COMPUTATIONAL SIMULATIONS OF ENZYME DYNAMICS AND THE MODELLING OF THEIR REACTION MECHANISMS

Jon Ainsley

PhD

2017

COMPUTATIONAL SIMULATIONS OF ENZYME DYNAMICS AND THE MODELLING OF THEIR REACTION MECHANISMS

Jon Ainsley

A thesis submitted in partial fulfilment

of the requirements of the

University of Northumbria at Newcastle for the degree of

Doctor of Philosophy

Research undertaken in the

Faculty of Health and Life Sciences

October, 2017.

Abstracts

Proteins and enzymes are large and complex biological molecules, characterized by unique three-dimensional structure are highly flexible and dynamic nature. Thorough understanding of protein and enzyme function requires studying of their conformational flexibility, because important physiological processes, such as ligand binding and catalysis rely on an enzyme's dynamic nature and their ability to adopt a variety of conformational states. Computational methods are widely applied in studying enzymes and proteins structure and function providing a detailed atomistic-level of resolution data about the dynamics and catalytic processes, mechanisms in biomolecules, therefore even more nowadays a term 'computational enzymology' has emerged. Experimental methods often have difficulty in predicting dynamic motions of proteins. Computational simulations techniques, such as Molecular Dynamics simulations, have proven successful in simulating the conformational flexibility of proteins in studying structure-function relationships.

Additionally, the binding events between two molecules, e.g. an enzyme and its substrate, can be computationally predicted with molecular docking methods. Enzymes are proteins that catalyse almost all biochemical reactions and metabolic processes in all organisms. In order to study the conformational flexibility of proteins we apply molecular dynamics simulations, and in order to simulate their reaction mechanisms we apply quantum mechanical simulations. Quantum mechanical simulations can also be used to predict the electronic structure of organic compounds, by calculating their electronic structures we perform orbital analyses and predict their optical properties. The results gained from our computational simulations can give new insights into explanation of experimental findings and data and can inspire and guide further experiments.

Focus of this thesis is computational investigation of the following systems:

Comparative Molecular Dynamics (MD) investigation of Tryptophan 7-halogenase and Tryptophan 5-halogenase – These two enzymes are responsible for the halogenation of tryptophan, they do this in a highly regioselective manner. Halogen substituted natural products, such as antibiotics, are a particular area of interest to the pharmaceutical industry, however efficiently producing these compounds in a regioselective manner poses a significant challenge. In contrast, halogenating enzymes can perform direct halogenation of aromatic organic compounds, using only halide salts and mild conditions they can produce greater yields than present synthetic methods. Therefore a detailed understanding of the enzymatic mechanism of regioselective halogenation of natural organic compounds and knowledge of the origin of their regioselectivity is important for inspiring new paths for the organic chemical synthesis of new halogenated compounds.

Our study focuses on structural analysis through extensive Molecular Dynamics (MD) simulations of two flavin-dependent tryptophan halogenases; tryptophan 7-halogenase (PrnA) and tryptophan 5-halogenase (PyrH). Long range atomistic MD simulations of the two enzymes were performed, for 1 microsecond, in order to elucidate structure-function relationships and mechanistic implications related to the origin of regioselectivity in both enzymes. The performed MD simulations show that the binding sites of the cofactor-flavin adenine dinucleotide (FAD) and the substrate tryptophan do not come into close proximity during the timescale of the simulations. MD simulations confirmed that the FAD "strap" region observed previously in the PyrH crystal structure also exists in PrnA and likely fulfils a similar role, and acts as a line of communication between the two distant binding sites. The two catalytically important active site residues glutamate and lysine proposed experimentally to be involved in binding and the reaction mechanism of chlorination of tryptophan are vital to the orientation and positioning of the proposed chlorinating agent hypochlorous acid. It is likely that the positioning of the hypochlorous acid supported by glutamate and lysine and the environment they create in close proximity to the to-be

halogenated carbon drives the reaction. In addition, the differences in the regioselectivity between PrnA and PyrH probably arises due to differences in the tryptophan binding site, positioning different sides of the tryptophan indole ring towards the active glutamate and lysine sidechains.

Quantum mechanical (QM) investigation of the electronic structure of the Fat mass and obesity associated protein (FTO) – FTO is a protein associated with the regulation of metabolism, certain mutant forms of the enzyme are found to cause an increased risk of obesity, cancer, diabetes, and Alzheimer's disease. FTO is part of a family of enzymes known as non-heme iron 2-oxoglutarate dependent oxygenases, it catalyses the removal of methyl groups from DNA and RNA nucleotides. It repairs DNA/RNA bases that have become damaged by external alkylating agents, and restores them to their natural states, it also has an as of yet unknown regulatory role in demethylating adenine in messenger RNA. Understanding the reaction mechanism of FTO can aid in the design of novel inhibitor molecules that may be used to treat a range of conditions related to FTO malfunction. The reaction mechanisms of 2OG dependent oxygenases are well conserved across the family of enzymes, and other enzymes of this family have been extensively studied with quantum mechanical simulation techniques. Our study focuses on the formation of the reactive Fe(IV)-oxo species, this species is responsible for substrate hydroxylation which leads to the formation of the demethylated substrate. To study the electronic structural changes of the reaction mechanism quantum mechanical calculations were performed on cluster models of the enzymes active site, as well as combined quantum mechanics/molecular mechanics methods. These calculations have given us insight into the electronic structure of the enzyme's active site and the effect that individual residues have on the proposed reaction mechanism.

MD simulations and docking of the mammalian fructose transporter protein GLUT5 – GLUT proteins are facilitator transporter proteins responsible for the movement of carbohydrate molecules into cells, they do this in a substrate specific manner. GLUT5 is an unusual GLUT transporter as it is responsible solely for the transport of fructose, and is found to be aberrantly over expressed in certain types of cancer. Cancer cells use altered metabolic pathways to fuel their rapid proliferation into tumours, understanding the mechanism of fructose transport by GLUT5 may assist in the development of new diagnostic and therapeutic techniques against cancer. To aid in this, we have collaborated with an experimental group focussed on the development of fluorescent probes that are selectively transported by GLUT5. These Mannose-Coumarin (ManCou) probes can be used as fluorescent markers against cells that have over expressed GLUT5 allowing for the non-invasive identification of these cancer cells. Molecular docking techniques have been used to predict the binding pose of the natural substrate fructose to the GLUT5 receptor due to the lack of experimental structures of this receptor-ligand complex. Based on this we were able to predict the binding mode of the experimentally developed ManCou probes and provide structural insight into how they interact with the GLUT5 receptor. Based on the predicted ligand bound structures we have created a membrane bound GLUT5 system and performed molecular dynamics simulations in order to understand its conformational flexibility and the interactions between the ligand and the receptor over time.

Docking analysis of Odorranalectin- This peptide is the smallest known protein found to exhibit lectin-like carbohydrate binding properties, this selective recognition and binding of specific carbohydrate residues make it a particularly interesting compound. Cancer specific glycans are aberrant glycosylated structures that are only produced by cancerous cells, these structures can be recognised on the surface of cancer cells. Lectin proteins can specifically recognise these cancer specific glycans, allowing for the creation of new diagnostic tests for

cancer that allow for early detection, and lead to better treatment outcomes. Lectins can also bind to cancer specific cell surface glycans and interfere with cell adhesion to prevent tumour cells from metastasising. Our computational simulations were used to compliment experiments performed by our experimental collaborators to predict the binding of various sugar molecules to Odorranalectin and predict their binding energies. The resultant energies were found to correlate with the experimental binding affinities of the sugars and provided insights into key experimental findings.

QM calculations of Fluorescent Probes – Fluorescent probes are important chemical compounds used to track the movement of tagged molecules through biological systems in non-invasive manner. The pH sensitive fluorescent probes show observable differences in their fluorescent properties in response to pH change, and thus can allow the visualisation of pH in incredibly small intracellular spaces. Cancer cells are known to have particularly acidic intracellular pH levels in comparison to those of healthy cells, being able to detect the pH changes associated with tumour formation could lead to new methods of cancer diagnosis. Our experimental collaborators have developed novel pH sensitive fluorescent probes for the purpose of imaging intracellular pH. Our computational studies have aided in explaining the difference in the electronic structures and properties of the developed probes. Using quantum mechanical simulation methods, we have provided atomistic details that will assist in the design and development of new fluorescent probes.

Acknowledgments

Firstly, I would like to express my sincerest gratitude to my supervisors Dr. Tatyana Karabencheva-Christova and Dr. Christo Christov for giving me the opportunity to undertake this research. They have always encouraged me to adapt and improve myself in the face of any challenge that the world of scientific research has presented me with.

I would like to thank my experimental collaborators for providing me with the exciting opportunity to put my research into wider contexts and practical applications.

I thank my friend and fellow researcher Dr. Warispreet Singh for sharing his wealth of practical knowledge with me. The motivation he has given me over the course of many discussions and cups of coffee has been instrumental in the completion of this degree.

I would also like to thank all my friends who have maintained and preserved my optimism over the course of my studies, their faith in me has been a constant source of reassurance I have drawn upon over the course of this doctorate degree.

Finally, yet most importantly, I would like to thank my parents and grandparents. Their hard work both past and present is an inspiration to me, without their help and support none of this would have been possible.

In appreciation of the emotional and moral support of my friends and family, I would like to dedicate this thesis to them.

Declaration

I declare that the work contained in this thesis has not been submitted for any other award and that it is all my own work carried out under the supervision of Dr Tatyana Karabencheva-Christova, Dr Christo Z Christov, and Professor Gary Black

I also confirm that this work acknowledges the opinions, ideas and contributions from the work of others.

Name: Jon Ainsley

Signature:

Date:

Contents

List of Abbreviations

Abbreviation	Name
µs	Microsecond
2D	Two Dimensional
2OG	2-Oxoglutarate
3D	Three Dimensional
3Me-T	3-Methyl-Thymidine
Å	Ångström
AlkB	A DNA demethylating enzyme identified in bacteria
AMBER	Assisted Model Building with Energy Refinement
Apo	Apoenzyme
B3LYP	The Becke, 3-parameter, Lee-Yang-Parr functional
BSA	Bovine Serum Albumin
CC	Coupled Clusters
CD	Circular Dichroism
CHARMM	Chemistry at HARvard Macromolecular Mechanics
CTD	C-terminal Domain
Dap	Diaminopropionic acid
DCCA	Domain Cross-Correlation Analysis
DFT	Density Functional Theory
FAD	Flavine Adenine Dinucleotide
fs	Femtoseconds
FTO	Fat mass and obesity associated protein
GAFF	General AMBER force field
GLUT	Glucose transporter proteins
GROMACS	GROningen MOlecular Simulation computer program package
Gromos	Forcefield developed for GROMACS
HF	Hartree-Fock
HOMO	Highest Occupied Molecular Orbital
IC	IntraCellular
LUMO	Lowest Unoccupied Molecular Orbtial
ManCou	Mannose Coumarin fluorescently probes
MCPB	Metal Centre Parameter Builder

MD	Molecular Dynamics
MM	Molecular Mechanics
MMGBSA	Mechanics/Generalized Born Surface Area
MMGBSA	The Generalized Born and Surface Area continuum solvation approach
NIR	Near-InfraRed
NMR	Nuclear Magnetic Resonance imaging
NOG	N-oxyl Glycine
NTD	N-terminal Domain
ONIOM	Our own N-layered Integrated molecular Orbital and Molecular Mechanics
OPLS	Optimized Potentials for Liquid Simulations
OPLS	Optimized Potential for Liquid Simulations
PBC	Periodic Boundary Conditions
PCA	Principal Component Analysis
PCM	Polarizable Continuum Model
PDB	Protein Data Bank
pdb	Protein Data Bank (also a file format for protein structures)
PES	Potential Energy Surface
PrnA	Tryptophan 7 halogenase
PRODRG	A webserver for the parametrization of small molecules
PROPKA	A webserver for predicting the protonation states of amino acids at a given pH
PyrH	Tryptophan 5 halogenase
QM	Quantum Mechanics
QM/MM	Quantum Mechanics/ Molecular Mechanics
RESP	Restrained Electrostatic Potential
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuations
RoG	Radius of gyration
SASA	Solvent Accessible Surface Area
SP2	Synthesized Peptide 2
SP3	Synthesized Peptide 3
TD-DFT	Time Dependent Density Functional Theory

TM	TransMembrane
TS	Transition State
VDW	Van der Waals forces
VMD	Visual Molecular Dynamics

1 – INTRODUCTION

Enzymes are complex biological molecules which catalyse biochemical reactions in the cellular environment, they are able to greatly accelerate the rate of a reaction and show high specificity towards a substrate. Enzymology is a mature field with a wealth of experimental research; however, this experimental methodology has limits regarding the study of structure function relationships at the atomistic level. When the first three-dimensional atomic structure of an enzyme was elucidated in 1965 [1] it was realised that the static structure provided by the atomic coordinates could not fully explain how the enzyme catalyses the reaction [2]. To understand an enzyme's reaction mechanism several factors must be accounted for; the steric and electronic environment created by the proteins active site as well as the way the enzymes internal dynamics can affect the structure of the active site. Quantum Mechanics (QM) simulations are an *ab-initio* method which allow us to investigate the electronic environment of a reaction by simulating electrons using the Schrödinger wave equation to better understand reaction mechanisms [3]. Molecular Dynamics (MD) simulations are an empirical method which allow us to sample the conformational dynamics of a protein over time, with the aim of identifying important motions that may influence the reaction mechanism and lead to a greater understanding of the proteins structure [4-6].

Empirical methods often referred to as Molecular Mechanics (MM), use parameters derived from experiments or high level *ab-initio* QM calculations and are computationally "cheap" allowing for large system sizes of thousands of atoms and long simulation timescales of microseconds (μs). This allows them to study phenomena such as protein dynamics and protein folding as well as the sampling of thermodynamic properties such as the free energy of binding. However, MM does not simulate electrons so cannot be used to study reactivity,

for this, QM simulation methods are necessary. By simulating electrons and nuclei with the Schrödinger wave equation QM methods can accurately model chemical reactions and study the factors that influence them. This high level of theory comes at the cost of large multifactorial calculations which take a long time to solve, this limits the size of systems to tens or hundreds of atoms depending on the exact computational cost of the method used. A technique which combines QM and MM exists [7]. Combined Quantum Mechanics/Molecular Mechanics (QM/MM) are a multi-scale method often used for the modelling of enzymes which treats the reactive subset of atoms in the system with a QM level of theory, without incurring the computational cost of applying this calculation method to the whole system by treating the remainder of the atoms with an MM level of theory [6]. The QM subset is influenced by the electronic and steric effects of the surrounding environment allowing for a more accurate representation of an enzymes active site. Molecular docking is another empirical technique used to predict the binding between two molecules, these are usually peptides or small drug like molecules being docked to receptor proteins [8]. Docking can be used to assist the design of novel compounds that target a desired receptor to produce an inhibitory or activating effect [9].

These computational techniques are often used in conjunction with and to compliment conventional lab based experiments, they can help explain experimental data and provide new insights to chemical phenomena that can drive more directed experimentation [10].

1.1 MOLECULAR MECHANICS

Molecular mechanics (MM) uses classical mechanical physics to model atoms and molecules, the models don’t rely on electronic structure simulation as is the case with QM but instead use empirical data gathered from known molecules or high level *ab-intio* QM calculations. Although they cannot explain the electronic structural changes of chemical

reaction processes, the MM method lends itself incredibly well to defining the structural properties of molecules such as shape, volume, flexibility and the change in these properties over time [11]. The measurement and observation of these properties is particularly important in the study of proteins and enzymes, when the first structures of proteins were resolved by experimental techniques it became apparent that these static models weren't adequate representations of the actual structure [2]. The dynamic movements of these biological molecules had to be accounted for to explain their biological functions [12]. MM methods are relatively computationally "cheap" compared to QM methods, they can incorporate systems sizes of hundreds of thousands of atoms as opposed to the hundreds of atoms offered by even the lower level QM theories. This allows them to simulate solvated protein systems, allowing the study of their dynamic nature in a physiological environment.

MM theory is based on force fields with most being specialised to a specific class of molecules e.g. protein and biological, hydrocarbons, polymers, etc. The forcefields are composed of parameters; each element can have several atom types each corresponding to how it is bonded to other atoms. For instance, a detailed MM force field for hydrocarbons will reflect all the different carbon-carbon bond types as well as the most common groups it may be bonded to. The atom types have defined parameters obtained through experimental observation with techniques such as X-ray crystallography, NMR, IR spectroscopy as well as high level QM calculation [13]. The parameters cover all the different atomic motions and interactions, these can be divided into the two categories bonding and non-bonding interactions. The bonding interactions can then be split into bond stretching, bond angle bending and dihedral torsion. Non-bonded interactions are split into van der Waals and electrostatic interactions. A typical force field to describe the energy of a system in its functional form is shown in Equation 1.

$$V(r^N) = \sum_{bonds} \frac{k_i}{2}(l_i - l_{i,0})^2 + \sum_{angles} \frac{k_i}{2}(\theta_i - \theta_{i,0})^2 + \sum_{torsions} \frac{V_n}{2}(1 + \cos(n\omega - \gamma))$$

$$+ \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(4\varepsilon ij \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \right)$$

Equation 1: A term for calculating the potential energy ($\boldsymbol{V}$) of a number ($\boldsymbol{N}$) of particles of positions ($\boldsymbol{r}$) and includes all the bonding and non-bonding interactions between these.

1.1.1 Bonded Interactions

The bond stretching and compression term calculates the potential energy associated with this action using Hooke's law, the bond is represented as a spring of length (l_i) and deviation from the ideal length ($l_{i,0}$) results in an increase in potential energy. Bond angles are calculated in the same fashion, instead of an ideal bond length there is a reference bond angle ($\theta_{i,o}$) parametrised for the angle between three atom types and any deviation from ($\theta_{i,o}$) to another angle (θ) coincides with an increase in the potential energy. Bond torsion or dihedral angle rotation occurs between four atoms and refers to the out of plane twisting of these. It is a more complex interaction than angle bending or bond stretching as there are several fluctuations in the potential energy when the bond is rotated 360° based on the steric hindrance and symmetry of the attached groups, hence it can't be expressed in a similar way to bond stretching and bending [14]. The angle ω is the dihedral angle between the 1st and 4th atom, V_n is the barrier height of the energy, n is the symmetry number i.e. the number of times the wave repeats in a 360° rotation and γ is used to determine whether the energy is at a minima of maxima when ω is 0°; if it is a maxima then γ=0 and if it is a minima then γ=180. Improper dihedrals are another type of torsion force that is used in some forcefields to ensure planarity is maintained for certain atoms such as aromatic rings.

1.1.2 Non-bonded Interactions

Non-bonded interactions all rely on charge relationships so the permittivity of the medium (ϵ) plays a role in modulating their strength. The Lennard-Jones 12-6 potential is one common way of approximating the Van der Waals (VDW) interactions between two particles [15]. The strength of the interaction is dependent on two parameters that can be most easily illustrated by a graph (Figure 1).

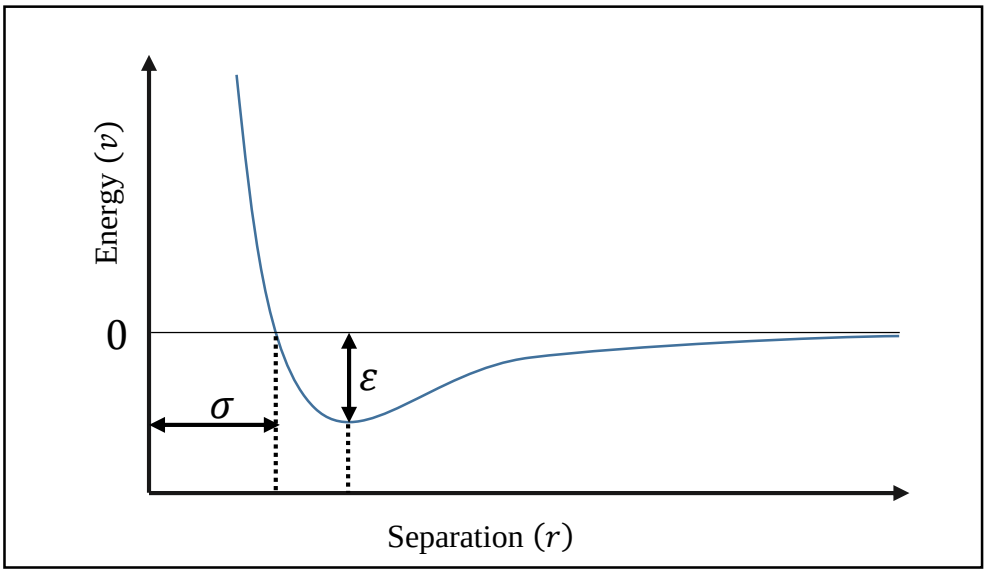

Figure 1: A graphical representation of Lennard-Jones potential.

