histogram of previously sampled regions of reaction coordinate space.

Equation 4-3

$$U(\xi) = k_B T \ln P(\xi)$$

An indication that the sampled surface has converged is when $U(\xi)$ is equal to the inverse of the free energy function (i.e. $U(\xi)=-W(\xi)$). Therefore, when applying biased forces from the converged surface in a simulation the resulting sampling from a long simulation of reaction coordinate space should be uniform. The two-dimensional puckering coordinates were gridded into 5329 (73^2) bins of area 2.5°*2.5°. The three-dimensional puckering coordinates were gridded into 389017 (73^3) bins of volume 2.5°*2.5°*2.5°.

The FEARCF module is included into the macromolecular program CHARMM[21] to calculate the effect of the perturbing forces generated from $U(\xi)$ for torsional and chemical reaction coordinates.[22] This method is generalised for multiple dimensions[18] as discussed in Chapter 3 and its effectiveness has been shown for calculating pucker free energies.[19] At each step of the simulation the biasing force for θ_i applied to the atoms involved in the rotatable plane is calculated from the gradient of the reaction coordinate potential for that puckering coordinate. Which is

Equation 4-4

$$-\frac{\partial U(\xi)}{\partial \theta_i} = F(\theta_i)$$

where the force $F(\theta_i)$ is calculated as the partial derivative $-\frac{\partial U(\xi)}{\partial \theta_i}$ of the applied biasing reaction coordinate potential. The result is the *i*th ring plane is biased away from the areas it has sampled with a force opposite to the accumulated sampling density as shown in Equation 4-2 and Equation 4-3. When implemented in a molecular dynamics (MD) routine the biasing force for each puckering coordinate is converted to an atomic force in Cartesian coordinates and added to the forces from the equilibrium Boltzmann dynamic forces applied to each atom. The biasing forces on each ring atom can be recovered from the reaction coordinate forces by recasting them in terms of the PMF

Equation 4-5

$$\frac{\partial W(\xi)}{\partial \theta_i} = F(\theta_i)$$

The force can be applied to the atoms that make up the puckering reaction coordinate by using the chain rule to get the component forces ($\frac{\partial W(\xi)}{\partial x_i}$) applied to the ring atoms (x_i). The total biasing force is calculated by summing the contributions from each reaction coordinate that is then applied to the ring atoms in Cartesian coordinates. This approach negates the need for a Jacobian correction.

At the start of the FEARCF simulations the reaction coordinate potential is zero. The resulting probability distribution is used as a first guess for $U(\xi)$ and the biasing forces derived from this is applied to the next simulation. This process continues iteratively until the pucker conformational space is uniformly sampled (i.e. reaching convergence). Figure 4-3a shows how the biasing forces can be applied to atoms that describe the θ_1 flap, to convert from a ^4E to a $^4C_1$ conformer of glucose. In this example, the angle between the reference plane and the θ_1 plane is initially 17.36° but is forced down to -35.26° by applying forces to the appropriate atoms of the reference and θ_1 planes.

The sampling rate of the reaction coordinate space was further improved by summing the histograms biased simulations after weighting them appropriately using the Weighted Histogram Analysis Method (WHAM).[23-25] The WHAM equations are applied iteratively until the maximum tolerance between the previous and current iteration weighting coefficients is less than 0.001.

4.3.2 Reducing dimensionality with Boltzmann-averaging

Multi-dimensional free energy hyper-volumes (or equivalent) can be reduced in dimensionality. For example, a free energy cubic volume could collapse into a free energy square surface by averaging over a chosen dimension. Simple averaging can be used but this does not weight the chosen data appropriately along the coordinate. Boltzmann-averaging is an alternative method where 3 spatial coordinates are reduced to 1 spatial coordinate by applying the following expression,

Equation 4-6

$$W(\theta_0)_{\theta_1,\theta_2} = \frac{\sum_{\theta_1,\theta_2} W(\theta_0,\theta_1,\theta_2) e^{-W(\theta_0,\theta_1,\theta_2)/k_B T}}{\sum_{\theta_1,\theta_2} e^{-W(\theta_0,\theta_1,\theta_2)/k_B T}}$$

where k_B is Boltzmann's constant, T is the temperature in Kelvin and $W(\theta_0, \theta_1, \theta_2)$ is the multi-dimensional free energy calculated from simulations. An analogous expression applies for reduction of 2 spatial coordinates to 1 spatial coordinate.

4.4 Simulation Details

For both the furanose and pyranose model sugars a series of simulations were carried out with AM1[9], PM3[10], PM3CARB-1[11] and SCC-DFTB[12] semi-empirical potentials. Both the AM1 and PM3 potentials have been used for simulating carbohydrates.[26-29] The PM3CARB-1[11] parameters were developed to more accurately model carbohydrates and have been used to investigate the reasons for glucose's and galactose's primary alcohol rotamer populations in water.[30] SCC-DFTB has recently become a popular method for QM/MM molecular simulations[28, 31, 32] including carbohydrates.

For both β-D-ribose and β-D-glucose 20 iterations of QM dynamics were carried out using CHARMM[21] v33b2. At each iteration 8 PMF calculations of 0.8 ns in length were run using velocity-verlet dynamics at 298.15K with group based cutoffs of 10, 12 and 14Å. The electrostatics were treated with force switching and the van der Waals potentials were shifted to account for long distance non-bonded interaction discontinuities. The non-bonded interactions were updated using CHARMM's built-in heuristic algorithm. The AM1 and PM3 modules (CHARMM version 33b2) were used while PM3CARB-1 was implemented based on the PM3 semi-empirical routine where the modified parameters were included with the EXTE function.[30] The REMO keyword was used to remove MM energies from QM atoms. SCC-DFTB calculations were run with the mio-0-1[33] parameters and an improvement to hydrogen bonding interaction was included with the HBON keyword.

4.4.1 DFT energies of selected conformers

Canonical conformers of ribose and glucose were extracted from FEARCF simulation trajectories. Bearing in mind that the primary alcohol hydrogen bonding interaction can dramatically stabilise the energy of a conformer[30] the "native" extracted coordinates were used in addition to coordinates where the primary alcohol (O5-C5-C6-O6) was rotated into each of the three (trans-gauche, gauche-gauche, gauche-trans) staggered rotamers. Each conformer was then optimised and the energy evaluated for AM1, PM3, PM3CARB-1 and SCC-DFTB. These energies were compared with the same conformations which were optimised with RB3LYP/6-311++G(d, p). We optimised the DFT structures using methods as prescribed by Momany et al.[34, 35] based on the suggestion of Csonka.[36]

4.5 Results

4.5.1 Ribose

For the furanose ring, the free energy is a function of (θ_0, θ_1). The central and rotatable planes were chosen as described in § 4.2 and shown previously in Figure 4-3b. The pucker coordinates can be mapped to the ribose (10 Envelope (E), 10 Twist (T) and 1 planar) canonical conformers as illustrated in Figure 4-4 and listed in Table 4-1. Figure 4-4 shows a star-shaped line diagram indicating the directions along which furanose conformers exist in relation to the rotatable plane angles θ_0 and θ_1. The nodes on the lines represent the θ_0 and θ_1 puckering coordinates that describe the ideal canonical envelope and twist conformers for cyclopentane. Each line represents varying degrees of pucker for that conformer starting from the planar conformer at the origin (0°, 0°). For example, the E_2 conformer has the C2 atom below the plane of ring atoms (C3-C4-O4-C1). The characterisation of this conformer as E_2 does not define the extent to which the C2 atom is beneath the ring plane. The conformers of the cyclopentane structures found describe geometrically "ideal" canonical conformers where each ring atom can distort above or below the plane equally. In heteroatomic cyclic systems, such as carbohydrates, the canonical conformers may be described in a series of possible θ_0 and θ_1 puckering coordinates represented by the labeled lines. In Figure 4-6 five contour plots of $W(\theta_0, \theta_1)$ are shown for a) AM1, b) PM3, c) PM3CARB-1, d) SCC-DFTB and e) HF/6-31G with contours drawn at 0.1 kcal/mol, in increments of 0.25 kcal/mol till 2 kcal/mol, after which they are contoured every 2 kcal/mol. The energy is represented in colour using a rainbow colour scale where blue represents low energies (starting at 0 kcal/mol) and red high energies (up to 15 kcal/mol).

The AM1 free energy surface (Figure 4-6a) has a global minimum at (θ_0=7.5°, θ_1=7.5°) which is an E_O conformation. At 0.01 kcal/mol the ${}^1T_O$ conformation, which borders the E_O conformation appears. These conformations and several others (${}^4T_O$, ${}^1E$, ${}^1T_2$, E_2, ${}^3T_2$, E_3, ${}^4T_3$, ${}^4E$, and planar) lie in a well 0.5 kcal/mol from the global minimum. There are no other stationary states on the free energy surface. Consequently, even though the deviations of the two planes C3-C4-O4 and O4-C1-C2 are greater for the ${}^4T_3$ and ${}^1T_2/E_2$ conformations compared with ${}^OT_1$ and ${}^OE$, all conformers are readily accessible at room temperature at the AM1 level of theory. The planar conformer is located only 0.41 kcal/mol above the global minimum.

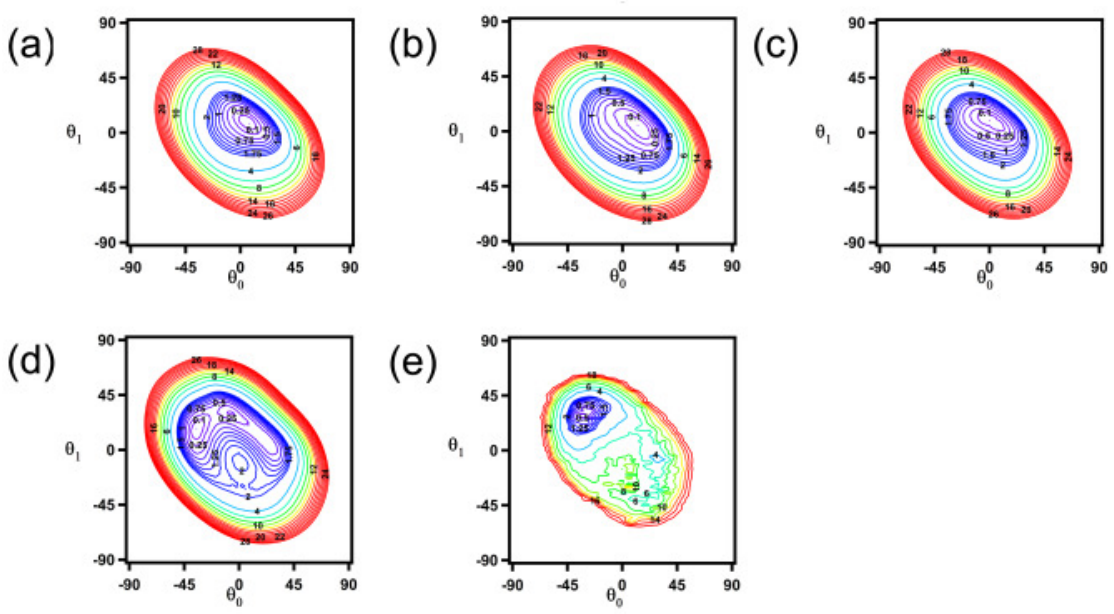

Figure 4-6 Furanose free energy of puckering shown as two-dimensional contour plots for (a) AM1, (b) PM3, (c) PM3CARB-1, (d) SCC-DFTB and (e) HF/6-31G. The Hartree-Fock surface is not completely converged and has been smoothed twice. Energy has been mapped to colour from 0 kcal/mol (blue) to 15 kcal/mol (red). Contours are shown at 0.1 kcal/mol; every 0.25 kcal/mol until 2 kcal/mol and every 2 kcal/mol thereafter.

Conversely, the global minimum of the PM3 free energy landscape (Figure 4-6b) is a ${}^{4}T_{O}$ conformation located at (12.5°, 5.0°). There are no other stationary points in the close vicinity of ${}^{4}T_{O}$. There are, as in the AM1 case, several conformers accessible within 0.5 kcal/mol of the global minimum, with ${}^{3}T_{4}$, E_{4}, ${}^{O}T_{4}$, ${}^{O}E$, ${}^{O}T_{1}$, E_{1}, ${}^{2}T_{1}$ *and* ${}^{2}E$ exhibiting the most distortion from the central planar structure. At 3 kcal/mol there is conformational interchangeability between all conformers. However, the energy required for puckering E_{1}, ${}^{O}T_{1}$, ${}^{O}E$, E_{O}, and ${}^{1}T_{O}$ away from the planar conformer is more than that for other conformers, which may imply that, the ring oxygen prefers geometries that are more planar. In fact, the planar conformer is located at 0.3 kcal/mol above the global minimum compared with AM1 where it is found at 0.41 kcal/mol above that free energy global minimum.

The shape of the free energy surface for PM3CARB-1 (Figure 4-6c) is very similar to that of PM3 (Figure 4-6b). However, the lowest energy PM3CARB-1 conformer is not a ${}^{4}T_{O}$ conformer as in PM3, but a ${}^{1}E$ found at (0°, 10°) with E_{O} at (5°, 7.5°) within 10^{-2} kcal/mol of it. Moreover the transition separating these two conformations is less than 0.1 kcal/mol making these structures energetically indistinguishable within the accuracy of these levels of theory. The planar conformer is found at 0.27

kcal/mol above the global minimum. Similar to the PM3 surface, all canonical conformers and the planar conformer can be accessed to some degree within an energy well of 0.5 kcal/mol. The conformers E_4, $^{O}T_4$, ^{O}E, $^{O}T_1$, E_1, $^{1}T_O$, E_O *and* $^{4}T_O$ are not puckered close to the ideal cyclopentane topologies but instead are closer to the central planar geometry (see Figure 4-4) relative to the rest of the conformers. The PM3 and PM3CARB-1 methods predict that the anomeric carbon may not easily tilt below the plane of the ring. While at 3 kcal/mol all conformers can be accessed, the shape of the minimum energy well dictates that the extent / magnitude of puckering of the structures in the E_3 and $^{3}T_2$ directions and adjacent to these is more than that of the $^{O}T_1$, $^{1}T_O$ conformers.

The furanose ring pucker free energy topologies calculated from AM1, PM3 and PM3CARB-1 (Figure 4-6a-c) are very similar to each other in so far that no significant distinction between the 21 different conformers exists. This is not the case for SCC-DFTB (Figure 4-6d). This semi-empirical method reveals a difference in free energy between a number of ring pucker conformations. The surface shows several distinct conformational minima surrounding the global minimum at (-35°, 17.5°). This global minimum is a ^{3}E conformer with some $^{3}T_4$ character. This conformer does facilitate hyperconjugation required to stabilise the TS oxocarbenium ion. These stationery points include a $^{1}T_2$ minimum conformer at (-5°, 30°) that is very similar in energy (0.08 kcal/mol higher) to ^{3}E which can be accessed via a E_2 / $^{3}T_2$ transition point (-20°, 27.5°) of 0.29 kcal/mol. There is a further minimum of 0.27 kcal/mol at (20°, 10°) which is a $^{4}T_O$ conformer. The $^{4}T_O$ can be reached from $^{1}T_2$ via an E_O transition of 0.31 kcal/mol at (15°, 15°). Unlike the NDDO semi-empirical methods, which allow the furanose ring to pucker into the planar conformer requiring less than 0.5 kcal/mol, the SCC-DFTB method produces a planar conformer (1.91 kcal/mol) that is more than 3kT above the global minimum. Therefore, while the planar conformer is accessible it should be rarely observed at T=298K.

AM1, PM3 and PM3CARB-1 all exhibit a large minimum energy well and no distinct global minimum. No significant energetic selectivity that discriminates between different puckering conformers is evident and all conformers are thermally accessible at 6kT. Nonetheless, while all the conformers lie within the low free energy envelope, it is energetically more difficult to pucker $^{O}T_1$, E_1, ^{O}E, E_O and $^{1}T_O$ into their ideal conformer geometries than for other conformers shown in Figure 4-4. The SCC-DFTB produced furanose free energy pucker surface is strikingly different from that generated from the NDDO methods since an energetically variegated minimum well exists within the 6kT free energy envelope. This implies that when treated with this level of theory, the furanose ring can be puckered relatively easily into discrete conformations at room temperature.

Each of the IUPAC defined canonical conformers were extracted from the FEARCF AM1, PM3, PM3CARB-1 and SCC-DFTB free energy simulations and optimised at both that level of theory and separately using DFT RB3LYP/6-311++G(d, p) and HF/6-31G. Table 4-3 lists conformers and energies obtained from these optimisations. Both DFT and HF show similar preferred conformers, although HF does not optimise ribose to stable ${}^{4}T_{O}$ and ${}^{3}E$ conformers. The same global minimum, ${}^{3}T_{2}$ is seen for both DTF and HF but the HF relative conformer energies are of greater magnitude than the DFT energies. The planar conformer was not stable for SCC-DFTB or DFT methods. While the DFT RB3LYP/6-311++G(d, p) treatment produces a ${}^{3}T_{2}$ minimum (the pucker in between E_2 and ${}^{3}E$), none of the semi-empirical free energy surfaces had this conformer as a global minimum. SCC-DFTB shows a ${}^{3}T_{4}$ global minimum that occupies the same North-Western quadrant of Figure 4-4 as does the ${}^{3}T_{2}$ of DFT. The ${}^{1}T_{O}$ of AM1 and ${}^{4}T_{O}$ of PM3 and PM3CARB-1 are not similar to ${}^{3}T_{2}$ occupying the North-Eastern quadrant of Figure 4-4. Here it is interesting to note, that an analysis of crystallographically observed puckering for ribose sugars found stable ${}^{2}E$, E_2, ${}^{3}E$, E_3, and E_4 conformers. The minimum energy conformer of ribose found by Jalbout[37] using CCSD(T)/6-31G**//MP2/6-31G** is an E_2 with an E_3 conformer located within 1 kcal/mol from it.

Table 4-3 A comparison between AM1, PM3, PM3CARB-1, SCC-DFTB, RB3LYP/6-311++G(d, p) and HF 6-31G potential energy optimisations (relative to lowest energy conformer in kcal/mol) for β-D-ribose. All conformers extracted from FEARCF dynamics simulations. The same canonical conformers used for optimisations at each level of theory. * do not exist after optimisation. ** the lowest energy conformer that the planar conformer optimises to and that is very close to planar. These are AM1 (9.7°, 5.6°); PM3 (6.1°, 3.9°); PM3CARB-1 (11.5°, 3.1°).

	AM1	PM3	PM3CARB-1	SCC-DFTB	DFT optimised	HF optimised
From Planar	0.52**	2.14**	0.00**	Optimises out of planar conformation into E_4, ^4E, ^4T_O	Optimises out of planar conformation into ^4E, $^3T_2$, E_2	Optimises out of planar conformation into $^3T_2$ and E_3
^OE	*	*	1.41	*	*	*
E_4	4.11	*	*	0.00	*	*
$^3T_4$	*	*	*	0.00	*	*
^3E	*	*	1.99	0.29	2.38	*
$^3T_2$	*	*	*	*	0.00	0.00
E_2	*	*	*	*	1.10	3.67
$^1T_2$	5.52	*	2.33	0	0.68	8.48
^1E	5.47	0.31	2.49	0.28	*	*
^1T_O	0.00	0.28	0.82	0.29	*	*
E_O	0.42	0.28	0.29	0.38	*	*
^4T_O	2.44	0.00	0.00	0.46	0.57	*
^4E	1.88	1.16	2.73	0.27	0.44	3.39

A partially converged *ab initio* free energy surface using Hartree-Fock theory (Figure 4-6e) has been used to calculate the free energy of ring pucker. Even though the *ab initio* surface is not as well sampled as the semi-empirical surfaces, it reveals distinct minima for canonical pucker conformer states. This is in agreement with the SCC-DFTB surface and contradicts the pucker free energy surfaces that were produced using AM1, PM3 and PM3CARB-1. The *ab initio* surface shows ^3E (-35°, 30°) as the global minimum with a second minimum $^4T_3$ at (27.5°, -7.5°) that is 1.02 kcal/mol higher on the free energy surface. Other stationary points that feature on the *ab initio* free energy surface are an 8.09 kcal/mol high energy $^OT_1$ conformer at (-7.5°, 17.9°) and a low energy $^OT_1$ conformer at (-12.5°, -32.5°). The overall shape of the free energy surface bears some resemblance to SCC-DFTB, particularly the global minimum conformer ^3E that is similar to that observed in the solid state for the substituted ribose ring in phenylalanine transfer RNA.[38] Despite the lack of electron correlation in the *ab initio* calculation this

remains a promising result implying that for furanose systems SCC-DFTB bears a closer resemblance to *ab initio* methods than do AM1, PM3 and PM3CARB-1.

Pucker free energies and potential energies are often studied in a single dimension to, for example, arrive at a "minimum transition pathway" connecting two equilibrium wells. To illustrate the effect that a reduction in dimensionality has on the conformational analysis of carbohydrate rings, Boltzmann-averaging is used to contract the two-dimensional (θ_0, θ_1) furanose surfaces to one dimension. Boltzmann-averaging over θ_1 yields a free energy coordinate as a function of θ_0. As an example of the effect of a reduction in dimension, when analysing AM1 more closely it is observed that the $W(\theta_0)_{\theta_1}$ minimum occurs at 10° and rises to 10 kcal/mol at θ_0 = +/-50°. On the other hand averaging over θ_0 gives $W(\theta_1)_{\theta_0}$, where the minimum occurs at 5° and the free energy rises to 10 kcal/mol at about θ_1 = +/-52.5°. Comparing both 1D curves to the 2D free energy surface it is clear the position of the global minimum shifts to either side of the 2D global minimum position observed at (7.5°, 7.5°). The more serious problem of dimension reduction is the collation of non-degenerate conformers into single averaged points in conformational space. This is apparent as shown here for the AM1 furanose surface, which by all accounts is relatively simple displaying very little character. The topologies of the 1D curves and 2D surfaces are listed in Table 4-4 to further illustrate the information penalty due to dimension reduction of complex pucker conformational space. This difference between 1D and 2D topologies is most severe for SCC-DFTB, which is the surface with the most character.

Table 4-4 The Hill-Reilly coordinates for Boltzmann-averaged minimum *vs* the actual 2D minimum for AM1, PM3, PM3CARB-1 and SCC-DFTB of ribose. The coordinate values at an energy of 10 kcal/mol (so called "maxima") are listed along with the minimum.

	AM1	PM3	PM3CARB-1	SCC-DFTB
Min(θ_0)	10.0	15.0	0.0	-35.0
Min(θ_1)	5.0	5.0	10.0	25.0
Min(θ_0, θ_1)	(7.5, 7.5)	(12.5, 5.0)	(0.0, 10.0)	(-35.0, 17.5)
Max(θ_0)	-57.5; 60.0	-62.5; 62.5	-60.0; -60.0	-70.0; -67.5
Max(θ_1)	-60.0; 60.0	-62.5; 60	57.5; 57.5	65.0; 67.5

A check for free energy surface convergence is usually done by evaluating a simulation that has biasing forces derived from the free energy surface. If the system has converged to the true free energy then it will be driven away from parts of the reaction coordinates that are local minima on the

equilibrium surface. The result is equal sampling across reaction coordinate space, if the space is closed (e.g., angles, distances etc.). Ring puckering has a closed canonical conformational space[19] with respect to Hill-Reilly coordinates that is bounded by unphysical geometries (e.g., rings are twisted or folded into shapes that resemble pretzels) that are not easily sampled as they are very high in energy. In this case, the extent of convergence can be gauged by comparing the sampling of high energy conformers with that of the global minimum pucker sampling. A ratio of less than 50:1 is considered good,[39, 40] and the surface is determined to be converged. In the case of ribose the most to least sampled conformer for AM1 is 4:1, PM3 is 6:1, PM3CARB-1 is 7:1 and SCC-DFTB is 8:1.

4.5.2 Glucose

The free energy landscape of glucose puckering is necessarily more complex than that of ribose. Its free energy is a function, of at least three degrees of freedom; the three puckering angles making up the reaction coordinate ξ_{glucose}. The choice to monitor the motion of the planes C4-C5-O5 (θ_0), O5-C1-C2 (θ_1) and C2-C3-C4 (θ_2) differ from those chosen by Hill and Reilly but allow observation of the contribution of the primary alcohol and anomeric carbon to the overall pyranose puckering free energy.

The more complex four-dimensional free energy landscape for hexoses can be visualised in three dimensions by representing the free energy as a color variable. Here the color scale represents low energies (0 kcal/mol) as blue and very high energies as red (15 kcal/mol). Although all simulations sampled regions up to and beyond ~30 kcal/mol, the structures in these regions are distorted high energy conformers that are not chemically interesting (e.g., tightly bent, twisted and scrunched up rings) and therefore not part of IUPAC canonical space definitions. The 38 canonical conformers categorising 6-membered rings are shown in a Hill-Reilly coordinate frame with respect to the chosen reference plane (Figure 4-5) with the values for θ_0, θ_1 and θ_2 listed in Table 4-2. While the complete volume is cubic in shape, i.e. $\theta_0, \theta_1, \theta_2 \in [-90°, 90°]$ the canonical conformers lie within a spherically shaped region. This spherical region is similar in appearance to the one used by Cremer and Pople (who mapped *q2, q3 and* ϕ to spherical coordinate space). The chairs (C) are the "poles" ($^1C_4$ North and $^4C_1$ South) while the equator contains all possible boat (B) and twist / skew boat (S) canonical conformers. The northern or axial tropic contains half chair (H) and envelope (E) conformations that are accessible from the $^1C_4$ chair; while the southern or equatorial tropic contains H and E conformations that neighbor the $^4C_1$ chair.

In each of the $W(\theta_0, \theta_1, \theta_2)$ landscapes (AM1 through SCC-DFTB), an inner isosurface (deep blue) at 3 kcal/mol represents the pucker conformers which can be accessed at ~6kT. Where the lowest energy

conformer is $^{4}C_{1}$ in all cases (Figure 4-7). The free energy volumes calculated using the original NDDO methods (AM1 and PM3) imply that the pyranose ring is easily deformed (Figure 4-7a and b). An exchange between $^{4}C_{1}$ and $^{1}C_{4}$ requires only about 4 kcal/mol for both levels of theory. The 3 kcal/mol energy isosurface for AM1 show all twist and boat conformers (i.e. $^{O,3}B$, $^{O}S_{2}$, $B_{2,5}$, $^{1}S_{5}$, $^{1,4}B$, $^{1}S_{3}$, $B_{O,3}$, $^{2}S_{O}$, $^{2,5}B$, $^{5}S_{1}$) at room temperature except $^{3}S_{1}$ and $B_{1,4}$. In the PM3 case a large number of boat and twist conformers ($B^{O,3}$, $^{2}S_{O}$, $^{2,5}B$, $^{5}S_{1}$, $^{3}S_{1}$, $^{O,3}B$, $^{O}S_{2}$, $B_{2,5}$, $^{1}S_{5}$ and a very restricted $B_{1,4}$) are seen within the 3 kcal/mol free energy volume well. There appears to be an easy exchange between $B_{O,3}$, $^{2}S_{O}$, $^{2,5}B$ *and* $^{5}S_{1}$ conformers. Similarly, transitions between $^{1}S_{5}$, $B_{2,5}$, $^{O}S_{2}$, $^{O,3}B$ and $^{3}S_{1}$ conformers are feasible.