The value of σ represents the distance at which the potential energy between the two particles is 0 being neither attractive nor repulsive, it is different for each pair of particles and is obtained through experimental observation. The Lennard-Jones 12-6 potential is widely used due to its computational simplicity even though more accurate models of van der Waals interactions exist [11].

Electrons and protons aren't modelled explicitly in the MM model so atoms have to be assigned partial charges based on their atom type parameters. The forces between two charged particles are then calculated using Coulomb's law, where q is the magnitude of the

charge expressed in coulombs for each point charge and r is the distance between the particles. Giving partial charges to atoms is an approximation, real charges rely on electron movement which is dynamic and polarizable. Although polarizable forcefields exist they present their own unique challenges and are mainly used for specialized systems where they are absolutely necessary [16].

Because non-bonded interactions don't decay to zero in theory every particle in a system has an effect on every other particle, this would be incredibly computationally expensive to implement. To simplify non-bonded interactions cut-off ranges are used, atoms outside of a predetermined distance range aren't considered when calculating the non-bonded energy of two particles [14]. The cut-off distance for long range interactions can have a significant effect on the results of an MD simulation, choosing the cut-off distance for the van de Waal's and electrostatic interactions is an important compromise between being large enough for accuracy and small enough for computational efficiency [17].

1.1.3 Solvent Modelling

Solvent effects play a key role in enzyme catalysis as well as overall protein structure, dehydrated enzymes are generally not active [18], therefore simulations done in vacuum aren't representative of the true nature of the protein. The majority of modern enzyme simulations use explicit solvation and depending on their size will usually contain thousands of water molecules which will be treated at the MM level of theory. A variety of forcefields exist to simulate water, commonly used ones include SPC[19], SPC/E[20], TIP3P and TIP4P[21]. These different water models vary in how they treat the distribution of point charge in the water molecule as well as minor changes in geometry and Lennard-Jones potential parameters, these are adjusted to better correlate with experimental results [22].

1.1.4 The Common MM forcefields

The AMBER forcefield was originally developed by the Weiner group [23] for the modelling of proteins and nucleic acids [24]. It used bonded parameters obtained from crystallographic structures of peptide fragments, VDW interaction parameters from earlier liquid state simulations and partial charges were obtained through QM calculations. The AMBER forcefield is updated often with each new package of AMBER. Related AMBER forcefields also exist for the simulation of lipids, DNA, zinc coordination and non-biological systems [25, 26].

The GROMOS forcefield was created alongside the GROMACS program for molecular dynamics simulations at the University of Groningen [27]. Containing parameters for a range of biological molecules the program is free and open source. The GROMOS forcefield exclusively uses a united atom system where aliphatic non-polar hydrogen atoms are subsumed into their constituent carbon atoms.

CHARMM was created as both a forcefield and software package by the Karplus group[28] for use in molecular dynamics simulations. It features parameters for DNA, RNA, proteins, lipids, sugars and other biological molecules of importance.

OPLS was originally created by Jorgensen et al[29] for the simulation of water and other liquids but grew to include proteins and other biological molecules. Like AMBER CHARMM and GROMOS the first versions of OPLS were united atom forcefields but now all atom versions are also available.

1.1.5 Parametrisation

To simulate new molecules containing atom types not available in a forcefield, parametrisation of the new molecule must be undertaken. Some examples of molecules generally not found in biological forcefields but that are important to include are: small molecule ligands, non-standard or modified amino acids, and coordinated metal ions. The

new parameters for the bonded and non-bonded interactions as well as atomic charges and assigned atom types are collected in a topology file, this topology can then be integrated into the simulation setup process and allows the parametrised molecule to be simulated along with other atoms in a simulation.

Parametrisation is generally achieved using a combination of identifying atom types in the new molecule for which parameters already exist and using these, as well as QM calculations to create new parameters for those which no usable analogous parameters exist. This can be done in an automated way using webservers like PRODRG[30] or SwissParam[31], or packages like Antechamber[32] or MCPB[33] for AMBER[34].

1.2 MOLECULAR DYNAMICS SIMULATIONS

Molecular Dynamics (MD) uses the MM forcefield to calculate the forces acting on a particle, this is a rearrangement of Newton's second law (Equation 2). In this equation, the subject is changed to calculate the acceleration (a) acting on the particles, where the forces (F calculated by the forcefield) and mass (m) are known. For more than one particle we can write the equation of motion with acceleration as a derivative of velocity v, and velocity as a derivative of position x, where x represents the coordinates of all atoms (Equation 3). The resulting differential equations can then be solved numerically, but in reality Verlet methods are employed along with discrete timesteps to improve their implementation into MD software.

$$a = F/m$$

Equation 2: The rearrangement of Newton's second law equation.

$$\frac{dx}{dt} = v \qquad \frac{dv}{dt} = \frac{F(x)}{m}$$

Equation 3: The derivative equation of motion.

The simulation proceeds in timesteps (usually 1-2 femtoseconds), for each timestep the forces are calculated and the atomic coordinates are moved and velocities are updated, this process is repeated with each new calculation updating the velocity and coordinates to produce a trajectory. Simulations are started from coordinates with each atom having an initial velocity that is randomly assigned according to a Boltzmann distribution for the temperature of the simulation.

MD simulations of proteins typically use coordinates derived from X-ray crystallography or NMR experiments, although homology modelling techniques can be used predict the 3D structures of proteins based on their amino acid sequences[35]. X-ray crystallography cannot accurately resolve the position of hydrogen atoms due to their small size, these are added to the structure at the start of the simulation setup process. The initial structure needs to be energy minimised to remove slight inaccuracies or artefacts in the initial coordinates that can lead to unnatural structures and ultimately simulation crashes. Energy minimisation is performed using the MM forcefield usually using different algorithms designed to move the atomic coordinates towards local energy minima. When using MD methods to study proteins it is usually desirable to study them under physiological conditions, this means solvated in water at ~300K and 1atm of pressure. Solvation effects can be simulated with either implicit or explicit methodologies, implicit solvation attempts to mimic the effects of solvent

molecules using an artificial polarizable homogenous medium to reduce the number of particles in the system, however they cannot model important atomistic protein-solvent interactions. Explicit solvation uses large numbers of solvent molecules, usually with fixed atomic charges, to model the effect of solvation on the protein structure. The inclusion of solvation brings about the problem of periodicity, that is, how are particles at the edges of the system treated? In a simulation box with a fixed size, particles near the edge will collide with the boundaries and experience different forces than those at the centre of the box, this would create irregularities in the density of the system. A solution to this problem is the inclusion of Periodic Boundary Conditions (PBC), under these conditions the box is surrounded by identical copies of itself in every plane, a particle moving out of the box on one side will re-enter the box on the opposite side.

1.3 MOLECULAR DOCKING

Molecular docking was primarily designed to predict the binding of small drug like molecules to receptor proteins, many diseases are caused by the malfunction of proteins and therapies are focused on the inhibition or activation of these specific target proteins. Traditional lead generation methods for drug discovery normally entail assaying a large variety of interesting compounds against a specific protein known to be a disease target and hoping to observe a binding interaction [36]. Modern high throughput screening uses robots to automate the both the handling of the experiment and observation of the results to greatly increase throughput. However, these experimental techniques are still limited by the need to have a constant supply of new compounds, proteins/enzymes and laboratory resources [36]. Docking can be used to virtually screen new compounds in a similar way to physical high throughput screening as well as offering atomistic level insight to aid in structure based drug design [37].

The aim of docking is to find stable binding conformations between two molecules, this is achieved by assessing all the possible conformations in a defined space [38]. Usually the spatial search is limited to a known binding site which is predicted by experimental observations, however specialised software can also be used to identify protein binding sites [39]. This reduction of spatial search area reduces the number of possible conformations to a more manageable search. Blind docking refers to the practice of not restricting the search area of the docking to a confined active site and instead considering the entire protein surface for possible docking conformations, this technique is much more computationally intensive but can be useful for finding previously unknown secondary binding sites [40, 41].

Aside from considering all possible conformations a docking process must also be able to rank these conformations to determine which pose produces the most chemically realistic and stable binding conformation. Evaluation of the different potential poses is also called scoring, most docking programs use an MM type forcefield to estimate the difference in free energy of binding ($\Delta G_{binding}$) between the free ligand and receptor and the ligand and receptor complex, a typical way of estimating this is shown in Equation 4 [42]. In this equation L is the ligand, R is the receptor, the potential energy difference terms are represented with V and ΔS_{conf} represents an estimation of the change in entropy upon ligand binding. Calculating the potential energy V has several constituent factors such as: van der Waals, electrostatic and hydrogen bonding interactions, any internal strain energy incurred by the docking pose as well as the desolvation energy needed to displace water molecules in the receptor site. One way of combining these factors for energy calculation is shown in Equation 5 where W is a weighting constant which is a parameter based on observed experimental binding constants. The energy contribution of the different types of interaction: *VDW*, *H-bond*, *desolvation*, *Electrostatic* and *torsional* are calculated via classical mechanics methods similar to that used for the typical MM forcefield in Equation 1.

$$\Delta G_{binding} = (V_{bound}^{L-L} - V_{unbound}^{L-L}) + (V_{bound}^{R-R} - V_{unbound}^{R-R}) + (V_{bound}^{R-L} - V_{unbound}^{R-L} + \Delta S_{conf})$$

Equation 4: A typical way of calculating the free energy of binding for a receptor ligand complex.

$$V = W_{VDW} + W_{H\text{-}bond} + W_{desolvation} + W_{Electrostatic} + W_{torsional}$$

Equation 5: An example of an empirical potential for calculating the potential energy between a ligand and recpetor

Another large factor to consider in the docking search is the flexibility of the ligand and receptor molecules, in reality molecules are never static so the consideration of flexibility and torsion of bonds is incredibly important in ensuring the accuracy of the docking process. Rarely is completely static docking appropriate, both ligand and receptor will change shape to accommodate one another [43]. Allowing bonds to rotate and flex increases the degrees of freedom and therefore the complexity of the conformational search. Allowing a small number of bonds to rotate in the ligand doesn't incur much of a penalty with modern computational power. However, simulating the flexibility of a protein receptor can be a more complex task due to the much larger number of rotatable bonds. Most docking software places limitations on this by only allowing a set number of user specified sidechains to be flexible, this in order to keep the possible number of conformations to a manageable number [42].

Different algorithms have been designed to more optimally search the conformational space thus reducing the computational time needed allowing for a more exhaustive search. One of the simplest algorithms for docking is the shape-complimentary method, the ligand and

receptor binding site are represented as two complimentary surfaces and the algorithm seeks to create poses with the best fit between them as well as considering those with the most complimentary interactions [44, 45]. Complimentary docking methods are relatively computationally inexpensive but cannot consider flexibility so must be used with caution.

Other docking algorithms are based on MD, by sampling the conformational flexibility of the ligand and receptor in multiple short MD runs followed by MM minimisations some of the flexibility of both the ligand and receptor can be simulated [46]. MD based docking algorithms tend to be relatively computationally expensive, however they have the advantage of being able to model flexibility. Another common method for docking involves the use of genetic theory to aid the conformational search, genetic algorithms represent each potential pose as a "gene" [47]. Pairs of genes are combined to produce new genes which are a random combination of their parents, in practice this means potential poses are combined to create a pose which include elements of the two parent poses [48]. In addition to the combination genes the algorithm implements a fitness function which discards poses which score poorly. In addition to regular combination, some random mutation of the pose geometry is introduced to improve variety in the gene pool. Genetic algorithms are very effective at finding low energy binding poses in docking situations with large conformational search areas, however, they may require a large number of evaluations to find the pose with the best binding affinity [46].

1.4 QUANTUM MECHANICS

Quantum mechanical methods for studying chemistry are primarily concerned with the Schrödinger wave equation (Equation 6), this equation seeks to calculate the wavefunction (ψ) for a particle [49]. The wavefunction of a particle is a property which is difficult to represent using classical physics and is a unique concept in quantum mechanics, however it can provide us with the means to calculate physical properties such as the velocity, position and spin of a particle. In the Born interpretation the square of the wavefunction is the electron density, calculating this gives us the probability of finding an electron in a given volume [50], and thus we can derive all the physical properties and measurements in the system [49]. The Hamiltonian operator ($\widehat{H}$) is the sum of the kinetic and potential energy of all the particles present, it acts on the wavefunction to produce the total energy (E) as an eigenvalue [51].

$$\widehat{H}\psi = \mathrm{E}\psi$$

Equation 6: The time independent Schrodinger wave equation.

$$-\frac{(h/2\pi)^2}{2m}\nabla^2\,\psi - \frac{Ze^2}{4\pi\varepsilon_0 r}\psi = \mathrm{E}\psi$$

Equation 7: The Schrodinger equation for the hydrogen atom.

In Equation 7, h is Planck's constant, m is the resting mass of an electron and the portion of the equation referring to the change in Cartesian coordinates is substituted by the Laplacian operator (∇^2). The coulomb interaction term describes the attractive energy between the orbiting electron and the nucleus; it is always a negative value to reflect this

being a potential energy quantity. In this term Ze represents the charge of the nucleus where Z is the atomic number of the atom and e is the unit of elementary charge (1.6022×10^{-19} Coulombs), ε_0 is the permittivity of a vacuum (a constant 8.8542×10^{-12} $J^{-1}C^2m^{-1}$) and r is the distance between the electron and the nucleus [52].

Somewhat simplifying the equation for atoms larger than hydrogen is the Born-Oppenheimer approximation, this states that since the electrons weigh around 1800 times less than the lightest nucleus (the proton at the centre of the hydrogen atom) the movements of the electrons will instantaneously adjust with any changes in position of the nuclei [53].

$$\left[-\frac{(h/2\pi)^2}{2m_e}(\nabla_A^2 + \nabla_B^2) + \left(-\frac{{Z_e}^2}{4\pi\varepsilon_0 r}\left(\frac{Z}{r_A} + \frac{Z}{r_B} - \frac{1}{r_{AB}}\right)\right)\right]\psi = \mathrm{E}\psi$$

Equation 8: The Schrodinger equation for a helium atom.

The Schrödinger wave equation for the helium atom is shown in Equation 8, in this equation there are two electrons *A* and *B*, each ∇_A^2 or ∇_B^2 represents the change in position of the Cartesian coordinates of one of the electrons. r_A and r_B represent the distance of the two electrons *A* and *B* from the nucleus and r_{AB} is the distance of the two electrons from one another. Since there are two electrons in the helium atom there are three particles all interacting with one another, two electrons and one nucleus, this is a three bodied problem for which no exact solution can be calculated [54]. Another problem arising with the Hamiltonian for the helium atom relates to it not including a term to take into account the electron spin. Electrons like most other sub atomic particles have an innate quantum mechanical property known as spin; electrons can either have a spin quantum number of

$+1/2$ (spin up) or $-1/2$ (spin down). The importance of spin becomes apparent when considering the Pauli Exclusion Principle; it states that no two electrons may occupy the same orbital unless they are of opposite spin. This lays down the foundations of atomic orbital structure, and helps defines the s, p, d and f orbitals chemists are familiar with [55].

Another facet of the Pauli Exclusion Principle proposes that since all electrons are equivalent so any two electrons regardless of their spin orientation must be able to be swap positions without changing the electron density of the system, this is the referred to as the antisymmetry principle [51]. In a system like Helium with two electrons the wavefunction must not change value if the coordinates of electron 1 and 2 are swapped as this would change the electron density of the system. The wavefunction is however allowed to change sign, since the electron density of the system can be described as the square of the wavefunction a change in the sign doesn't affect the probability of locating an electron in the given volume. An alternative way to express this so that the wavefunction does not change value is shown in Equation 9.

$$\psi(1,2) = -\psi(1,2)$$

$$\psi(1,2) = \frac{1}{\sqrt{2}}[\psi_1(1)\psi_2(2) - \psi_1(2)\psi_2(1)]$$

Equation 9: A way to express the antisymmetry criteria of a two electron system e.g. helium

In Equation 9, $\frac{1}{\sqrt{2}}$ is a normalisation factor that ensures the wavefunction stays as a continuous probability value, ψ_1 and ψ_2 are the spin states of the electrons and (1) and (2) are the orbital positions of the two electrons in the helium atom. However, this method of writing out all the possible combinations of orbitals and positions becomes unwieldy with

systems with more than two electrons. A more general form is needed, one that can be extended to any number of electrons, this is a Slater determinant and is shown in Equation 10.

$$\psi(1,2,\dots \mathrm{N}) = -\psi(1,2,\dots \mathrm{N})$$

$$\psi(1,2,\dots \mathrm{N}) = \frac{1}{\sqrt{N!}}\begin{vmatrix} \psi_1(1) & \psi_2(1) & \cdots & \psi_N(1) \\ \psi_1(2) & \psi_2(2) & \cdots & \psi_N(2) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_1(N) & \psi_2(N) & \cdots & \psi_N(N) \end{vmatrix}$$

Equation 10: A Slater determinant for N number of electrons in a system.

Atomic systems containing more than one electron cannot be exactly solved by the Schrödinger wave equation, this is in part due to the larger number of variables involved, and also due to it being an n-bodied problem which are not possible to solve exactly with analytical methods [52]. Diverse approximation methods have been developed over the years to estimate the Schrodinger wave equation for systems with more than one electron, by utilising iterative calculation methods they can arrive at solutions that provide insight into chemical phenomena that can't be observed with experimental methods [56]. Originally QM calculations were performed by people, however computer algorithms can more rapidly perform iterative calculations so produce results more rapidly calculations than those done by hand. A brief survey of the most used calculation methods is given below.

1.4.1 Hartree-Fock (HF)

Originally published in 1928 [57] this approach was formulated by Douglas Hartree and further developed by Vladimir Fock leading to it being named the Hartree-Fock method (HF). It was found to offer a good *ab-initio* approximation of the Schrödinger wave equation that allowed for the calculation of a relatively large number of electrons in a system. Also

called the self-consistent field approach, this name gives a good description of the calculation process. To begin, all the atomic orbitals are approximated for each electron, and then a single electron is selected and its wavefunction is calculated while the influence of the other electrons is modelled as a single source of potential. The position of the electron is adjusted according to this potential whilst keeping all the other electrons frozen in place. The process is repeated for each electron in the system, the calculated single electron wavefunctions are combined to produce the multi-electron wavefunction. The calculated wavefunction is consequently inputted into the next cycle of calculations. The calculation process is repeated until the wavefunction is no longer improved, in this way the calculations are said to be self-consistent [58]. The way in which the other electrons are frozen in place whilst only one is calculated leads to problems, electrons in the HF method are repelled by an average field of electron repulsion rather than by individual electrons. In reality electrons interact and repel each other instantaneously and neglecting to calculate these correlation effects can mean the energies differ between model and experiment significantly[59]. Post Hartree-Fock methods attempt to fix this problem with the HF method; this is mainly addressed by creating a way in which electron correlation effects can be integrated into the HF method. Coupled Cluster (CC) methods promise to be the one of the most accurate QM methods to date, they are however restricted to very small systems due to their high computational cost [60].

1.4.2 Semi-empirical

In the early days of Quantum chemical modelling the HF method was restricted to only small molecules due to the large number of integrals involved in their calculation. Researchers began to calibrate their results against experimental findings and replace some portion of the *ab-initio* calculation with parameters from experimental results, these methods offered the ability to study larger molecules and polyatomic systems with the computational power

available at the time [61]. The combination of *ab-initio* methodology with parameters calibrated with experimental results is referred to as a semi-empirical method. Semi-empirical methods attempt to simplify larger systems by splitting the electrons into two groups, valence and non-valence electrons. The majority of the electrons in most larger systems aren't situated in the valence shell, hence aren't directly involved in bonding[62]. The core electrons, or non-valence shells, contribution can be approximated by combining their charge into the nuclear core to give a reduced nuclear charge. The core potential of a given atom has its properties estimated based on its similarity to other atoms in an empirical force field which is derived from experimental observations or high level QM calculations. The valence electrons are then treated with *ab initio* methods and combined with the core potentials; this process hopes to produce an answer comparable to that of a higher level of calculation theory at only a fraction of the computational cost [63]. The use of parameters to simplify calculations means the method is inherently inaccurate, if the electronic environment surrounding the atom in the simulation differs significantly from the one in which parameter is based on then the results will be negatively impacted[64]. Semi-empirical methods are most used where total accuracy isn't the prime concern and instead the calculation needs to be very quick, this makes it ideal for the optimisation and parametrisation of organic molecules for the purposes of docking and MD simulation [32].

1.4.3 Density Functional Theory (DFT)

In HF methods the wavefunction is constructed of a combination of single electron wavefunctions, Density Functional Theory (DFT) methods do also calculate single electron wavefunctions, however they do not use these to calculate a multi-electron wavefunction. Instead of calculating a multi-electron wavefunction DFT methods use the single electron wavefunctions to calculate the total electronic energy, as well as the density and distribution of electrons in the system, from these all the other properties of the system can be derived

[65]. DFT has the advantage of accounting for electron correlation effects and this means in some select systems it can produce results closer to experimental values. By not calculating the multi-electron wavefunction it can also do this at a fraction of the computational cost of wavefunction based *ab initio* methods [66]. The excellent accuracy/effort ratio of this technique has made it one of the most popular methods in computational chemistry studies, particularly QM/MM studies, owing to its efficient treatment of a relatively large number of atoms such as those composing an enzymatic active site region. Like every method, DFT has its short-comings and has trouble in describing weak non-bonded interactions (Van der Waals, dipoles and hydrogen bonding), it can also have trouble describing excited states although solutions are becoming increasingly available to these problems and its performance continues to improve [67, 68].

1.4.4 Hybrid Functionals

By relying on the strengths and advantages of both, DFT and wavefunction methods can be combined to produce a method that is more accurate whilst also being more computationally efficient than its constituent methods. The popular Becke, 3-parameter, Lee-Yang-Parr (B3LYP) method named after its creators was developed in 1993 [69] as an attempt to offer an alternative to the very computationally expensive Post Hartree Fock Møller-Plesset perturbation theory methods. To achieve this, elements of HF methodology (known as HF exact exchange) parameters are added to DFT methodology to improve the description of electron correlation effects. Hybrid functionals such B3LYP are very popular as they give structural and chemical properties that agree well with experimental results, however they can perform poorly at describing weak interactions like van der Waals forces [70]. More recently double hybrid functionals (such as B2-PLYP) have been developed to resolve this, these methods intergrate two additional parameters into the DFT method [71]. In addition to HF exact exchange, double hybrid functionals incorporate elements from post-HF

methodologies (such as MP2) to achieve greater chemical accuracy [72]. This methodology hopes to bridge the gap between the chemically accurate but difficult to implement "gold standard" that is coupled cluster calculation methods, and the widely used DFT methodology [71, 73].

1.5 QUANTUM MECHANICS / MOLECULAR MECHANICS

QM/MM simulations of enzymatic reactions have developed significantly as a technique since their first implementation nearly 40 years ago, the overall concept however has remained the same [7]. A detailed electronic structure simulation of a relatively small number of atoms that make up the reacting portion of the system, this QM region often encompasses the substrate, cofactor as well as surrounding reacting amino acid residues. This QM region is embedded in a larger region that is simulated using an MM forcefield, the MM region consists of the remainder of the protein structure and the solvent environment. The purpose of this approach is to treat the important reactive group of atoms in the QM system with the scrutiny of a high level of calculation theory such as *ab initio* or density functional theory, without incurring the computational cost of applying this calculation method to the large number of atoms that represent the rest of the system. In this way the bond formation/breaking that occurs at the active site can be accurately described whilst also accounting for the bonding and non-bonding interactions contributed by the rest of the protein and solvent system which is simulated at the MM level [74]. The combined QM/MM simulation of enzymes in this way can provide insights into the mechanistic details of the enzymatic reaction. The simulation can be used to model the reaction path [75], define the structure of transition states [76], and calculate energetic transitions of the reaction [77, 78].

1.5.1 Calculating the QM/MM Energy of the System

The total energy for the system cannot be summed up as a simple addition of the energies of the QM region and MM region, another term needs to be introduced to account for the interaction between the two regions. In the additive method for QM/MM calculations (Equation 11) the total energy of the QM region (E_{QM}) and the MM region (E_{MM}) are calculated separately and then added together. This addition alone does not fully define the total energy of the whole system so a coupling term ($E_{QM/MM}$) is needed, this represents the bonding, electrostatic and Van der Waals interactions between the two regions [79]. The additive method is the most popular from the point of view of biomolecular simulations because it considers the influence of the MM region on the QM region, this is often vital when understanding how the protein environment contributes to enzyme reactivity [80]. A distinct disadvantage with the additive method is the need for a coupling term and the calculation of this term can create difficulties of its own.

$$E_{all} = E_{QM} + E_{MM} + E_{QM/MM}$$

Equation 11: The additive method for calculating the energy of a QM/MM simulation system.

The subtractive method (Equation 12) starts with the energy of the whole system ($E_{MM\ whole\ system}$) being first calculated at the MM level, the energy contribution of the QM region calculated at the MM level ($E_{MM\ QM\ system}$) is then subtracted from this. To this subtracted amount the energy contribution of the QM region as calculated using QM methods ($E_{QM\ QM\ region}$) is added, this gives the total energy of the whole system as calculated with two distinct levels of theory [81].

$$E_{total} = (E_{MM\ whole\ system} - E_{MM\ QM\ system}) + E_{QM\ QM\ region}$$

Equation 12: The subtractive method for calculating energy of a QM/MM simulation system.

The subtractive energy calculation method has some advantages over the additive method in that it can process more than two levels of calculation, and can in fact be extended to "*n*" number of layers each represented by a different level of theory [82]. Unlike the additive method, the QM region of the subtractive method must also be calculated at the MM level. Problems can arise as to the treatment of non-parametrised atom types such as metals or transition states, without a successful calculation of this region with an MM force field the total energy of system cannot be calculated using the subtractive method.

1.5.2 The Treatment of Bonds Spanning Across the QM/MM Partition

In systems where the QM region is covalently bound to the MM region, for example this can occur with attempts to include a smaller reactive portion of a larger molecule such as a protein into the QM region. Such partitions are particularly necessary when active site residues directly react with the substrate, therefore need to have their electronic structure accurately modelled [66]. Creating a QM/MM partition along a covalent bond to include one portion of a molecule in the QM and other in the MM region requires the bond to be broken; this leaves the two boundary atoms with unsatisfied valencies. Not all bonds can be broken to form partitions in this way, conjugated polar or aromatic bonds should not be cut across as the uneven distribution of electrons in these will lead to strange behaviours in the simulation [83].

A simple solution originally suggested in a paper by Singh and Kollman[84] proposed saturating the unoccupied valencies with link atoms, this method is relatively easily implemented. The link atom method can be a hydrogen atom, a methyl group or even

pseudo-halogen atoms to the partitioned atoms. However, the addition of these extra atoms into the QM region can cause problems, it changes the thermodynamics of the system by adding to the number of total atoms. In studies seeking to simulate the thermodynamics of a reaction this can cause complications in defining the energy of the QM region[85]. These extra atoms can also negatively affect geometry optimisation processes[86].

Another alternative method for treating boundary atoms is the use of pseudoatoms, instead of a link atom being used a new atom with only a single valence is added to cap the system [87]. This atom is added to the boundary the QM region and through parametrisation of model systems be adjusted to have the same bond length, strength and electronic properties as the bond that it is replacing. It is "visible" to both the QM and MM region, in the QM region it replicates the bond which has been severed and in the MM region it functions as a normal MM atom [66]. The pseduoatom method suffers from some of the same problems as the link atom method in that it introduces artificial atoms into the simulation. Another problem is the parametrisation process for the creation of the pseudoatom, if done incorrectly it could result in a QM system that isn't representative of the real system.

The frozen local orbital approach is another method for treating the partition of bonds, an orbital is introduced to the unoccupied valence of the boundary atom to close the QM system [88]. The orbital is excluded from calculations and acts as lone pair pointing towards the severed MM atom. In this way it does not affect the other orbitals in the QM region and can be specifically designed and adjusted in such a way as to not affect the charge distribution, position or interactions of the boundary atoms [89]. Although the frozen orbital method is more difficult to implement it does provide a truer description of the QM/MM boundary as it doesn't introduce artificial atoms. The generalized hybrid orbital approach is a modification of other methods; the QM atom remains attached to the MM atom. The MM atom then has multiple orbitals added to it, the bonds between the MM atom and other MM atoms are frozen orbitals and the orbital between QM and MM atom remains open. Because

of this the electron density across the partition can readjust over the course QM/MM calculation rather than remaining fixed as in other methods [90].

The choice of the treatment of boundary regions is a crucial part of the QM/MM simulation process and is often a balance between simplicity and accuracy.

1.5.3 Embedding Techniques

How the electrostatic interactions are treated across the QM and MM region defines the embedding technique that is used. Originally the charges from the MM region were evaluated and their influence added to the atoms of the QM region as point charges, this is known as mechanical embedding because the interaction is described without the need for quantum terms [91]. However this can create problems as both regions must be parameterised in an MM force field so that the partial charges can be combined, transition state atom types and metals are not normally included in MM force fields and therefore create problems with this method of embedding [86].

The newer and more complex alternative to this is electrostatic embedding, instead of MM electrostatic interactions being added to the QM system as partial charges, the charge distribution of the MM region is directly allowed to influence the wavefunction in the calculation of the QM region [66]. Although requiring a more accurate sampling of the electrostatic environment of the MM region electrostatic embedding has distinct advantages in that it does not require the parametrisation of the QM system in an MM forcefield, additionally it allows the QM systems charge distribution to constantly adjust with any changes in the MM system [92].

1.6 AIM OF THE PROJECT

Aims:

1A Study the similarities and differences in the conformational dynamics of the two FAD-dependent tryptophan halogenase enzymes tryptophan 7-halogenase (PrnA, *Pseudomonas fluorescens*) and tryptophan 5-halogenase (PyrH, *Streptomyces rugosporus*) that lead to their differences in regioselectivity. **1B** Investigate the effect of point mutations on the structure of PrnA and compare the results with those obtained experimentally in order to offer atomistic insight to these effects. **1C** Study the interactions between the substrate tryptophan, hypochlorous acid FAD and the protein at the atomistic level to provide insights into the enzyme's proposed reaction mechanism.

2A Create a model of the Fat mass and obesity associated enzyme's (FTO) active site in order to study the electronic environment surround the catalytic iron centre. **2B** Using the active site cluster model, simulate key points in the FTO reaction mechanism using a range of conditions to examine the effects that cluster size and calculation method have on the geometry and electronic environment of the model system. **2C** Create a QM/MM model of the FTO enzyme substrate complex and compare the electronic structure and geometry with that of the various cluster models. **2D** Compute the energy barrier of key reaction steps with varying cluster sizes and computational methods to compare against previous experimental and computational models from other publications.

3A Elucidate reasonable bound structures for a range of ligands to the sugar facilitator transporter protein (GLUT5 protein encoded by the SLC2A5 gene) and attempt to correlate these results with those structures obtained via experimental methods. **3B** Study the conformational dynamics of the GLUT5 structure in a model lipid membrane to assess the contribution of the lipid membrane to the protein structure. **3C** Study the interaction between

the ligands and receptor over time to assess the influence of the ligands on the dynamic motions of the protein. **3D** Estimate the free energy of binding between the GLUT5 receptor and a range of ligands to compare the computationally calculated values with those obtained with experimental methods.

4A Predict the bound structures of a range of sugar ligands to the experimentally obtained odorranalectin (no currently identified gene) structure as well as two lactam bridged modified derivative lectin structures. **4B** Estimate the free energy of binding between the sugar ligands and the three lectin structures and compare these results with those obtained experimentally. **4C** Use the predicted bound structures to explain the preferred binding of some sugars as opposed to other at an atomistic level.

5A Predict the atomic structure and molecular orbital structures and energies of the various fluorescent probes in both the protonated and deprotonated. **5B** Computationally predict the fluorescence spectra of the probes and compare the results with those obtained by experiment. **5C** Use the predicted structures of the probes along with the computed orbital structure to explain the differences in the fluorescence between the different probes.

Methodology: To simulate and study the conformational dynamics and structural flexibility of the studied enzymes, molecular dynamics simulations were performed using the Gromacs [93] and Amber [25] packages. To predict the binding poses of ligands to proteins and peptides docking techniques were utilised with the Autodock4 [42] software. To study the geometry and electronic structure of enzymatic active sites, cluster models were created and calculations were performed using the Gaussian09 [94]. The influence of the protein environment on the geometry and electronic structure of the active site was studied using a QM/MM model, calculations on the model were performed using the ONIOM method [82] in Gaussian09 [94].

2 STRUCTURAL INSIGHTS FROM MOLECULAR DYNAMICS SIMULATIONS OF TRYPTOPHAN 7-HALOGENASE AND TRYPTOPHAN 5-HALOGENASE

2.1 PREFACE

This chapter concerns MD studies of two tryptophan halogenase enzymes namely tryptophan-7-halogenase (PrnA) and tryptophan-5-halogenase (PyrH), in order to study their conformational flexibility and remarkable regioselectivity. All the computational simulations and analysis are completed and a manuscript is in the final editing stages before submission for publication. Additional figures, graphs and tables can be found attached in the appendix section as “Supporting Information for Structural Insights from Molecular Dynamics Simulations of Tryptophan 7-Halogenase and Tryptophan 5-halogenase”.

2.2 INTRODUCTION

Many pharmaceutically important natural organic compounds (including antibiotics, such as chlortetracycline [95] and vancomycin [96], the antifungal compound pyrrolnitrin [97] and chemotherapeutics - salinosporamide A [98] and rebeccamycin [99, 100]) are chlorinated. Halogenating enzymes perform regioselective halogenation of aromatic compounds efficiently in solution using only chloride ions at pyshiological temperatures and atmospheric pressure. Selective non-enzymatic chlorination of C-H bonds remains a chemical synthesis challenge, for example the halogenation of tryptophan in solution lacks regioselectivity and produces a mixture of products with chlorine added at 1^{st}, 5^{th} and 7^{th} carbon of the indole ring [101]. From and industrial point of view this is unacceptable as the desired isomer is produced with a lower yield and expensive to separate from the others. At the same time many natural products with pharmaceutical relevance contain halogen atoms at a range of different positions, which would be difficult to synthesise and rely on the use of protecting groups and metal based catalysts, these strategies introduce extra reaction steps to the synthesis; increasing financial costs and lowering yields [102, 103]. Hence a detailed understanding of the enzymatic mechanism of regioselective chlorination/halogenation of natural organic compounds and knowledge of the origin of the regioselectivity is of importance to organic chemical synthesis. Halogenating enzymes are attractive as biocatalysts because they can be engineered to suit different synthetic purposes [104], not only adjusting their regioselectivity but their ability to accept a range of different substrates [105]. The indole ring of tryptophan is chlorinated at different positions - respectively at the 5^{th}, 6^{th} or 7^{th} carbon atom by distinct flavin-dependent halogenases, these include: PyrH-*Streptomyces rugosporus* [106] (trypthophan 5-halogenase), Thal-*Streptomyces albogriseolus* [107] and SttH-*Streptomyces toxytricini* [108] (tryptophan 6-halogenases) and RebH-*Lechevalieria aerocolonigenes* [109] ,and PrnA-*Pseudomonas fluorescens* [110]

(tryptophan 7-halogenase). All of these enzymes exhibit high level of regio- and stereo-selectivity.

Our study focuses on the simulation of two flavin-dependent halogenases - tryptophan 7-halogenase PrnA and tryptophan 5-halogenase PyrH to study their conformational flexibility and possible structure function relationships. In particular PrnA catalyzes chlorination of free tryptophan to 7-chlorotryptophan, the first step in the biosynthesis of the antibiotic and antifungal compound pyrrolnitrin [97]. PyrH catalyzes chlorination of free tryptophan to 5-chlorotryptophan, part of the biosynthesis of the antibiotic compound pyrroindomycin B [106]. It is important to analyze the reason for the regioselectivity, with focus towards the differences at the active sites of these structurally similar enzymes. X-ray crystallographic structures, of the two halogenases exist and provide the basis for our MD simulations [110, 111]. The catalytic turnovers of tryptophan 7-halogenase and tryptophan 5-halogenase differ, in one experiment PyrH was found to convert 100% of tryptophan, whereas PrnA converted only 59%, the origin of this difference has not yet been elucidated [105].

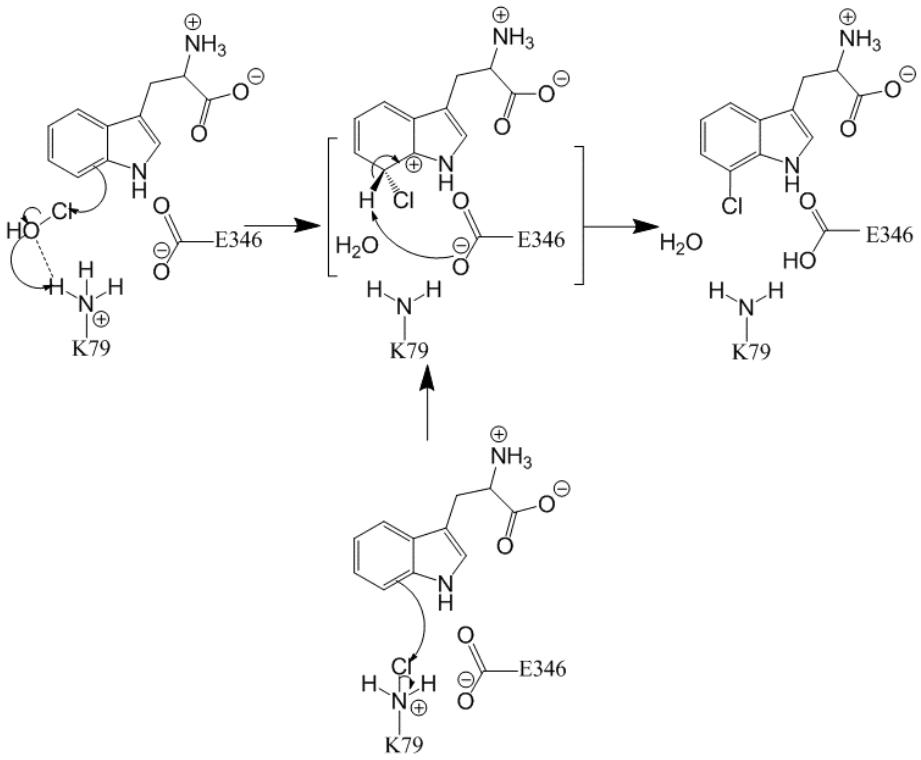

Figure 2: Two proposed mechanisms for the halogenation of tryptophan by hypochlorous acid.

Several reaction mechanisms were proposed for the enzymatic chlorination of tryptophan [112], all the presented alternate mechanisms agree that FAD reacts with dioxygen to produce to produce a reactive 4α-hydroperoxyflavin intermediate. In Figure 2**,** mechanism A and B 4α-hydroperoxyflavin reacts with a chloride ion to produce hypochlorous acid, the free hypochlorous acid travels through a "tunnel" in the protein to the bound tryptophan substrate. In Mechanism A electrophilic aromatic substitution of tryptophan occurs at C7 facilitated by the interaction of hypochlorous acid with the active lysine sidechain [110]. In Mechanism B, hypochlorous acid reacts with a deprotonated form of the active lysine residue to produce a Chloro-Lysine intermediate, this then reacts with the C7 of tryptophan in an electrophilic aromatic substitution reaction [113]. Mechanism A and B both agree that hypochlorous acid is produced but differ in the way hypochlorous acid reacts with tryptophan, both mechanisms also specify that deprotonation of the chlorotryptophan intermediate by the active glutamate to produce 7-chlorotryptophan. Other mechanisms were hypothesised before the X-ray crystal structure of PrnA was solved, they are based on similar monooxygenase enzymes like PHBH [114, 115]. These mechanisms are based on direct contact between FAD and the substrate tryptophan. Presently this type of mechanism seems unlikely, inspection of the crystal structures of the known FAD dependent halogenase enzymes shows a >10Å distance between the FAD and tryptophan binding sites [106-110]. The separation between the ligands would be too far for direct interaction. However, close contact between the cofactor and the substrate, whilst not observable in the crystal structure, is a possibility that cannot be entirely excluded. A large conformational change could take place in the protein structure, bringing the two binding sites into close proximity and allow direct reaction between the two ligands.

Because the FAD and tryptophan binding sites are distant, the chlorinating agent hypochlorous acid is thought to travel through a channel in the protein[110]. Two amino acids - K79 and E346 in PrnA (analogous to K75 and E354 in PyrH) are positioned in close

proximity to the reactive carbon of tryptophan's indole ring, they are thought to be involved in the activation of the hypochlorous acid for the halogenation step of the reaction [116]. The role of K79 and E346 in PrnA is supported by an experimental mutagenesis showing that the K79A mutant had no detectable activity, and in the E346Q mutant the Kcat values for the halogenation are decreased by two orders of magnitude [110].