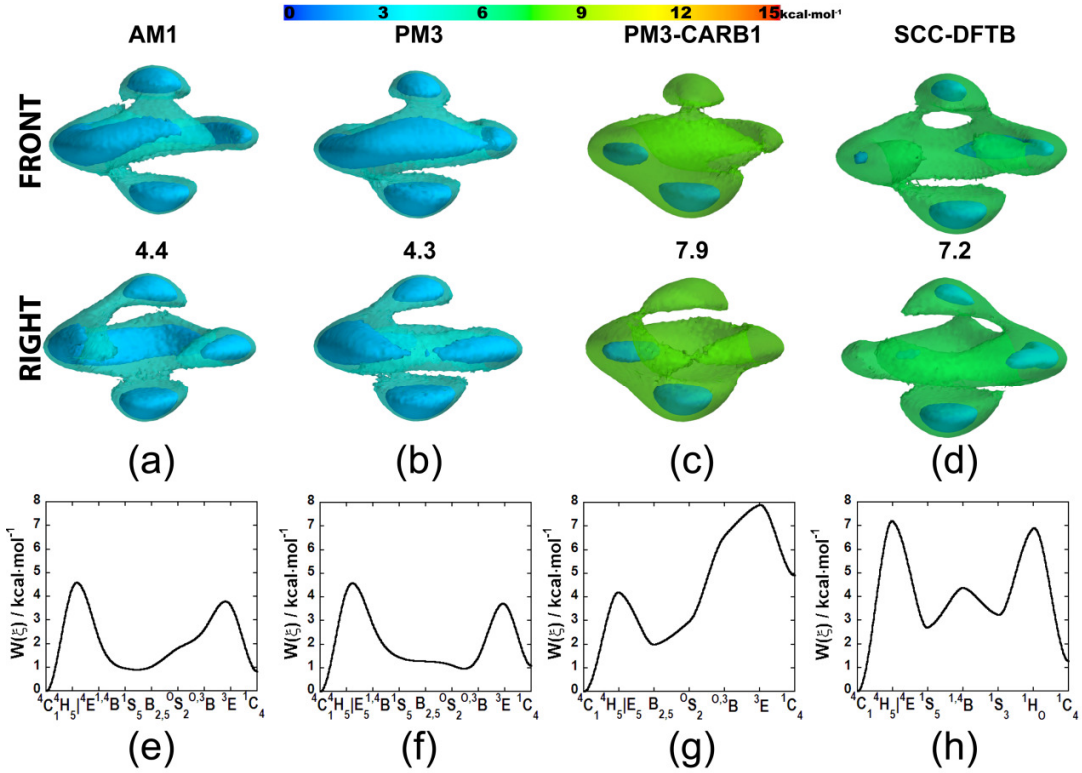

Figure 4-7 The free energy of puckering colour mapped to three dimensional volumes for (a) AM1, (b) PM3 (c) PM3CARB-1 and (d) SCC-DFTB (been mapped to colour from 0 kcal/mol (blue) to 15 kcal/mol (red). The inner isosurface is at 3 kcal/mol and the outer isosurface indicating the minimum free energy to connect the "poles" ($^{1}C_{4}$, $^{4}C_{1}$) occurs at (a) 4.4, (b) 4.3, (c) 7.9 and (d) 7.2 kcal/mol respectively. The minimum free energy paths between the equatorial chair ($^{4}C_{1}$) and axial chair ($^{1}C_{4}$) have been extracted from the free energy volumes (5a-d) and represented for (e) AM1, (f) PM3, (g) PM3CARB-1 and (h) SCC-DFTB.

Using PM3CARB-1, the parameter set that was especially modified to model carbohydrates, a greater restriction of pyranose pucker (Figure 4-7c) is observed when compared with AM1 and PM3. While $^4C_1$ is the global minimum, contouring at 3 kcal/mol shows only this conformer and a pathway between two other conformers ($^1S_5$ and $B_{2,5}$) at the equator to be possible. There is no path to the other chair ($^1C_4$) observable at room temperature since more than 6kT separate them.

While SCC-DFTB does not display the almost floppy pyranose ring seen in AM1 and PM3 treatments it is not nearly as conformationally restricted as the PM3CARB-1 model at 3 kcal/mol. At this contour both chair conformers are accessible but not equally so, as the volume about $^1C_4$ is marginally smaller than that about $^4C_1$. Five canonical conformers at the equator are observed with $^1S_5$ showing up as being isolated at this contour where easy interchange between $B_{O,3}$, ^2S_O, $^{2,5}B$ and $^5S_1$ appears to be possible.

The conformational path from the $^4C_1$ conformer, where all the glucose hydroxyls are equatorial, to the $^1C_4$, where all the hydroxyls are axial, goes via strained conformers that differ in energy from one semi-empirical method to another. The barrier heights separating $^4C_1$ from $^1C_4$ for AM1 and PM3 are quite similar (i.e. 4.4 and 4.3 kcal/mol respectively) whilst the PM3CARB-1 and SCC-DFTB methods have barrier heights that are nearly twice as high with values of 7.9 and 7.2 kcal/mol respectively. The isosurfaces corresponding to this energy are shown as free energy volume plots that can be projected onto the schematic diagram for pyranose rings (Figure 4-5) and are shown in Figure 4-7a-d.

Minimum free energy paths have been extracted and plotted as line diagrams (Figure 4-7e-h). In the line diagrams, the minimum energy paths are shown while the isosurfaces detail all conformers available for free energies at and below the contour level shown. Starting from $^4C_1$ using the AM1 level of theory and driving northward the chair distorts into $^4H_3$, ^4E and $^4H_5$ after which these ring conformers can easily alter into the $^1S_3$, $^{1,4}B$ and $^1S_5$ conformations. At the equator, $^1S_5$ is lowest in energy and all conformers but $B_{1,4}$ are accessible. The connection to the $^1C_4$ conformer can be made from the $^OS_2$, $^{O,3}B$ or $^3S_1$ structure via the $^3H_2$ / ^3E bridge where $^OS_2$ is lowest in energy. The minimum energy path is shown in Figure 4-7e.

In the PM3 case, the route from $^4C_1$ to $^1C_4$ occurs via the E_5 / $^4H_5$ conformers through the $B_{2,5}$ or $^1S_5$ conformers with $B_{2,5}$ being slightly lower in energy than $^1S_5$ (a difference of less than 1 kcal/mol). The $^1C_4$ conformer is reached via ^3E or $^3H_4$ from the $^{O,3}B$ conformer. Here, all conformations located on the equator (Figure 4-5) can be reached from the chair conformations requiring no more than 3 kcal/mol.

PM3CARB-1 shows a distinctly different minimum path between the $^{4}C_{1}$ and $^{1}C_{4}$ conformers. The lowest energy conformer that can be used to access the equator from $^{4}C_{1}$ is $^{4}H_{5}$ with some E_{5} character. Following this, the $^{1}S_{5}$, $B_{2,5}$ and $^{O}S_{2}$ can be accessed with the lowest energy conformer being $^{1}S_{5}$. The route to $^{1}C_{4}$ is through the ^{3}E / $^{3}H_{2}$ conformers via $^{O,3}B$. This free energy volume is very different to the AM1 and PM3 volumes as all envelopes and half chairs near to $^{4}C_{1}$ are energetically accessible when using original NDDO parameter sets as implemented in CHARMM. Furthermore all but the E_{1}, $^{2}H_{1}$ and ^{2}E envelopes and half chairs provide a direct path to the skew-boats and boat conformers. For PM3CARB-1 it is easier to transform $^{4}C_{1}$ to conformations found at the equator of the pyranose spherical pucker volume than it is for $^{1}C_{4}$. There are sometimes multiple paths of similar energy connecting the poles to the equator. For example in SCC-DFTB, the $^{1}C_{4}$ to equator transition can occur via $^{1}H_{O}$ or $^{5}H_{O}$ which are similar in energy at ~7 kcal/mol.

Using the SCC-DFTB level of theory, the $^{4}C_{1}$ conformer can be transformed into $^{1}S_{5}$ and $^{1,4}B$ conformers via the ^{4}E state. Once in the S/B conformations the $^{1}C_{4}$ conformer can be reached via a $^{5}H_{O}$ or $^{1}H_{O}$ conformation. While the $^{1}H_{O}$ conformer presents the shorter conformational route a transformation via $^{5}H_{O}$ is slightly lower in energy. Most boats and twist boats are energetically accessible within the 7.2 kcal/mol free energy volume. It is not possible to reach the canonical $^{O,3}B$ located at the equator of the pucker sphere as well as ^{4}E, $^{5}H_{O}$ and $^{1}H_{O}$ that are positioned at the "tropics" of the sphere shown in Figure 4-5. There is only one possible transition at this contour from $^{4}C_{1}$ to the boat and twist boat conformers while two such paths exist when starting at the $^{1}C_{4}$ chair.

There have been several theoretical studies investigating the energetic differences of ring puckers for glucose and selected derivatives.[1, 41] While it would be ideal to employ *ab initio* methods to construct the free energy volumes, this is not within reach of current computational methods and hardware. However, the semi-empirical methods used here can be evaluated against calculations done for selected conformers. An example of such a case is the work of Ionescu et al.[41] who carried out dynamical density functional theory (DFT) calculations with the projector augmented-wave (PAW) method and found that the transition state for per-methylated glucose from the $^{1}C_{4}$ chair to $^{1,4}B$ was an ^{1}E envelope conformation. Consequently, from the envelope-like transition state, the system evolved into the $^{1,4}B$ minimum (this is a TS for the SCC-DFTB and other semi-empirical methods investigated here). They found a $^{1}S_{5}$ skew-boat minimum and discovered that pucker pseudo rotation stopped before the $^{O}S_{2}$ skew-boat was reached. The transition state to the inverted $^{4}C_{1}$ chair was a mix of envelope and half-

chair conformations. The energy barrier to leave the $^{4}C_{1}$ chair minimum was 1.488 kcal/mol (7.2 kJ/mol), nearly twice the free-energy barrier required to make a transition from the $^{1}C_{4}$ chair minimum.

The lowest optimised potential energy conformer for the AM1 and PM3 parameter set is the $^{1}C_{4}$ conformer where secondary and primary hydroxyls are axial. This result does not correspond to RB3LYP/6-311++G(d, p) optimisations that show the lowest potential energy is due to the $^{4}C_{1}$ conformer (Table 4-5). However, optimisations on the SCC-DFTB and PM3CARB-1 potential energy surface do correspond to this result. These two methods match the DFT potential energy surface more closely. While, as with DFT, the PM3CARB-1 parameter set shows the $^{1}C_{4}$ chair to be much higher in energy compared with the most favoured $^{4}C_{1}$ conformer, SCC-DFTB presents a small energy difference (1.24 kcal/mol) between the two chairs. Optimisations with DFT do not yield half-chair, envelope, $^{2,5}B$, $B_{2,5}$ or $B_{1,4}$ conformers suggesting these are not minimum energy stationary points. However, AM1 yields $B_{2,5}$, $^{1}H_{2}$, ^{5}E, $^{2,5}B$; PM3 yields $B_{2,5}$, $^{2,5}B$, $B_{1,4}$, $^{1}H_{2}$, $^{3}H_{2}$; PM3CARB-1 yields $^{3}H_{2}$, E_{2}, $^{5}H_{4}$, $B_{2,5}$ and SCC-DFTB yields $B_{1,4}$ in addition to the conformers shown in Table 4-5. In terms of the stable conformers that it optimises to, the performance of SCC-DFTB is similar to the DFT methods used, while NDDO methods are not.

Table 4-5 A comparison between AM1, PM3, PM3CARB-1, SCC-DFTB and RB3LYP/6-311++G(d, p) potential energy optimisations (relative to lowest energy conformer in kcal/mol) for β-D-glucose. * Optimised from $^{1,4}B$, indicating it is not a minimum on the potential energy surface. ** Optimised from $^{1}S_{5}$.

Conformer	AM1	PM3	PM3CARB-1	SCC-DFTB	DFT Optimised
$^{4}C_{1}$	0.14	1.12	0.00	0.00	0.00
$^{1}C_{4}$	0.00	0.00	4.68	1.24	7.14
$^{1}S_{5}$	1.31*	3.61*	5.46*	4.08*	8.16*
$^{5}S_{1}$	5.36	3.99	7.90	5.14	9.96
$^{1}S_{3}$	1.92	3.93	1.88	2.98	6.67
$^{3}S_{1}$	3.70	3.05	6.62	3.24	7.20
$B_{O,3}$	1.11	2.63	1.59	3.48	5.46
$^{O,3}B$	1.54	2.82	8.30	3.47	7.95
$^{O}S_{2}$	1.79	3.40	5.59	1.29	9.73
$^{2}S_{O}$	0.59**	1.46	1.94	3.90	3.87

Static conformational optimisations have severe limitations since they do not include the contributions of the many degrees of freedom (e.g., primary and secondary alcohol rotations, hydrogen bonding etc.) to the pucker conformational space. The free energy volume is a useful measure of the

relative stability of various ring puckers. To evaluate our results we have to rely on the few reports[42, 43] of complete free energy of the pucker volumes for pyranose rings, that exist. Biarnés et al.[42] were able to sample only the southern hemisphere of the Cremer-Pople pucker volume using a metadynamics procedure. Sega et al. analysed the metadynamics method used to calculate the Cremer-Pople pucker volume and concluded that when Biarnés et al. attempted to traverse free energies using the Cremer-Pople Cartesian coordinates they were limited to free energy barriers that were lower than the thermal energy at the equatorial (boat and skew conformers) region.[43] When analysing the sampled hemisphere produced by the metadynamics simulation it becomes apparent that the transition ${}^4C_1 \rightarrow {}^1S_5 / B_{O,3}$ goes via a ${}^4H_3 / E_3$. This result is similar to that seen in all the semi-empirical free energy volumes. Using static B3LYP/6-311++G** calculations Kurihara et al.[44] proposed the transition to be via $E_3 / {}^2H_3$ conformations. However, Davies et al. suggest that there is good evidence for the existence of ${}^4H_3$, ${}^3H_4$, ${}^{2,5}B$, and $B_{2,5}$ TS puckers of the pyranose ring in glycosidases.[45]

It was argued above, that a reduction in dimension of free energy results in a significant loss of information. For a pyranose ring, a Boltzmann-average over the motion of a single rotatable plane results in a two dimensional free energy surface. Such surfaces are plotted as contour maps $W(\theta_0, \theta_1)_{\theta_2}$, $W(\theta_0, \theta_2)_{\theta_1}$ and $W(\theta_1, \theta_2)_{\theta_0}$ for AM1, PM3, PM3CARB-1 and SCC-DFTB in Figure 4-8. Both ${}^4C_1$ chair and ${}^1C_4$ chair structures can be distinctly seen for $W(\theta_0, \theta_1)$ and $W(\theta_0, \theta_2)$ and $W(\theta_1, \theta_2)$, at ~(-35°, -35°) and ~(35°, 35°) for all methods, although the boat structures are included in the averaging. A cursory inspection of the 2D surfaces leaves one with the incorrect impression that all semi-empirical levels of theory model the pyranose ring in the same way, with the only differences being some changes in barrier heights and minima of a couple of kcal/mol. Furthermore, in all of these plots the planar structure appears unresolved because it is averaged to a lower energy along with boat, envelope, twist boat and half-chair structures. This is contrary to observations made in the 3D free energy hypersurfaces.

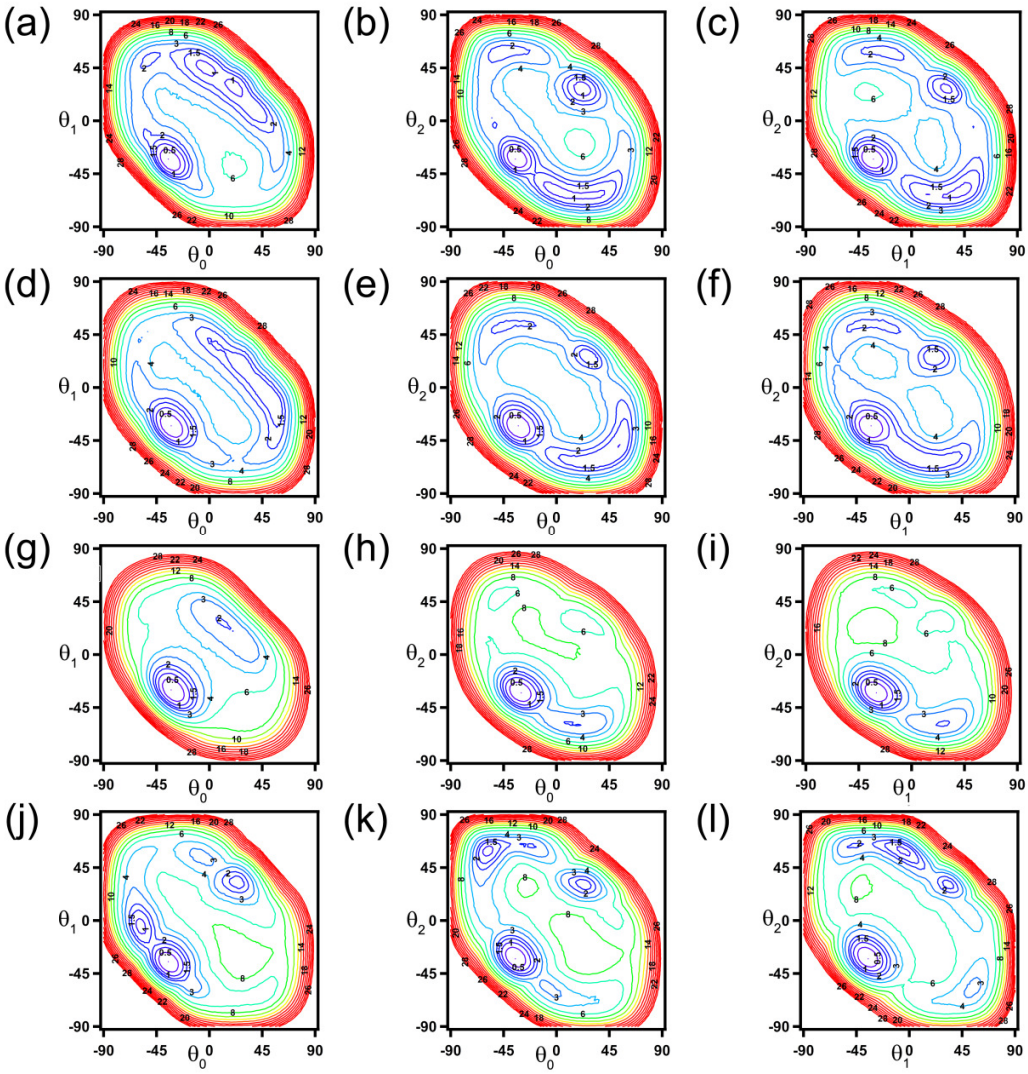

Figure 4-8 The 2D free energy contour plots that are Boltzmann averaged over a single dimension for β-D-glucose. For AM1, averaging the free energy *W(θ₀, θ₁, θ₂)* yields (a)$W(\theta_0,\theta_1)_{\theta_2}$, (b) $W(\theta_0,\theta_2)_{\theta_1}$ and (c) $W(\theta_1,\theta_2)_{\theta_0}$. The same plots are shown for PM3 (d, e, f), PM3CARB-1 (g, h, i), and SCC-DFTB (j, k, l).

As in the complete free energy volumes for AM1 and PM3 (Figure 4-7a and b) there are strong similarities in the surfaces (Figure 4-8a-f). The sugar specific parameter set, PM3CARB-1 (Figure 4-8g-i), produces surfaces that resemble AM1 (Figure 4-8a-c) and PM3 (Figure 4-8d-f) surfaces but with much higher energy barriers separating the minima. The ring pucker for the *W(θ₀, θ₂)* and *W(θ₁, θ₂)* cases appears more limited to the $^4C_1$ chair. This minimum well is smaller than AM1, PM3, SCC-DFTB, indicating that the PM3CARB-1 potential allows pyranose rings far less conformational freedom compared with the other methods. However, the 2D surfaces do preserve the underlying differences in

topologies for SCC-DFTB compared with the other methods. The typical minimum (~-35°, ~-35°) that includes the $^{4}C_{1}$ chair for all methods (Figure 4-8j-l) is observed but an additional minimum at (θ_0=-57.5°, θ_1=-5°) showing possible E_5 envelope character (Figure 4-8j) is present. Minima at (θ_0=-6°, θ_1=57.5°), (θ_0=25°, θ_1=32.5°) suggest $^{3}H_{2}$ and a mixture of $^{1}C_{4}$, $B_{1,4}$, E_4 respectively. Similar minima are seen in Figure 6k (θ_0=-58.5°, θ_2=57.5°); (θ_0=-20°, θ_2=65°); (θ_0=20°, θ_2=32.5°) and Figure 6l (θ_1=-40°, θ_2=65°); (θ_1=0°, θ_2=57.5°); (θ_1=35°, θ_2=32.5°) ; (θ_1=52.5°, θ_2=-60°), which are suggestive of the more intricate volume enjoyed by SCC-DFTB.

A Boltzmann-average over the motion of two rotatable plane results in one dimensional plots $W(\theta_0)_{\theta_1,\theta_2}$, $W(\theta_1)_{\theta_0,\theta_2}$ and $W(\theta_2)_{\theta_0,\theta_1}$. These plots provide very little information on the conformational preferences for pyranose rings (Figure 4-9) as there is not a 1:1 mapping between the energy and the conformer. These line drawings give the false impression of a one dimensional free energy 'path'. However, on comparison with the extracted lowest free energy paths, as shown in Figure 4-7e-h, it is clear that they bear no resemblance to a minimum free energy path. This illustration should serve as cautionary evidence when attempting to reduce complex ring puckering to lower dimensions, particularly, in the cases of reduction to a single dimension. These line graphs conflate puckers sharing coordinates in a reduced dimension even though they are energetically very different. Once reduced, the contributing conformers and their associated free energies cannot be recovered from the average energy and averaged conformers making up the line graphs of Figure 4-9.

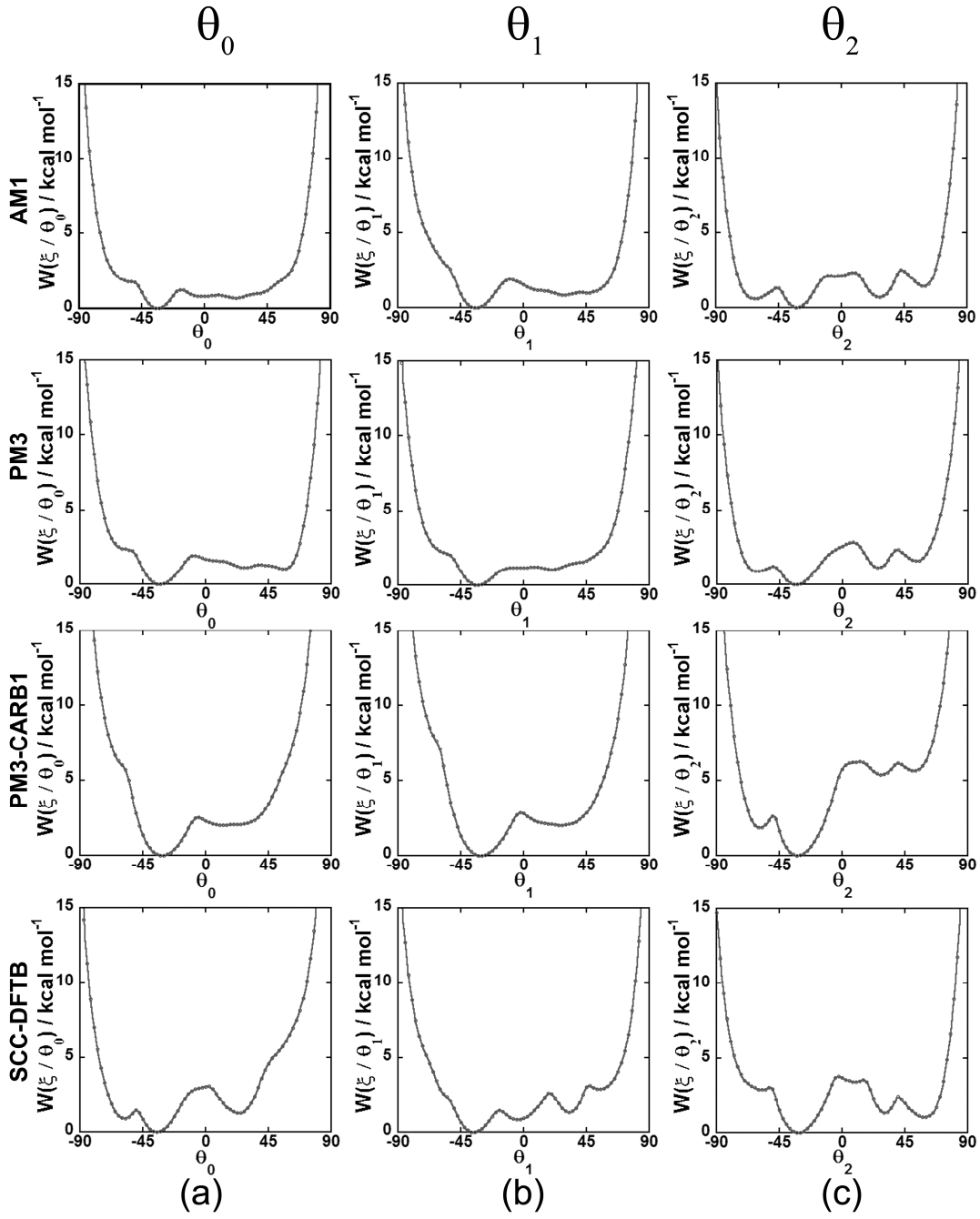

Figure 4-9 The 1D free energy contour plots that are Boltzmann averaged over two rotatable planes for β-D-glucose. The free energy W(θ_0, θ_1, θ_2) averaged over θ_1, θ_2 yields (a) $W(\theta_0)_{\theta_1,\theta_2}$, over θ_0, θ_2 yields (b) $W(\theta_1)_{\theta_0,\theta_2}$ and over θ_0, θ_1 yields (c) $W(\theta_2)_{\theta_0,\theta_1}$

Despite, the fact that they add little value to any investigation of puckering pathways and preferences, the $W(\theta_0)_{\theta_1,\theta_2}$, $W(\theta_1)_{\theta_0,\theta_2}$ and $W(\theta_2)_{\theta_0,\theta_1}$ line graphs do maintain some character observed in the complete $W(\theta_0, \theta_1, \theta_2)$ free energy volume. The graphs for AM1(Figure 4-9a) and PM3 (Figure 4-9b) are very similar. AM1 shows a preference for θ_0 to be at -35° and PM3 for θ_0 at -32.5°. This is indicative of boat or chair structures. The energy increases rapidly to the left of this region with a slight plateau / inflection at θ_0=-55°. At 0° there is a slight barrier. This includes the average of planar, boat, envelope and even half-chair structures. Between 0° and 60°, there is a broad, poorly resolved region. PM3 shows a slight minimum at θ_0=50° due to twist boat conformers.

As with $W(\theta_0)_{\theta_1,\theta_2}$ and $W(\theta_1)_{\theta_0,\theta_2}$, the $W(\theta_2)_{\theta_0,\theta_1}$ line graphs for AM1 and PM3 are similar with a minimum at θ_2=-32.5°. However, the PM3CARB-1 line graph is more asymmetric, as was seen in the full 3D $W(\theta_0, \theta_1, \theta_2)$ volume. PM3CARB-1 is similar to AM1 and PM3, with the important difference that with this potential there is a sharper increase in energy when deviating from the minimum conformer. This indicates that the ring is less flexible with the PM3CARB-1 potential. The minima for the three NDDO methods all lie at θ_0=-30°, θ_1=-30° θ_2=-32.5°. SCC-DFTB shows similar minima at θ_0=-35° θ_1=-35° θ_2=-32.5°.