PyrH and PrnA have a 40% sequence identity and 58% sequence similarity making their amino acid sequencing quite similar[117, 118]. In the crystal structure of the PyrH FAD binding domain a "strap" region was identified. The FAD binding strap is hypothesized to be a structural feature that allows for "communication" between the two binding sites, in addition it is also hypothesized to be involved in the regulation of FAD binding [111]. The crystal structures of PyrH and PrnA reveal that the FAD binding sites are basically identical, the FAD binding site of the RebH crystal structure [109] was also found to overlay almost identically to PrnA and PyrH (Figure 3). PyrH possesses a structurally different tryptophan binding site to PrnA, the substrate tryptophan in the PyrH crystal structure is bound in a way that is upside down in respect to tryptophan in the PrnA crystal structure (Figure 4). However, the active to-be halogenated carbon of tryptophan in PrnA and PyrH superimpose when the two protein structures are aligned. The positioning of the active carbon is located between the active lysine and glutamate residues (Figure 4), these two residues show similar orientation between the two enzymes, this suggests that the reason for the two enzymes producing different halogenated products lies in the way that tryptophan is bound in the active site. The tryptophan and FAD binding domains are highlighted in Figure 5.

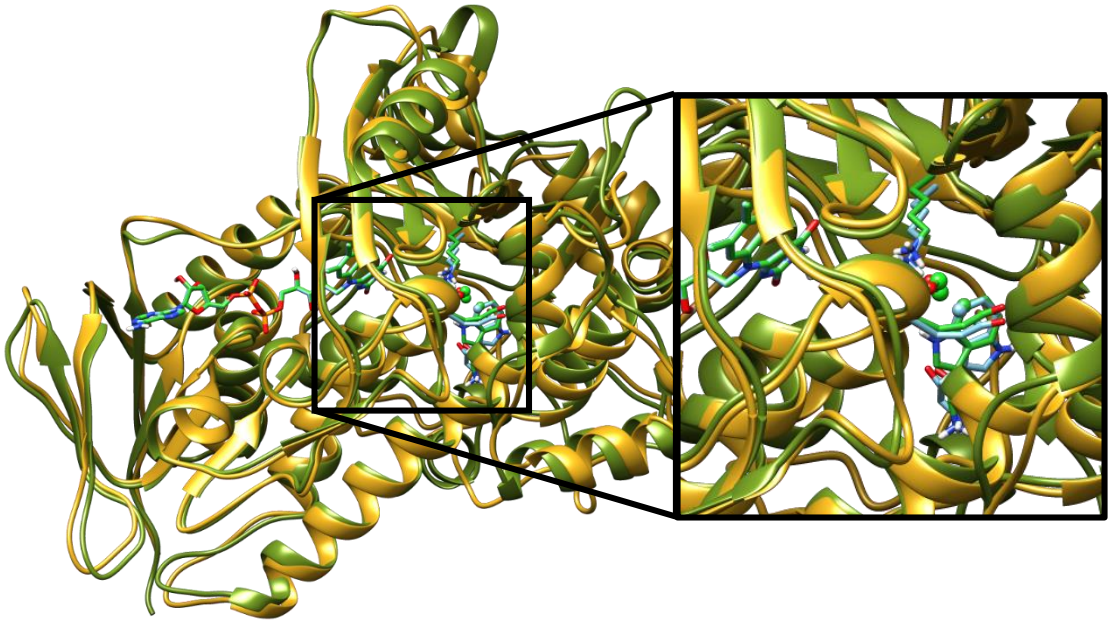

Figure 3: The ribbon structure of PrnA (dark green) and PyrH (gold) structurally aligned with one another. Stick representations of the bound ligands FAD and tryptophan as well as the active glutamate and lysine residues. Hypochlorous acid and the to-be halogenated carbon of tryptophan are respresented as spheres. Light blue carbons represent PrnA and green carbons represent PyrH. Additonally element colours are as follows; nitrogen is dark blue, oxygen is red and hydrogen is white.

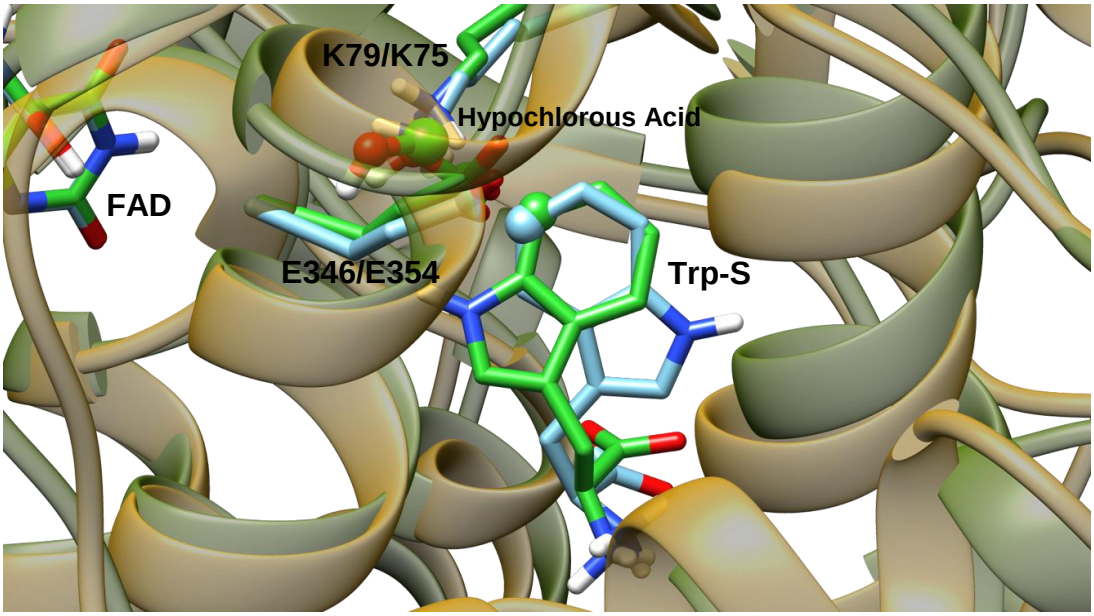

Figure 4: A view of the aligned crystal structures of PrnA and PyrH rendered with transparent protein ribbons, with PrnA with yellow ribbons and PyrH with green ribbons. The substrate tryptophan, hypochlorous acid and the active lysine and glutamate residues are rendered as tubes with carbon atoms coloured according to the protein, PrnA in bright green and PyrH in light blue. The to-be halogenated carbon (C7/C5) of the substrate tryptophan is rendered as a sphere.

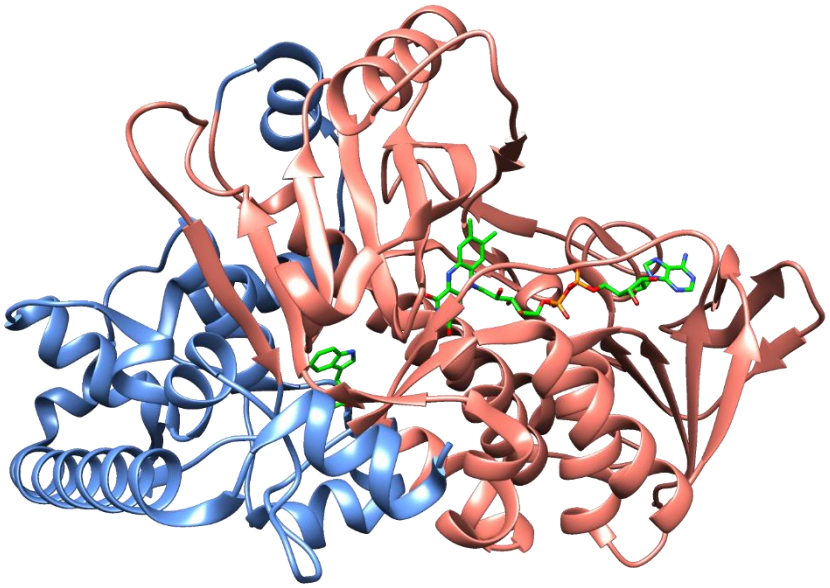

Figure 5: The structure of PrnA highlighting the FAD and tryptophan binding domains, the FAD binding domain ribbons are coloured oink while the tryptophan binding domain ribbons are coloured light blue. The FAD and substrate tryptophan ligands are rendered as tubes with carbon atoms coloured bright green and heteroatoms coloured according to: nitrogen is dark blue, oxygen is red, phosphorus is orange and hydrogen is white.

Enzymes are large, flexible and dynamic molecules that naturally undergo a wide range of conformational changes and molecular motions, these range in timescale from femtoseconds to hours[5, 119]. Many of these motions are functionally important and relate enzyme structure to function [2, 120]. Experimentally determined protein structures (e.g. by X-ray crystallography) provide valuable structural information however, limited to only a snapshot or static structure, averaged over the number of molecules in the crystal lattice, and the duration of the experiment [121]. In addition, steric effects can arise due to the compactness of the crystal environment [122]. Enzyme conformational flexibility can play a substantial role in stabilizing the protein interactions vital in facilitating ligand binding and unbinding events [123]. In addition enzymatic molecular plasticity is involved in assisting the migration of ligands to the binding site, the diffusion of gases and small molecules through the protein [124]. Mutations of key residues, involved in catalysis, and binding, can influence not only locally the structure, but can also exercise a long range structural effect on the protein conformation as a whole. Exploration of dynamic events in proteins, using experimental methods, can be a challenge; thus computer-based experiments, for example Molecular Dynamics (MD) simulations can be applied to study this [119-121, 125]. Long range atomistic molecular dynamics (MD) simulations of the two enzymes were performed in order to elucidate the structure-function relationships and mechanistic implications related to the origin of regioselectivity in both enzymes.

2.3 METHODS

An initial structure for the MD simulations of the wild type full complex PrnA was created from the pdb structure of the enzyme (PDBID: 2AR8) [110]. The product 7-chlorotryptophan was separated to create the substrate tryptophan and hypochlorous acid, in addition the chloride ion bound at the FAD binding site was removed and FAD modified to create hydroxyl-FAD (from this point forward FAD will be refer to hydroxy-FAD) . These

changes were made with the aim of replicating the active full complex of mechanism A and B prior to the halogenation of the substrate tryptophan. Modification of the atomic coordinates was performed using Maestro 9.9.013 [126]. The site-mutants K79A and E346Q were prepared by mutating the respective residues in the wild-type full complex structure using Maestro [126]. The initial structure of PyrH for MD simulations was prepared by superimposing the PyrH crystal structure (PDBID: 2WET) with that of the wild-type full complex PrnA, the coordinates of hypochlorous acid from this were then added to the PyrH crystal structure and the sulphate and chloride ions from the crystal structure were removed. The parameters for FAD-OH and hypochlorous acid were generated by the PRODRG web server [127] for the GROMOS96 43a1 forcefield [128] with atomic partial charges for hypochlorous acid supplemented from QM calculations performed by the Automated Topology Builder web server [129]. The missing coordinates of the two loop regions in the PyrH crystal structure (PDBID: 2WET) structure were modelled using the Modeller [130] plug-in for Chimera 1.10.2 [131], the setup for PyrH then followed the same protocol as the one for PrnA.

The hydrogen atoms missing from the X-ray crystal structure were added using Gromacs 4.5.5 [93]. To remove unfavourable steric clashes in the starting structure *in vacuo* energy minimization was performed using the steepest descent algorithm until the maximum force was less than 100 KJ/mol^{-1}/nm^{-1}, the protein was then placed in a box with periodic boundary conditions. The energy minimized protein structure was then solvated using the Single Point Charge [19] (SPC) model for water. The total charge of the system was neutralized by adding the correct number of Na^+ or Cl^- ions, to make the overall charge of the system zero. Another energy minimisation (using the same conditions as described for *in vacuo* energy minimisation) was then performed, to reduce close contacts between solvent molecules or ions that may be unfavourably close to the protein structure. The energy minimized structure was then subject to position restrain MD for 50ps at 300K, during that, the protein structure

was restrained and the water was allowed to equilibrate. The position restrained dynamics simulations are performed in NVT ensemble, a constant Number of particles, Volume, and Temperature with a time step of 2fs. The productive MD was then carried out with the output structure from position restrained dynamics providing the initial structure for 1μs as in NPT ensemble at a temperature of 300K. The MD trajectories, were analysed over the time period of 100ns-1000ns, after equilibration phase was reached, using tools provided in Gromacs and visualisation and inspection of the trajectories was made with Visual Molecular Dynamics (VMD) [132]. Dynamic Cross Correlation Analysis (DCCA) and Principle Component Analysis (PCA) were performed using the Bio3D package [133] for Rstudio [134]. DCCA is used to visualise which residues play a role in the correlated motions that occur between different components of the protein structure [135]. The level of correlation between each c-α atom can be quantified and visualised on a plot, with correlations ranging from +1 to -1 indicating strong positive to negative correlation. This allows the identification of regions in the protein showing correlated motion in the simulation that could be overlooked by visual inspection. All images of molecular models for this chapter were created using UCSF Chimera 1.11.2 [131].

2.4 RESULTS AND DISCUSSION

2.4.1 Conformational Dynamics of Full Complex Wild-Type PrnA

In total five 1μs MD simulations were performed: the full complex of wild-type PrnA, apoenzyme PrnA, two single point mutant forms K79A and E346Q of PrnA as well as the full complex wild-type PyrH.

The RMSD profile for all Cα atoms for the 1 μs MD simulation of the wild-type full complex PrnA is 3.5Å. The RMSD profiles of all five of the 1μs simulations (Figure 6) indicated that the initial equilibration phase was completed after 100ns. As well as the 1μs simulations three additional independent 200ns MD simulations of the wild-type full complex of PrnA

were performed, these used the same initial structure but different initial velocities (SI Figure 1). These repeat simulations were created to evaluate the effect of statistical error on the quality of the simulations. The RMSDs for the each of these trajectories was consistent with the 1μs wild-type full complex PrnA simulation indicating the good quality of the simulations. The average RMSD from the 200ns runs (calculated between 100ns-200ns) was 3.3Å[136] and their radius of gyration (SI Figure 2) was 22.9Å, showing that the protein remains compact and stable during the simulation timescale.

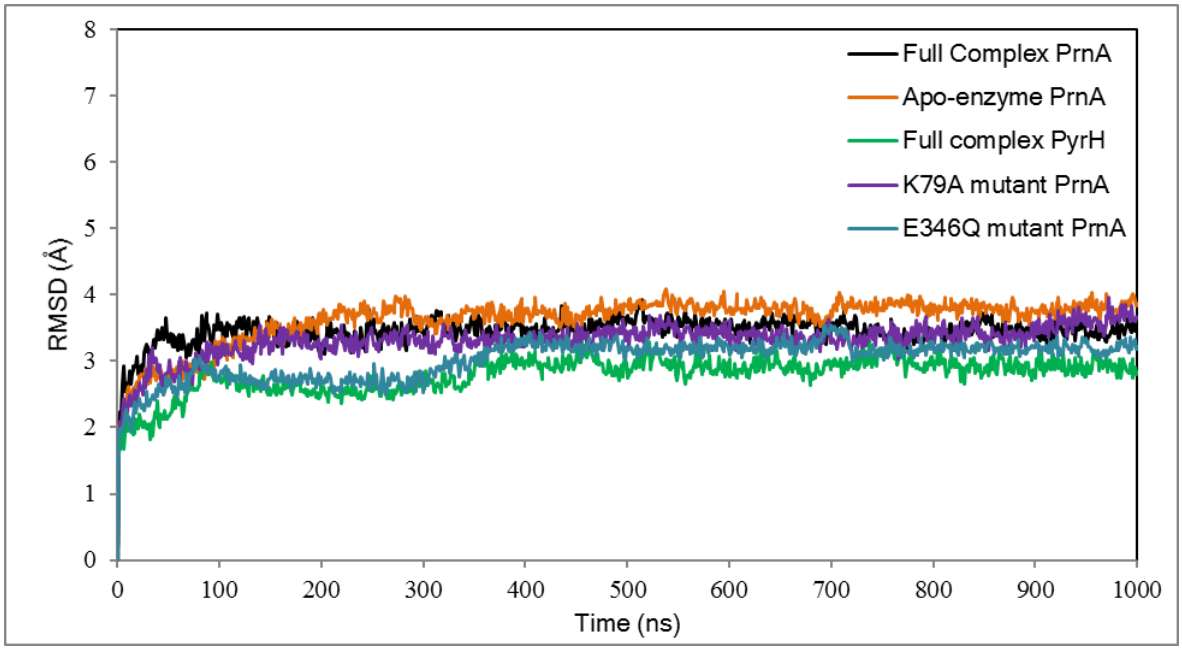

Figure 6: The RMSD plot of all five 1µs MD simulations; the PrnA full complex, Apo-enzyme PrnA, PyrH full complex as well as the K79A and E346Q mutant forms of PrnA.

The average RMSD of the full complex wild-type PrnA was 3.5Å and of the apoenzyme PrnA was 3.6Å, this shows that the ligand bound structures have similar levels of structural flexibility. The solvent accessible area (SAS) of the apoenzyme PrnA is slightly lower than the SAS of the full complex wild-type PrnA (SI Figure 3). This is likely due to structural change associated with ligand binding, in which the protein structure of PrnA becomes less compact to accommodate the bound ligands.

The RMSF profiles of the full complex PrnA and apoenzyme PrnA are presented in Figure 7. For the full complex PrnA, The peak region at residue P93 exhibits a high RMSF reflecting its position at a particularly flexible point of the loop that precedes the key tryptophan interacting residues: H101, F103, G104 and N105, these residues are involved in the binding of the tryptophan substrate. It is therefore possible that the flexibility of this loop is functional in the binding and orientation of the tryptophan substrate.

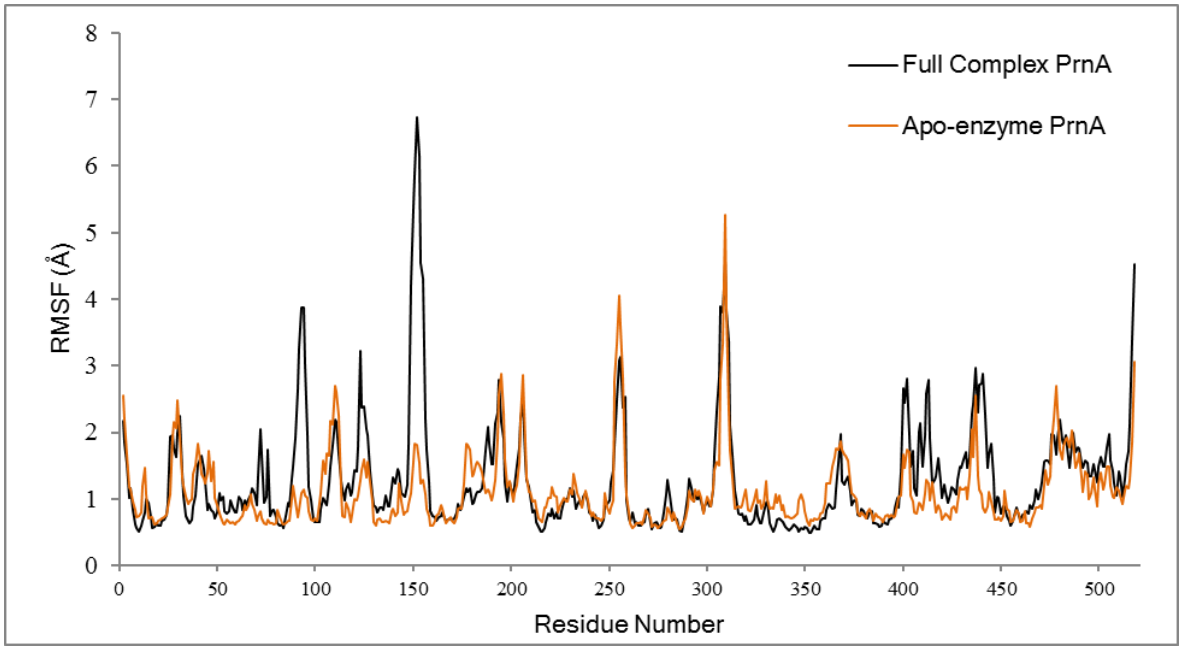

Figure 7: The RMSF plot of the PrnA full complex and PrnA Apo-enzyme simulations.

A flexible loop consisting of residues from 147-159 has a maximum RMSF value of 6.73Å centred on residue G152 in the full complex PrnA (Figure 7**)**, the same region has an RMSF value of only 1.81Å in the apoenzyme form of PrnA. The loop in questions is found on the exterior of the protein and is solvent, intra-loop hydrogen bonds are found between the residues 147-159 of the loop. The loop immediately precedes S157, a hydrogen bonding residue of the tryptophan substrate. The RMSD plot of the loop region 147-159 (Figure 8) shows the loop adopting a stable orientation after 300ns, it is upon adopting this conformation that S157 forms a hydrogen bond with the carboxylate of the substrate tryptophan. The dynamics of the loop differ greatly between the full complex wild-type PrnA and the apoenzyme PrnA, suggesting that a conformational change occurs in the loop upon binding of the substrate tryptophan. In the apoenzyme the S157 sidechain forms hydrogen bonds with the neighbouring residues A80, M156 and Y443 instead. These protein-protein hydrogen bonds stabilize the 149-159 loop of the apoenzyme and it maintains a more compact conformation which is reflected in the lower RMSD of the loop in the apoenzyme.

The DCCA of the full complex wild-type PrnA (SI Figure 5), indicates about strong positive correlation (the red region) between the strap region (residues 39-54) and the flexible loop region 149-157 in the tryptophan binding domain (roughly residues 154-170 in the DCCA plot). In the DCCA plot of apoenzyme PrnA (SI Figure 6) this region of positive correlation is much smaller suggesting that the correlated motion may be associated with the binding of the tryptophan substrate. A correlated motion between the portion of the FAD strap region closest to the substrate tryptophan (residues 50-54) and the important catalytic residue E346 was also found and supports the idea of the strap region being an important link between the two binding sites.

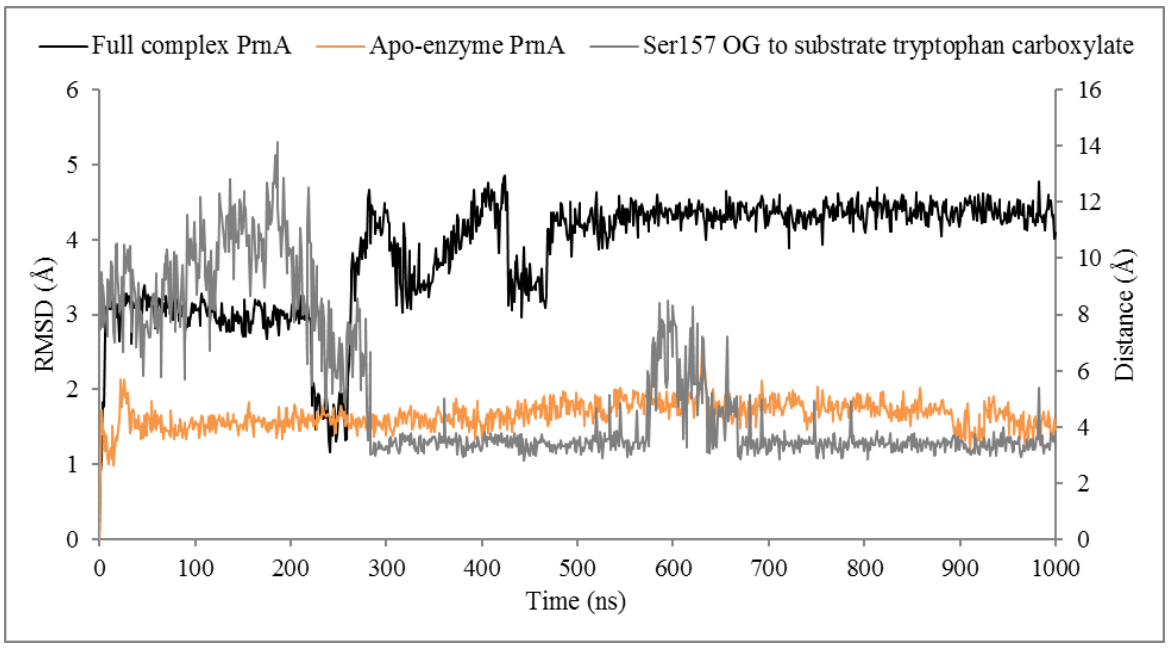

Figure 8: A plot showing the relationship between the RMSD of the flexible loop region in both the full complex PrnA and the apo-enzyme PrnA spanning residues 154-170 (left y axis), and the hydrogen bonding interaction distance between S157 side chain oxygen and the substrate tryptophan (right y axis).

The regions of residues 250-262 correspond to a flexible loop positioned between the FAD and tryptophan binding sites, this region shows correlation with the catalytic residue K79 in the DCCA plot of the full complex wild-type PrnA (SI Figure 5). Upon inspection of the trajectory this correlated motion is likely due to the close proximity of the two protein strands forming hydrogen bonds along their respective backbones.

The region of residues from 355 to 380 that show fluctuation in both the RMSF plot of the full (Figure 7**)** complex wild-type PrnA and apoenzyme PrnA correspond to a long α-helix that intersects the FAD and tryptophan binding sites, it contains the tryptophan hydrogen bonding residue Tyr351. In the DCCA plot of the full complex wild-type PrnA (SI Figure 5) this region 355-380 shows correlation with the important tryptophan binding residue W455. The same region in the DCCA of the PrnA apoenzyme -enzyme (SI Figure 6) shows a less intense correlation, this likely indicates that the motion is related to tryptophan binding. A relatively large span of residues from 396-456, shows relatively more fluctuation in the full complex than the apoenzyme (Figure 7**)**, possibly due to its proximity to many important tryptophan binding residues. This area 396-456 is composed of several helices joined by loops, intermolecular interactions create a compact hydrophobic cluster stabilising the structure of the area, and the region directly precedes the important tryptophan binding residues: Y443, Y444, W455 E450 and F454 and N459.

2.4.2 Tryptophan binding site interactions of Wild Type PrnA

The high level of regioselectivity of FAD-dependent halogenating enzymes is thought to depend on the proper orientation of the substrate tryptophan [113]. Tryptophan position and orientation allows only for the respective carbon from the indole ring (C7 in PrnA or C5 in PyrH) to be most favourably oriented for reaction. In order to accomplish stable binding of tryptophan, an extensive network of hydrogen bonds, electrostatic interactions, and Van der Waals interactions are found around it. The measured distances of the significant interactions

of the tryptophan substrate observed in the X-ray crystal structure and our wild type full complex PrnA MD simulation are recorded in Table 1, Table 2 and Table 3. K79 and E346, thought to be important for hypochlorous acid activation are also involved in a network of hydrogen bonding and electrostatic interactions that maintain their orientation in the active site relative to the substrate tryptophan and the chlorinating agent hypochlorous acid.

Residue 1	Atom 1	Residue 2	Atom 2	% of the simulation time <3.5Å	Average distance (Å)	Distance in Crystal Structure (Å)
H395	NE2	E346	OE2	53	3.5	2.5
E346	OE1	HYP	O1	91	3.0	3.5
E346	OE2	HYP	O1	90	3.1	5.2
G104	N	Trp	O	82	3.0	6.0
G104	N	Trp	OXT	81	3.1	8.0
Y443	OH	Trp	N	95	3.1	3.1
F454	O	Trp	N	69	3.4	2.8

Table 1: Hydrogen bonding interactions for the tryptophan-binding site in the wild-type full complex PrnA. Measurements are made between the donor and acceptor atoms.

The substrate tryptophan (Table 1) can act as both a hydrogen bond donor and acceptor with its amino, carboxylate and indole ring nitrogen. For example, the backbone nitrogen of G104 participates in a hydrogen bond to tryptophan's carboxylate moiety. The amino group of tryptophan is hydrogen bonded to the sidechain phenolic oxygen of Y443 and the backbone carbonyl oxygen of F454. The amino group of tryptophan can also makes electrostatic interactions (Table 2) with the carboxylate of E450 with an average distance of 4.2Å. E450 in turn interacts with the sidechain amino group of K57

distance 5.8Å, this kind of electrostatic interaction network ensures strong binding of the substrate tryptophan. The proposed electrostatic interactions of the substrate tryptophan are shown in Figure 9.

Residue 1	Atom/Group 1	Residue 2	Atom/Group 2	Average distance during MD (Å)	Crystal Structure Distance (Å)
H395	NE2	E346	Carboxylate	3.7	4.5
E346	Carboxylate	Hypochlorous acid	H	2.5	4.6
K79	NZ	Hypochlorous acid	O1	6.7	3.2
E450	Carboxylate	Trp	N	4.2	3.8
K57	NZ	E450	CD	5.8	10.3

Table 2: Electrostatic interactions for the tryptophan binding site of the PrnA full complex. Measurements are made from the centre of the charged group.

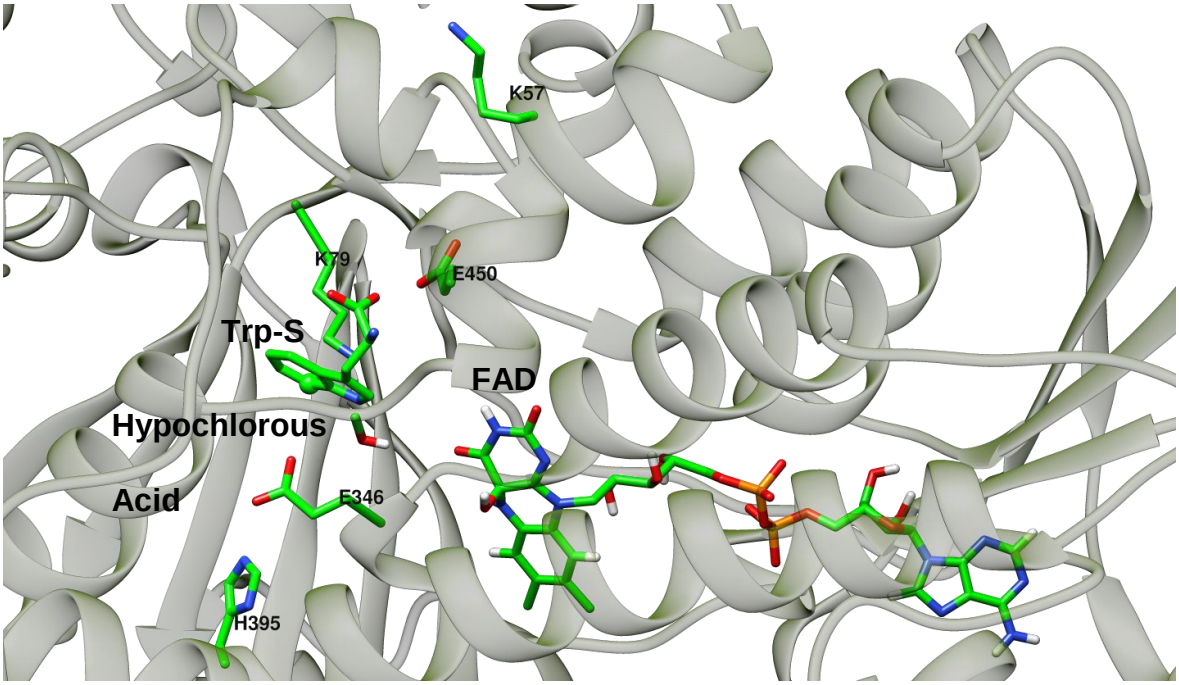

Figure 9: The potential electrostatic interactions around the substrate tryptophan binding site in PrnA. The distances between the interacting atoms are found in Table 1.

Hypochlorous acid can participate in hydrogen bonds as well as interacting with charged residues in the enzyme active site. For example the hydrogen atom of hypochlorous acid has a partial positive charge (0.455*e*) and forms a strong hydrogen bond with the sidechain of E346, at a distance of 2.5Å (Table 2**)**. In our initial structure of PrnA K79 is in close proximity to hypochlorous acid and seems a likely candidate for hydrogen bonding, however in the MD simulation hypochlorous acid moves away from K79, reflected in the average distance of 6.7Å in the MD (Table 2). The carboxylate sidechain of E346 has two oxygen atoms: OE1 and OE2 with which it is possible to form potential hydrogen bonds (Table 1). The E346 carboxylate forms an electrostatic interaction with the NE2 nitrogen atom of the protonated H395 sidechain. Hypochlorous acid makes a strong hydrogen bonding interactions with the carboxylate sidechain of E346, the average distance between the hydrogen and two carboxylate oxygen atoms was measured and found to be 3.1Å. The E346 carboxylate side chain also interacts with the positively charged doubly protonated H395

3.7Å (Table 2). In the crystal structure the indole nitrogen atom of tryptophan forms a hydrogen bond with the backbone carbonyl oxygen of E346 [105], however in the MD simulation the backbone carbonyl of E346 moves away from tryptophan to make other hydrogen bonding interactions with Thr348 and the hydroxyl oxygen of hydroperoxyflavin moiety of FAD. The hydrogen bonding interactions of the substrate tryptophan are shown in Figure 10.

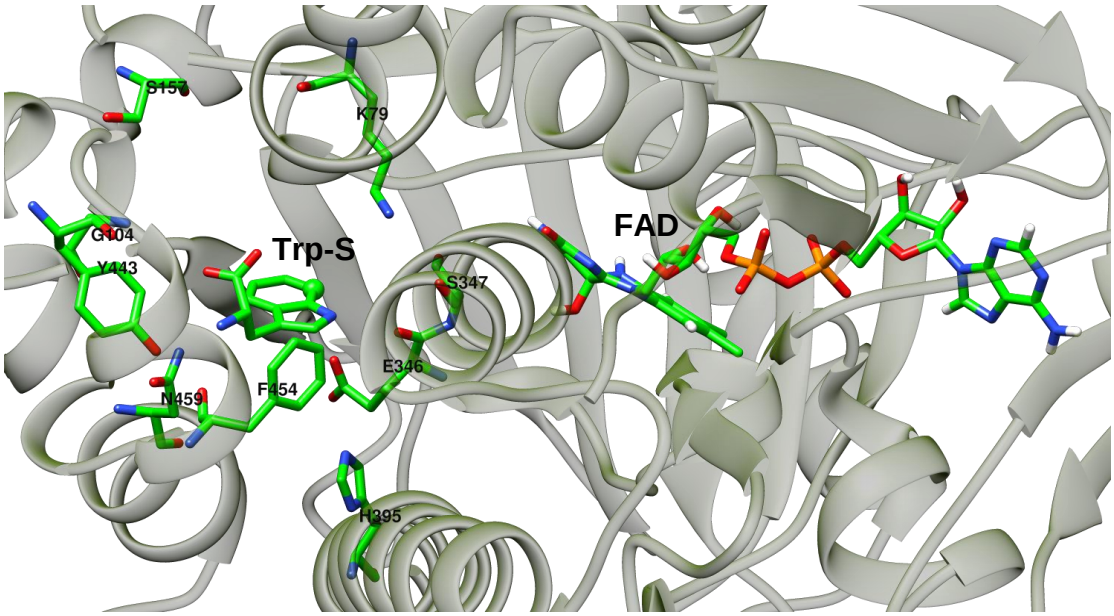

Figure 10: The hydrogen bonding interactions surrounding the substrate tryptophan in PrnA, the distances between the donors and acceptors are shown in Table 1.

Residue name and number	Average distance (Å)	Distance in Crystal Structure (Å)
Ile52	6.2	5.0
His101	6.3	5.4
Phe103	4.7	5.6
W455	5.1	5.9

Table 3: Distances between the Centres of Mass between the hydrophobic sidechains and the indole ring of the substrate tryptophan.

F103, W455 and H101 form π-π stacking interactions with tryptophan (average distances between the centres of mass respectively 4.71Å, 5.07Å and 6.27Å). Strong positive correlation between residues 100-130 and residues 475-505 is found in the DCCA plot of the full complex wild-type PrnA (SI Figure 5) The region of Residues 475-505 form a long alpha helix running perpendicular to His101/Phe103 region. The correlation is likely caused by hydrophobic interactions between the two features as the loop portions of the His101/Phe103 region are seen in SI Figure 7, from the structure it is clear to see that the two regions are interwoven and any movement in one will affect the other region as well. The intensity of the correlation is much less pronounced in the DCCA plot of the apoenzyme (SI Figure 6), suggesting that this correlation only arises in when tryptophan is bound. Throughout the MD simulation W455 remains close to the substrate tryptophan participating in a stable π-π stacking interaction with the substrate (Table 3). E346 and hypochlorous acid are also located in close proximity to the sidechain of W455, why then doesn't W455 become halogenated? K79 is not found in proximity to W455, this reinforces the idea that suitable proximity to both K79 and E346 are needed for the halogenation reaction to occur as otherwise W455 would become halogenated. The potential hydrophobic interactions in the tryptophan binding site are shown in Figure 11.

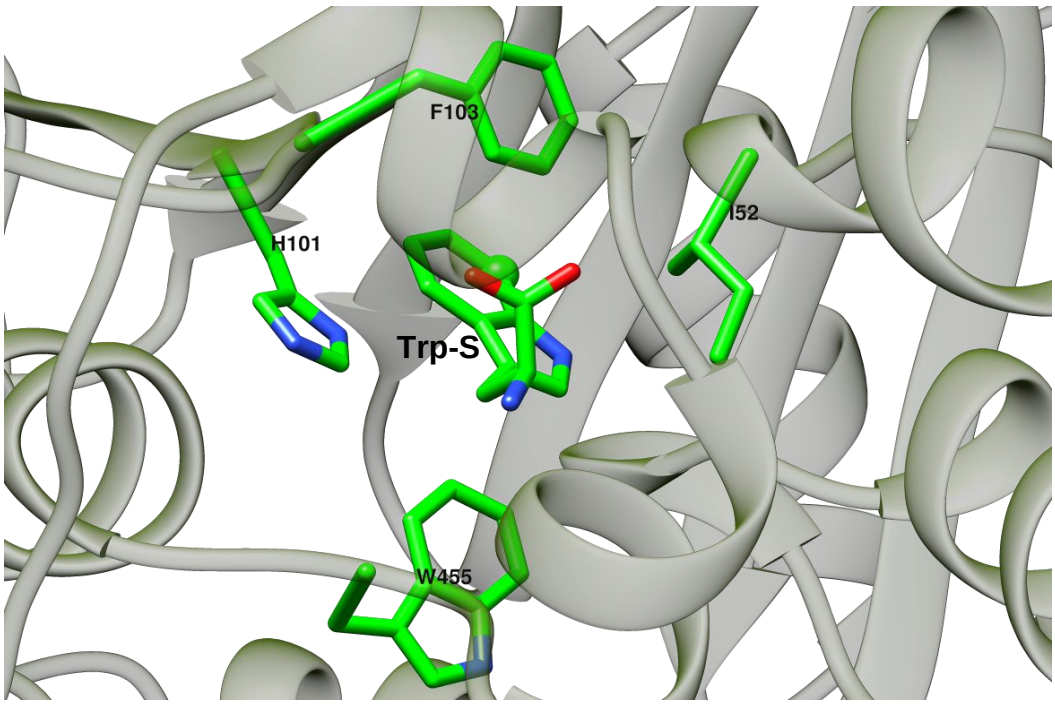

Figure 11: The hydrophobic contacts surrounding the substrate tryptophan in PrnA, the distances between the centers of mass of the substrate tryptophan's indole ring and the hydrophobic amino acid sidechains are shown in 3.

2.4.3 FAD binding site in PrnA

A structural feature previously observed in the crystal structure of PyrH, is the FAD binding "strap". This strap region is thought to control the binding of FAD, and is also hypothesised to act as a line of communication between the FAD-binding and tryptophan-binding sites in PyrH [111]. In the crystallographic study of PyrH the authors remarked that the electron density of the strap region is relatively low, implying that it is a particularly flexible region of PyrH [111] Through inspection of the crystal structure, we found that a similar strap region superimposes with that of the Pyrh crstal structure so likely exists in PrnA. The high flexibility of the strap region, along with its probable influence on both FAD and tryptophan as shown in SI Figure 8 and SI Figure 9 points towards the strap region fulfilling a similar role to one it is hypothesised to in PyrH. In both enzymes, the strap region consists of a long straight section of residues running parallel to FAD with no secondary structure elements (SI Figure 9). The region of the strap that is in close contact to FAD forms several hydrogen bonds, electrostatic interactions, Van der Waals and cation-π interactions with FAD, these are evident in the crystal structure as well as the MD simulations of the PrnA wild type full complex (SI Table 1) [112]. The interactions of FAD in the binding site can be seen in Figure 12 and Figure 13.