It has been shown previously, that a comparison of the sampling of high energy conformers, such as the planar conformer, to the lowest energy conformer is a useful measure of the convergence of the pyranose free energy volume calculation.[19, 20] A ratio of less than 50:1 has been proposed to declare convergence[39, 40] and thus that the volume is resolved. The free energy volumes shown here are well resolved, since, the sampling of the planar conformer to the canonical conformers in the final iteration of calculation of the PMF, are: AM1 11:1; PM3 8:1; PM3CARB-1 15:1; SCC-DFTB 10:1.

4.6 Conclusions

Carbohydrate ring puckering plays a central role in several biochemical processes such as hydrolysis, phosphorylation and glycosylation reactions. Here, it has been shown that semi-empirical methods commonly used to simulate these reactions, are mostly unsuitable when modelling the ribose or glucose derivative ring puckering. In the case of ribose, the NDDO methods (AM1, PM3 and PM3CARB-1) make little distinction between canonical ring conformers, displaying no stationary points on the free energy surface for any one pucker simulated at room temperature. It seems that simulations employing these methods and that derive conformational minima and transition states from them, may lead to incorrect

structures. In addition these structures cannot be corrected by adjusting the energy with a calculation at a higher level of theory. However, the second-order expansion of the density functional theory method, the so-called SCC-DFTB method produces a free energy surface that is very different from the NDDO methods and similar to the surface obtained with *ab initio* methods.

The glucose free energy volume for the AM1 and PM3 NDDO parameter sets result in an unrealistically flexible pyranose ring. At room temperature the preferred $^{4}C_{1}$ conformer is easily distorted into half chairs and boats. In contrast, the recently introduced PM3CARB-1 parameter set presents a rigid pyranose ring that shows very high barriers between the $^{4}C_{1}$ conformer and the half chairs and boats located at the equator of the free energy volume. The SCC-DFTB method compromises between the two extremes. None the less, the SCC-DFTB free energy volume results in a $^{4}C_{1} \rightarrow {}^{1}C_{4}$ minimum path that has barrier heights of the same magnitude (near 7 kcal/mol) as the path extracted from the PM3CARB-1 free energy volume. Of note is the wide range of minimum free energy pathways seen for the semi-empirical methods investigated here. The $^{4}C_{1}$, $^{1}C_{4}$, $^{1}S_{5}$, $^{2}S_{O}$, $^{2,5}B$, $B_{O,3}$ and $^{5}S_{1}$ conformers are likely to be observed at room temperature when described by the SCC-DFTB method.

The FEARCF method in combination with the Hill Reilly pucker coordinates allow for complete exploration of the ribose and glucose free energy space. Analyses of reduced free energy space will lead to severely approximated puckers, with free energies that bear no relation to the actual values calculated using a minimum of N-3 pucker coordinates. Finally, although not perfect, the SCC-DFTB performs better for prototypical furanose and pyranose systems when the energies of semi-empirically optimised conformers and DFT optimised conformers are compared (the relative energy of the optimised canonical conformers found from each semi-empirical method are compared with the energies of the same conformers optimised using DFT methods).

4.7 References

1. S. Fushinobu, B. Mertz, A. D. Hill, M. Hidaka, M. Kitaoka and P. J. Reilly, *Carbohydrate Research*, 2008, **343**, 1023-1033.
2. G. P. Johnson, L. Petersen, A. D. French and P. J. Reilly, *Carbohydrate Research*, 2009, **344**, 2157-2166.
3. L. S. Devi-Kesavan and J. Gao, *Journal of the American Chemical Society*, 2003, **125**, 1532-1540.
4. P. J. Berti and J. A. B. McCann, *Chemical Reviews*, 2006, **106**, 506-555.
5. D. E. Sunko, I. Szele and W. J. Hehre, *Journal of the American Chemical Society*, 1977, **99**, 5000-5005.
6. A. Vasella, G. J. Davies and M. Böhm, *Current Opinion in Chemical Biology*, 2002, **6**, 619-629.
7. G. Sulzenbacher, H. Driguez, B. Henrissat, M. Schulein and G. J. Davies, *Biochemistry*, 1996, **35**, 15280-15287.
8. V. L. Schramm, *Annual Reviews in Biochemistry*, 1998, **67**, 693-720.
9. M. J. S. Dewar, E. G. Zoebisch, E. F. Healy and J. J. P. Stewart, *Journal of the American Chemical Society*, 1985, **107**, 3902-3909.
10. J. J. P. Stewart, *Journal of Computational Chemistry*, 1989, **10**, 209-220.
11. J. P. McNamara, A.-M. Muslim, H. Abdel-Aal, H. Wang, M. Mohr, I. H. Hillier and R. A. Bryce, *Chemical Physics Letters*, 2004, **394**, 429-436.
12. Q. Cui, M. Elstner, E. Kaxiras, T. Frauenheim and M. Karplus, *Journal of Physical Chemistry B*, 2001, **105**, 569-585.
13. H. B. F. Dixon, *European Journal of Biochemistry*, 1980, **111**, 295-298.
14. D. Cremer and J. A. Pople, *Journal of the American Chemical Society*, 1975, **97**, 1354-1358.
15. A. D. Hill and P. J. Reilly, *Journal of Chemical Information and Modelling*, 2007, **47**, 1031-1035.
16. A. Bérces, G. Enright, T. Nukada and D. M. Whitfield, *Journal of the American Chemical Society*, 2001, **123**, 5460-5464.
17. A. Bérces, D. M. Whitfield and T. Nukada, *Tetrahedron*, 2001, **57**, 477-491.
18. J. Strümpfer and K. J. Naidoo, *Journal of Computational Chemistry*, 2010, **31**, 308-316.
19. C. B. Barnett and K. J. Naidoo, *Molecular Physics*, 2009, **107**, 1243 - 1250.
20. C. B. Barnett and K. J. Naidoo, *American Institute of Physics: Conference Proceedings*, 2009, **1102**, 214-218.
21. B. R. Brooks, C. L. Brooks, 3rd, A. D. Mackerell, Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York and M. Karplus, *Journal of Computational Chemistry*, 2009, **30**, 1545-1614.
22. R. Rajamani, K. J. Naidoo and J. Gao, *Journal of Computational Chemistry*, 2003, **24**, 1775-1781.
23. S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman and J. M. Rosenberg, *Journal of Computational Chemistry*, 1992, **13**, 1011-1021.
24. S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen and P. A. Kollman, *Journal of Computational Chemistry*, 1995, **16**, 1339-1350.
25. S. Kumar, P. W. Payne and M. Vasquez, *Journal of Computational Chemistry*, 1996, **17**, 1269-1275.
26. T. Rungrotmongkol, A. J. Mulholland and S. Hannongbua, *Journal of Molecular Graphics & Modelling*, 2007, **26**, 1-13.
27. K. Byun, Y. Mo and J. Gao, *Journal of the American Chemical Society*, 2001, **123**, 3974-3979.
28. H. Zhou, E. Tajkhorshid, T. Frauenheim, S. Suhai and M. Elstner, *Chemical Physics*, 2002, **277**, 91-103.

29. C. Pereira da Silva and A. Araujo de Souza, *Quimica no Brasil*, 2007, **1**, 57-66.
30. C. B. Barnett and K. J. Naidoo, *Journal of Physical Chemistry B*, 2008, **112**, 15450-15459.
31. H. L. Woodcock, M. Hodošcek and B. R. Brooks, *Journal of Physical Chemistry A*, 2007, **111**.
32. D. Xu, H. Guo and Q. Cui, *Journal of the American Chemical Society*, 2007, **129**, 10814-10822.
33. M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai and G. Seifert, *Physical Review B*, 1998, **58**, 7260.
34. M. Appell, G. Strati, J. L. Willett and F. A. Momany, *Carbohydrate Research*, 2004, **339**, 537-551.
35. S. N. Steinmann, G. Csonka and C. Corminboeuf, *Journal of Chemical Theory and Computation*, 2009, **5**, 2950-2958.
36. G. I. Csonka, *Journal of Molecular Structure-Theochem*, 2002, **584**, 1-4.
37. A. F. Jalbout, L. Adamowicz and L. M. Ziurys, *Chemical Physics*, 2006, **328**, 1-7.
38. S. C. Harvey and M. Prabhakaran, *Journal of the American Chemical Society*, 1986, **108**, 6128-6136.
39. C. Bartels and M. Karplus, *Journal of Computational Chemistry*, 1997, **18**, 1450-1462.
40. C. Bartels and M. Karplus, *Journal of Physical Chemistry B*, 1998, **102**, 865-880.
41. A. R. Ionescu, A. Bérces, M. Z. Zgierski, D. M. Whitfield and T. Nukada, *Journal of Physical Chemistry A*, 2005, **109**, 8096-8105.
42. X. Biarnés, A. Ardévol, A. Planas, C. Rovira, A. Laio and M. Parrinello, *Journal of the American Chemical Society*, 2007, **129**, 10686-10693.
43. M. Sega, E. Autieri and F. Pederiva, *Journal of Chemical Physics*, 2009, **130**.
44. Y. Kurihara and K. Ueda, *Carbohydrate Research*, 2009, **344**, 2266-2269.
45. G. J. Davies, V. M. A. Ducros, A. Varrot and D. L. Zechel, *Biochemical Society Transactions*, 2003, **31**, 523-527.

5 Outlook

This thesis has been primarily targeted at elucidating the puckering conformational character of pyranose rings in glycosidases using a QM/MM free energy approach and a novel implementation of a force based free energy puckering code. With these goals in mind, the introductory chapters explored: methods that can been used to elucidate the structure, function and active site of enzymes; the current understanding of the role of puckering and computational methods to simulate enzymes and descriptions of ring pucker and free energy methods. As was discussed in Chapter 1, the puckered conformations of furanose and pyranose carbohydrate rings are central to analysing the action of enzymes on carbohydrates. Enzyme reaction mechanisms are generally inaccessible to experiments and so have become the focus of QM (semi-empirical)/MM simulations. In the previous chapter, a comparison with DFT optimised structures and an *ab initio* free energy surface, revealed that SCC-DFTB provides the best semi-empirical description of 5- and 6-membered carbohydrate ring deformation. In subsequent chapters, it is applied to the study of a glycosidase and a phosphorylase.

Understanding carbohydrate ring pucker is critical to rational design in materials and pharmaceuticals. Using FEARCF SCC-DFTB QM/MM non-Boltzmann simulations, it possible to calculate multidimensional ring pucker free energies of conformation. Here, this strategy is applied to the six-membered glucopyranose ring located in an eight-membered 1-4 linked octaose oligosaccharide (cellooctaose). The cellooctaose was built following the conformation of the saccharides bound to a glycosidase, cellobiohydrolaseI (CBHI) of *Trichoderma reesei.* The free energy of ring puckering of the glucopyranose ring was calculated at the -1 position in vacuum, in water, and bound to the protein. The protein induces ^{4}E and $^{4}H_{3}$ conformations that are much more stable than the usually preferred $^{4}C_{1}$ conformer. Furthermore, for the $^{4}H_{3}$ conformation in the catalytic binding domain, there is significant electronic rearrangement that drives the structure toward the transition state of the glycosylation reaction. The glycosidation reaction, the cleavage of the glycosidic linkage and covalent addition to a glutamate from the protein active site, follows a similar FEARCF protocol but for covalent bond reaction coordinates. CBHI follows the expected retaining Glycoside Hydrolase family 7 mechanism. It uses the active site to favour the ^{4}E conformer of the -1 position glucose of the cellulosic strand. Protonation of the glycosidic linkage, in the mode of acid catalysis, occurs as the linkage breaks with concurrent formation of the oxocarbenium ion which is predominantly $^{4}H_{3}$. The oxocarbenium ion, which is more like a transition state than intermediate, is attacked by an active site nucleophile Glu212. The free

energy of the glycosidation reaction TS is found to be 17.48 kcal/mol. As Glu212 covalently bonds to the anomeric position of the -1 glucose, the pucker reverts to the $^4C_1$ conformer.

As for cellulase, free energy of ring pucker FEARCF SCC-DFTB QM/MM non-Boltzmann simulations are applied to a furanose sugar in Bovine Purine Nucleoside Phosphorylase(PNP). PNP catalyses the reversible phosphorolysis of β-nucleosides to free purine base and ribose-α-1-phosphate. The free energy of ring puckering is calculated for a β-nucleoside, guanosine, in vacuum, in water and in the active site of PNP. When guanosine is placed in the active site of PNP, the enzyme dramatically restricts the available conformational space when compared to vacuum and water simulations. While the interactions of the QM amino acids are imperative for stabilising the incoming phosphate and leaving base, it is not clear what their role is in inducing a preferred puckering conformation. Instead, the entering nucleophile induces the sugar ring conformational preference. The charged phosphate uses its oxygen atoms to interact strongly with the hydroxyls of C2 and C3. This interaction stabilises/induces a preference for the ^4E conformation.

Computer simulations and free energy methods, the FEARCF method in particular, are valuable for investigating the complexities of biomolecules. These methods are especially useful in situations where “traditional” experiments cannot provide all the details needed to understand these systems.

6 Cellulase*

6.1 Introduction

Cellulose, the β-1-4 linked straight chain homo-polymer of glucose, found in all plants, is ideal for the construction of stress-resistant, fibrous materials. As such, it serves a key structural role in plants as a component of lignocellulose (along with hemicellulose and lignin) and is extremely abundant. This high availability means that it is an excellent candidate for use as a source for bio-fuels such as ethanol. However, perhaps due to its important structural role, this material is not easily degraded.[1] *Homo sapiens* can degrade plant derivatives such as α-glycosidically linked branched glucopolymers (starch), but not the linear β-1-4 linked glucopolymer, cellulose. Currently, lignocellulose is converted to fuel via gasification (treatment with controlled amount of oxygen gas and steam) to generate syngas. Alternatively, dilute acid[2, 3] or superheated steam[2, 3] can be used to degrade the crystalline nature of the cellulose to make it amenable to hydrolysis. The ability of ionic liquids to dissolve cellulose suggest their use in the bio-fuel process. Enzymatic hydrolysis by cellulase follows and then a fermentation process is used to produce ethanol. The difficulty of this process arises due to the highly crystalline nature of lignocellulose.

Cellulose is degraded by catalytic hydrolysis via the cellulase class of enzymes, particularly the cellulose β 1-4 cellobiosides (EC 3.2.1.91). There are two types; endoglucanases break internal bonds in amorphous[4, 5] cellulose chains while cellobiohydrolases (CBH) act on crystalline cellulose. These hydrolases are glycosidases as they hydrolyse glycosyl compounds. Specifically they cleave the glycosidic linkages of the polymeric cellulose via glycosylation and deglycosylation. In each of these reactions the glucose ring puckers away from the commonly occurring $^4C_1$ solution chair conformation to take on strained conformers in the transition state. Degradation of cellulose in this manner is ideal as it is catalytically controlled and does not generate hazardous by-products. However, the lignocellulosic biomass has evolved to resist enzymatic assisted breakdown[1] (partly due to the interference of hemicelluloses[6]) making enzymatic cellulose conversion a slow process compared with the breakdown of amylase. A **solely** enzymatic hydrolysis[7] industrial process has not yet been *realised*. Answering the

* Parts of this chapter have been published. Barnett, C. B., Wilkinson, K. A. & Naidoo, K. J. Pyranose Ring Transition State Is Derived from Cellobiohydrolase I Induced Conformational Stability and Glycosidic Bond Polarization. Journal of the American Chemical Society 132, 12800-12803 (2010).

questions of how and why the sugar rings are puckered in hydrolysis reactions may be pivotal in understanding the relative inefficiency of the cellulases.

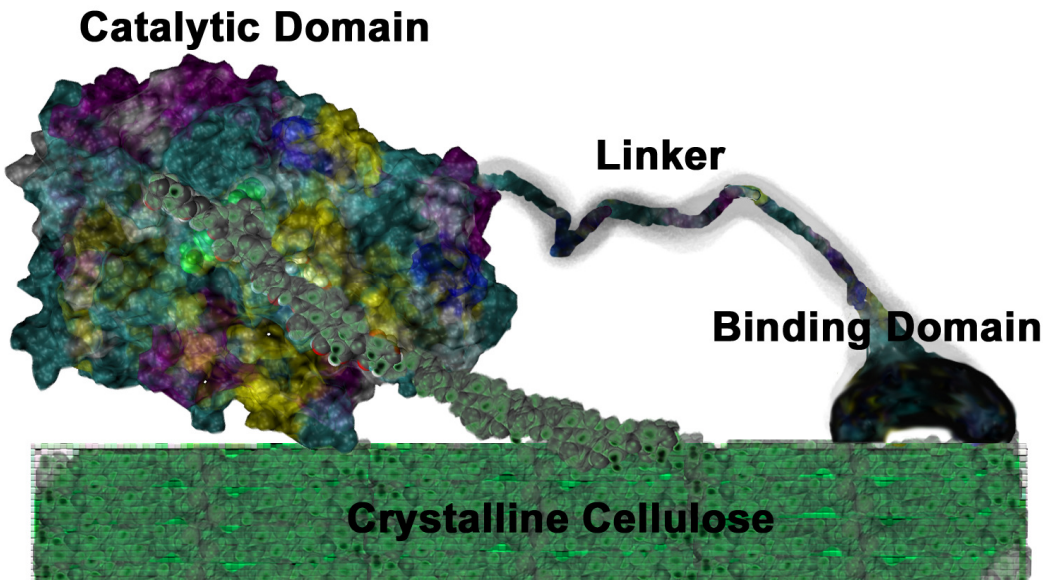

Figure 6-1 An artistic impression of the necessary domains for a functional cellulase (adapted from [8])

The primary action of cellobiohydrolases is to extract a single chain of cellulose (see Figure 6-1) and to cleave the β-1, 4 glycosidic linkage between the main cellulose chain and two terminal residues (Figure 6-2). The glycosidic linkage between position -1 and +1 of the cellulosic strand is cleaved and a shortened cellulosic strand and a cellobiose dimer result.

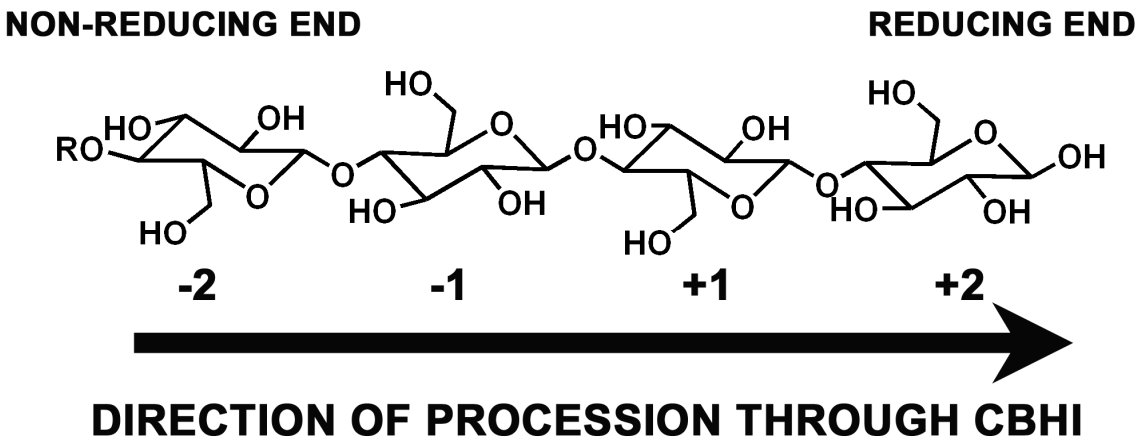

Figure 6-2 Labelling the cellulosic strand. Residues after the glycosidic linkage to be cleaved are labelled positively (+) while those before it are labelled negatively (-). R indicates the remainder of the cellulosic strand. (adapted from [9])

The cellulase acts processively along the cellulosic strand cleaving off cellobiose units which are accordingly ejected from the catalytic domain. Cellulases extract a single chain of cellulose from the cellulose fibril with the aid of the cellulose binding domain [8, 10, 11] (CBD, also sometimes referred to as the carbohydrate binding module (CBM)). The CBD is bound to the cellulase catalytic region by a flexible linker (Figure 6-1) and has conserved tyrosine residues that bind it to the surface of crystalline cellulose.[12] The CBD processes along the surface until it overlaps the reducing end of a broken cellulosic strand. The broken cellulosic strand induces conformational changes[12] in the CBD such that it gains traction in the broken cellulose surface and can direct the cellulose chain towards the catalytic binding channel. The cellulosic strand is directed past the catalytic binding domain where the glycosylation/deglycosylation reactions takes place.

CBH's may be inverting or retaining. CBH I's cleave from the reducing end of a cellulose chain to yield β-cellobiose and CBH II' s act from the opposite end of the chain yielding α-cellobiose (these enzymes have been renamed as Cel7A and Cel6A respectively[13] but will be referred to as CBHI/II) . The enzymes secreted by the high cellulase producing filamentous fungus *Trichoderma reesei* (*T. reesei, Hypocrea Jecorina* is the sexual stage) are known to be highly effective against native, crystalline cellulose. *T. reesei* produces CBHI and CBHII. CBHI is able to act on the inaccessible, crystalline regions of cellulose[14, 15] and it cleaves off cellobiose from the chain ends in a processive manner.[16-18] These enzymes are believed to act synergistically to solubilise cellulose, as shown for native cotton cellulose.[19]

The primary aims of this thesis are to evaluate the free energies of pyranose pucker, gauge how glycosidases influence substrate puckering conformation and evaluate the relationship between the puckering and the mechanism of glycosidases. CBHI from *T. Reesei* fits the guidelines as it is a glycosidase and the substrate is a polymer of a pyranose subunit.

The proposed mechanism for CBHI is shown in Figure 6-3 and from this it is apparent that the first step, the glycosylation, is a general acid mediated mechanism. Overall, the glycosylation step is a substitution reaction in which the oxygen of a glutamate residue (Glu212) substitutes the glycosidic linkage at the anomeric carbon of the -1 position. The cleavage of the glycosidic linkage is assisted by protonation of the glycosidic oxygen by glutamic acid (Glu217). The resultant cellobiose unit is able to leave the binding site.

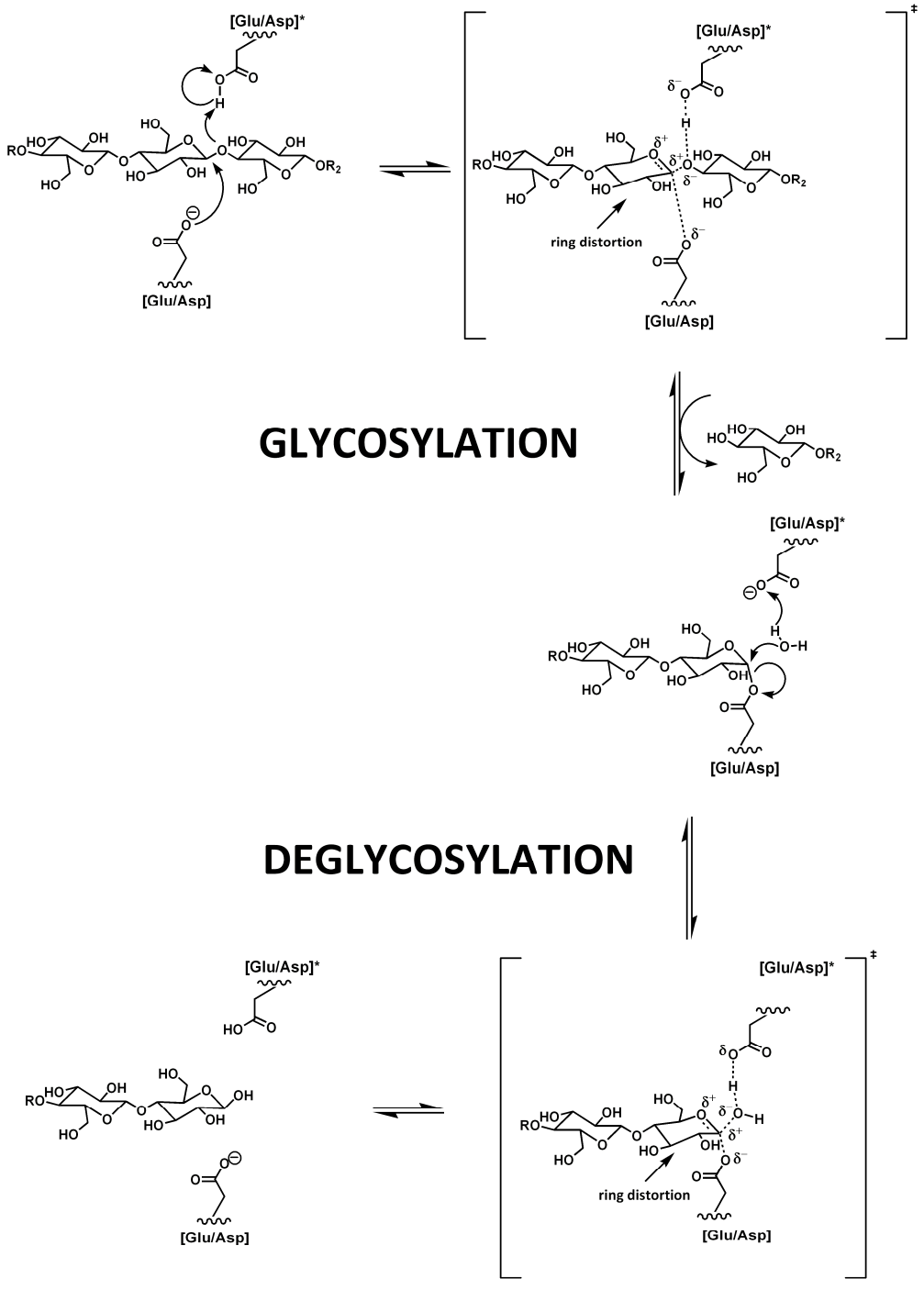

Figure 6-3 The proposed general mechanism for a retaining glycosidase(adapted from [20, 21]). R indicates the remainder of the cellulosic strand while R_2=H or glucose (depending on the specificity of the glycosidase). * indicates protonation and depending on the glycosidase in question the specific catalytic residue may be Glu or Asp

The second part of the mechanism, which is not considered in this thesis, is proposed to be water mediated deglycosylation. Glu217 abstracts a proton from a nearby water molecule reinstating its catalytic activity and also positioning a hydroxyl anion above the anomeric position of -1 ready for attack. On nucleophilic attack, the glutamate-carbon bond is broken and the anomeric carbon stereochemistry is retained. The remaining cellulosic strand is released which subsequently proceeds through the catalytic channel allowing this process to repeat.

The catalytic domain where glycosylation takes place is a specific part of the lengthy binding channel which is surrounded by the β-jelly roll scaffolding of CBHI (Figure 6-4). The binding channel[22] is 40Å in length, has capacity for 9 glycosyl residues[9] and has a bent shape straining the conformations of the glycosyl chain. Mutagenesis studies[23] and crystallographic [22]studies have indicated the importance of Glu212, Asp214 and Glu217 for activity. The Glu-X-Asp-X-X-Glu motif is considered key in this family of enzymes. These residues occur in the latter part of the channel and are the catalytic residues responsible for glycosylation/de-glycosylation. X-ray crystallographic studies have shown the key residues in relation to short oligosaccharide strands.[9]

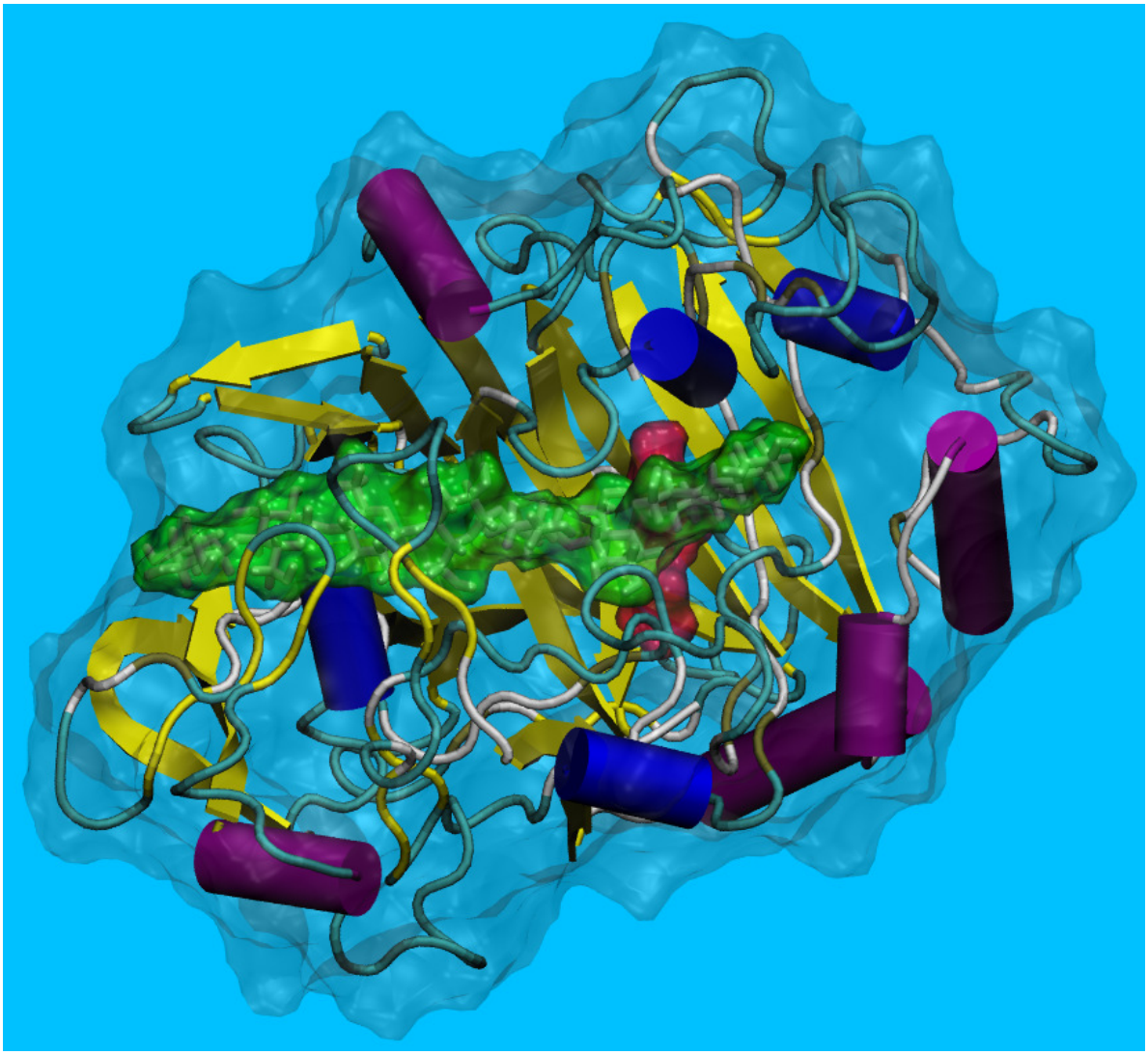

Figure 6-4 The tertiary β-jelly roll structure of CBHI indicated in carton style and coloured to show α helices in purple and deep blue with β sheets being yellow. The outer globular CBHI shape is transparently outlined and the position of a short cellulosic strand in the binding channel is indicated using a green globular surface. Glu, Asp and Glu of the Glu-X-Asp-X-X-Glu motif are indicated using a pink globular surface.

The key interactions along the binding channel were fully detailed by Divne.[9] Amino acids within 4Å of the shortened cellulosic substrate and possible hydrogen bonding interactions are indicated in Figure 6-5. Of particular importance is the interaction of Glu217 to the glycosidic linkage to be cleaved. Divne showed that this interaction depends on the orientation of the substrate in the binding channel.[9]

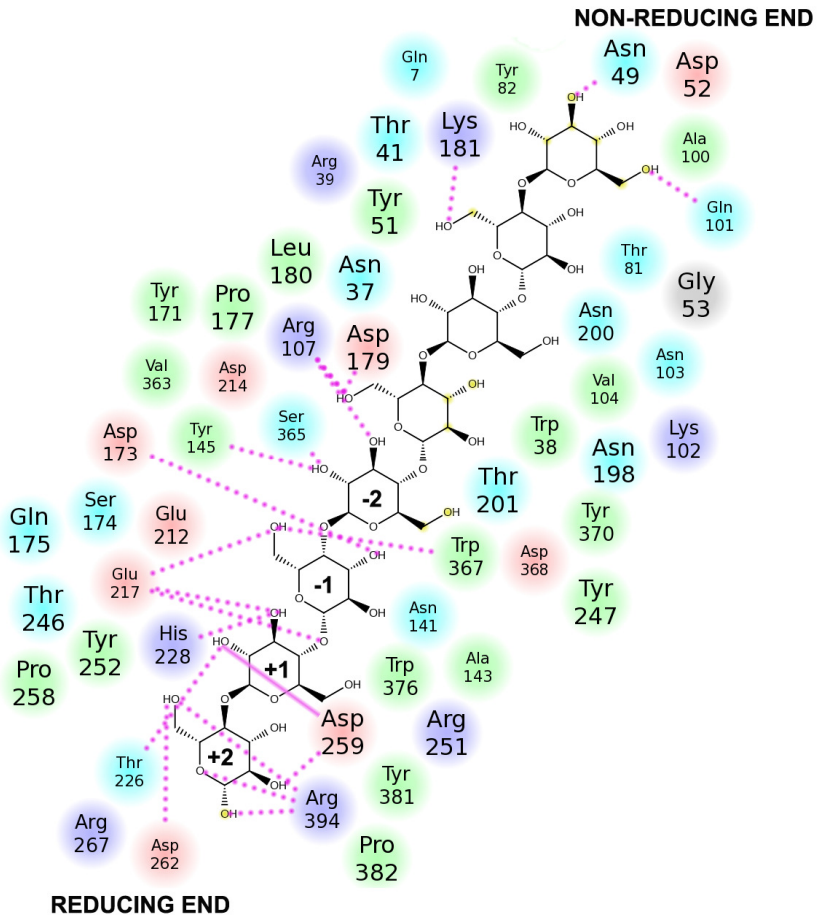

Figure 6-5 2D plot of short cellulosic strand in the binding channel. Amino acids within 4Å are indicated. Hydrophobic residues are indicated in green, cationic residues are deep blue, anionic (although may be protonated) residues are pink. Light blue indicates the Thr, Gln, Ser and Asn amino acids.

Here, the nature of glucose ring puckering at the site in cellulose where cellobiose is being cleaved by CBHI is investigated. This puckering is considered at the stage preceding glycosylation. The role of the protein environment in preparing the substrate for reaction will be considered. Subsequently, glycosylation is simulated to investigate its mechanism.

Application of the FEARCF approach – a versatile free energy approach that provides extensive sampling of potential energy surfaces – can aid in determining transition state structures and barriers. It is applied to CBHI of *T. Reesei* with an 8 membered β 1-4 linked octaose oligosaccharide (cellooctaose) substrate. The influence of the active site amino acids on the puckering free energy conformational space are to be gauged by considering the pucker of the glucose ring in the -1 position. In addition, the

mechanism of the glycosylation reaction is investigated by considering reaction free energies for the incoming Glu212 to the anomeric carbon at position -1 and the breakage of the requisite glycosidic linkage.

6.2 Simulation Details

This section details the setup of the free energy volume calculations in vacuum, water and CBHI. Certain theoretical details have already been reviewed and the reader should refer to Chapter 2 for details. Cellooctaose was built (Figure 6-6) following the conformation of the saccharides bound to CBHI reported in the 7CEL.pdb crystal structure (resolution : 1.90Å).[9] To preserve the linkage conformation, simulations of cellooctaose were performed in vacuum and water with weak harmonic force constants of 1.0N (tethers – see Figure 6-6) applied to either end of the oligosaccharide (Preliminary calculations and this reference[24] showed that saccharides exhibit a complexity of random coil structures, especially folded over self-associated conformers, whereas in the active site this is not possible). The glucose units on either side of the breaking glycosidic bond were treated quantum mechanically (SCC-DFTB/d) and GHO[25] were used to bridge the parts of the sugar that were treated quantum mechanically and those that were treated using the CSFF forcefield.[26] Calculations were run with the mio parameters and hydrogen bonding and dispersion parameters were included.

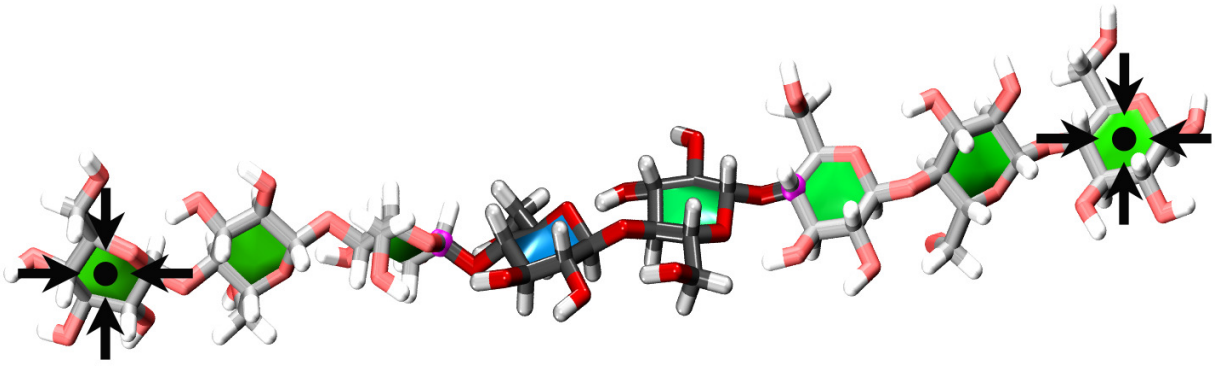

Figure 6-6 Cellooctaose indicating the tethering forces. The reducing and nonreducing ends are loosely held in place as indicated by the focused arrows. The QM region includes the two central sugars (shown in a darker colour) and the GHO atoms are indicated in magenta.

In vacuo leap frog langevin dynamics of the cellooctaose was carried out at 298.15K with group based cutoffs of 10, 12 and 14 Å force shifting for electrostatics and van der Waals potentials. The non-bonded interactions were updated using CHARMM's built-in heuristic algorithm. The SHAKE algorithm

was applied to MM atoms. Eight 300ps QM/MM FEARCF calculations were run for each iteration of the reaction coordinate force simulation. At the end of each iteration the histograms from each simulation were combined with those from the previous iterations. This was done using the weighted histogram analysis method (WHAM[27]) implemented for multidimensional systems.[28] This results in an improved estimate of the free energy volume of puckering. Eleven iterations were required to reach convergence.

The minimised coordinates of cellooctaose from the vacuum calculations were solvated in a 38 Å sphere of TIP3P[29, 30] water. Dynamics were carried out as described for the vacuum simulation but with the addition of a spherical boundary potential to maintain the integrity of the solvent sphere. The system was then heated for 30ps and equilibrated for 100ps with the same dynamics specification as for vacuum. Four QM/MM cellooctaose in water FEARCF calculations of 150ps were run for each iteration of the reaction coordinate force simulation. At the end of each iteration, the histograms from each simulation were combined with those from the previous iterations. This was done as described above. Eleven iterations were run to reach convergence.

The 7CEL.pdb structure as reported by Divne et al[9] was protonated following a pK_a analysis.[31-33] Aspartates and glutamates greater than 18Å far from the active site were protonated to achieve an overall neutrally charged MM part of the system. Three actives site amino acids that were previously found to be significant for transition state *stabilisation* (glutamic acid 217, aspartic acid 214 and glutamate 212) were treated quantum mechanically and placed in the correct orientation in the binding site.[9] The cellooctaose was placed inside the catalytic domain (Figure 6-7) and glucoses at the -1 and +1 position of the substrate were treated quantum mechanically. The QM and MM regions were joined using the GHO method. The quantum region, with an overall charge of –1, was modelled using SCC-DFTB/d.

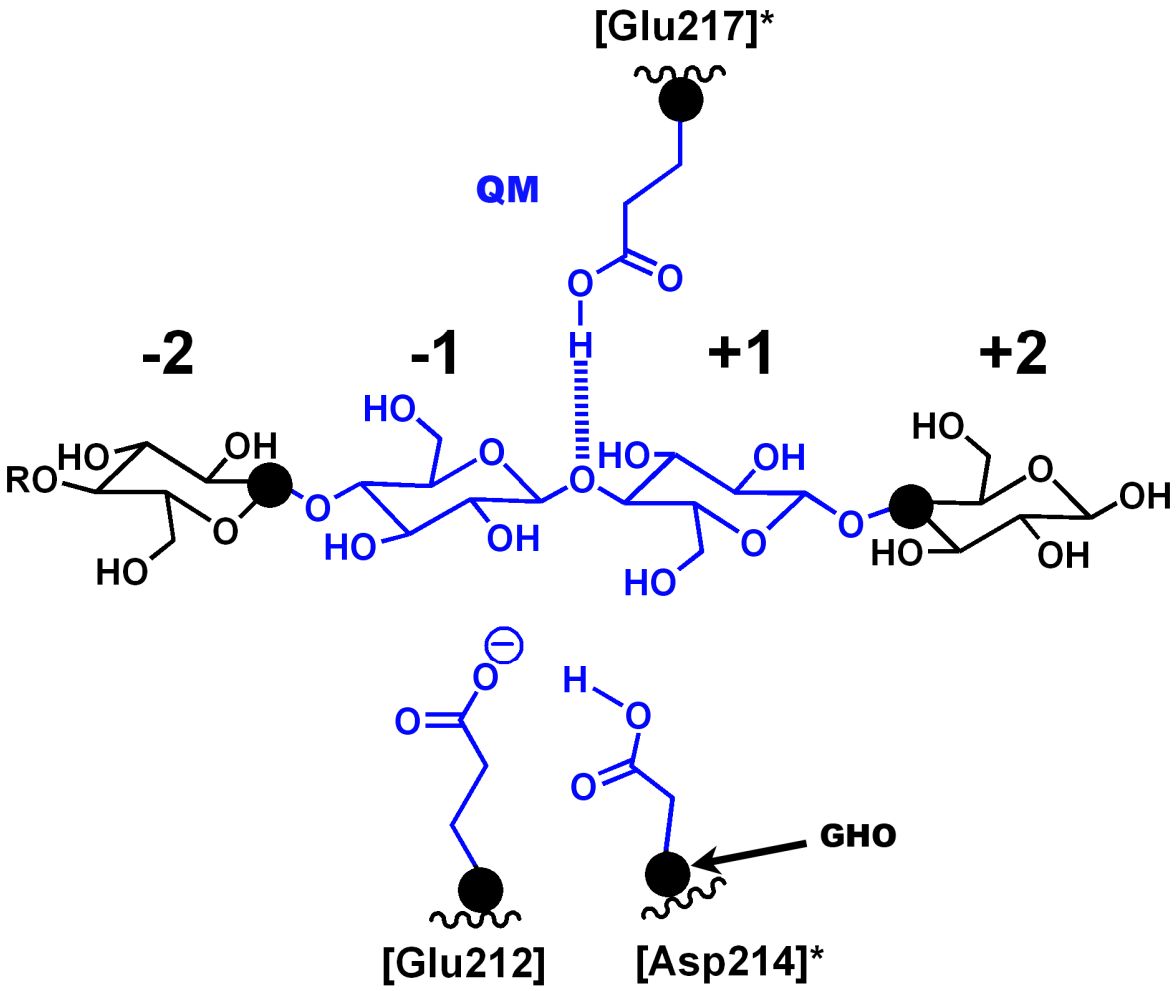

Figure 6-7 The active site indicating the QM (highlighted in blue) and GHO atoms (black spheres) chosen.

After initial ABNR *minimisation* for 500 steps the protein was solvated in a 38 Å sphere of TIP3P water centered on the glucose at the -1 site. Heating and equilibration were carried out under the same conditions as for the water case (the generalised protein setup discussed in Chapter 2 was followed). Eight QM/MM protein FEARCF calculations of 200 ps were run for each iteration of the FEARCF pucker simulation. WHAM was used as described previously. Convergence of the free energy volume was reached after 10 iterations of QM/MM dynamics. Similarly, QM/MM FEARCF reaction simulations were performed, reaching convergence of the free energy surface after 10 iterations of QM/MM dynamics.

6.3 Puckering Free Energies[34]

The free energy analysis of a glucose monomer in vacuum, and for an internal glucose of a short cellulose strand comprising eight glucose monomers (cellooctaose) in vacuum, water and at the catalytic domain (CD) of the CBHI from *T. reesei* was performed. The free energy volumes for these systems are shown in Figure 6-8b, c, d, and e respectively with the available canonical pucker of a 6-membered ring indicated in Figure 6-8a (a full size figure is shown in Chapter 4).

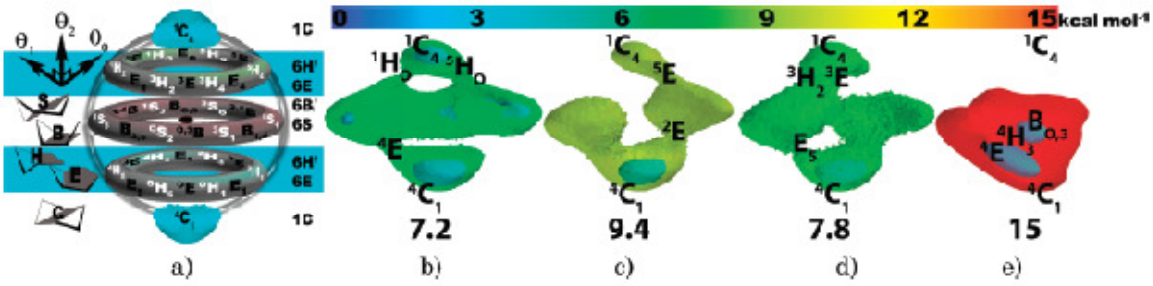

Figure 6-8 a) Pucker definition shown on a reaction coordinate sphere where the planar conformer is at the centre (θ_0, θ_1, θ_2)=(0°, 0°, 0°). The ring pucker free energy volumes for b) the glucose monomer in vacuum, and cellooctaose in c) vacuum, d) water and e) CBHI.

At room temperature (298.15K), pucker conformations separated by low energy barriers (3-5 kT) can be interchangeably accessed. The relationship between these conformers can be observed from the free energy iso-surface at 3kcal/mol (the inner iso energy contours shown in Figure 6-8b-e). It is immediately obvious for all but one of the pucker volumes (the one determined for glucose in the CD of CBHI) that the $^4C_1$ chair and conformers topologically closely related to it are the most favoured. This has been known for some time for glucose in water. Boat and skew-boat conformers along with the less favoured $^1C_4$ can be found for glucose *in vacuo* (inner iso energy contours Figure 6-8b). However, when a glucose monomer is glycosidically linked to other monosaccharides within an oligosaccharide, it appears unable to pucker out of the preferred $^4C_1$ chair conformer. This is the case whether the oligosaccharide is in vacuum (Figure 6-8c) or in water (Figure 6-8d). To pucker the glucose ring from the favoured chair form to the less favoured $^1C_4$ requires forces considerably greater than those to be had from the thermal energy available at room temperature. Therefore, the minimum free energy connecting the $^4C_1$ and $^1C_4$ chair conformers is observed at higher energy levels (outer contours in Figure 6-8). For glucose in vacuum (Figure 6-8b) this is at 7.2 kcal/mol while cellooctaose in vacuum (Figure 6-8c) requires a free

energy of 9.4 kcal/mol to flip the ring and cellooctaose in water (Figure 6-8) requires a free energy of 7.8 kcal/mol.

The free energy volumes mapped out at thermally accessible pucker volumes for cellooctaose bound to CBHI reveals that the sugar ring is puckered out of the favored chair conformer and is restricted to a region of pucker space where half-chair and envelope conformers are preferred. Even when contouring at a higher energy level i.e. 15 kcal/mol (red) the sugar ring at the -1 position of the active site is unable to access the vast majority of the 38 canonical conformers defined by IUPAC.

The glucose monomer is able to traverse several pathways between the $^4C_1$ and $^1C_4$ conformers (Figure 6-8) where for example $^4C_1 \rightarrow {}^4H_5/{}^4E \rightarrow {}^{1,4}B \rightarrow {}^1S_3 \rightarrow {}^1H_O \rightarrow {}^1C_4$ and $^4C^1 \rightarrow {}^4H_5/{}^4E \rightarrow {}^{1,4}B \rightarrow {}^1S_3 \rightarrow B_{O,3} \rightarrow {}^2S_O \rightarrow {}^5H_O \rightarrow {}^1C_4$ are two of the lowest energy pathways that have a similar energy profile but follow different conformational paths. This was extensively elaborated upon in Chapter 4. When exploring the pucker free energy volume of the cellooctaose glucose in vacuum it is immediately apparent that the path linking the $^4C_1$ and $^1C_4$ conformers ($^4C_1 \rightarrow {}^2E \rightarrow {}^2S_O/{}^{2,5}B \rightarrow {}^5E \rightarrow {}^1C_4$) becomes much more restricted (Figure 6-8c) compared with that of the single glucose monomer (Figure 6-8b). The glycosidic bonds to neighboring sugar monomers tether the puckering glucose, limiting the ring flexibility compared with the monomer.

When that same cellooctaose strand is hydrated the resulting glucose ring pucker free energy volume is very different (Figure 6-8d), as is the minimum energy path which is now $^4C_1 \rightarrow E_5 \rightarrow {}^OS_2 \rightarrow B_{2,5} \rightarrow {}^3E/{}^3H_2 \rightarrow {}^1C_4$. Of the conformers in the minimum energy pathway just indicated, the $B_{2,5}$ conformer shows the most strained glycosidic bond (1.54 Å) yet the electron density (ρ, calculated from an Atoms in Molecules[35] analysis and not equivalent to the DFT expressions for ρ in Chapter 2) changes only slightly (0.2104 a.u) compared with the equilibrium $^4C_1$ case. Although the $^3H_2$ conformation is not on the minimum energy path for the cellooctaose vacuum PMF volume, it is an interesting conformer since significant strain on the glycosidic bond (1.56 Å) is observed. The electron density (ρ) in the bond decreases from 0.2540 au in the $^4C_1$ conformer to 0.1922 au in the $^3H_2$ half-chair form.

In the presence of the CBHI enzyme, cleavage of the glycosidic bond was observed and it was surmised that the glucose would assume a high energy boat conformation.[17] In general, for glycosidases an oxocarbenium ion forms in the Michaelis complex that is accompanied by the distortion of the

saccharide ring. The nature of the distortion is complex and poorly understood. This is mainly because enzyme substrates often contain saccharide rings, which exhibit several forms of puckering in the transition state. Using X-ray crystallography, Davies et al. commented that there is good evidence for enzymes to utilize four possible transition-state conformations $^{4}H_{3}$, $^{3}H_{4}$, $^{2,5}B$ and $B_{2,5}$.[36] The puckering conformation of saccharides in the transition state of various glycosidases results in selective reaction pathways and specific reaction products. In the case of pyranoses, competing ring puckers have been observed for several members of the glycoside hydrolase family.[37]

The complete exploration of the puckering free energy landscape of the glucose at catalytic site -1 of the CBHI enzyme reveals that there is a dramatic restriction of pyranose ring flexibility (Figure 6-8e). In the enzyme, both pyranose chair conformations are highly unfavorable. The $^{4}C_{1}$ conformer, that was previously the most favored in all other environments, is now 5.97 kcal/mol higher in energy than the most favored conformation (^{4}E) on the CBHI pucker free energy volume (Table 6-1). The ring puckering conformational space for the glucose at the CBHI catalytic domain is very restricted with most puckers being unachievable. The simplest path through the free energy pucker volume that traverses from the $^{4}C_{1}$ conformer (5.97 kcal/mol) to a stable boat conformer (1.56 kcal/mol) is $^{4}C_{1}\rightarrow{}^{4}E\rightarrow{}^{4}H_{3}\rightarrow B_{O,3}$ where ^{4}E is the most stable pucker conformation overall. When the glucose in the oligosaccharide chain is puckered in the CD, an envelope conformation appears that is 11.96 kcal/mol and 9.42 kcal/mol lower in energy compared with the $^{4}C_{1}$ conformer in the vacuum and aqueous environments respectively. The solvent stabilises the ^{4}E conformer by 2.5 kcal/mol more than the vacuum case. However, the protein makes ^{4}E the most preferred pyranose ring pucker, *stabilising* it by more than 15 kcal/mol compared with the hydrated oligosaccharide. The $^{4}H_{3}$ and ^{4}E conformers are closely related to each other by the movement of C3. The $^{4}H_{3}$ conformer can be accessed via ^{4}E by distorting the C-3 atom below the ring plane from that found in the envelope conformation.

Table 6-1 The bond lengths (Å) and corresponding electron densities (ρ in au.) in the breaking glycosidic bond observed for an internal glucose of cellooctaose *in vacuo* (v) water (w) and at the -1 position of the CBHI i.e., the CBHI catalytic site (e). Free energies are reported in kcal/mol (The maximum error due to sampling is 1 kcal/mol).

Conformer	Cellooctaose (v)			Cellooctaose (w)			Cellooctaose (e)		
	C1---O4	ρ	ΔA	C1---O4	ρ	ΔA	C1---O4	ρ	ΔA
$^4C_1$	1.45	0.2540	0.00	1.40	0.2480	0.00	1.48	0.2213	5.97
^2E	1.43	0.2808	9.60	1.42	0.2619	7.20	—	—	—
^2S_O	1.44	0.2527	6.50	1.51	0.2145	3.10	—	—	—
$^{2,5}B$	1.47	0.2391	6.50	1.42	0.2645	5.50	—	—	—
^5E	1.49	0.2230	8.90	1.44	0.2473	9.60	—	—	—
E_5	1.41	0.2635	8.25	1.46	0.2467	7.82	—	—	—
$^OS_2$	1.43	0.2531	7.00	1.44	0.2254	3.38	—	—	—
$B_{2,5}$	1.51	0.2200	7.00	1.54	0.2104	3.91	—	—	—
^3E	1.41	0.2156	12.48	1.50	0.2183	7.30	—	—	—
$^3H_2$	1.56	0.1922	12.48	1.41	0.2621	7.30	—	—	—
^4E	1.45	0.2503	11.96	1.48	0.2332	9.42	1.48	0.2208	0.00
$^4H_3$	1.49	0.2328	11.16	1.44	0.2498	8.83	1.61	0.1657	3.40
$B_{O,3}$	1.50	0.2199	11.42	1.43	0.2509	6.50	1.56	0.1900	1.56

In these calculations, only pucker forces were applied and no deliberate forces were used to break or form any bonds especially not those participating in the hydrolysis reaction. The difference in bond lengths and electron densities (Table 6-1) of the glycosidic bond (C1—O4) in the $^4C_1$ conformation(both for water and for the enzyme) and the $^4H_3$ conformation in the enzyme environment is therefore remarkable. CBHI *stabilises* the ^4E conformer leading to a $^4H_3$ transition state conformer where the anomeric carbon takes on an sp^2 character making it susceptible to attack from the glutamate 212 (GLU212) nucleophile in the glycosylation step proposed for the CBHI *catalysed* reaction. Here, the $^4H_3$ conformation is just 3.4 kcal/mol above the ^4E global minimum and 2.57 kcal/mol more stable than the normally preferred $^4C_1$ chair conformer. Compared with the chair, the C1-O4 bond lengthens from 1.40 to 1.61 Å. This is accompanied by a decrease in bond electron density from 0.2480 to 0.1657 au.