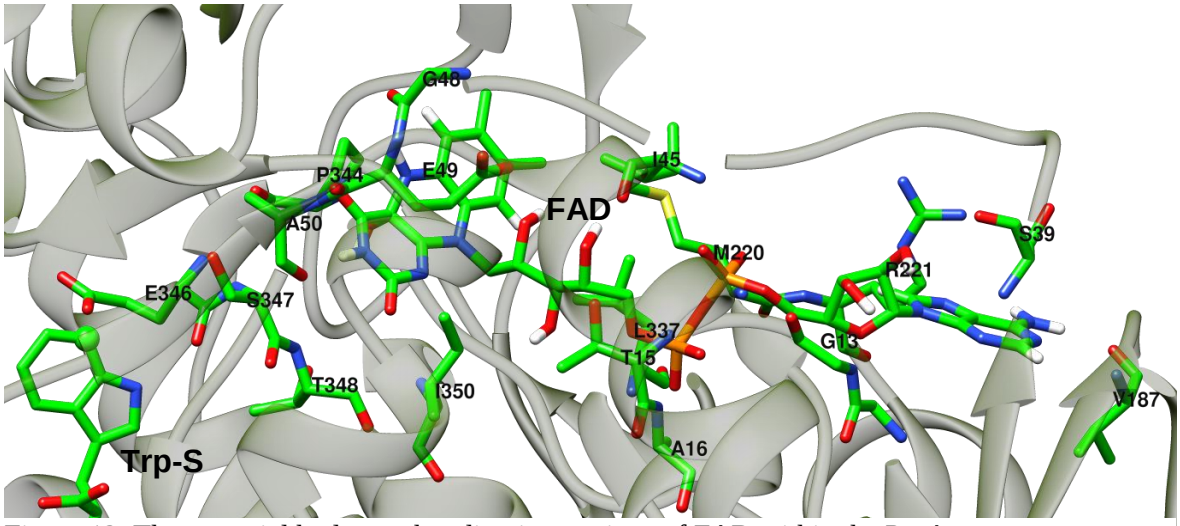

Figure 12: The potential hydrogen bonding interactions of FAD within the PrnA enzyme.

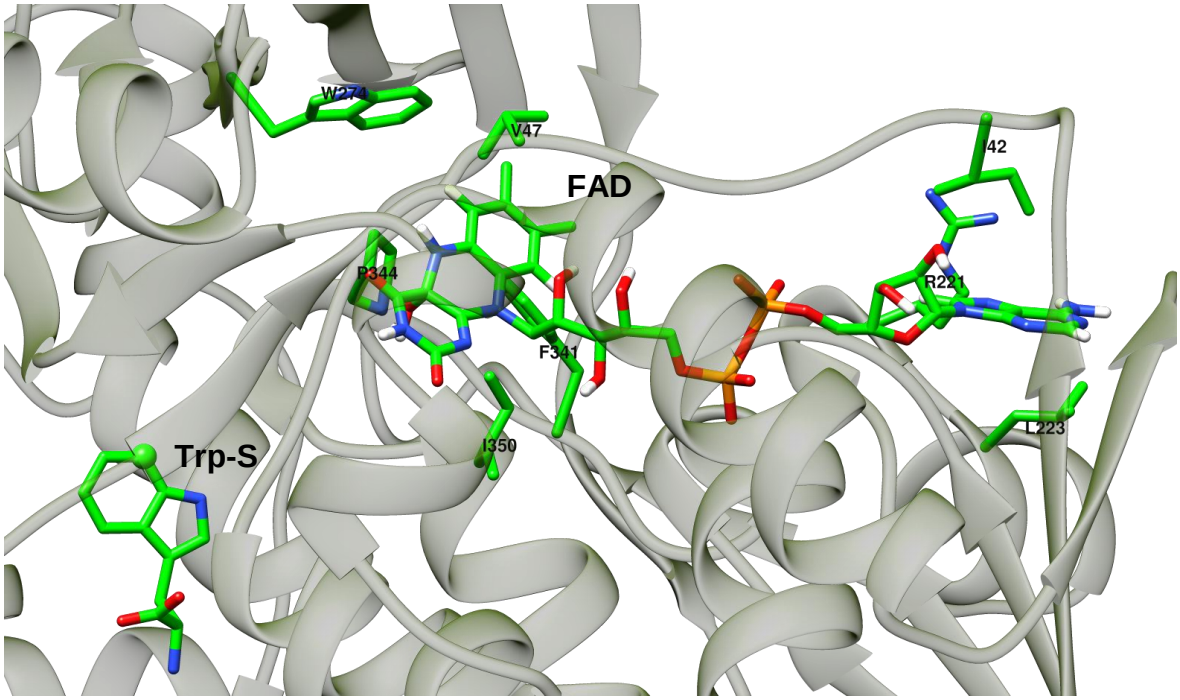

Figure 13: The potential hydrophobic and cation-π interactions of FAD within the PrnA enzyme.

FAD is a flexible molecule, in the crystal structure it adopts a linear extended conformation, in the MD we see it undergoing a structural transition around 200ns to adopt a more bent and compact conformation (SI Figure 19 and SI Figure 21). We see this structural transition of FAD in all MD simulations of PrnA and PyrH. In the PrnA full complex simulation the change in the conformation of FAD happens concurrently with a structural transition seen in the FAD strap region (SI Figure 8), the strap adopts a conformation after around 200ns that shows reduced structural flexibility. In comparison, the RMSD of the strap region in the apoenzyme form of PrnA shows higher flexibility as well as larger fluctuations (SI Figure 8). The increased flexibility of the strap region in the apoenzyme PrnA MD in contrast to the full complex PrnA MD suggests that the strap region is heavily involved with the binding of FAD and becomes more stable in the presence of FAD. This is most obvious when comparing hydrogen bonds of the X-ray crystal structure to those from the MD (SI Table 1).

The RMSF profile of the strap region (SI Figure 10**)** shows that although the apoenzyme form possesses higher flexibility this is mainly due to residues 45-49 located in close proximity to the flavin moiety of FAD and forms stable hydrogen bonding interactions with FAD (SI Table 2**)**. The region of residues from 50 to 53 connects the FAD binding residues to those of the tryptophan binding site. S54 hydrogen bonds E450, is a key salt bridge interaction for the binding of the amino group of tryptophan. The equilibration of the strap region and FAD causes a conformational change that brings the sidechain of S54 into the proximity of E450 to form a hydrogen bond. This movement brings E450 into the proximity of the tryptophan amino group **()**. This movement shows a structural basis for the predicted communication between the FAD and tryptophan-binding sites, the binding of FAD can directly influence the binding of tryptophan through the strap region that connects both domains.

Most of the hydrophobic interactions of FAD in the MD simulations and X-ray crystal structure (SI Table 1**)** are made with the flavin moiety. The adenine moiety of FAD has the potential of forming a cation-π interaction with the sidechain of R221. Most FAD binding hydrogen bonds are formed with the backbone carbonyl oxygen and nitrogen atoms of the surrounding residues (SI Table 2). Residues, A50, S347, T348, and I350 are key residues found to form stable hydrogen bonding interactions with the flavin ring moiety of FAD. The residues E346 and P344 interact with the hydroxyl group of FAD; this was added to more accurately depict the reactant complex after formation of hypochlorous acid so its interactions in our MD cannot be compared with the crystal structure. The interaction between hydroxy-FAD and E346 could have interesting implications for the mechanism of the reaction.

In the DCCA plot of the full complex wild-type PrnA (SI Figure 5), one the largest regions of positive correlation corresponds to two areas of protein from residues 205 to 255 and 305 to 350 (Figure 14). The area makes up a large part of the FAD binding site and contains many important FAD binding residues (SI Table 1 and SI Table 2). It also contains important residues from the tryptophan binding site such as E346 and S347. The intensity of the correlation is reduced in the DCCA plot of the apoenzyme PrnA (SI Figure 6) suggesting that the correlated motion relies on the protein's interactions with FAD.

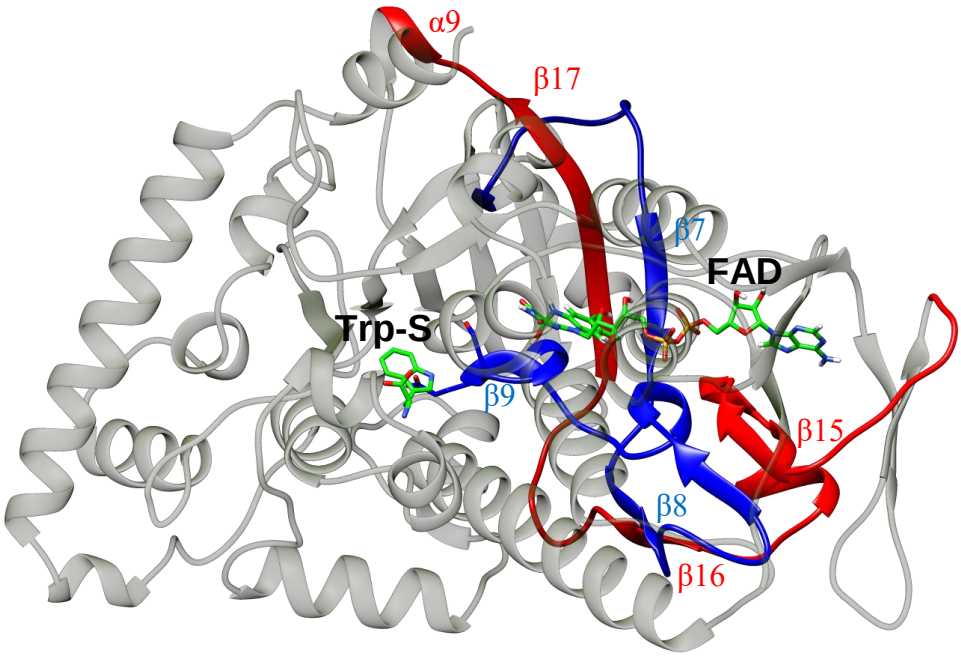

Figure 14: The two DCCA correlated regions described as spanning residues 205 to 255 coloured in red and the region spanning from residues 305 to 350 coloured in blue. The important residues E346 and S347 are displayed with blue tubes and the FAD and the substrate tryptophan are rendered with green coloured carbon tubes.

2.4.4 The possibility of direct contact between FAD and tryptophan binding site/module

The 1μs MD simulations show that close contact between FAD and tryptophan does not occur at this timescale, the distance between the FAD binding site and the substrate binding site remains relatively high during MD. The average distance between the Flavin and tryptophan binding sites was found to be 12.08Å, 13.45Å and 14.23Å respectively for PrnA full complex, apoenzyme PrnA and PyrH full complex MD simulations (SI Figure 11). The distance between the proposed site that chloride reacts with FAD, the FAD C4A atom, and the to-be halogenated carbon of tryptophan (CZ2/CZ3) remains prohibitively high throughout the MD simulations of PrnA and PyrH (Table 4). The sidechains of residues K79 and E346 remain distant from flavin at the 1μs timescale of the MD simulations (Table 4). Some hydrogen bonding is found between hydroxyl group of flavin and the backbone carbonyl oxygen of E346, this would not however allow for direct halogenation of the E346 or K79 sidechains (SI Table 1). These observations support the main catalytic mechanism in which the intermediary halogenating hypochlorous acid is created at the FAD binding sire and travels through a channel between the FAD and tryptophan binding sites [110].

	Average distance PrnA full complex (Å)	Average distance PyrH full complex
Flavin C4A – Substrate tryptophan CZ2/CZ3	11.6	11.0
Flavin C4X – Active Lysine NZ	7.0	6.9
Flavin C4X – Active Glutamate CD	8.3	10.8

Table 4: Average distances between the proposed reactive atoms of the tryptophan binding site: Lys-NZ, Glu-CD and Tryptophan CZ2/CZ3 and the proposed reactive atom of FAD-C4X in the MD simulations.

2.4.5 Effects of Mutations on Binding

Mutational studies show that both K79 and E346 in PrnA, and K75 and E354 in PyrH, play a vital role in the reaction of chlorination of tryptophan. The residues are conserved across the known FAD-dependent halogenases: PyrH [106] (trypthophan 5-halogenase), Thal [107] and SttH [108] (tryptophan 6-halogenases) and RebH [109] ,and PrnA [110] (trypthophan 7-halogenases), indicating their key roles in catalysis [137]. In PrnA the mutation of K79 leads to a complete loss of activity, and when E346 is mutated it is found that activity drops by orders of magnitude to a level where it is barely detectable[110]. Although possessing no formal charge hypochlorous acid has a strong dipole moment (oxygen -0.456 and hydrogen 0.445 calculated by the ATB [129]) so dipole interactions with K79 and E346 will have an influence on its position and orientation. In order to test the stabilising effect of the two charged residues and explain experimental effects of mutations we performed MD simulations on the *in silico* mutated forms of PrnA K79A and E346Q. In the absence of the

electronic environment created by both K79 and E346, hypochlorous acid moves away from tryptophan and back along the proposed channel towards FAD. In this position, hypochlorous acid is too distant from tryptophan and would be unable to carry out halogenation (SI Figure 12). In the MD simulation of K79A mutant the hypochlorous acid remains closer to the flavin ring and forms hydrogen bonds with the O4 atom of FAD. In the E346Q mutant form MD simulation, the hypochlorous acid moves away from the tryptophan binding site along the channel towards FAD where it forms a hydrogen bond with T263, which although technically close to K79 is separated by internal protein structure and not accessible for interaction. The simulations of the two mutant forms show hypochlorous acid Cl to Trp-CZ2 distances which indicate that both residues are in key importance for the positioning hypochlorous acid in proximity to tryptophan and ensuring that the movement of hypochlorous acid only occurs from FAD to tryptophan.

2.4.6 Comparison of PyrH to PrnA

With PyrH and PrnA being structurally similar enzyme carrying out similar reactions it is surprising that the enzymes have different kinetics, PyrH is the more efficient enzyme at chlorinating tryptophan, it converts 100% conversion to 5-chlorotryptophan, whereas under the same conditions PrnA converts only 59% of tryptophan to 7-chlorotryptophan [105]. The binding mode of FAD in both enzymes is almost identical (Figure 3 and SI Figure 9) so the differences in their regioselectivities and kinetics probably has a structural basis found in the tryptophan binding domain. The way the two enzymes bind the substrate tryptophan in the active site is quite different, the substrate tryptophan in PyrH is oriented in an upside down position relative to its binding orientation in PrnA. The benzene moieties of the indole rings for both PrnA and PyrH are near superimposable in the crystal structure (Figure 3), the slight rotational difference means that the C5 atom in PyrH is in an almost identical place to that of the C7 atom in PrnA. During the MD simulation of PyrH we observe a rotation of the tryptophan substrate to a slightly different orientation, after equilibration this orientation

remains relatively stable (SI Figure 13). Despite this movement the orientation of the C5 atom of the indole ring of tryptophan in relation to K75 and E354 is preserved. The position of hypochlorous acid in the PyrH MD simulation also remains more stable relative to the PrnA MD, this is evidenced by the lower levels of fluctuation in the RMSD plot of hypochlorous acid in SI Figure 14.

We believe that the higher efficiency of PyrH may correlate with the reduced flexibility of the PyrH tryptophan binding site, with the more rigid binding site making it an overall more efficient enzyme for halogenating tryptophan [105]. The averaged distances from the MD trajectories between K79/K75-NZ, hypochlorous acid Cl and Trp (CZ2 in PrnA and CZ3 in PyrH), clustered and shown in SI Figure 15 and SI Figure 16. K79/K75-NZ to hypochlorous acid-Cl and tryptophan CZ2/CZ3 to hypochlorous acid distances in PyrH are shorter indicating a more compact binding site. We propose that our MD simulations of PyrH showing hypochlorous acid making more stable interactions with relatively lower levels of fluctuation between the active lysine and glutamate residues. The stability of these interactions could mean that a more energetically favourable version of the reaction mechanism is taking place in PyrH as opposed to PrnA, this could be one of the contributing factors to the experimentally observed greater catalytic turnover of PyrH compared to PrnA.

The average RMSD of PyrH was 2.8Å, this is significantly lower than the PrnA full complex and PrnA apoenzyme simulations and means that PyrH is less flexible than PrnA (Figure 6). In the RMSF plot comparing the PyrH full complex and PrnA full complex MD simulations (SI Figure 17) the region around G37, which immediately precedes the FAD binding strap that runs from residues 37-50, we see a similar feature seen in the RMSF plot of PrnA. However, in PyrH this region shows lower levels of flexibility. However, In PyrH both the FAD strap region (SI Figure 18) and FAD show more conformational fluctuations than that of PrnA (SI Figure 19). Due to the differences in tryptophan binding between PyrH

and PrnA we find no analogous interaction between S50 and E450 that is found in PrnA. Instead a direct hydrogen bonding interaction between the sidechain of S50 and the carboxylate moiety of tryptophan is found (SI Table 3). This implies that despite the differences in tryptophan binding between PyrH and PrnA, the role of the "strap" region between the two enzymes remains the same. The function of providing a direct means of communication between FAD and the tryptophan binding site is conserved even though the mechanism in which it does this is not identical to that of PrnA.

Another important structural difference hypothesised by us to be of key importance to tryptophan binding, in PrnA is the flexible loop region which spans residues 147-159. In this region there is a hydrogen bond between the sidechain of S157 and the carboxylate moiety of tryptophan seen to be affected by the dynamics of the 147-159 loop which we propose is important for tryptophan binding (SI Figure 4). In PyrH there is a similarly positioned flexible loop which had to be modelled due to its lack of coordinates in the crystal structure. This loop in PyrH is similar to the loop in PrnA in that G153 acts as a hinge residue with several internal hydrogen bonding interactions forming within the loop during the MD simulation such as between: T156 and D149, S151 and D149, and R154 and E150. This loop could play a similar role in PyrH as the equivalent loop 157-59 does in PrnA, acting as a structural link between the FAD binding strap and the substrate tryptophan. However, the residue S157 from PrnA is not found in PyrH, instead F164 is found in a similar spatial position, probably fulfilling a similar role in hydrogen bonding the substrate tryptophan. After equilibration, Phe164 forms a stable hydrogen bonding interactions with the substrate tryptophan amino and carboxylate moieties (Figure 15).

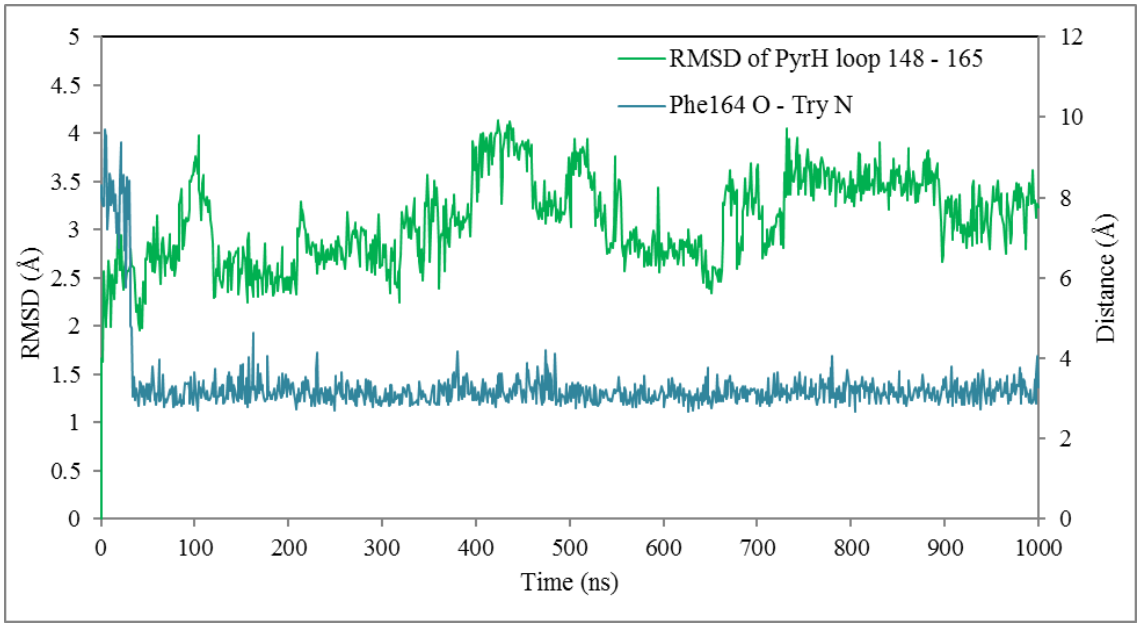

Figure 15: A plot showing the relationship between the RMSD of the flexible loop region spanning residues 148-165 in the PyrH simulation and the hydrogen bonding interaction between F164 and the substrate tryptophan.