To determine the extent to which the enzyme has induced an electronic change on the $^4H_3$ conformer, the Merz-Singh-Kollman (MK) scheme[38] was used to calculate the charges on the O5, C1 and O4 atoms (Table 6-2). When comparing the $^4H_3$ half-chair in the enzyme with that of the $^4H_3$ chair conformer in water it is clear that the glycosidic oxygen and anomeric carbon in the former case have increased in negative and positive charge respectively.

Table 6-2 Selected atomic charges for cellooctaose in water (w) and in the CBHI catalytic site (e).

Atom	Cellooctaose (w)		Cellooctaose (e)	
Conformer	$^4C_1$	$^4H_3$	$^4C_1$	$^4H_3$
Ring : O5	-0.51	-0.40	-0.41	-0.26
Anomeric: C1	0.51	-0.16	0.51	0.44
Glycosidic: O4′	-0.39	-0.25	-0.42	-0.55

The free energy volumes (Figure 6-8) show that CBHI induces conformational deformations in the pyranose ring at position -1 of the catalytic site. This, in part, results from a bend in the binding channel at the -1 position. The bent channel causes the cellulosic substrate (Figure 6-9a) to kink away from the $^4C_1$ topology as it passes through the enzyme.

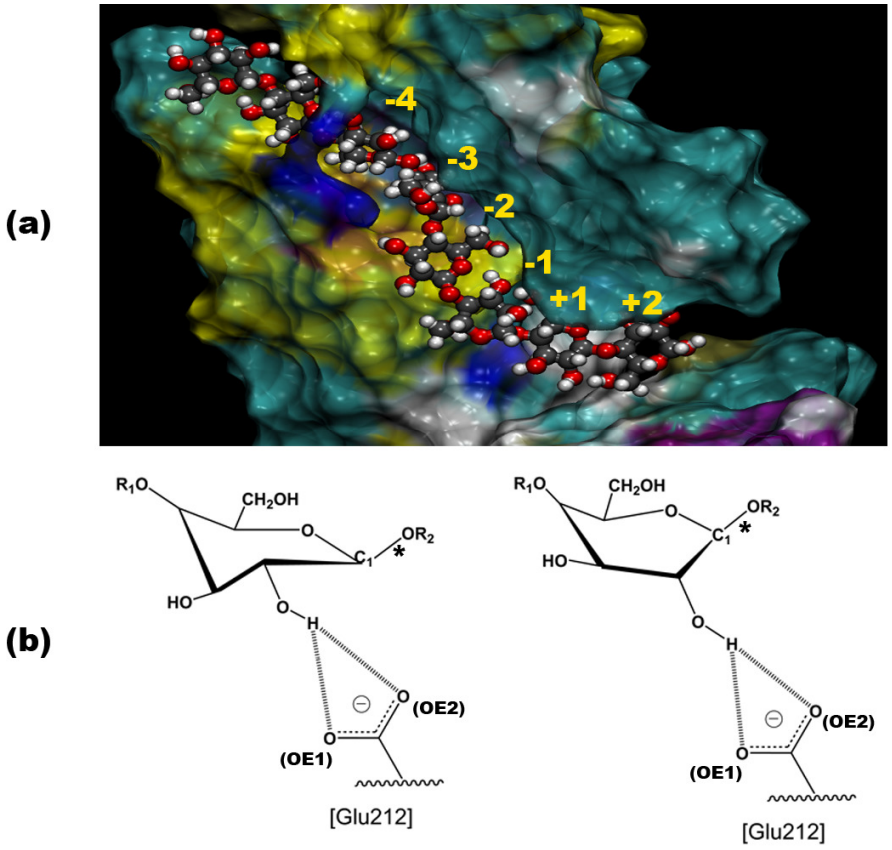

Figure 6-9 (a) A surface cut-away of CBHI of CEL7 reveals the bent shape of the long catalytic binding tunnel. Cellooctaose is drawn as a CPK model and the catalytic -1 location along with others in the CD are shown. b) At -1 the half-chair (left) and envelope (right) conformers are *stabilised* by the glutamate residue (Glu212) forming hydrogen bonds with the hydroxyl at the C2 ring carbon. R_1 is the -8 to -2 part of the cellulosic strand and R_2 is the +1, +2 part of the cellulosic strand. * denotes the glycosidic bond that is catalytically cleaved.

Stabilisation of the envelope and half-chair conformers occurs through hydrogen bonding to a glutamate (Glu212). The C2 ring carbon is levered via its hydroxyl hydrogen bonding with Glu212, which is located on the underside of the -1 sugar (Figure 6-9b). Through a dynamic "push-pull" mechanism it librates the ring between the ^4E envelope and $^4H_3$ half-chair. It does this by sustained hydrogen bonding interactions (between its oxygens, OE1 and OE2, and the hydrogen of O2, Figure 6-9b) which pull the C2 carbon into a plane with respect to the other pyranose ring atoms, except C4, resulting in an ^4E conformation. A time series analysis (included electronically as supplementary data) indicates that only 2.8% of the distorted pyranose conformers (as seen in Figure 6-8e) are not *stabilised* by this hydrogen bond interaction. The electron sharing character of the hydrogen bond (specifically Glu212 OE1 to H of O2) is seen from the high electron densities calculated for it in the ^4E and $^4H_3$ conformers, which are 0.0456 a.u. and 0.0329 a.u. respectively (included electronically as supplementary data). This movement "opens up" the bottom face of the anomeric carbon assisting attack of the nucleophilic Glu212. Furthermore, the distortion of the ring compromises the tetrahedrality at C1, weakening the glycosidic bond and so lowering the reaction barrier.

6.4 Glycosylation Reaction

Much of the knowledge on enzymes stems from a number of prominent studies on lysozyme.[39-44] This enzyme which was serendipitously discovered by Fleming[45, 46], can be *crystallised* without difficulty[44] and was sequenced in 1963.[40] It is found in chicken egg white and so is readily available. This enzyme has been extensively studied, but even so, it was only recently shown that the existence of a covalently bonded glycosyl-enzyme[41] in the mechanism is more likely than the previously postulated, long lived oxocarbenium intermediate. The covalent glycosyl-enzyme forms part of the general mechanism used for glycosyl hydrolases that retain the anomeric configuration of the substrate (Figure 6-3).

CBHI, also being a glycoside hydrolase, has certain similarities with lysozyme. For example, the hydrolysis of the glycosidic bond is catalysed by two amino acids. In the case of CBHI these are Glu217 and Glu212 which act as a general acid/base and a nucleophile respectively while in chicken egg-white lysozyme, Glu35 and Asp52 fulfill these roles. Lysozyme, however, has specificity for peptidoglycans especially those of gram positive bacteria cell walls, while CBHI acts on crystalline cellulase.

The IUBMB specification of enzymes mentioned in Chapter 1 classes enzymes based on their substrate specificity but this is not necessarily related to their structure, as structure-function is not

necessarily a 1:1 mapping. An improved classification scheme of glycoside hydrolases is to use amino acid sequence similarities as amino acid sequence is directly related to the folding (tertiary structure) of proteins.[13, 47-50] The tertiary structure is better conserved than the actual sequence and so it sometimes is the case that different sequences generate similar tertiary structures. In such cases, families are defined which also take mechanism into their specification. Where families share traits, these are grouped into clans. Using this classification CBHI is part of the Glycoside Hydrolase B clan (GH-B) and family GH7 having a β jelly roll structure. GH16 which includes licheninase (EC 3.2.1.73) and xyloglucanase (EC 3.2.1.151) are also part of the GH-B clan. Lysozyme on the other hand (depending on the species it is extracted from) may reside in several families (GH22, GH23, GH24, GH25, GH108). Hen Egg-white Lysozyme (for example 1HEQ) is part of the GH22 family, has been classified as having a tertiary structure of "lysozyme type" and as yet has not been classified to share similar tertiary structures with other families.[13, 47-49] Being in distinctly different families, the tertiary structure ("scaffolding"), primary sequence and certain mechanistic feature of CBHI and lysozyme are different.

The important covalent interactions in the glycosylation reaction are the glycosidic linkage protonation, breakage of the glycosidic linkage and the attack of the nucleophilic glutamate 212 on the anomeric position. To place this into context the active site is sketched in Figure 6-10 and these interactions are highlighted. The reaction coordinates for this are the O-glycosidic linkage-anomeric carbon distance and the Glu212-anomeric carbon distance. Protonation of the glycosidic linkage occurs without additional force assistance.

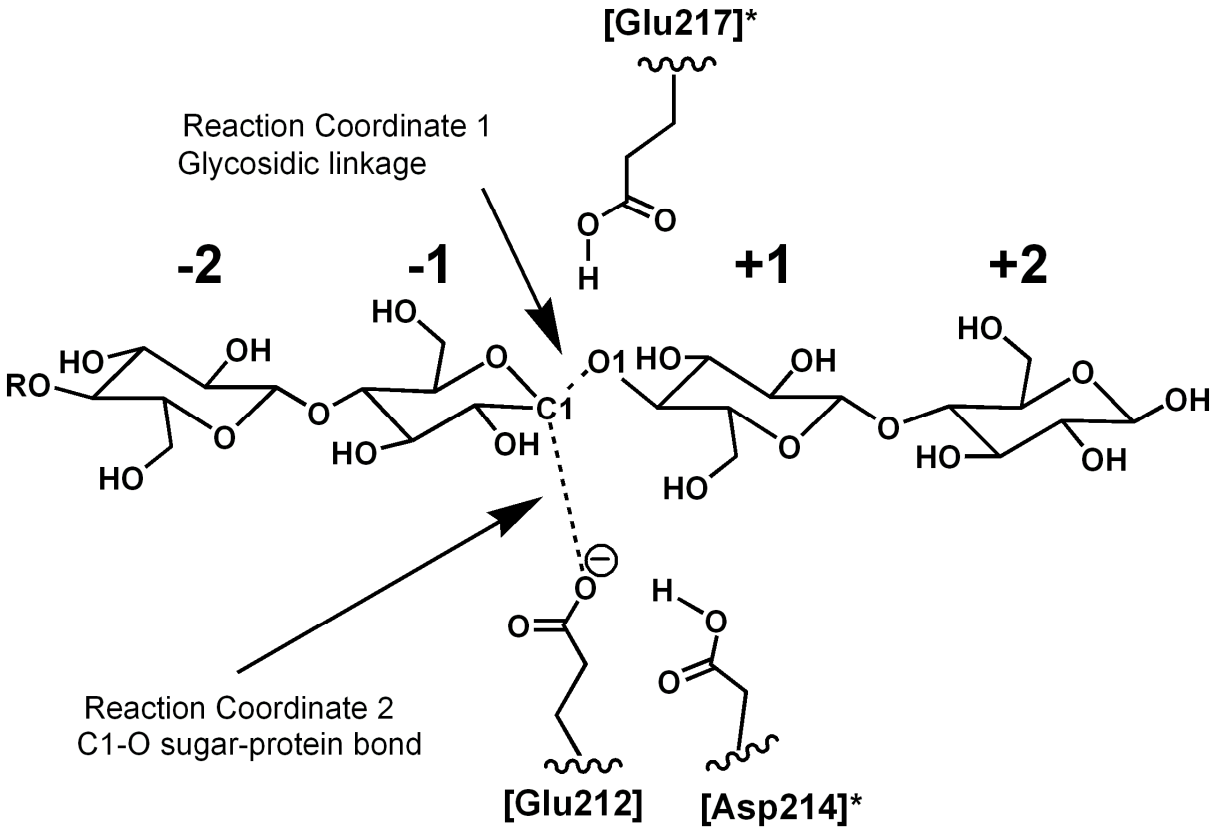

Figure 6-10 Active site of CBHI with the chosen reaction coordinates labelled. R indicates the remainder of the cellulosic strand.

The free energy surface (2D pmf) of glycosylation has been determined using the FEARCF method (Figure 6-11). There are two broad wells which represent the reactant (Figure 6-11a) and the product (Figure 6-11c). These are at (1.5 Å, 3.0 Å) and (3.05 Å, 1.5 Å) respectively. The broad wells are separated by a single transition state barrier and not an intermediate. The reactant well resides at 0 kcal/mol and requires 17.48 kcal/mol to cross the transition state and fall into the product well with a free energy of 0.89kcal/mol.

Having just evaluated the free energy of ring pucker for the reactant in the previous section, it is known that the substrate exhibits ring puckering as CBHI induces preferential puckering of the -1 position of the cellulosic strand. In particular, CBHI *destabilises* the $^4C_1$ conformer and maximally *stabilises* the ^4E conformer with the aid of Glu212 electrostatic and hydrogen bonding interaction to the C2 hydroxyl.

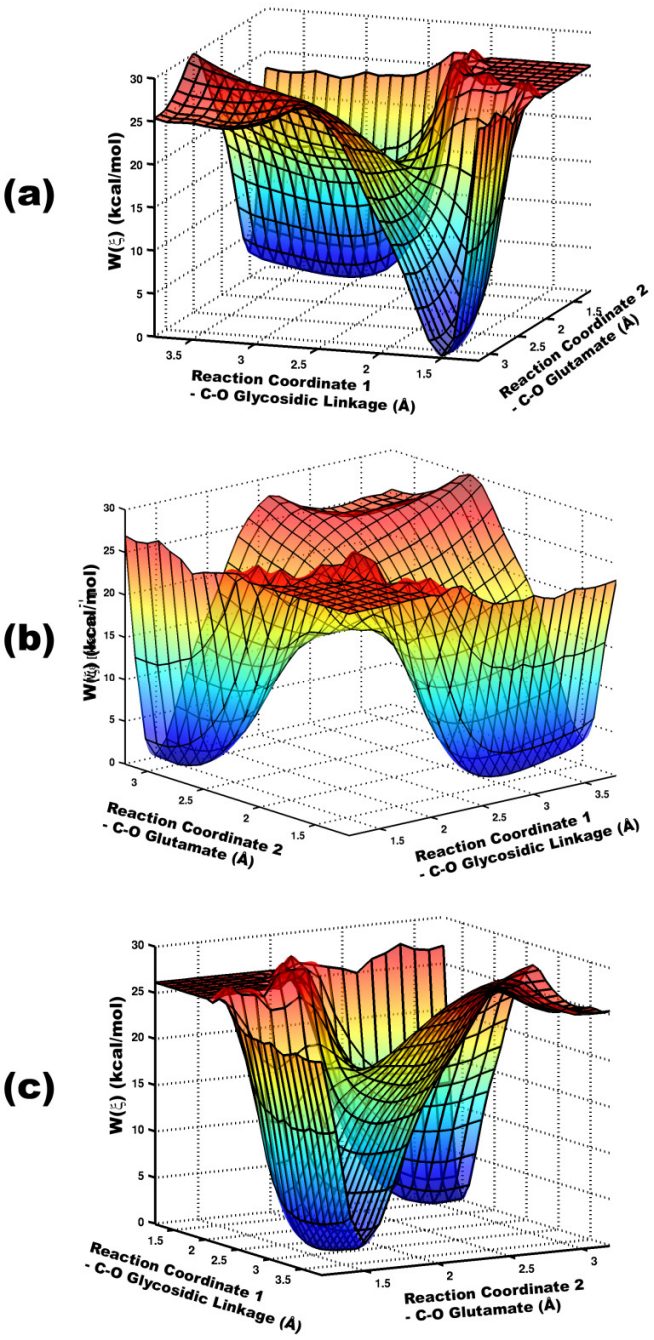

Figure 6-11 2D Free energy surface for the chosen reaction coordinates of CBHI. Viewed from the reactant (a), the TS (b) and the product (c).

It is likely that the ring pucker evolves as the reaction progresses and this is illustrated in Figure 6-12. Scatter plots of certain pucker conformers are shown in the region of the pmf below 18kcal/mol and have been overlaid onto a free energy surface. ^4E, $^4H_3$, E_3 and $^4C_1$ conformers were chosen and the interactions that these conformers may have with the Glu212 are shown in Figure 6-12 (see Figure 6-18 later for evidence of this).

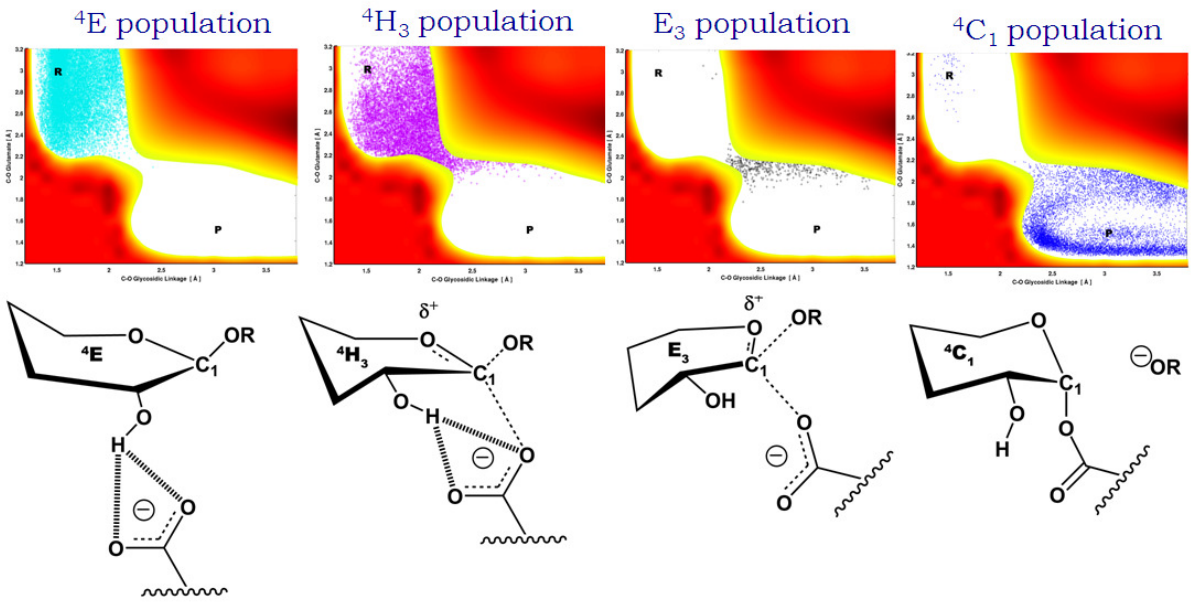

Figure 6-12 The ^4E conformer is prevalent in the reactant region, the $^4H_3$ conformer is prevalent in the reactant and TS region, The E_3 conformers is seen in the TS region. The $^4C_1$ conformer is noted in the TS and product regions.

In the reactant region, a strong Glu212 –HO2 interaction predominantly *stabilises* the ^{4}E conformer with preference for the ^{4}H$_3$ conformer as the reaction proceeds. At the TS the Glu212-HO2 is less strong as the Glu212 has moved to attack the anomeric carbon, thus the $^4H_3$ conformer dominates but some E_3 also exists. On the approach to and in the product region the Glu212 is covalently bonded to the cellulosic strand (now cellohexaose) and does not induce a pulling effect on the C2 carbon.

A minimum free energy path can be calculated on the 2D free energy surface. Figure 6-13a views the free energy surface shown earlier (Figure 6-11) from above and gradient vectors have been etched onto the surface and aimed "downhill". From this, the minimum energy path has been extracted and plotted separately (Figure 6-13b). This simple free energy curve shows the shape/profile of the reaction and it resembles a side on view of the transparent free energy surface (Figure 6-11b) as expected.

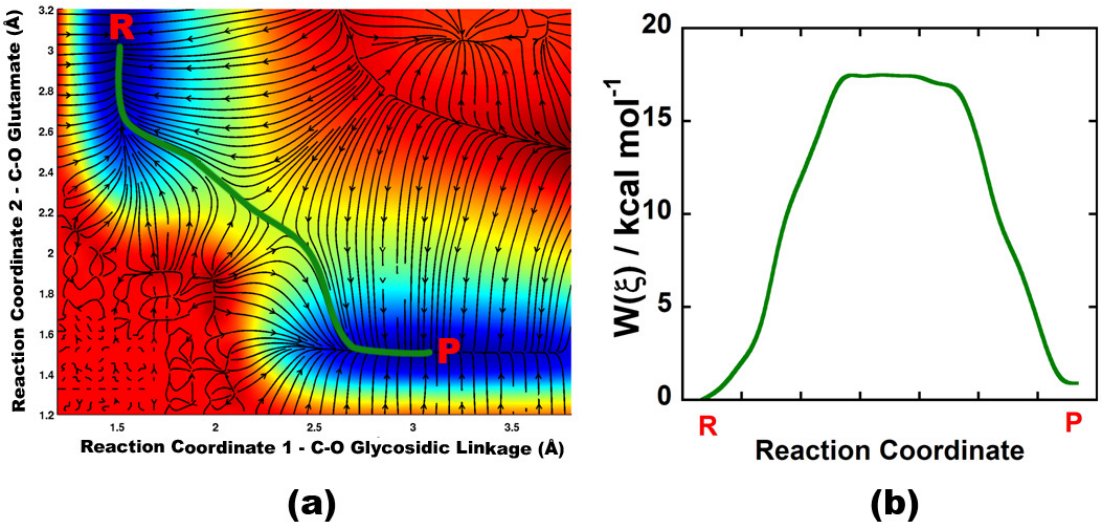

Figure 6-13 (a) A plot of the 2D free energy surface showing the vector gradient paths. The minimum energy path is highlighted. The minimum free energy path is extracted as a (b) 1D graph. R represents the reactant and P the product.

Using the minimum free energy surface as a guide, sample populations can be extracted from the FEARCF dynamics and the puckers states and their fraction of total states assessed as the reaction proceeds along the minimum free energy path. The pucker populations along the minimum free energy path are shown in Figure 6-14. ^4E dominates for the reactant with a good number of $^4H_5$ conformers (a slight deviation from ^4E) but as the reaction progresses the population of $^4H_3$ rises while that of ^4E decreases to zero. $^4H_3$ predominates at the transition state but E_3 and distorted $^4C_1$ are also observed. As the reaction extends past the transition state in the direction of the product, all the conformers become $^4C_1$.

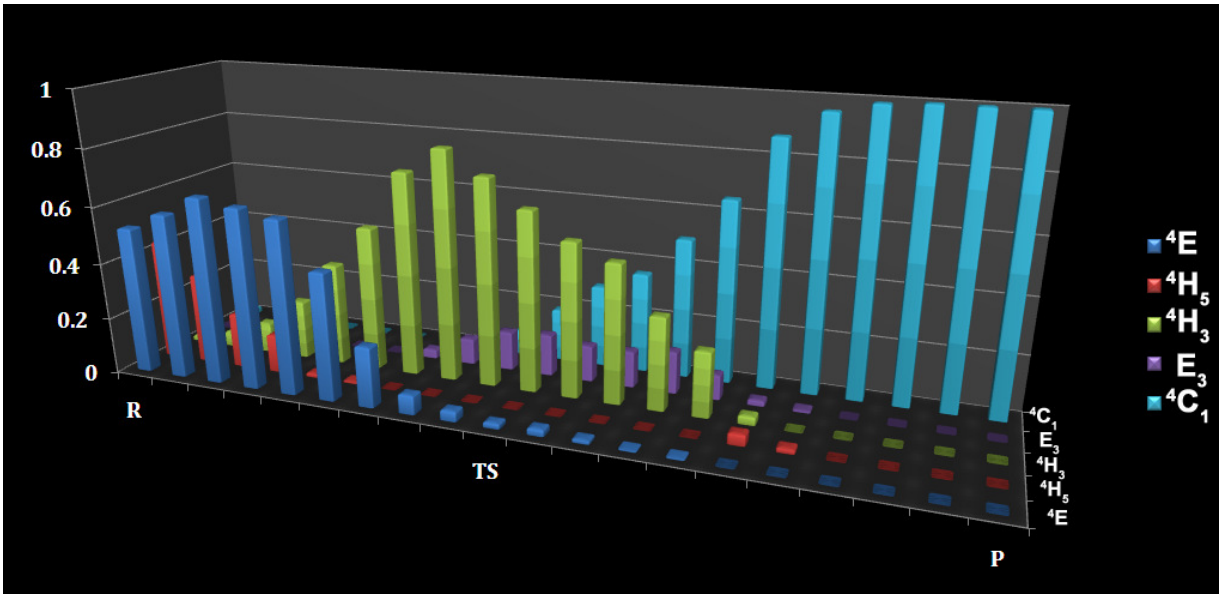

Figure 6-14 Pucker distributions at chosen points along the minimum free energy path. At each point the populations sum to unity.

The nature of the TS has not been fully understood. It is clear that a single TS state exists (there is a single barrier between reactant and product). The lifetime of this state (see later) and its shape (it is not a local minima) indicates that it is not an intermediate. The TS free energy points each represent ensemble averaged FEARCF sampled structures. It is possible that slightly different paths through this region exist – as illustrated by the subtly different conformers ($^4H_3$, E_3 and distorted $^4C_1$) observed at the TS.

The existence of an oxocarbenium ion has been postulated[51] and the oxocarbenium character is taken to be an integral part of the reaction mechanism (Figure 6-3). The TS exists when the reaction coordinates are at (2.3Å, 2.2Å). The glycosidic bond has dissociated slightly and the glutamate has aligned itself for nucleophilic attack. In the TS, the nucleophile and leaving group are approximately the same distance from the anomeric carbon.

Although the reaction coordinate distances are almost equal at the TS, observing the minimum free energy path aids further in elucidation of this mechanism. Following the path from the reactant to just before the TS, in particular from (1.6Å, 2.5Å) to (2.0Å, 2.3Å), it is clear that the glycosidic bond length changes more than the glutamate carbon distance. This suggests that the leaving group (LG – the oxygen of the glycosidic linkage) dissociates more quickly than the nucleophile (Nu – Glu212) associates. Taking

the bond distance between the LG and Nu into account, the nature of the electrophilic carbon centre in this reaction seems to contain both D_NA_N (Dissociation nucleophile, association nucleophile) and D_N*A_N mechanistic character (Figure 6-15). Taking the distance of the LG away from the anomeric carbon and the oxocarbenium like character into account, the D_N*A_N is most true to the simulation results.

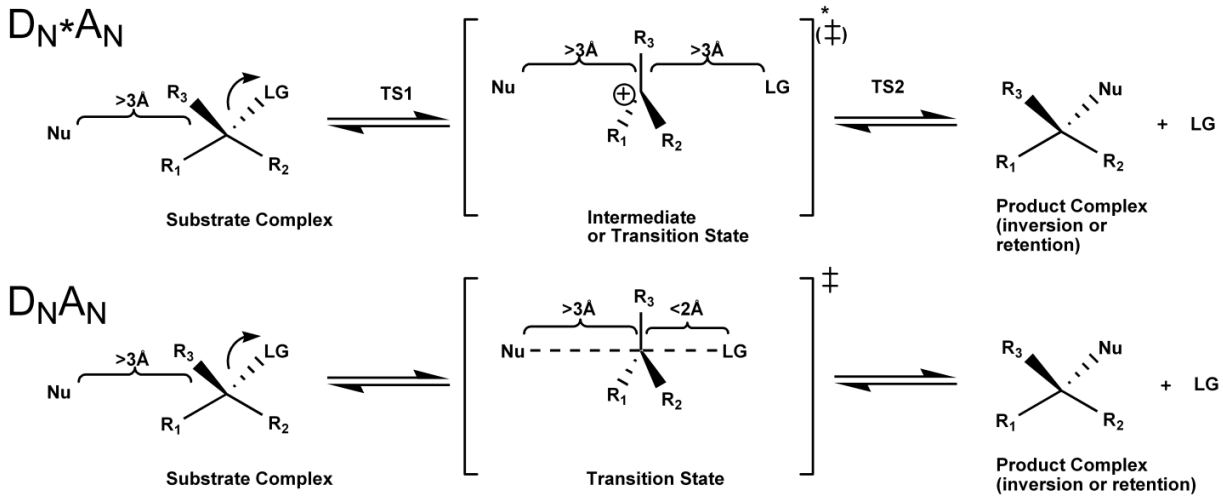

Figure 6-15 The D_N*A_N and D_NA_N mechanisms (adapted from [52])

Monitoring the reaction coordinate distances over part of a FEARCF trajectory (Figure 6-16) shows that this transition occurs in less than 1ps, as expected.[51] Crossing of the reaction coordinates in this trajectory occurs at (2.2Å, 2.2Å) as seen in the 2D pmf (Figure 6-12).

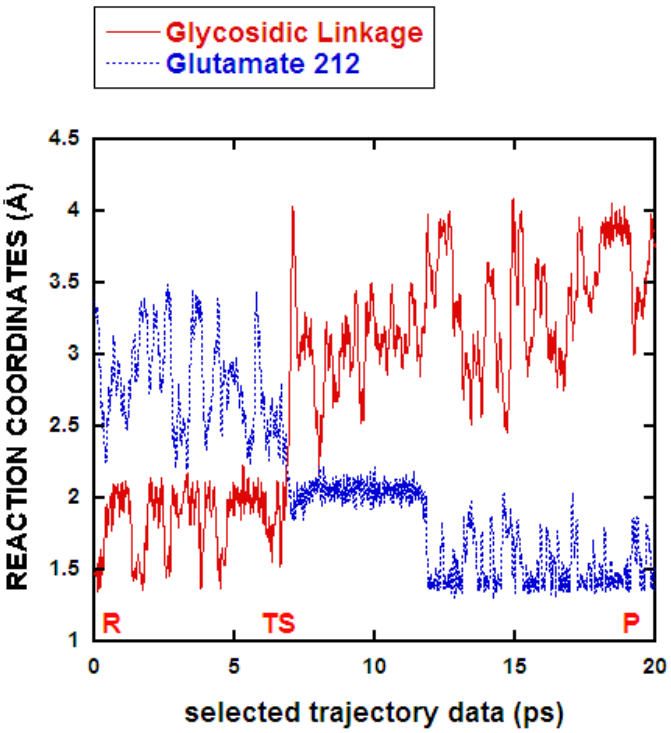

Figure 6-16 The reaction coordinates plotted for a selected FEARCF trajectory going from reactant to product.

The glycosidic dissociation occurs in conjunction with protonation of the glycosidic oxygen by glutamic acid (Glu217). This protonation occurs readily as the hydrogen of Glu217 is already hydrogen bonded to the glycosidic linkage encouraging its dissociation (Figure 6-17).

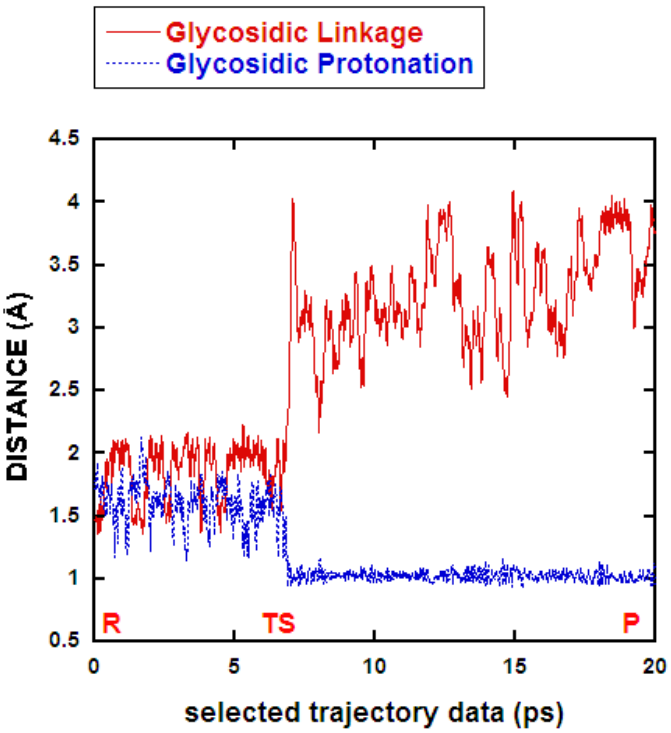

Figure 6-17 Protonation distance *vs* glycosidic distance.

To further investigate the role of the nucleophile (Glu212) in the observed puckering of the reaction, the change in glutamic acid reaction coordinate, pucker conformation and HO2-glutamic acid hydrogen bonds interaction can be plotted. This is done for a selected FEARCF trajectory (Figure 6-18). At least one of the oxygens of the nucleophile (Glu212) is responsible for a strong interaction with the hydroxyl of C2 during the existence of the reactant. However, when the product is formed this interaction is not possible. This confirms the proposed structures in Figure 6-12.

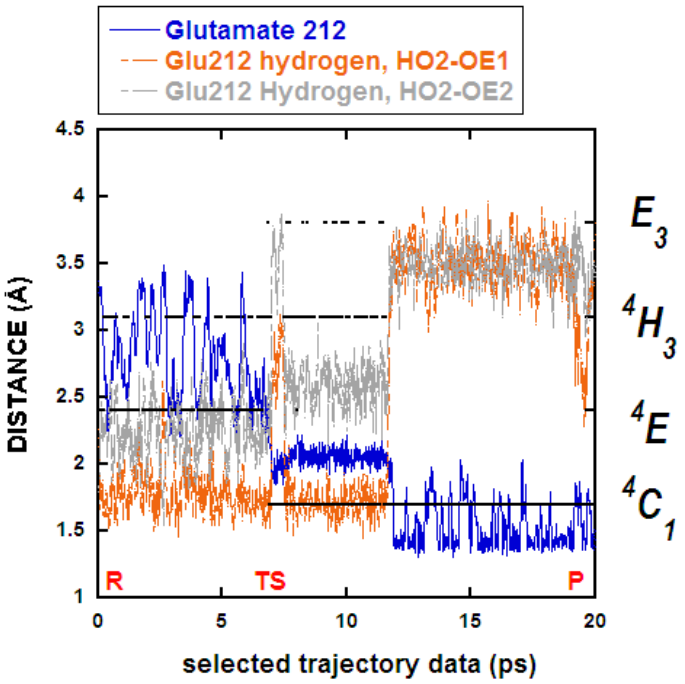

Figure 6-18 The extent of reaction (as a function of the glutamic acid) is plotted together with the selected pucker conformers. The distance between the glutamic acid oxygens (OE1, OE2) and the HO2 hydrogen is also plotted.

Although the puckers at the TS are distorted and certain parts of the ring are planar, this does not guarantee that the anomeric carbon is actually sp^2 hybridised. The anomeric carbon will be planar and sp^2 like in character if the ring oxygen O5, the anomeric carbon, C1, the carbon adjacent to it C2 and the anomeric hydrogen H1 are in the same plane. Using the H1-C2-C1-O5 dihedral angle the planarity of these atoms can be observed. In Figure 6-19 the extent of reaction and planarity of the dihedral are compared. In the reactant well, the hydrogen points down (it is α), at the transition stateit is planar, while in the product well it is β (pointing up).

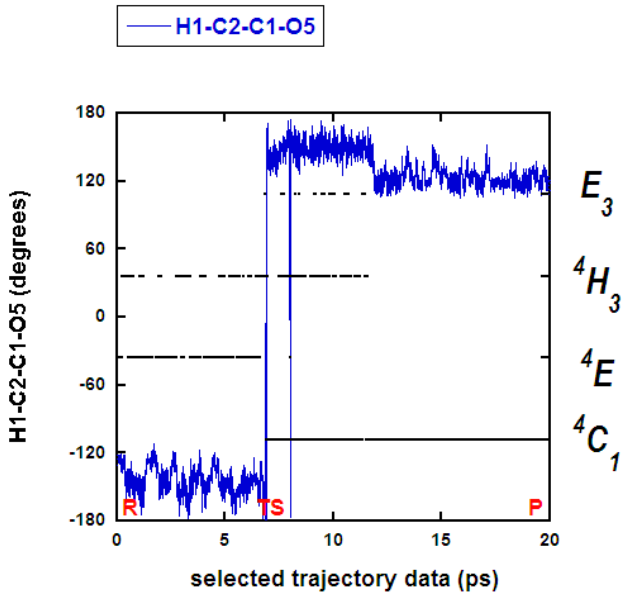

Figure 6-19 The H1-C2-C1-O5 torsional angle is used to observe the planarity of the hydrogen of C1 and this is compared with the puckering conformation over a short trajectory during which glycosylation occurs.

The planarity does not confirm that the anomeric carbon is sp^2 hybridised (think of the trigonal bipyramidal transition state of S_N2 reactions (D_NA_N)). An observation of the C1-O5 bond length, and shortening of this bond indicates the oxocarbenium-like character of the TS. Comparing the C1-O5 bond length and pucker in Figure 6-19 and reaction coordinates in Figure 6-16 confirms that the C1-O5 bond shortens at the TS. That is, when the hydrogen is in the plane of C1, C2 and O5, the C1-O5 bond gains double bond character and the anomeric carbon is sp^2 hybridised – the oxocarbenium character exists.

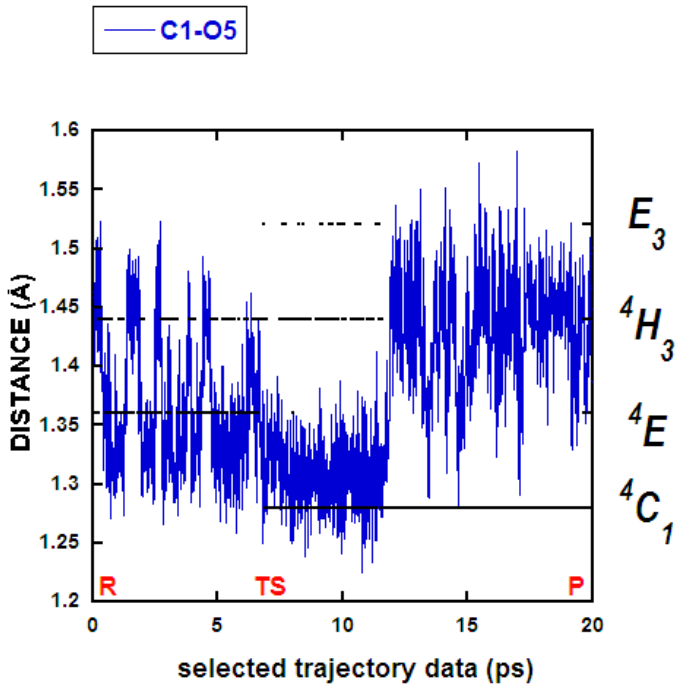

Figure 6-20 The C1-O5 bond distance and the puckering plotted for a short trajectory

6.5 Conclusions

In summary, CBHI follows the expected retaining GH7 family mechanism. It uses the active site nucleophile Glu212 to pull the C2 hydroxyl and favour the ^{4}E conformer of the -1 position glucose of the cellulosic strand. Protonation of the glycosidic linkage in the mode of acid catalysis is done by the glutamic acid Glu217 which encourages dissociation of the linkage. As the linkage breaks, cellobiose is released with concurrent formation of the oxocarbenium ion which is predominantly $^{4}H_{3}$. The oxocarbenium ion which is more like a transition state than intermediate is attacked by the nucleophile Glu212. At its pinnacle, this reaction requires 17.48 kcal/mol to overcome the barrier to glycosylation. As Glu212 covalently bonds to the anomeric position its influence on the C2 hydroxyl ceases and so the $^{4}C_{1}$ conformer is preferred.

6.6 References

1. M. E. Himmel, S.-Y. Ding, D. K. Johnson, W. S. Adney, M. R. Nimlos, J. W. Brady and T. D. Foust, *Science*, 2007, **315**, 804-807.
2. http://www.iogen.ca/index.html, http://www.iogen.ca/cellulose_ethanol/what_is_ethanol/process.html, http://www.iogen.ca/index.html, http://www.iogen.ca/cellulose_ethanol/what_is_ethanol/process.html, Accessed 10-09-2010.
3. http://www.forfuelfreedom.com/ethanol_process.htm, http://www.forfuelfreedom.com/ethanol_process.htm, Accessed 10-09-2010.
4. E. Rudnik, *Compostable Polymer Materials*, Elsevier, 2008.
5. *Biocatalysis and Biomolecular Engineering*, Wiley, 2010.
6. S. Y. Ding and M. E. Himmel, *Journal of Agricultural and Food Chemistry*, 2006, **54**, 597-606.
7. J. Sheehan and M. Himmel, *Biotechnology Progress*, 1999, **15**, 817-827.
8. G. T. Beckham, J. F. Matthews, Y. J. Bomble, L. Bu, W. S. Adney, M. E. Himmel, M. R. Nimlos and M. F. Crowley, *The Journal of Physical Chemistry B*, 2010, **114**, 1447-1453.
9. C. Divne, J. Ståhlberg, T. T. Teeri and T. A. Jones, *Journal of Molecular Biology*, 1998, **275**, 309-325.
10. L. Zhong, J. F. Matthews, P. I. Hansen, M. F. Crowley, J. M. Cleary, R. C. Walker, M. R. Nimlos, C. L. Brooks Iii, W. S. Adney, M. E. Himmel and J. W. Brady, *Carbohydrate Research*, 2009, **344**, 1984-1992.
11. C. Mulakala and P. J. Reilly, *Proteins*, 2005, **60**, 598-605.
12. M. R. Nimlos, J. F. Matthews, M. F. Crowley, R. C. Walker, G. Chukkapalli, J. W. Brady, W. S. Adney, J. M. Cleary, L. Zhong and M. E. Himmel, *Protein Engineering Design and Selection*, 2007, **20**, 179-187.
13. B. Henrissat, T. T. Teeri and R. A. J. Warren, *FEBS letters*, 1998, **425**, 352-354.
14. M. Vrsanska and P. Biely, *Carbohydrate Research*, 1992, **227**, 19-27.
15. B. K. Barr, Y. L. Hsieh, B. Ganem and D. B. Wilson, *Biochemistry*, 1996, **35**, 586-592.
16. C. Divne, J. Stahlberg, T. Reinikainen, L. Ruohonen, G. Pettersson, J. Knowles, T. Teeri and T. Jones, *Science*, 1994, **265**, 524-528.
17. C. Divne, J. Stahlberg, T. T. Teeri and T. A. Jones, *J. Mol. Biol.*, 1998, **275**, 309-325.
18. G. T. Beckham, J. F. Matthews, Y. J. Bomble, L. T. Bu, W. S. Adney, M. E. Himmel, M. R. Nimlos and M. F. Crowley, *J. Phys. Chem. B*, 2010, **114**, 1447-1453.
19. K. M. Kleman-Leyer, M. Siika-Aho, T. T. Teeri and T. K. Kirk, *Applied and Environmental Microbiology*, 1996, **62**, 2883-2887.
20. G. J. Davies, L. Mackenzie, A. Varrot, M. Dauter, A. M. Brzozowski, M. Schulein and S. G. Withers, *Biochemistry*, 1998, **37**, 11707-11713.
21. G. J. Davies, M. Dauter, A. M. Brzozowski, M. E. Bjornvad, K. V. Andersen and M. Schulein, *Biochemistry*, 1998, **37**, 1926-1932.
22. C. Divne, J. Stahlberg, T. Reinikainen, L. Ruohonen, G. Pettersson, J. K. Knowles, T. T. Teeri and T. A. Jones, *Science*, 1994, **265**, 524-528.
23. J. Ståhlberg, C. Divne, A. Koivula, K. Piens, M. Claeyssens, T. T. Teeri and T. A. Jones, *Journal of Molecular Biology*, 1996, **264**, 337-349.
24. S. Queyroy, F. Müller-Plathe and D. Brown, *Macromolecular Theory and Simulations*, 2004, **13**, 427-440.
25. J. Gao, P. Amara, C. Alhambra and M. J. Field, *The Journal of Physical Chemistry A*, 1998, **102**, 4714-4721.

26. M. Kuttel, J. W. Brady and K. J. Naidoo, *Journal of Computational Chemistry*, 2002, **23**, 1236-1243.
27. S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen and P. A. Kollman, *Journal of Computational Chemistry*, 1995, **16**, 1339-1350.
28. K. J. Naidoo and J. Strümpfer, *Journal of Computational Chemistry*, 2010, 308-316.
29. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *Journal of Chemical Physics*, 1983, **79**, 926-935.
30. W. L. Jorgensen and C. Jenson, *Journal of Computational Chemistry*, 1998, **19**, 1179.
31. J. E. Nielsen, *Journal of Molecular Graphics and Modelling*, 2007, **25**, 691-699.
32. J. E. Nielsen and J. A. McCammon, *Protein Science*, 2003, **12**, 1894-1901.
33. C. R. Sondergaard, L. P. McIntosh, G. Pollastri and J. E. Nielsen, *Journal of Molecular Biology*, 2008, **376**, 269-287.
34. C. B. Barnett, K. A. Wilkinson and K. J. Naidoo, *Journal of the American Chemical Society*, 2010, **132**, 12800-12803.
35. R. F. W. Bader, in *Handbook of Molecular Physics and Quantum Chemisty*, John Wiley and sons, Chichester, 2003, pp. 770-791.
36. G. J. Davies, V. M. A. Ducros, A. Varrot and D. L. Zechel, *Biochemical Society Transactions*, 2003, **31**, 523-527.
37. S. Fushinobu, B. Mertz, A. D. Hill, M. Hidaka, M. Kitaokac and P. J. Reilly, *Carbohyd. Res.*, 2008, **343**, 1023-1033.
38. U. C. Singh and P. A. Kollman, *J. Comp. Chem.*, 1984, **5**, 129-145.
39. L. N. Johnson, *Nature Structural and Molecular Biology*, 1998, **5**, 942-944.
40. R. E. Canfield, *Journal of Biological Chemistry*, 1963, **238**, 2698-2707.
41. D. J. Vocadlo, G. J. Davies, R. Laine and S. G. Withers, *Nature*, 2001, **412**, 835-838.
42. C. B. Post and M. Karplus, *Journal of the American Chemical Society*, 1986, **108**, 1317-1319.
43. X. M. Cherian, S. A. Van Arman and A. W. Czarnik, *Journal of the American Chemical Society*, 1988, **110**, 6566-6568.
44. C. C. F. Blake, D. F. Koenig, G. A. Mair, A. C. T. North, D. C. Phillips and V. R. Sarma, *Nature*, 1965, **206**, 757-761.
45. D. M. Chipman and N. Sharon, *Science*, 1969, **165**, 454-&.
46. A. Fleming, *Proceedings of the Royal Society of London Series B - Containing Papers of A Biological Character*, 1922, **93**, 306.
47. B. Henrissat, *Biochemistry Journal*, 1991, **280 (Pt 2)**, 309-316.
48. B. Henrissat and A. Romeu, *Biochemistry Journal*, 1995, **311 (Pt 1)**, 350-351.
49. G. Davies and B. Henrissat, *Structure*, 1995, **3**, 853-859.
50. B. L. Cantarel, P. M. Coutinho, C. Rancurel, T. Bernard, V. Lombard and B. Henrissat, *Nucleic Acids Research*, 2009, **37**, D233-238.
51. V. L. Schramm, *Current Opinion in Structural Biology*, 2005, **15**, 604-613.
52. V. L. Schramm and W. Shi, *Current Opinion in Structural Biology*, 2001, **11**, 657-665.

7 Purine Nucleoside Phosphorylase

7.1 Introduction

The homotrimeric enzyme, Purine Nucleoside Phosphorylase(PNP), catalyses the reversible phosphorolysis of β-nucleosides to free purine base and ribose-α-1-phosphate. Although the formation of nucleoside is usually thermodynamically favoured (Figure 7-1) the phosphorolysis direction is favoured when the PNP reaction is coupled to purine base oxidation or phosphoribosylation (by xanthine oxidase or hypoxanthineguanine phosphoribosyltransferase respectively) due to rapid metabolic removal of purines.[1, 2]

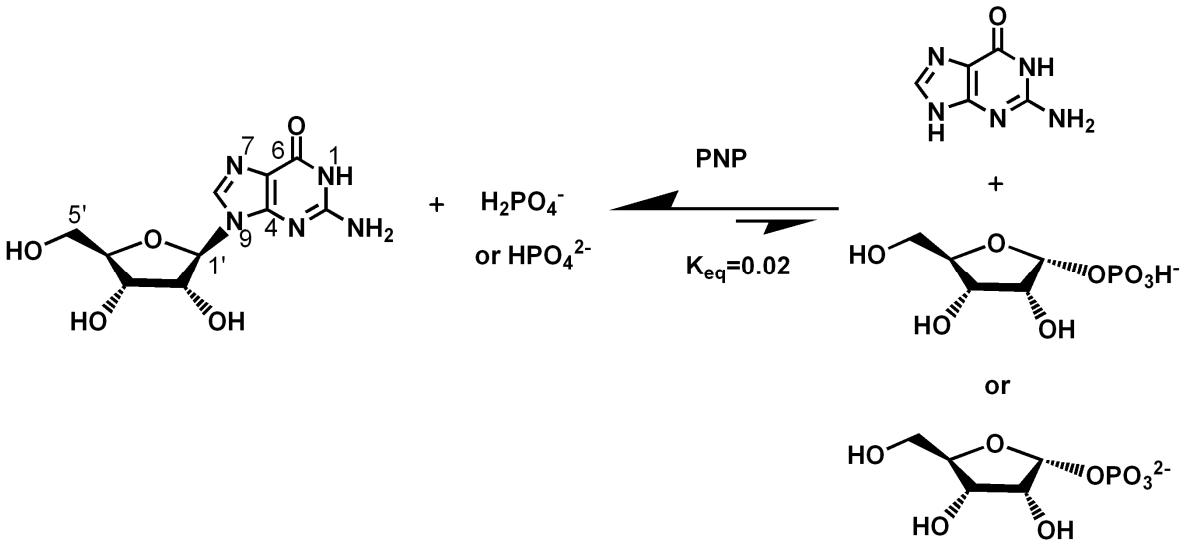

Figure 7-1 The generic mechanism of PNP (adapted[1 3])

Deficiency of PNP[4] reduces the immune effect of T-cells causing developmental disorders and autoimmune disease. This is as a result of build-up of deoxy-guanosine-triphosphate (dGTP) in tissues especially in the lymph causing T-cell apoptosis.[5] Over-active T-cells can cause certain autoimmune disorders, tissue transplant rejection and several cancers. Inhibition of PNP can be used to induce T-cell apoptosis and so PNP has been targeted for rational drug design.[1, 5]

Bovine PNP is a homotrimer with P213 symmetry and the active site is situated at the frontier between subunits (Figure 7-2). Mammalian PNPs also share this symmetry, but not all PNPs do.[6] Figure 7-2a highlights individual subunits of PNP while Figure 7-2b emphasises the secondary structure. The active site in a particular monomer would be very exposed if it were not capped by the adjacent monomer.

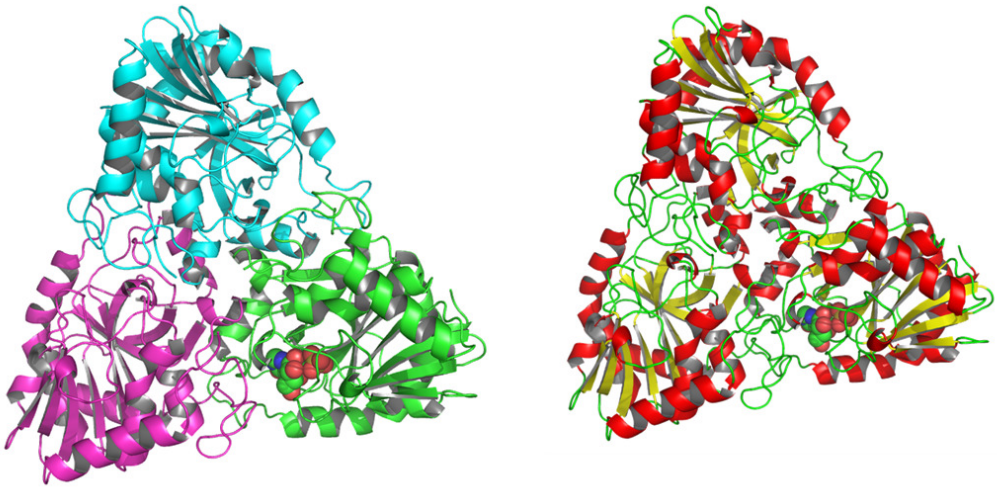

Figure 7-2 The trimeric structure of 1A9S[3] PNP. (a) the subunits of the protein are highlighted. The secondary structure highlighted with alpha helices in red, beta sheets in yellow and random coil in green. The van der Waals spheres in one of the monomers indicates the position of one of the active sites.

The exact nature of the PNP scaffold for various PNPs and the location of the binding sites have been elucidated with crystallographic techniques. The binding site location for various substrates has been established using this technique. In structure –function studies, where the steady-state kinetics of amino acid mutations[7] of PNP are studied, crystallography techniques validate the mutation and are able show modifications to the substrate binding mode.

To rationally design inhibitors, the catalytic activity of amino acids and their mechanistic role must be determined. Methods for design of inhibitors have included using a combination of crystallographic and computer modelling.[8] Erion et al.[1] studied the mechanism by implementing a model of the active site that included only the amino acids that interacted closely with the substrate. Using this site model, specific active site mutants were created. The effect of the mutation was determined by the change in k_{cat} for inosine hydrolysis. His86Ala Glu89Ala mutations showed 10-25 fold reduction in catalytic activity while the Asn243Ala mutant decreased k_{cat} by 1000-fold, suggesting its importance in stabilisation of the

TS. Figure 7-3 illustrates Erion's site model, the substrate in the active site and the surrounding amino acids.

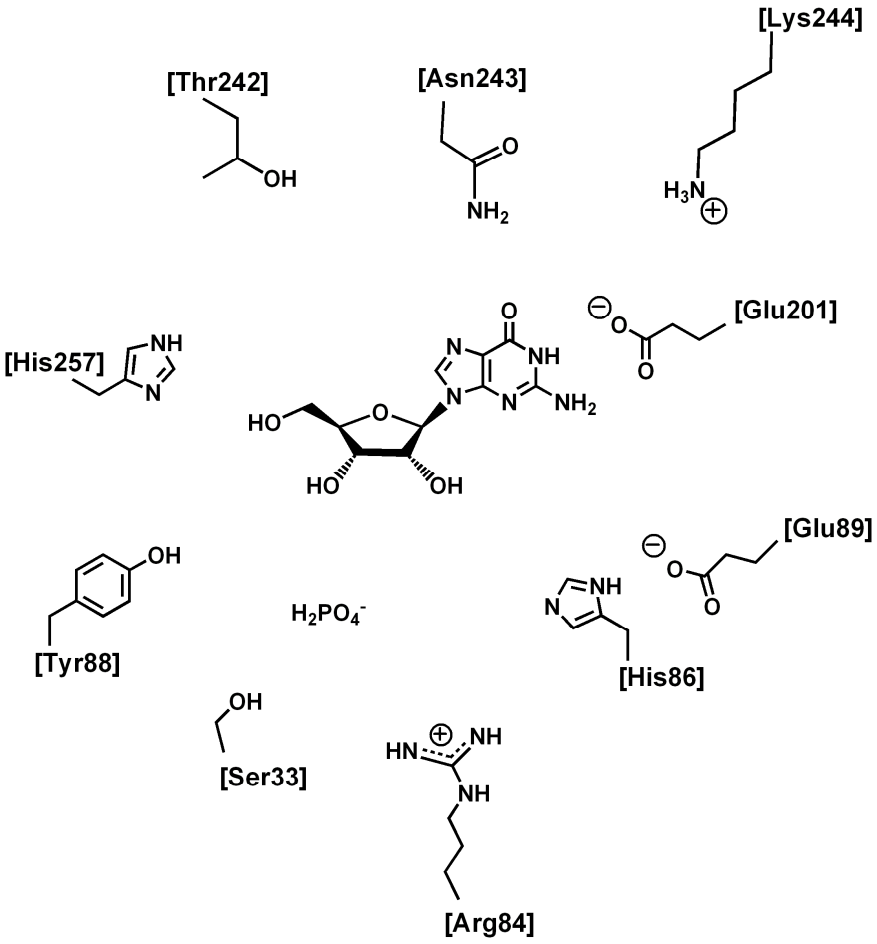

Figure 7-3 The PNP active-site model used by Erion and based on the coordinates of a Human PNP guanine complex PNP4 (adapted[1])

Further elucidation of active site contacts has been done by nuclei labeling NMR studies. Such studies yield the preferred protonation states of the purine base. For instance, Schramm[2] concluded that hypoxanthine binds to human PNP using the N7-tautomeric form.