When comparing the DCCA plot of the PyrH full complex (SI Figure 20) with of the PrnA Full complex (SI Figure 5) we see a lot of similar correlations in similar regions indicating that the two enzymes share a lot of correlated motions.

In general, most of the binding interactions of the substrate tryptophan and the cofactor FAD in the PrnA have analogous interactions in PyrH. For example tryptophan is bound in a similar way between hydrophobic sidechains making π-π stacking interactions with their indole rings (SI Table 4). In PrnA the residue W455 is replaced by the similarly positioned residue F451 in PyrH, although the distance is greater (average distance 6.95Å) making the interaction weaker and less significant to tryptophan binding. F49 in PyrH occupies a similar position to F454 in PrnA, with an average distance of 5.34Å to the substrate tryptophan (SI Table 4). H101 and F103 from PrnA are conserved in PyrH as H92 and F94 and fulfil a

similar role as hydrophobic residues in proximity to the substrate tryptophan (SI Table 4). A comparison of the tryptophan binding sites of PrnA and PyrH is shown in Figure 16.

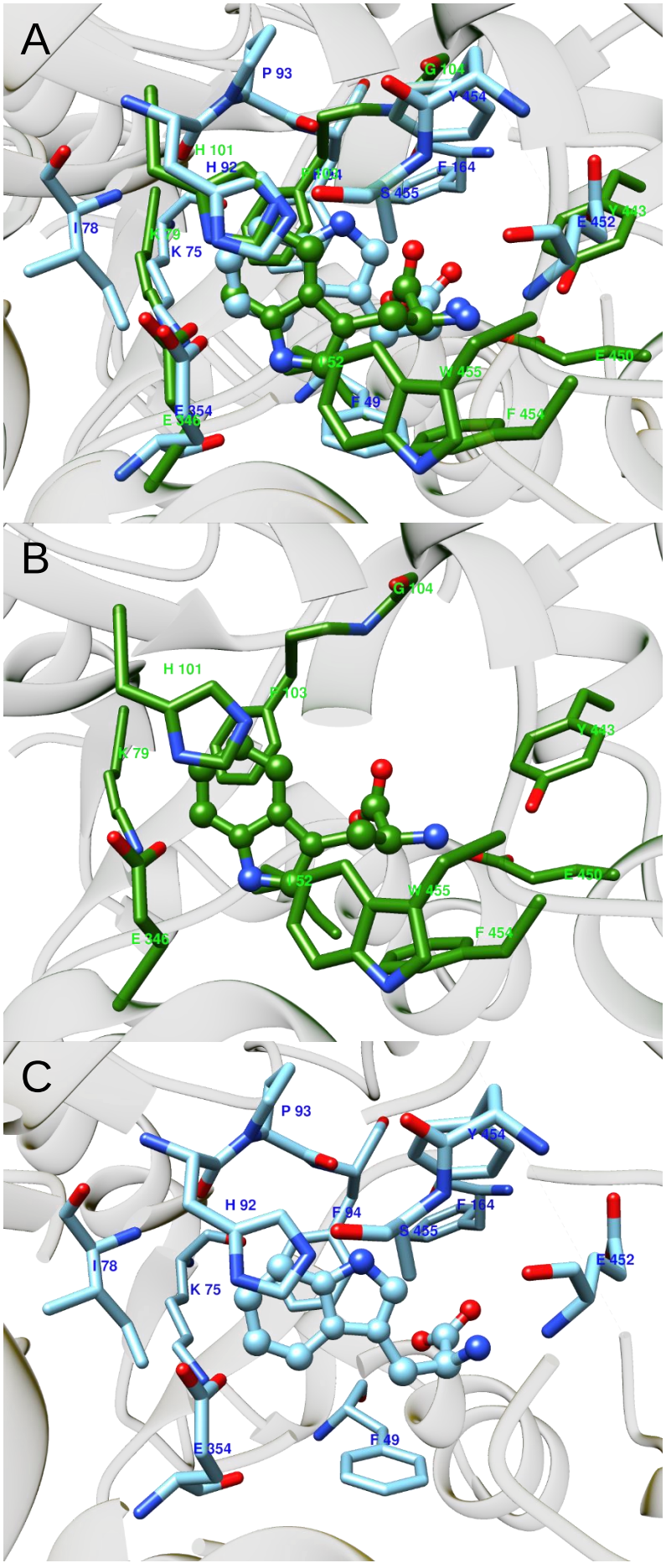
A
P 93
G 104
Y 454
H 101
H 92
F 164
S 455
I 78
K 75
E 452
E 450
W 455
F 49
F 454
E 354
B
G 104
H 101
K 79
E 346
C
P 93
Y 454
H 92
F 164
S 455
I 78
K 75
E 452
E 354
F 49

Figure 16:A - The overlaid tryptophan binding sites of PrnA and PyrH. B - The tryptophan binding site of PrnA. C - The tryptophan binding site of PyrH. In all parts of the figure the substrate tryptophan is rendered with balls and sticks whereas the interacting residues are rendered as tubes.

In PyrH the amino group of the substrate tryptophan interacts electrostatically with the sidechain of E452 (distance of 3.92Å) (SI Table 5). In the crystal structure of PyrH a similar electrostatic interaction supports E452 and appears to be created by R96, however the average distance of this interaction during the MD simulation is much greater and therefore weaker than the analogous interaction between K57 and E450 observed in the MD simulation of PrnA. In PyrH the important catalytic residue E354 interacts with the protonated H40 with a similar distance to that of H395 to E346 seen in PrnA (SI Table 5). Hypochlorous acid forms strong interactions with both the active site K75 and E354, low average distances indicate that these interactions are stronger and more stable than in PrnA (SI Figure 15 and SI Figure 16). This may be another contributing factor to the increased efficiency of PyrH as an enzyme.

The greatest differences in tryptophan binding between PyrH and PrnA are found in the hydrogen bonding of the substrate tryptophan. The different orientations of the substrate between PrnA and PyrH lead to very different hydrogen bonding. In PyrH the relative upside down positioning of tryptophan means that the tryptophan NE1 atom no longer points towards hypochlorous acid channel, this means it is more able make hydrogen bonds with the surface residues of the tryptophan binding pocket (Figure 3). In PyrH S345 forms a hydrogen bond with the substrate tryptophan NE1 atom, S50 is the main residue responsible for the hydrogen bonding of the substrate tryptophan carboxylate moiety, and Phe164 forms hydrogen bonds with both the amino and carboxylate moiety of tryptophan (SI Table 3).

Similarly, to the PrnA MD simulation we find that FAD undergoes a structural transition from a linear to a bent form (SI Figure 21). The transition occurs much more rapidly in PyrH than PrnA, this is seen in the loss of several hydrogen bonds in the crystal structure compared to the MD (SI Table 5). The binding mode of FAD in PyrH shows a high level of similarity to that of PrnA, in that it is mainly bound by backbone hydrogen bonds. Some of the Van

der Waals interactions between hydrophobic sidechain residues and the adenine and flavin moieties are conserved by similar residues to those in PrnA in both the crystal structure and MD simulations (SI Table 6).

2.5 CONCLUSIONS

Through the creation and analysis of several MD simulations we have gained insight into the structure function relationships of the two halogenases PrnA and PyrH and the origin of their selectivity. The study of the tryptophan binding site in our MD simulations have identified the key residues for the positioning of tryptophan in the active site, the specific positioning of tryptophan is likely the key factor in the regioselectivity of the enzymatic reaction. By comparing the crystal structure to that gained from the MD simulation, we can see that important structural differences created during the crystallisation process. The MD simulations have identified several dynamic regions that have implications for substrate binding and the proposed function of the "strap" in the mechanics of the enzyme. We have provided suggestions for the structural basis of the previously proposed line of communication that links the tryptophan and FAD binding sites. The study of the two mutant forms of PrnA has reaffirmed the results experimental mutagenesis studies and helped to understand the likely structural basis for reduced activities observed for these two mutant forms of the enzyme. Comparison of the PyrH and PrnA has led to the suggestions of why PyrH is the more efficient halogenating enzyme. The atomistic insights gained from the simulations provide important structural information that can be related to enzyme function.

3 COMPUTATIONAL INSIGHTS IN THE REACTION MECHANISM OF THE 2-OXOGLUTARATE DEPENDENT OXYGENASE - THE FAT MASS AND OBESITY-ASSOCIATED PROTEIN (FTO)

3.1 PREFACE

This chapter represents an ongoing computational investigation into the reaction mechanism of the fat mass and obesity associated protein (FTO), utilising QM and QM/MM methodology, to specifically investigate the electronic and geometric changes that occur during the formation of the Fe(IV)-oxo intermediate. The continued direction of the research is highly dependent on planned calculations and the scope of the project will be expanded as more results become available for analysis. The results and discussion sections of this chapter represent the interpretation of preliminary findings. Suggestions for how the project will be developed are described in the future work section and when completed the research detailed in this chapter will lead to one or more publications.

3.2 INTRODUCTION

The fat mass and obesity-associated protein (FTO) is an Fe (II) 2-oxoglutarate (2OG) dependent oxygenase found in humans that has several variants that are strongly correlated with obesity susceptibility in humans [138]. The FTO gene variants that are linked to obesity are thought to exert their body mass increasing effects by increasing overall appetite, increasing the intake of high fat and protein foods and reducing satiety [139]. As well as the increased risk of obesity in both children and adults [140], FTO malfunction has been implicated as a risk factor for certain forms of cancer [141-144]. However, the problem is twofold because the risk of cancer is greatly increased in overweight individuals, it has been shown that there is a link between the state of altered energy metabolism regulation through FTO activity and the formation of cancerous cells in humans [145]. The effects of FTO function are long reaching with certain variants of FTO also being implicated as risk factors for Alzheimer's disease [146] and type 2 diabetes [147]. In healthy individuals, FTO functions as a demethylase enzyme, catalysing the removal of methyl groups from methylated DNA (3Me-T) and RNA (6Me-Ad) bases (a simplified mechanism of this reaction is shown in Figure 17) to fulfil roles in DNA/RNA repair and epigenetic control. In the case of DNA/RNA repair, bases that have become damaged through external toxins can be demethylated by FTO to return the base to its natural demethylated state [148]. In the case of epigenetic control, bases are modified by the addition and removal of methyl groups to control whether a gene is expressed or not, FTO is expressed to remove these methyl groups to allow the gene to be expressed [149]. FTO is also part of a family of human AlkB analogue enzymes, including ALKBH1-ALKBH8. AlkB was first identified in *E. coli*, where it is expressed as a response to alkylating damage of DNA for the purpose of DNA repair.

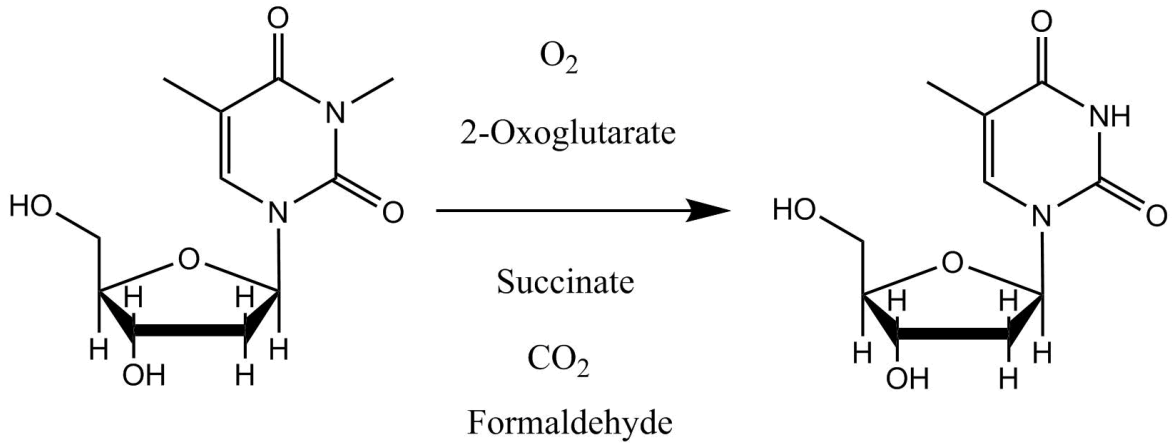

Figure 17: A simplified mechanism for the demethylation of the methylated base 3-MeT by FTO. A more detailed mechanism is shown in Figure 18.

Belonging to the family of non-heme iron 2OG dependent oxygenases, these enzymes all utilise a conserved non-heme iron centre. In the resting state iron (II) is coordinated by a conserved 2 Histidine and 1 Aspartate/Glutamate motif (aka the HXD/E...H triad) and 3 water molecules. 2OG binds in a bidentate fashion at the equatorial position and 2 water molecules are displaced. Upon substrate binding the remaining bound water molecule dissociates from iron, this creates the 5-coordinate complex with the free axial position for oxygen binding. Iron binds the dioxygen molecule and donates an electron to form the iron (III) super-oxo intermediate. The distal atom of the bound dioxygen attacks the C2 atom of 2OG to produce a peroxo bridged bicyclic intermediate. The intermediate complex collapses and the O-O bond breaks, the C1 atom of 2OG leaves as carbon dioxide and the remaining 4-carbon molecule (succinate) is monodentately bound by its newly formed terminal carboxylate group. In this state iron (IV) is a 5-coordinate complex with a single oxygen atom bound in the axial position. It has been shown by QM/MM study of AlkB that at this stage oxo-rotation occurs to bring the axial oxygen to an equatorial position, this is necessary to bring the oxo group into proximity of the methylated base [150]. The highly reactive iron

(IV) oxo then abstracts a hydrogen atom from the methyl group of the methylated base substrate. The methyl group radical of the substrate reacts with the iron (III) bound hydroxyl group to produce a hydroxylated substrate. The hydroxylated substrate collapses with the target methyl group leaving as formaldehyde, in this way the base is left demethylated, 3 water molecules displace succinate and the catalytic iron (II) centre is ready for further reaction [151]. The proposed generic catalytic cycle for this reaction is shown in Figure 18.

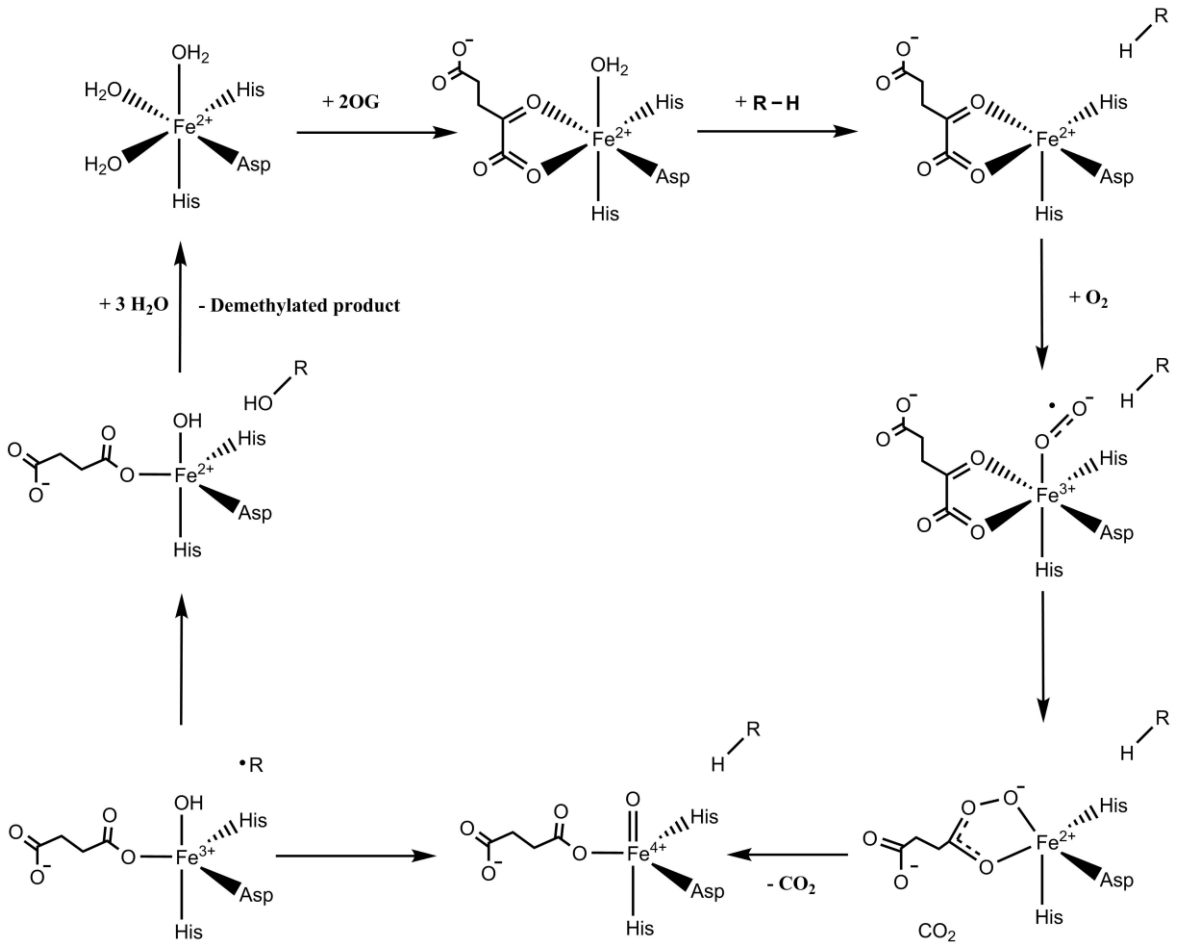

Figure 18: The conserved mechanism for non-heme iron 2OG dependent oxygenases.

FTO is composed of two distinct domains, the N-Terminal Domain (NTD) shares a lot of structural similarity with other Fe (II) 2OG-oxygenase enzymes, particularly AlkB and ALKBH1-ALKBH8, through a conserved "jelly-roll" motif. This arrangement of eight beta strands split into two anti-parallel sheets forms a distorted barrel shape which is open at one end [152]. The NTD contains the catalytic site; iron, 2OG as well as room to accommodate the methylated base and oxygen. Despite sharing this conserved structural feature, Fe (II) 2OG-oxygenase enzymes are found throughout nature where their oxidating abilities are employed in a diverse range of roles such as; the synthesis of secondary metabolites, lipid metabolism and collagen biosynthesis [153, 154]. With such diverse substrates and conserved active sites, the enzymes mostly differ in how they recognise and bind their substrates, in FTO the CTD is thought to play a key role in this process and does not show

significant homology with any other protein. Even though the CTD has no clear catalytic role, the isolated NTD was found to be inactive, this supports the idea that the CTD has a necessary but indirect role in catalysis [152]. Initial *In-vitro* experiments of FTO showed that it had a preference towards 3-methylthymidine (3-MeT) in single stranded DNA or 3-methyluradine in single stranded RNA with no reaction to double stranded DNA or RNA [140, 155]. However, more recently *in-vivo* experiments have shown that FTO more efficiently demethylates 6-methyladenosine, a methylated base that is mostly found in transfer and messenger RNA [156]. When preparing FTO for crystallisation the protein is incubated with the free base form of the methylated base, allowing the observation of the active site interactions of the methylated base substrates without the complications of crystallising the DNA/RNA bound complex [152]. However currently no crystal structure of FTO in complex with 6-methyladenosine exists, the publication of such a structure would aid in the explanation of the substrate preferences and specificity of FTO [151]. The 2OG cosubstrate analogue, N-oxylglycine (NOG), is also commonly used as it acts as an inhibitor allowing the enzyme-substrate complex to be more easily crystallised and studied [157].

FTO is a desirable drug target for the treatment and prevention of obesity, diabetes and cancer. Inhibitors that emulate 2-oxoglutarate are not effective due to the large number or other Fe (II) 2-oxoglutarate dependent enzymes in human body, with over 80 currently identified examples specific inhibition of FTO would be difficult [158]. Molecules that instead emulate the binding mode of the substrate should be the preferred inhibition strategy, each substrate is highly specific to only a small number of enzymes, therefore its corresponding inhibitor would have reduced off-target effects [159]. A proven method for discovering inhibitor molecules involves emulating the structure of transition states [160]. Transition states by their very nature cannot be directly observed by experimental means due to their instability leading to incredibly short half-lives [161]. The structures of transition

states can be determined by QM and QM/MM methodology, by modelling the enzymes reaction mechanism we can characterise the structure of transition states [160, 162]. These could be used in the design of novel inhibitor molecules for the treatment of diseases associated with the malfunction of FTO.