Theoretical understanding of the TS is vital for designing active site inhibitors.[9] Allosteric inhibitors are known to exist.[10] The nature of the TS has been shown to be oxocarbenium-like in character.[11, 12] Kinetic isotope studies in conjunction with density functional studies suggest ^3E and E_3 conformations[13] of the sugar ring at the TS and attempts have been made at capturing the ribo-carbocation.[14] The stabilising amino acid substrate interactions seem to be directed mainly towards the incoming

phosphate and the leaving base and not so much at the developing positive charge at C1'. This has been noted[12] and it is unlike the case for glucosyl hydrolases (example CBHI , previous chapter) where the catalytic sites have "highly conserved pairs of carboxylates that sandwich and stabilise glucose oxocarbenium transition states".[12] In glucosyl hydrolases the catalytic Glu212 is highly conserved as it is intimately involved in the reaction (it is the negatively charged nucleophile for the glycosylation step), while in PNP, the doubly negative character of the incoming phosphate is capable of effective transition state stabilisation.[3]

Reference to multiple forms of the inorganic phosphate (PO_4^{3-}, [15] HPO_4^{2-}, [1, 16] $H_2PO_4^-$) exist and the exact nature of the inorganic phosphate species and the phosphate substituent on ribose was not clear until recently.[17] FTIR spectroscopy was used to conclude that the bound phosphate is dianionic and exists in two distinct binding modes with similar binding affinities. The phosphate of bound ribose-1-phosphate is also dianionic. It has been suggested that a histidine in the binding site is integral in extracting the hydrogens of phosphate. The observed crystallographic P-O bond lengths of several molecules of ribose-1-phosphate of 3FB1[18] agree with both the dianionic (HPO_4^{2-}) and anionic ($H_2PO_4^-$) form.

Inihibitory drug studies originally focused on use of substrate analogue of bovine PNPs. Inhibitors like the immucillin's shown in Figure 7-4 proved effective for bovine PNP at concentrations between 36 and 71 pM, however these are not as effective for Human PNP.[15] A second generation of immucillin, specifically DADMe immucillin (Figure 7-4), is more potent than the first generation TS analogues requiring as little as 6 pM for activity. Human PNP is proposed to have a more dissociated TS than bovine PNP.[15] The reaction proceeds via $D_N{*}A_N$ and the extra linkage of the second generation immucillins is proposed to better mimic the dissociated TS.

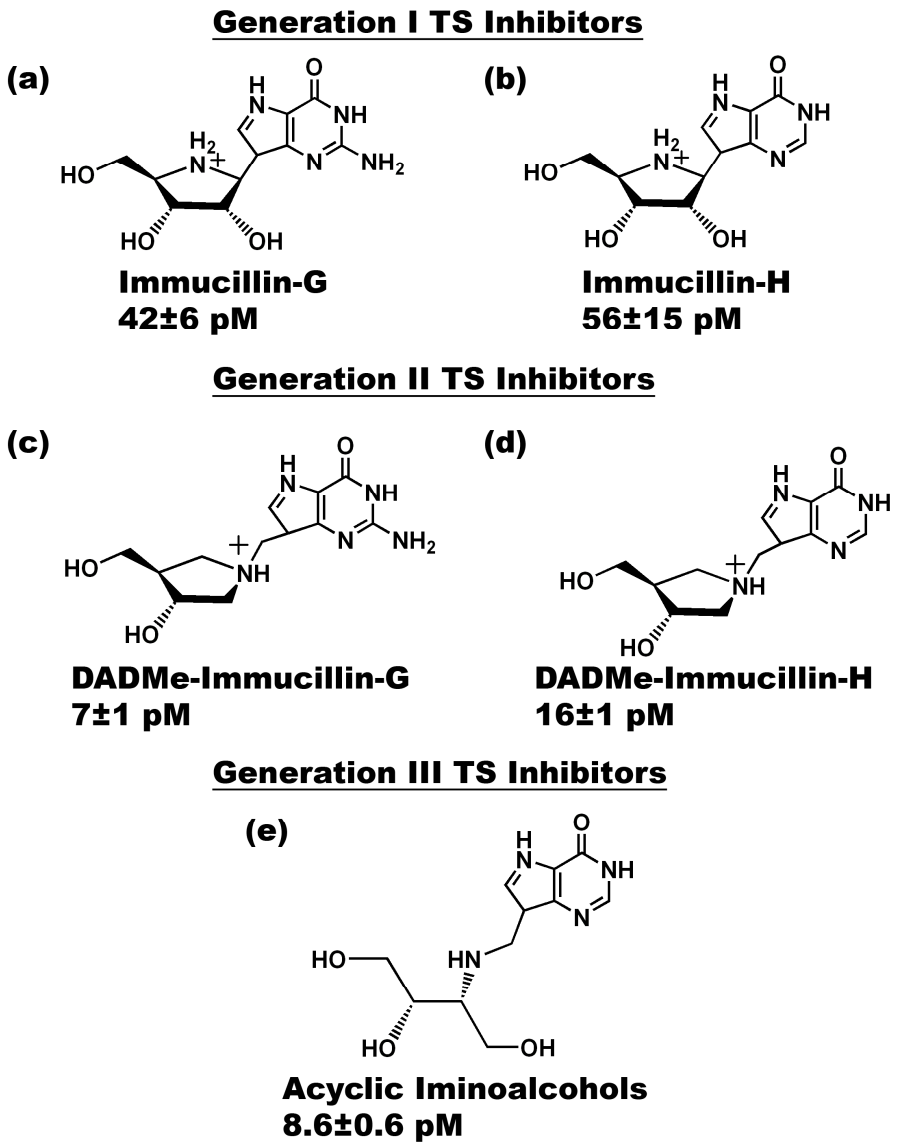

Figure 7-4 TS analogue inhibitors of PNP. The first generations inhibitors (a) Immucillin-G and (b) Immucillin-H are effective for bovine PNP but less so for human PNP. The Human PNP TS is more dissociated than that of bovine PNP and the second generation inhibitors (c) DADMe-Immucillin-G and (d) DADMe-Immucillin-H have an added -CH_2- linker which improves their inhibitory activity. The third generation of inhibitors are not cyclic and the example here (e) is similar in activity to (c).[15, 19]

A third generation of inhibitors has been designed that contains an acyclic iminoalcohol to replace the cyclic mimic of the ribooxocarbenium ion at the transition states of PNPs. The best third generation inhibitor is equivalent to the best inhibitors found in the previous generations TS analogues.[19] It is clear that the aim of PNP studies have centred around finding inhibitory drug targets in order to combat T-cell

mediated autoimmune diseases.[1, 16] The aim here is to evaluate the role the enzyme may play in prearranging the puckering conformation of the ribose sugar ring prior to the activation of the TS.

The nucleoside rather than the ribose-phosphate is modelled in the active site using SCC-DFTB (chosen based on the study of free energy of ring pucker in Chapter 4). There were several reasons for choosing the nucleoside over the phosphorylated sugar. Active site Inhibitors of PNP are based on a guanine or guanosine derivatives. Such inhibitors can enjoy the same stabilisation interaction as the purine base of guanosine. Moreover there is not yet a validated phosphorous parameter set for sugars in SCC-DFTB, although a set best for describing proton affinity of phosphates (SCC-DFTBPA) and a set best for phosphate hydrolysis (SCC-DFTBPR) have been derived by Cui et al. (Based on their[20] study of phosphate and methylated phosphate derivatives.)

A series of FEARCF calculations were carried out to pinpoint the role of the enzyme environment on the ring puckering of guanosine. Firstly, the preference shown by guanosine *in vacuo* and water are identified. These are compared to the already calculated ribose *in vacuo* puckering surface. Guanosine is placed in one of the active sites of trimeric bovine PNP and the pucker induced by the enzyme is determined by analysing the puckering free energy surface.

7.2 Simulation Details

This section details the setup of the free energy surface calculations in vacuum, water and PNP. Details of the theories used can be found in Chapter 2. Guanosine was built with the CHARMM Protein[21] forcefield and the entire system was modelled quantum mechanically with SCC-DFTB/d. Calculations were run with the mio parameters and hydrogen bonding and dispersion effects were included. Guanosine was minimised using the steepest descent (SD) algorithm followed by a short 100ps velocity-verlet dynamics simulation. The final coordinates and protein structure files were saved and used to start free energy calculations.

Eight *in vacuo* FEARCF calculations of 500ps in length were run for each iteration of the reaction coordinate force simulation. *In vacuo* velocity-verlet dynamics of the guanosine was carried out at 298.15K with group based cutoffs of 10, 12 and 14 Å. The electrostatic and van der Waals potentials were treated with force shifting. The non-bonded interactions were updated using CHARMM's built-in heuristic algorithm. At the end of each iteration the histograms from each simulation were combined with those from the previous iterations. This was done using the weighted histogram method

(WHAM[22]). Twenty iterations were run to reach convergence. The simulations were judged well-converged as the comparison of the sampling of lowest to highest energy regions of the free energy map within the canonical puckering region were 1:5.

The minimised coordinates from the vacuum calculations were solvated in a cube of side 33.9772Å with 1315 TIP3P[23, 24] waters. The system was then heated for 30ps and equilibrated for 100ps with the same dynamics specification as for vacuum. The final equilibrated coordinate and protein structure files were employed to start the free energy calculations. Four QM/MM guanosine in water FEARCF calculations of 500ps in length were run for each iteration of the reaction coordinate force simulation. . *In vacuo* velocity-verlet dynamics of the guanosine was carried out at 298.15K with atom based cutoffs of 10, 12 and 14 Å. The electrostatic and van der Waals potentials were treated with force shifting. The non-bonded interactions were updated using CHARMM's built-in heuristic algorithm. SHAKE was applied to all MM atoms. Periodic boundary conditions were applied and the cutoff was chosen at 14 Å. At the end of each iteration, the histograms from each simulation were combined with those from the previous iterations using WHAM (as described in Chapter 3). Sixteen iterations were run to reach convergence. The simulations were judged well-converged as the comparison of the sampling of lowest to highest energy regions of the free energy map within the "puckering zone" were 5.6:1.

The 1A9S structure as reported by Ealick[3], with a resolution of 2.00Å, was protonated following a pK_a analysis. Several amino acids are conserved across PNP from different organisms.[16] The trimeric form was built using the SYMMETRY records in the PDB and the atoms were placed using VMD.[25] A particular active site of the three available was chosen. The waters of crystallisation were not removed.[7] Depending on the dynamic structure of the binding pocket and movement of the substrate, different 'poses' (positions and conformations) of the substrate may be identified. Figure 7-5 demonstrates possible hydrogen-bonding motifs for Glu201, Asn243, Met219, Tyr88 and π-stacking interactions for Phe200.

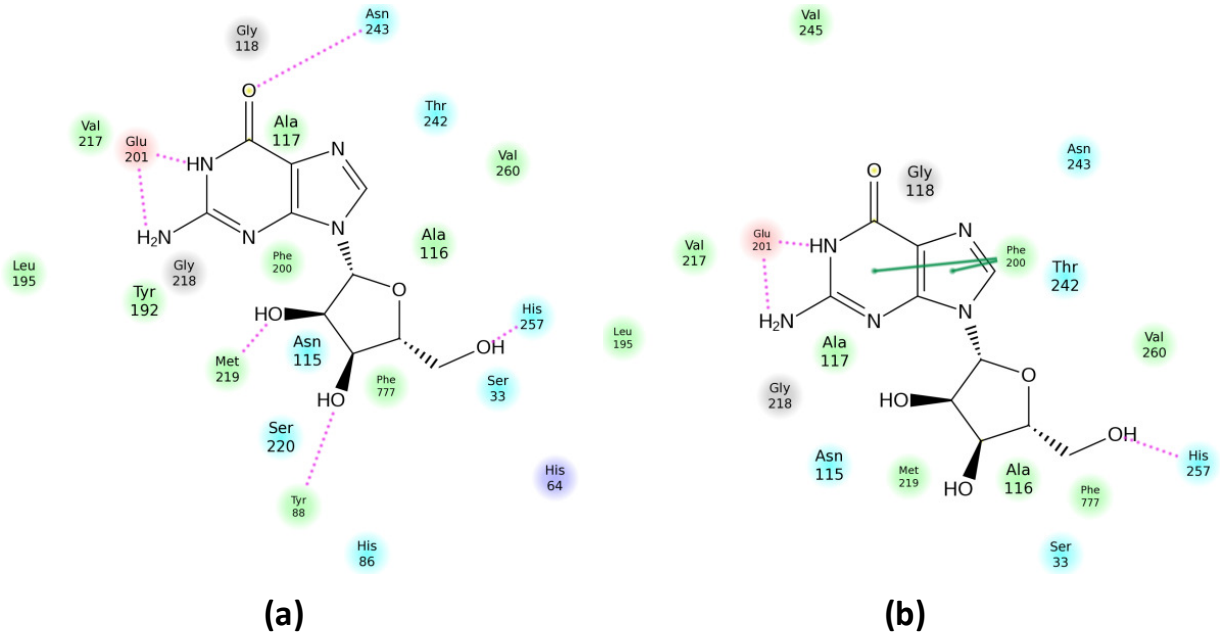

Figure 7-5 Different poses of guanosine within PNP condensed into a 2D active site map. Possible hydrogen bonding interactions are indicated with red stippled lines while hydrophobic interactions are indicated with green lines.

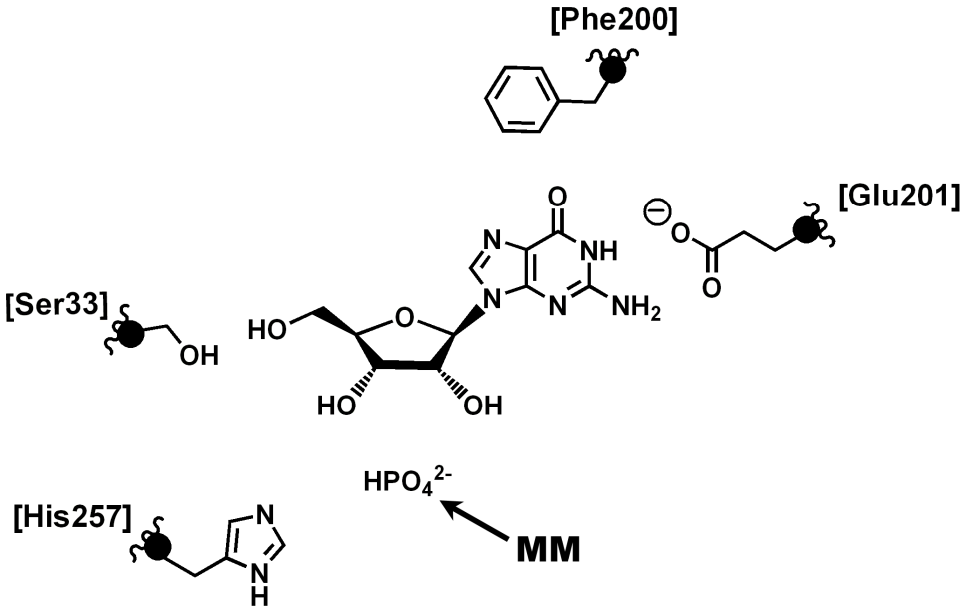

Figure 7-6 The chosen active site of PNP with QM atoms indicated. GHO atoms are indicated with black spheres. HPO_4^{2-} is modelled MM.

With reference to earlier considerations on phosphate protonation, HPO_4^{2-} is modelled in the binding pocket. Guanosine, Ser33, Phe 200, Glu201, and His257 make up the quantum region of the

active site in Figure 7-6 and the GHO approach (discussed in Chapter 2) joins the QM and MM regions. The amino acids are selected based on mutation studies, their conservation across species and on the ability to interact with both the sugar and base parts of guanosine. After initial minimization, a 24.5 Å TIP3P water sphere was positioned over the chosen active site. The general protein setup described in Chapter 2 was followed. After heating and equilibration, the equilibrated coordinate and protein structure files were employed to start the free energy calculations. Eight 200 ps in length QM/MM protein FEARCF calculations were run for each iteration of the FEARCF pucker simulation. Leap frog langevin dynamics were carried out as discussed in Chapter 2. WHAM was used as described in Chapter 3.

7.3 Puckering Free Energies

Guanosine, being a furanose derivative, can be analysed similarly to ribose as done in Chapter 4. The Hill-Reilly puckering planes obtained by triangular decomposition are indicated for ribose and guanosine in Figure 7-7a and b respectively. Here the pucker analysis is carried out for guanosine *in vacuo*, in TIP3P waters and in the PNP binding pocket.

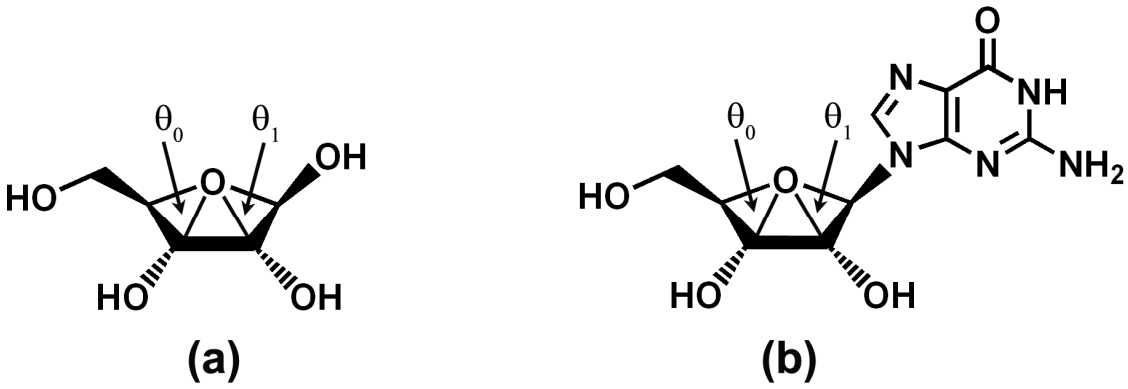

Figure 7-7 The triangular decomposition of the (a) ribose and (b) guanosine furanose rings

In vacuo it is likely that conformers that contain the best hydrogen bonding patterns will dominate. When considering the free energy of puckering for guanosine *in vacuo* (Figure 7-8) it is noted that the global minimum is an E_4 conformer with a little $^3T_4$ character residing at (-32.5°, 5.0°). Within 0.1 kcal/mol, a mixture of $^3T_4$ and E_4 may coexist. An E_1 with some $^2T_1$ character exists at (-13.75°, -17.5°) which is just 0.10 kcal/mol above the E_4 global minimum. There is barely a barrier between these two

minima, just 0.25 kcal/mol is needed to traverse between them. Guanosine can easily interchange between these conformers (E_4, E_1) *in vacuo* at 298.15K.

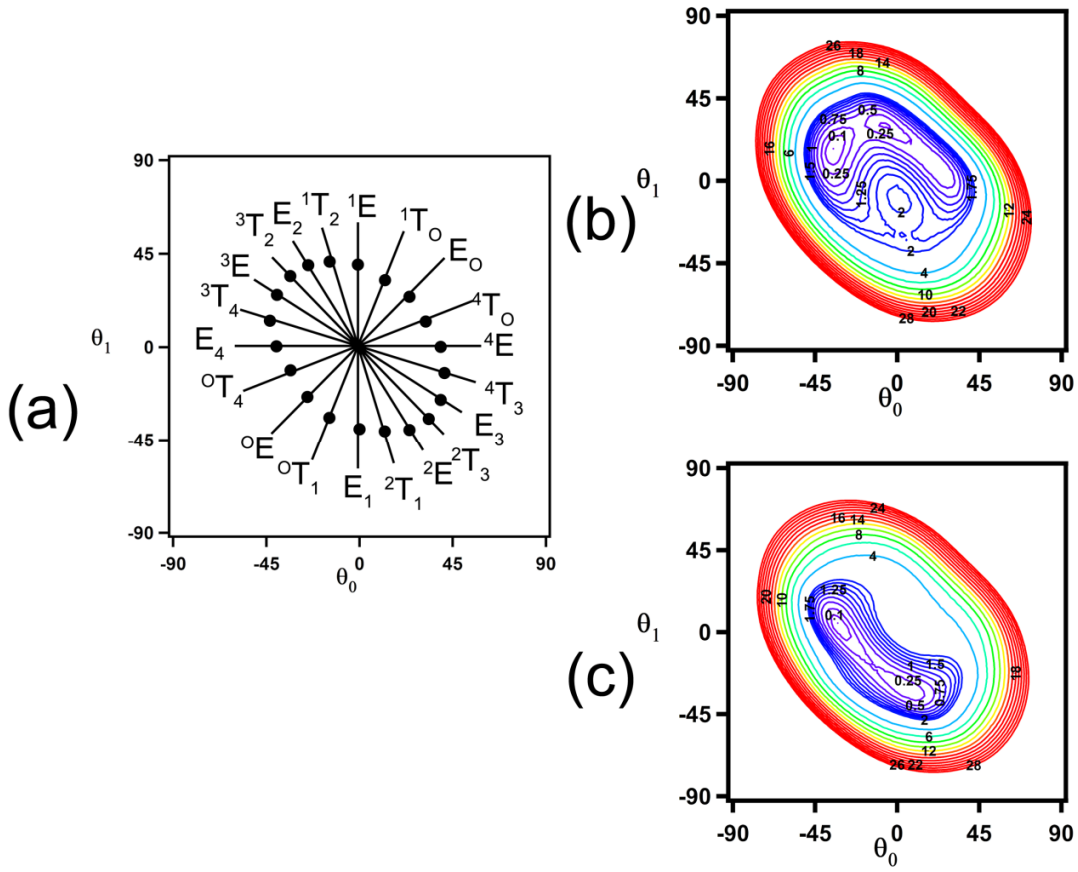

Figure 7-8 (a) The triangular decomposition pucker space for 5-membered rings as shown previously in Chapter 4.The free energy surface of puckering shown as a two dimensional contour plot for the (b) final iteration of ribose *in vacuo* (c) final iteration of guanosine *in vacuo*. Both of these, (b) and (c), are contoured where energy has been mapped to colour from 0 kcal/mol (blue) to 15 kcal/mol (red). Contours are shown at 0.1 kcal/mol; every 0.25 kcal/mol until 2 kcal/mol and every 2 kcal/mol thereafter.

The minimum path follows E_4-$^{O}T_4$-^{O}E-$^{O}T_1$-E_1 where the TS of 0.25 kcal/mol at (-13.75°, -17.5°) is between ^{O}E and $^{O}T_1$ in character. Within 2kcal/mol just on half of the available canonical conformers are accessible to guanosine. The planar is not thermally accessible at 298.15K needing 2.62 kcal/mol. The accessible puckers are those not in the Northerly, North-Easterly and Easterly directions namely $^{3}T_2$, ^{3}E, $^{3}T_4$, $^{O}T_4$, ^{O}E, $^{O}T_1$, E_1, $^{2}T_1$, ^{2}E, $^{2}T_3$, E_3. Within 4kcal/mol all of the canonical conformers can be accessed.

The guanosine *in vacuo* surface is notably different to the ribose surface. Puckering space is not only more constrained but a different region of the Hill-Reilly pucker map is preferred for guanosine. It was shown previously that ribose prefers North-Westerly and Northern parts of pucker space. In contrast, Guanosine prefers the Westerly, South-Westerly and Southern conformers.

In the aqueous phase, water molecules can hydrogen bond to the solute. There is competition between intra-molecular and inter-molecular hydrogen bond formation and the pucker of guanosine is less restricted (Figure 7-9). The angle between the defined puckering planes can be greater than *in vacuo* and this costs less free energy than *in vacuo*. Also high energy (>15 kcal/mol) physically unreasonable puckers are more easily accessed, possibly due to bridging interactions with water.

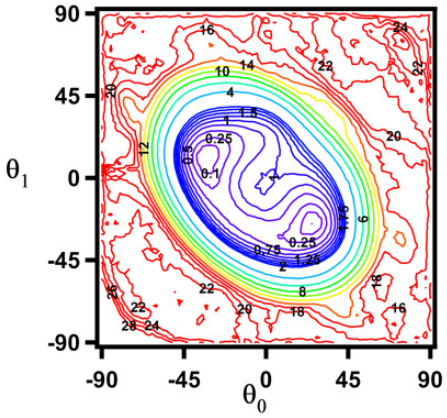

Figure 7-9 The free energy surface of puckering shown as a two dimensional contour plot for the final iteration of guanosine in water. Energy has been mapped to colour from 0 kcal/mol (blue) to 15 kcal/mol (red). Contours are shown at 0.1 kcal/mol; every 0.25 kcal/mol until 2 kcal/mol and every 2 kcal/mol thereafter.

Two global minima exist in the water free energy puckering surface of guanosine. A $^{3}T_{4}$ conformer at (-32.5°, 10°) and a $^{2}T_{3}$ conformer at (26.5°, 25.0°). These are interchangeable by following the minimum free energy path $^{3}T_{4}$-E_{4}-$^{O}T_{4}$-^{O}E-$^{O}T_{1}$-E_{1}-$^{2}T_{1}$-^{2}E-$^{2}T_{3}$ where ^{O}E is the transition state with a barrier of 0.43 kcal/mol at (-20°, -17.5°). As for the case of guanosine *in vacuo,* there are two low energy conformers of guanosine and there is a pucker conformational path between them. *In vacuo* these are E_{4} and E_{1}, but here these are $^{3}T_{2}$ and $^{2}T_{3}$ conformers. These minimum conformers are different and further apart than *in vacuo,* and a longer conformational free energy path is traversed between them. Guanosine pucker is much less restricted in water.

The 'perfect' planar conformer at (0.0°, 0.0°) requires 1.01 kcal/mol. Just adjacent to the perfect planar conformer are a maxima of 1.04 kcal/mol at (0°, -2.5°) and a TS of 0.48 kcal/mol at (5.0°, 0.0°). In water, guanosine can pucker relatively easily into the planar conformer, requiring less energy than when *in vacuo*. Unlike the vacuum case, when placed in water, guanosine can thermally access all canonical conformers at ambient temperature. Pucker is less restricted in water, however, only $^{3}T_{4}$ and $^{2}T_{3}$ are stable conformers.

When guanosine exists in the active site of PNP, the enzyme dramatically restricts the available conformational space. There is a ^{4}E global minimum on the puckering free energy surface (Figure 7-10) which lies at (25°, 2.5°).This is the single point of interest on the free energy surface.

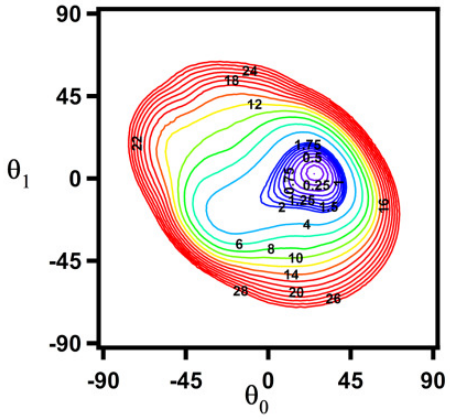

Figure 7-10 The free energy map obtained for guanosine in the active site of PNP. Energy has been mapped to colour from 0 kcal/mol (blue) to 15 kcal/mol (red). Contours are shown at 0.1 kcal/mol; every 0.25 kcal/mol until 2 kcal/mol and every 2 kcal/mol thereafter.

The planar conformer at (0°, 0°) requires 2.06 kcal/mol of energy and is not thermally accessible. However, PNP stabilises the planar conformer of guanosine more than *in vacuo* (2.62kcal/mol) but less so than in water (1.01 kcal/mol). Most pucker conformers are less easily accessed in PNP than *in vacuo* or water. Only the Eastern conformers can be thermally accessed (are less than 3kT or about 1.8 kcal/mol), in particular E_{O}, $^{4}T_{O}$, ^{4}E, $^{4}T_{3}$, E_{3}, $^{2}T_{3}$, ^{2}E.

$^{4}T_{O}$ and $^{4}T_{3}$ are adjacent to the ^{4}E and are easily accessible requiring more or less 0.5 kcal/mol. The $^{4}T_{O}$ conformer requires 0.5 kcal/mol in energy near its canonical position (the black sphere at $^{4}T_{O}$ in Figure 7-8a). The $^{4}T_{3}$ requires 1.5 kcal/mol energy near its canonical position (the black sphere at $^{4}T_{3}$ in

Figure 7-8a) but 0.5kcal/mol when it is puckered as a $^4T_3$ but is between the planar and $^4T_3$ canonical position. To access all the canonical pucker conformers requires just on 12kcal/mol. It is clear that the binding site of PNP severely restricts the available puckering conformational space.

In vacuum, puckers that favour intramolecular hydrogen bonding interactions are likely to dominate. In water, conformers of the 5-membered sugar ring of guanosine can be stabilised by intermolecular hydrogen bond interactions between the sugar hydroxyls and the solvent. Solvent bridging interactions also help stabilise the pucker space conformers, making it less restricted than in vacuum. In the protein, the guanosine does not induce large scale movements in the binding site; instead it is constrained by the size of the binding site cavity and specific amino acid interactions.

Of the chosen quantum amino acids, the Ser33 (Figure 7-11) and His257 (Figure 7-12) stabilise the phosphate (which was modelled using MM). In a selected 150ps trajectory, Ser33 alternates between interacting strongly with the hydrogen of phosphate and the oxygen of HO3'. Initially, Ser33 interacts with the phosphate oxygen, O4, but at 70 ps it interacts with the oxygen of HO3'. Although the interaction between His86 and a hydrogen of phosphate fluctuates between 2 and 4 Å (because of a competing interaction with Ser33), it is also important for stabilising the phosphate in the binding pocket (Figure 7-11).

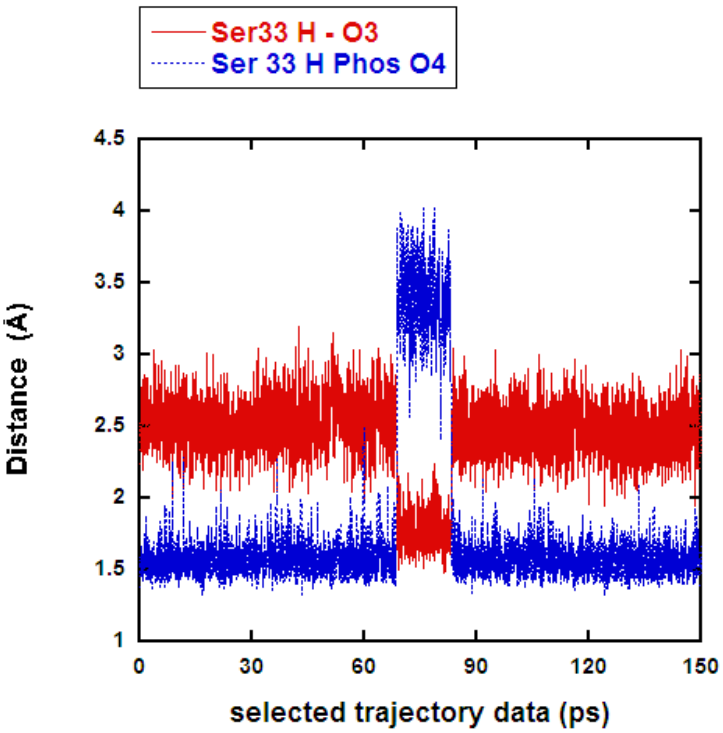

Figure 7-11 Time series showing distance interaction of Ser33 with O3 of guanosine and a hydrogen of HPO_4^{2-}

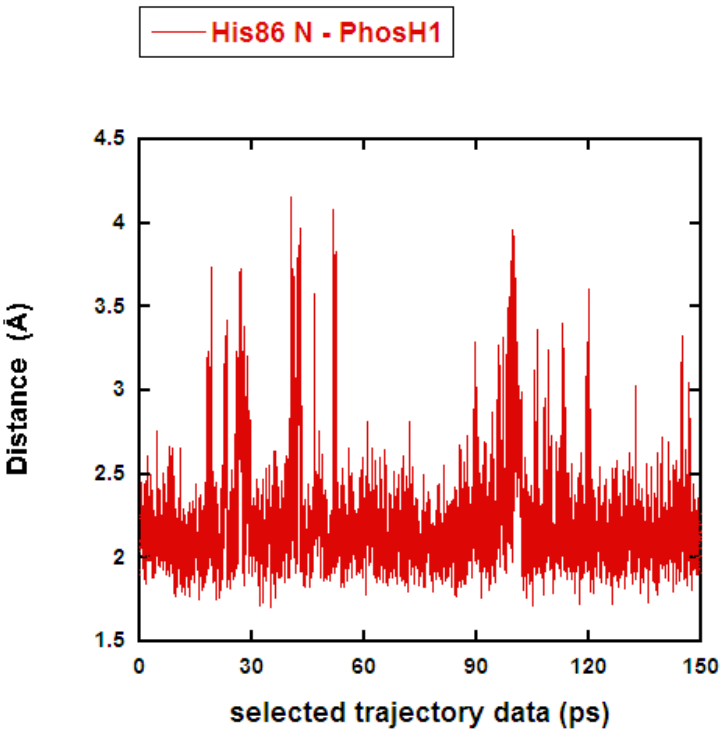

Figure 7-12 Time series showing distance interaction of His86 and a hydrogen of HPO_4^{2-} for a selected trajectory.

Glu201 neatly pairs with both the N-H hydrogens of the purine base as shown in Figure 7-5 and Figure 7-13, while Phe200 hydrophobically π-stacks in the face-on or in the T-stacked form with the purine base moiety of guanosine as indicated in Figure 7-5 and Figure 7-14. The interactions between the N-H hydrogens of the purine base and Glu201 are strong, these are close to 1.65 Å on average. The Phe200 to purine base distance fluctuates about 5Å, suggesting that it stacks in a "T-shape".

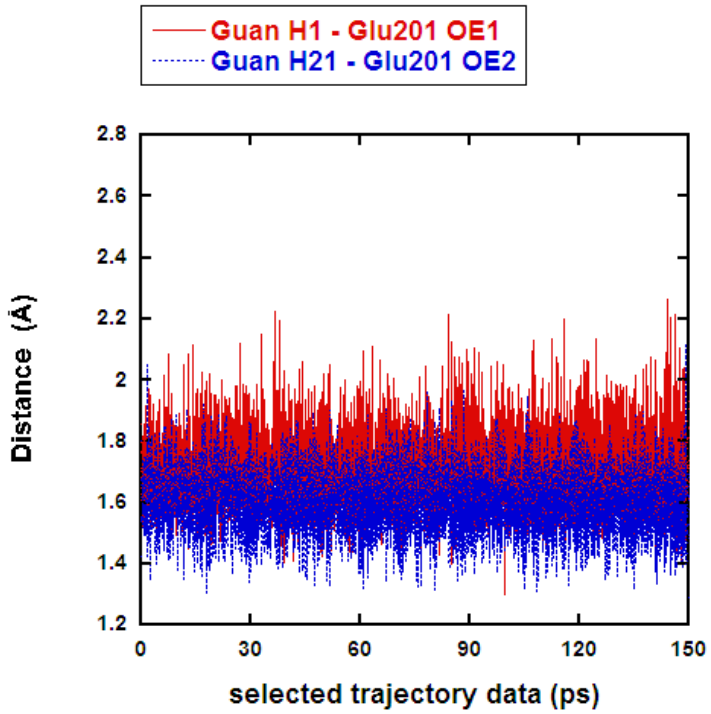

Figure 7-13 Time series showing distance interaction of Glu201 and hydrogens of the guanosine base.

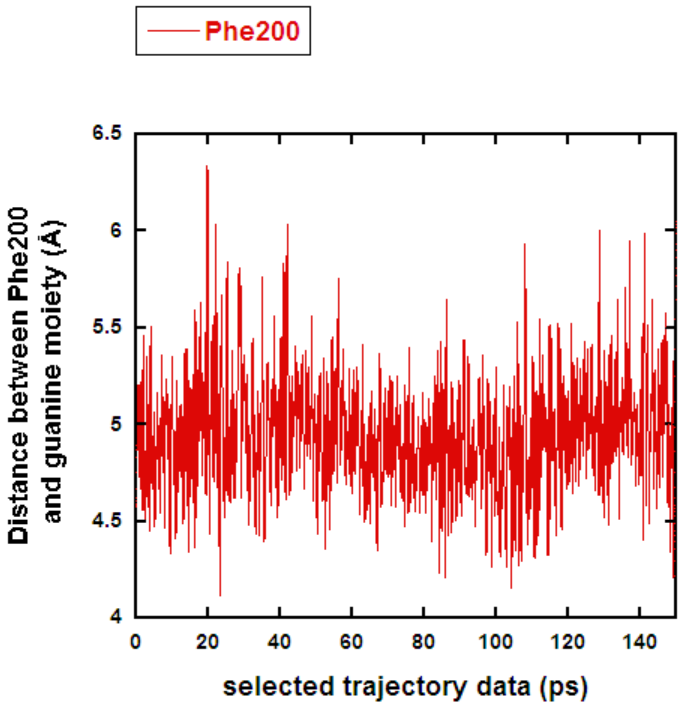

Figure 7-14 Time series showing the distance interaction of Phe200 and the base moiety of the guanosine base (the distance between a carbon of Phe200 (the carbon para to the substituted position of Phe200) and C4 of guanosine is computed).

In general, the C1'-O4' fluctuates about 1.47Å not showing any significant oxocarbenium double bond character (Figure 7-15a). The N-glycosidic linkage is also characterless, fluctuating about 1.45Å (Figure 7-15b). Using the H-C2'-C1'-O4' torsional angle as a measure of planarity shows that in this particular trajectory, the hydrogen at the anomeric position stayed pointing downwards in the αposition (Figure 7-15c). The pucker conformers sampled in this trajectory (Figure 7-15d), are those near to the global puckering minimum, ${}^{4}E$. These are ${}^{4}E$, ${}^{4}T_{O}$ and E_{O} with some ${}^{4}T_{3}$ and E_{3} and a small amount of ${}^{1}T_{O}$ sampled. Between 68 and 83 ps, the pucker conformer changes from a preference for ${}^{4}E$, ${}^{4}T_{O}$ and E_{O} conformers to ${}^{4}E$, ${}^{4}T_{3}$ and E_{3} conformers. In this time frame, the conformers above ${}^{4}E$ on the conformational map (Figure 7-8a) are less preferred, while conformers below ${}^{4}E$, in the direction of the observed water ${}^{2}T_{3}$ pucker minimum (Figure 7-9) become more likely. At present, the notable interaction that changes in concert with the ring pucker is the Ser33 to phosphate interaction. The Ser33 to phosphate hydrogen bonding interaction changes to a Ser33-O3' hydrogen bonding interaction between 68 and 83ps.

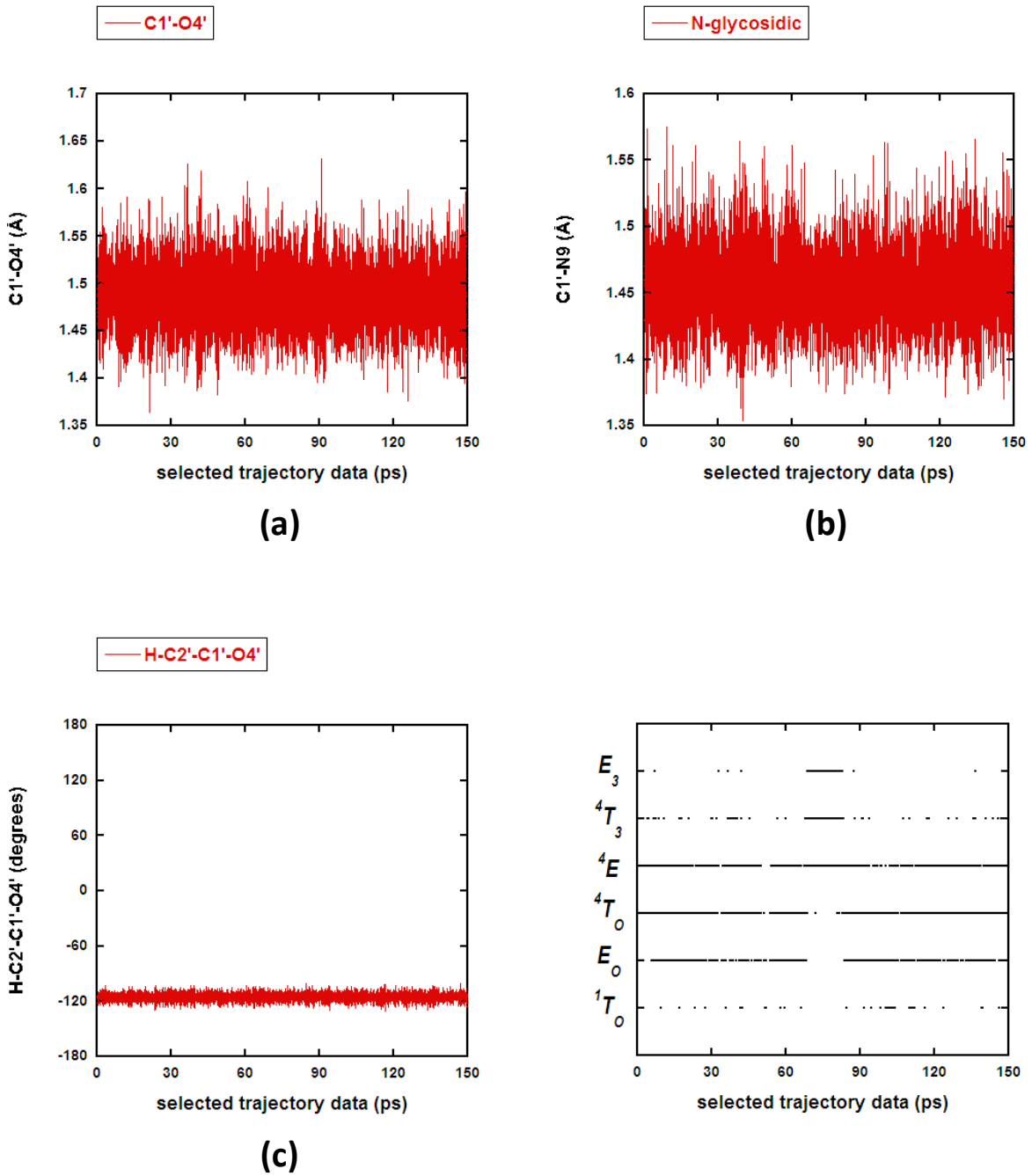

Figure 7-15 (a) Time series of the C1'-O4' bond distance, (b) the N-glycosidic linkage (C1'–N9) and (c) the H-C2'-C1'-O4' dihedral angle for a selected trajectory.

While the interactions of the QM amino acids are imperative for stabilising the incoming phosphate and leaving base, it seems that only the Ser33 played a role in inducing a change in puckering conformation. The major contributor for inducing a preferred sugar ring conformational is the entering nucleophile. The charged phosphate uses it oxygen atoms to interact strongly with the hydroxyl of C2

and C3 (Figure 7-16). From Figure 7-16a and b, it is clear that a specific phosphorus oxygen (O4) interacts strongly with the hydrogen of HO2 and HO3 for the entire trajectory analysed. Another of the phosphate oxygens (O3) doesn't interact with the hydroxyl groups. O2 takes an intermediate stance with HO2 sometimes being just 2Å away, and not interacting with HO3. The interaction of certain of the phosphate oxygens with the 2- and 3-hydroxyl groups of the sugar stabilises / induces a preference for the ^4E conformation. The phosphate oxygen, labelled O4, strongly interacts with both the C2' and C3' hydroxyl groups while the phosphate oxygen, labelled O2, when able to interact with the C2' hydroxyl induce a conformational preference for ^4E, ^4T_O and E_O conformers. When this interaction is temporarily disrupted (Figure 7-16a) and at the same time Ser 33 interacts with HO3 (Figure 7-11), the ^4E, $^4T_3$ and E_3 conformers are preferred.

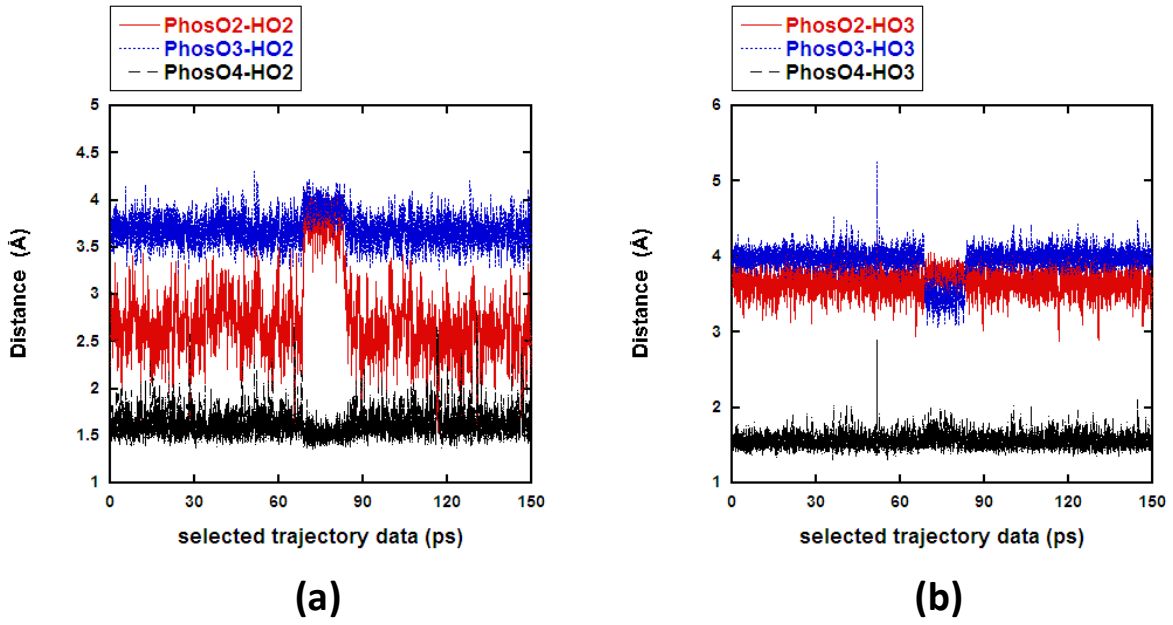

Figure 7-16 (a) Time series of the HPO_4^{2-} oxygen atoms interacting with the hydroxyl HO2 and (b) Time series of the HPO_4^{2-} oxygen atoms interacting with the hydroxyl HO3 for a selected trajectory.

The existence of the ^4E conformer is initially rather surprising as the ^4E conformer cannot form an oxocarbenium ion. Out of the 21 conformers accessible to a 5-membered ring only the planar, E_3 and ^3E can form a full character oxocarbenium according to a geometrical argument.[26] KIE coupled with QM optimisation clearly indicated that ^3E and E_3 exist at the TS for the PNP reaction for inosine.[13] It is plausible that PNP does not prepare the substrate for reaction (in this case guanosine) or that the ring pucker conformation shows distinct preferences during the course of the reaction as shown for CBHI in the previous chapter . The latter is more likely.

The C2' and C3'hydroxyls strongly interact with the phosphate while it is in the vicinity of the guanosine. It is encourages the formation of the ^4E conformer. There is crystallographic evidence for the ^4E conformational preference of guanosine. Several PDBS containing guanosine or similar (adenosine, inosine, 3DG) 1JE1[27], 1RFG[28], 3IEX[6], 1PK7[29], 1A9S[3] (the one used in this study), 1PR0[29], 1V45[28] (with 3DG) exhibit ^4E conformers. Although other conformers can exist for nucleosides, for example inosine adopts E_4 in 1RCT[30] (human PNP). As expected, guanosine does not pucker in the same manner as ribose-1-phosphate in the active site. Crystal structures containing ribose-1-phosphate, 1A9T[3] and 3FB1[18], do not exhibit ^4E. Rather 3FB1 has E^3 conformers and E_2 conformers. In contrast to the guanosine substrate, the conformers of transition state analogue inhibitors, such as immucillin-H mentioned earlier (Figure 7-4), adopt a planar conformer in 1RT9 (Human PNP).[31]

7.4 Conclusions

In summary, the guanosine substrate in PNP prefers the ^4E puckering conformer. This is in contrast to the two E_4 and E_1 puckering minima *in vacuo* and the two $^3T_4$ and $^2T_3$ puckering minima in the aqueous phase. The PNP active site significantly constrains the available puckering phase space, while in water pucker conformational space was more accessible than in vacuum simulations. The ^4E conformer found is not the TS pucker and cannot form the oxocarbenium ion. However, this conformer is consistent with the pucker conformer of nucleosides observed in PNP using X-ray crystallography. Likely candidates for the TS include: ^3E, E_3 or the planar conformer (observed for certain transition state analogues). The preference for the ^4E conformer is strongly dependent on the hydrogen bonding interactions with the incoming phosphate nucleophile.

7.5 References

1. M. D. Erion, K. Takabayashi, H. B. Smith, J. Kessi, J. Wagner, S. Hönger, S. L. Shames and S. E. Ealick, *Biochemistry*, 1997, **36**, 11725-11734.
2. H. Deng, S. M. Cahill, J.-L. Abad, A. Lewandowicz, R. H. Callender, V. L. Schramm and R. A. Jones, *Biochemistry*, 2004, **43**, 15966-15974.
3. C. Mao, W. J. Cook, M. Zhou, A. A. Federov, S. C. Almo and S. E. Ealick, *Biochemistry*, 1998, **37**, 7135-7146.
4. M. L. Markert, *Immunodeficiency Reviews*, 1991, **3**, 45-81.
5. G. A. Kicska, V. L. Schramm, L. Long, C. Fairchild, P. C. Tyler, R. H. Furneaux, H. Hãrig and H. L. Kaufman, *Proceedings of the National Academy of Sciences*, 2001, **98**, 4593-4598.
6. M. S. Castilho, M. P. Postigo, H. M. Pereira, G. Oliva and A. D. Andricopulo, *Bioorganic Medicinal Chemistry*, **18**, 1421-1427.
7. S. Saen-oona, S. Quaytman-Machledera, V. L. Schramm and S. D. Schwartz, *Proceedings of the National Academy of Science*, 2008, **105**, 16543-16548.
8. S. E. Ealick, Y. S. Babu, C. E. Bugg, M. D. Erion, W. C. Guida, J. A. Montgomery and J. A. Secrist, *Proceedings of the National Academy of Sciences of the United States of America*, 1991, **88**, 11540-11544.
9. V. L. Schramm, *Journal of Biological Chemistry*, 2009, **284**, 32201-32208.
10. P. A. Ropp and T. W. Traut, *Archives of Biochemistry and Biophysics*, 1991, **288**, 614-620.
11. P. C. Kline and V. L. Schramm, *Biochemistry*, 1995, **34**, 1153-1162.
12. A. Fedorov, W. Shi, G. Kicska, E. Fedorov, P. C. Tyler, R. H. Furneaux, J. C. Hanson, G. J. Gainsford, J. Z. Larese, V. L. Schramm and S. C. Almo, *Biochemistry*, 2001, **40**, 853-860.
13. A. Lewandowicz and V. L. Schramm, *Biochemistry*, 2004, **43**, 1458-1468.
14. M. Ghanem, A. S. Murkin and V. L. Schramm, *Chemical Biology*, 2009, **16**, 971-979.
15. A. Lewandowicz, P. C. Tyler, G. B. Evans, R. H. Furneaux and V. L. Schramm, *Journal of Biological Chemistry*, 2003, **278**, 31465-31468.
16. M. D. Erion, J. D. Stoeckler, W. C. Guida, R. L. Walter and S. E. Ealick, *Biochemistry*, 1997, **36**, 11735-11748.
17. H. Deng, A. S. Murkin and V. L. Schramm, *Journal of the American Chemical Society*, 2006, **128**, 7765-7771.
18. H. M. Pereira, R. C. Garratt and G. Oliva, *To be published 10.2210/pdb3fb1/pdb*.
19. E. A. Taylor, K. Clinch, P. M. Kelly, L. Li, G. B. Evans, P. C. Tyler and V. L. Schramm, *Journal of the American Chemical Society*, 2007, **129**, 6984-6985.
20. Yang Yang, Haibo Yu, Darrin York, Marcus Elstner and Q. Cui, *Journal of Chemical Theory and Computation*, 2008, **4**, 2067-2084.
21. B. R. Brooks, C. L. Brooks, 3rd, A. D. Mackerell, Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York and M. Karplus, *Journal of Computational Chemistry*, 2009, **30**, 1545-1614.
22. S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen and P. A. Kollman, *Journal of Computational Chemistry*, 1995, **16**, 1339-1350.
23. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *Journal of Chemical Physics*, 1983, **79**, 926-935.
24. W. L. Jorgensen and C. Jenson, *Journal of Computational Chemistry*, 1998, **19**, 1179.
25. W. Humphrey, A. Dalke and K. Schulten, *Journal of Molecular Graphics*, 1996, **14**, 33-38.
26. P. J. Berti and J. A. B. McCann, *Chemical Reviews*, 2006, **106**, 506-555.

27. T. C. Appleby, I. I. Mathews, M. Porcelli, G. Cacciapuoti and S. E. Ealick, *Journal of Biological Chemistry*, 2001, **276**, 39232-39242.
28. F. Canduri, R. G. Silva, D. M. dos Santos, M. S. Palma, L. A. Basso, D. S. Santos and W. F. de Azevedo, Jr., *Acta Crystallographica D Biological Crystallography*, 2005, **61**, 856-862.
29. E. M. Bennett, C. Li, P. W. Allan, W. B. Parker and S. E. Ealick, *Journal of Biological Chemistry*, 2003, **278**, 47110-47118.
30. F. Canduri, D. M. dos Santos, R. G. Silva, M. A. Mendes, L. A. Basso, M. S. Palma, W. F. de Azevedo and D. S. Santos, *Biochemical and Biophysical Research Communincations*, 2004, **313**, 907-914.
31. W. Shi, A. Lewandowicz, P. C. Tyler, R. H. Furneaux, S. C. Almo and V. L. Schramm, *To be Published (10.2210/pdb1rt9/pdb)*.