Although a QM/MM study of FTO was published in 2015 [163], this study primarily focused on the reaction path of the hydroxylation of the 3Me-T and 6Me-Ad substrates and their subsequent hydrolysis to their demethylated forms. Our study will focus on the part of the reaction mechanism concerned with the formation of the Fe(IV)-oxo species from the Fe(III)-superoxo complex, this reaction step has been the focus of other studies of other non-heme iron 2OG dependent oxygenases but not of FTO [148, 150, 164, 165]. In addition to modelling the reaction path we will attempt to investigate what effects the choice of methodology and cluster size has on the calculated structures. The results of these calculations on the different QM cluster structures will be used to assess the individual contributions that each key residues has on the important molecular orbitals, geometric features, spin densities and charges of the enzyme substrate complex of FTO. In addition the reaction path for the formation of the Fe(IV)-oxo species will be modelled with different cluster sizes, functionals, and basis sets to assess the effect these have on the resultant transition state and intermediate structure geometries, the effects on the reaction barrier and reaction energy will be investigated. The QM cluster results will be compared to those of the QM/MM model, as well as existing published results of other computational studies of FTO and similar non-heme iron containing 2OG dependent enzymes [163, 164].

3.3 METHODS

In order to study the reaction mechanism of FTO, several different cluster models in varying size as well as a QM/MM model were created. To create the desired reactant complex the crystal structures of 3LFM [152] and 4IDZ [157] were combined using chimera [131] to overlay the two protein structures. The 3LFM structure contains the methylated base 3MeT but has an unusual distorted active site, 4IDZ has the correct active site geometry but doesn't contain 3MeT so its coordinates are copied from the 3LFM structure. NOG was also modified to produce 2OG to represent the natural enzyme substrate complex. For the QM and QM/MM iron oxygen bound model the axial bound water molecule was replaced with dioxygen bound in the end on bent conformation with geometry that is consistent with experimental spectroscopic studies of model compounds [165]. Hydrogen atoms were added to the structure and the titratable amino acid protonation states were determined with the PROPKA web server [166], solvent molecules were added using the leap module for Amber14 [25]. The reactant complex was modelled with high spin iron (S=2), oxygen is antiferromagnetically coupled so has no net influence of on the quintet spin state of the overall system, this is based on past experimental and computational studies [163-165].

The QM cluster models were created by modifying the full enzyme structure in GaussView [167] and the calculations were performed using Gaussian09 [94]. In the X-ray crystallised structures of FTO 2OG/NOG is found to adopt an extended conformation, this is supported by other structural studies of Fe (II) 2OG dependent oxygenases [168, 169]. In the full enzyme structure the C5 carboxylate of 2OG interacts electrostatically with R316. To preserve this extended linear conformation of 2OG in the QM cluster models, we first employed a distance restraint to the C5 atom of 2OG, however this posed problems when scanning for the formation of succinate as the newly formed succinate was not able to move to re-bind to iron. To overcome this a dihedral restraint was used to freeze the dihedral angle between the C2 and C3 bond, this maintained the linear extended conformation whilst giving

more realistic results for the binding of succinate following the scan to form Fe (IV)-oxo. The iron coordinating histidine and aspartate residues were truncated at their β-carbon atoms and a hydrogen atom was added to fill the open valency, distance restraints between these β-carbon atoms were added to preserve the structure of the facial triad. Different cluster sizes incorporated additional amino acids to study the effect of cluster size on the calculations, these were treated in the same method as the iron coordinating amino acids. Different optimisations of the clusters were performed with the BP86 functional [170, 171] with 10% HF exchange. To simulate the hydrophobicity of the enzyme active site the implicit solvent model conductor-like polarisable continuum model was used, with a dielectric constant of ε=4.3 and diethyl ether solvent molecule was used[172], this method was applied to all the clusters. Frequency calculations of the optimised structures were performed to check for negative frequencies which might indicate an unnatural structure. The systems were treated with two different basis sets 6-311G* for iron and its immediate coordinating atoms and 6-31G* for the other atoms in the system. Potential energy scans were performed to simulate the reaction mechanism step for the formation of the Fe (IV)-oxo intermediate from the Fe (III)-superoxo complex. Multiple scans with varying conditions were made to assess the effect of cluster size and functional on the potential energy landscape.

The QM/MM model of FTO was created from the aforementioned FTO model, the Amber parameters for 2OG and 3MeT were created using the General Amber Force Field (GAFF)[26] in AntechAmber for Amber 14 [25]. Their atomic charges were amended using the Restrained Electrostatic Potential (RESP) method for charge fitting following a single point calculation with the HF method and a 6-31G* basis set using Gaussian09 [94]. The parameters for iron and its coordinating residues were created using the Metal Centre Parameter Builder (MCPB) tool, by designating the metal coordinate bonds as being covalent. The force constants for bond stretching and bending were derived using the Seminario method and charges were assigned using the charge model B method in MCPB.

The system setup in GaussView [167] and was truncated at 20Å with all atoms outside the 20Å limit being frozen, atoms within 20Å were allowed to move freely during optimisation. The whole system was initially optimised at the MM level using Amber14 for 10,000 steps using a combination of the steepest descent and conjugate gradient algorithms, during this process the active site residues were restrained with a potential of 100 kcal mol^{-1} $Å^2$ to preserve active site geometry. The system was split into two layers according to the ONIOM method [173], the QM region was defined as being iron and its bonded histidine and glutamate residues (cut-off at the β-carbon and link atoms used to saturate the open valency) as well as 2OG, 3Me-T and oxygen, the remainder of the system was treated with an MM level of theory. The optimisation was performed with ONIOM[173] in Gaussian09[94], with BP86 10%HF exchange/Amber method [25, 170, 171] and two basis sets, 6-311G* for iron and its coordinating atoms and 6-31G* for the rest of the QM layer. All images of molecular models for this chapter were created using GaussView [167].

3.4 RESULTS

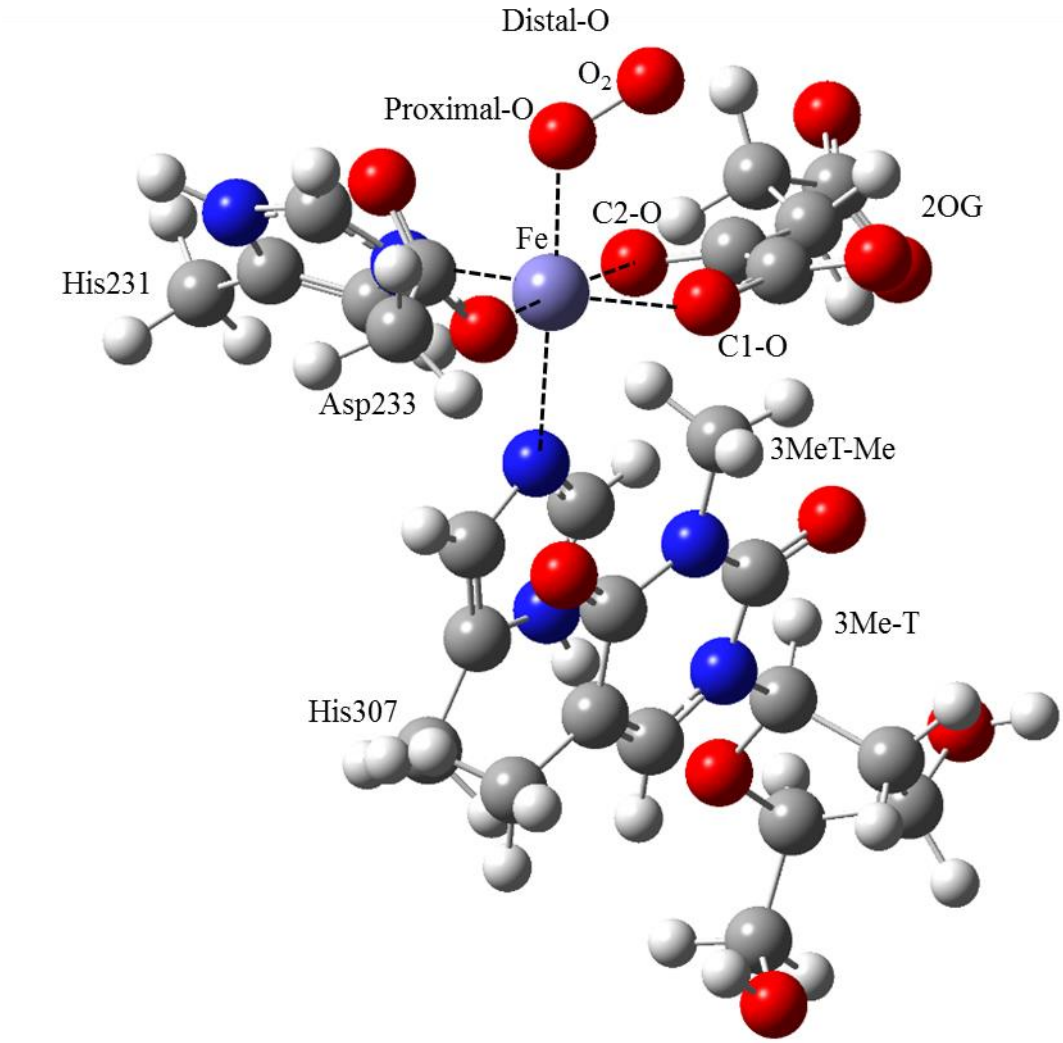

Figure 19: The optimised QM cluster structure of FTO rendered in ball and sticks rendering with the iron binding ligands labelled. Atom colours are as follows: carbon-grey, hydrogen-white, oxygen-red, nitrogen-blue and iron-purple. S=2, Charge=-1

The optimised geometry of the standard sized Fe(III)-superoxo complex QM cluster model is shown in Figure 19. Reduced versions of this standard cluster size are shown in: Figure 20, where 2OG is cut at the C3 atom to create the "cut 2OG" model; Figure 21, full 2OG without 3Me-T. The QM/MM optimised oxygen bound Fe(III)-superoxo complex is shown in Figure 32 and Figure 33, the measured distances from the QM/MM optimisation are compared against those from the initial pdb structure as well as the optimised structures of the different QM models in Table 5.

In addition to these reduced versions of the standard sized QM cluster model, several larger clusters were created. These are shown in Figure 22, the standard cluster with the sidechain of R316 added; Figure 23, the standard cluster with the sidechain of R322 added; Figure 24, the standard cluster with the sidechains of R322 and R316 added; Figure 25, the standard cluster with the sidechains of R96, R316 and R322 added. These complexes were all optimised from the pdb model created by combining the crystal structures of 3LFM and 4IDZ (exact procedure described above in the methods section). The measured distances for the extended cluster structures can be found in Table 6. All the above described structures were optimised with the BP86 functional with 10% HF exchange, the 6-311G* basis set was used for iron, O_2 and the iron coordinating atoms, the rest of the atoms in the system used the 6-31G* basis set.

Potential Energy Surface (PES) scans using different reaction coordinates were created from the optimised oxygen bound clusters, the PES scans used the reaction coordinate the C1-C2 bond of 2OG. The PES scans were performed by increasing the C1-C2 bond of 2OG by 0.05Å for ten steps, for the standard QM cluster (O_2 bound, full 2OG, with 3Me-T), the full 2OG without 3Me-T cluster, and the cut 2OG with 3Me-T cluster. All scans were performed using the same BP86/10% HF exchange functional with the basis sets applied to the atoms being the same as for the geometry optimisation calculations.

The PES scan plots for the three clusters are shown in Figure 26, the measured distances for the global maximum (potential TS) and global minimum structures can be found in Table 7. Additionally the final structures from the PES scans were optimised with CO_2 removed and the measured distances can also be found in Table 6.

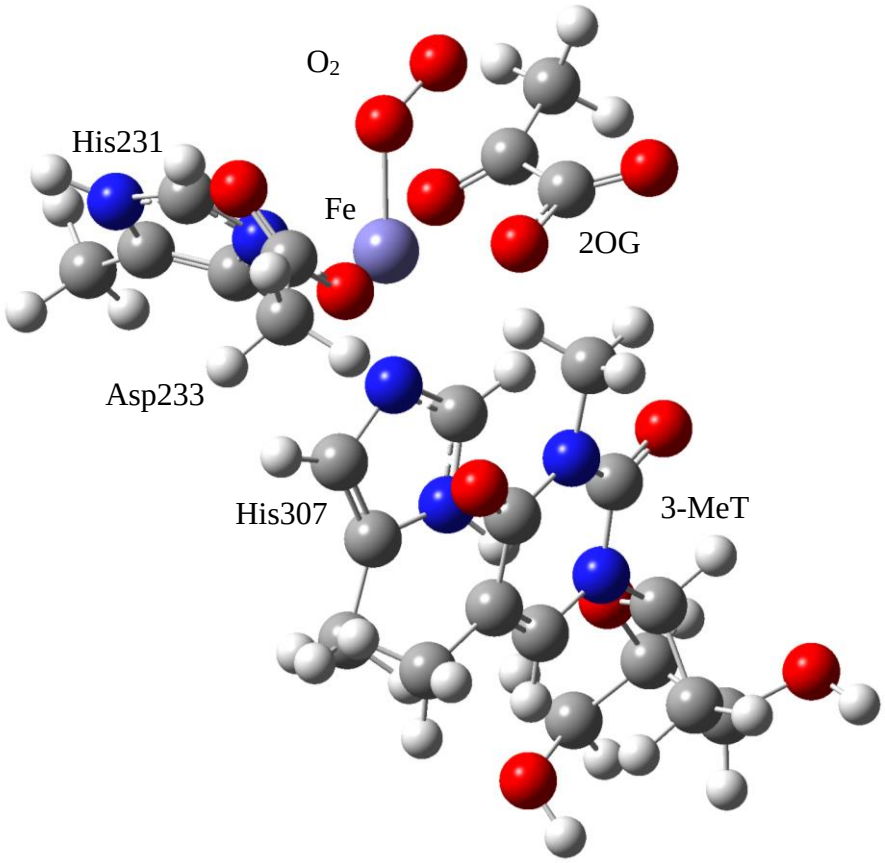

Figure 20: The optimised structure of the cut 2OG with 3Me-T QM cluster model. S=2, Charge=0

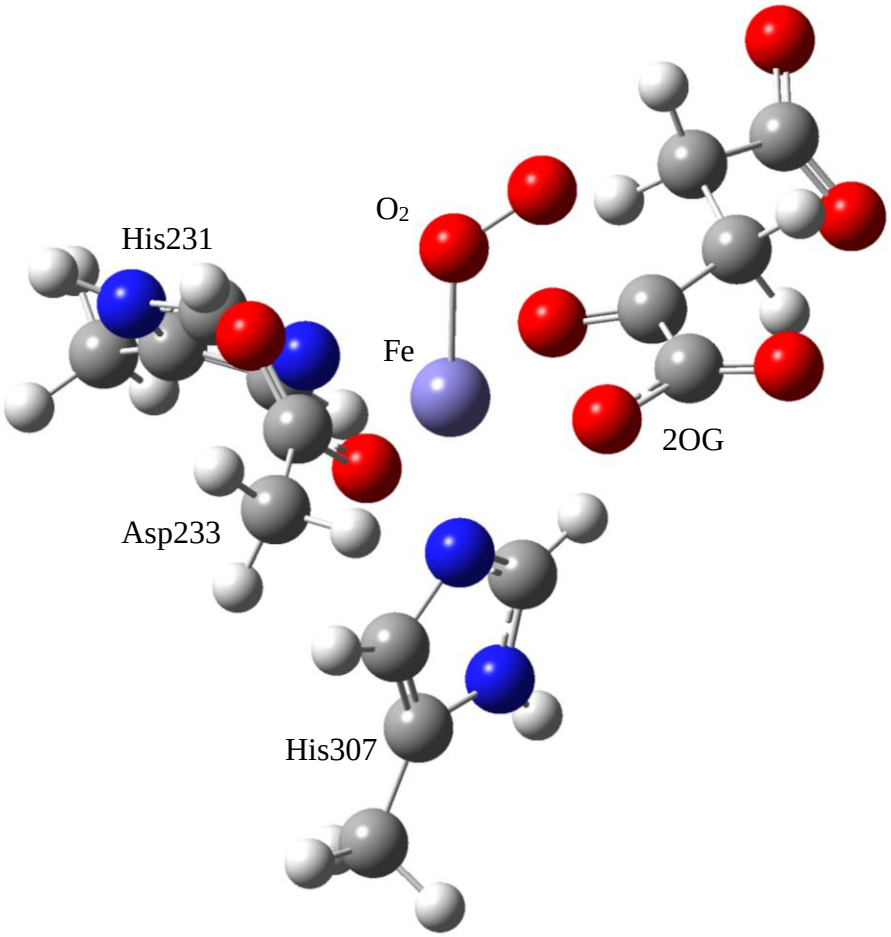

Figure 21: The optimised structure of the cut 2OG without 3Me-T QM cluster model.

S=2, Charge=-1

	H307	H231	D233	2OG (C1-O)	2OG (C2-O)	Fe-O (prox.)	O(prox.)-O(dist.)	2OG C2-O(dist.)	3Me-T(CH3)-Fe
PDB structure (4IDZ/3LFM)	2.02	2.04	2.07	2.08	2.07	n/a	n/a	n/a	4.28
Standard QM cluster	2.20	2.10	1.96	1.96	2.22	1.93	1.28	3.12	4.23
Full 2OG no 3DT optimised structure	2.15	2.13	1.96	1.99	2.20	1.92	1.28	2.93	n/a
Cut 2OG with 3DT optimised structure	2.04	1.99	1.95	1.93	2.32	1.88	1.29	2.43	4.20
QM/MM optimised structure	2.13	2.13	2.09	2.03	2.43	2.21	1.22	2.56	4.30

Table 5: Measured distances comparing the geometry of the initial pdb structure, optimised QM/MM structure and optimised QM cluster model structures.

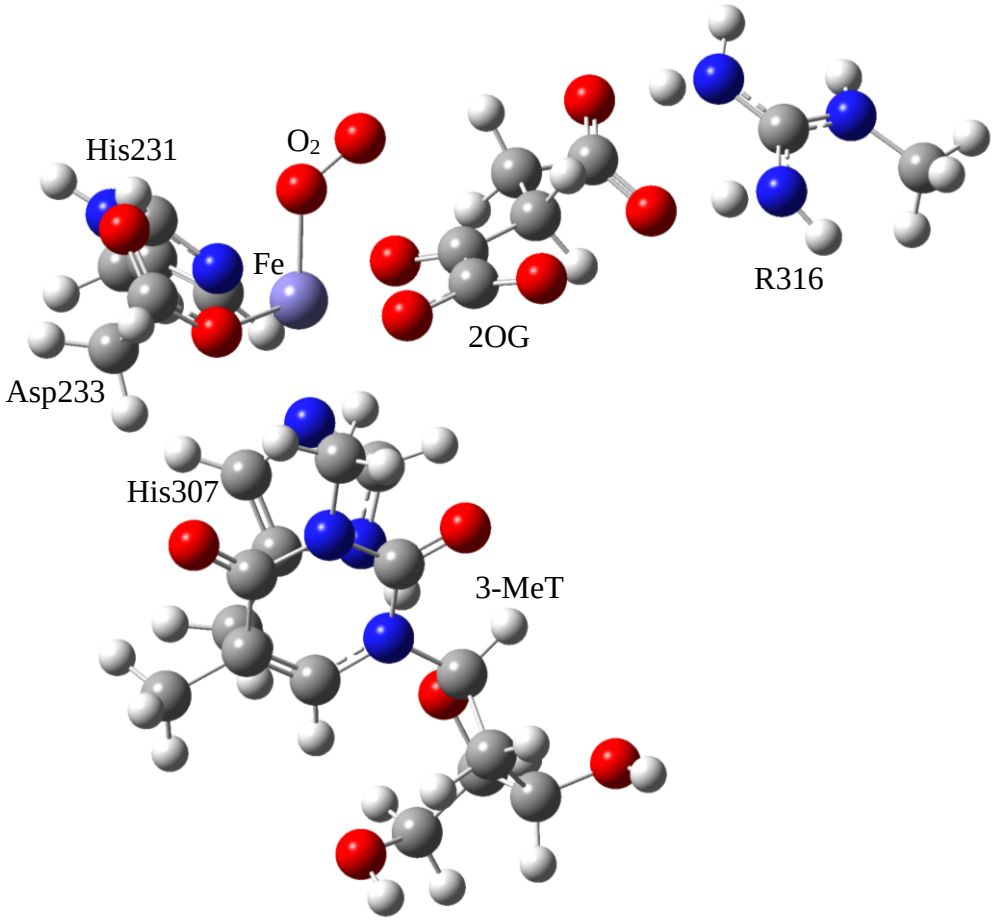

Figure 22: The extended QM cluster including the sidechain of R316. S=2, Charge=0

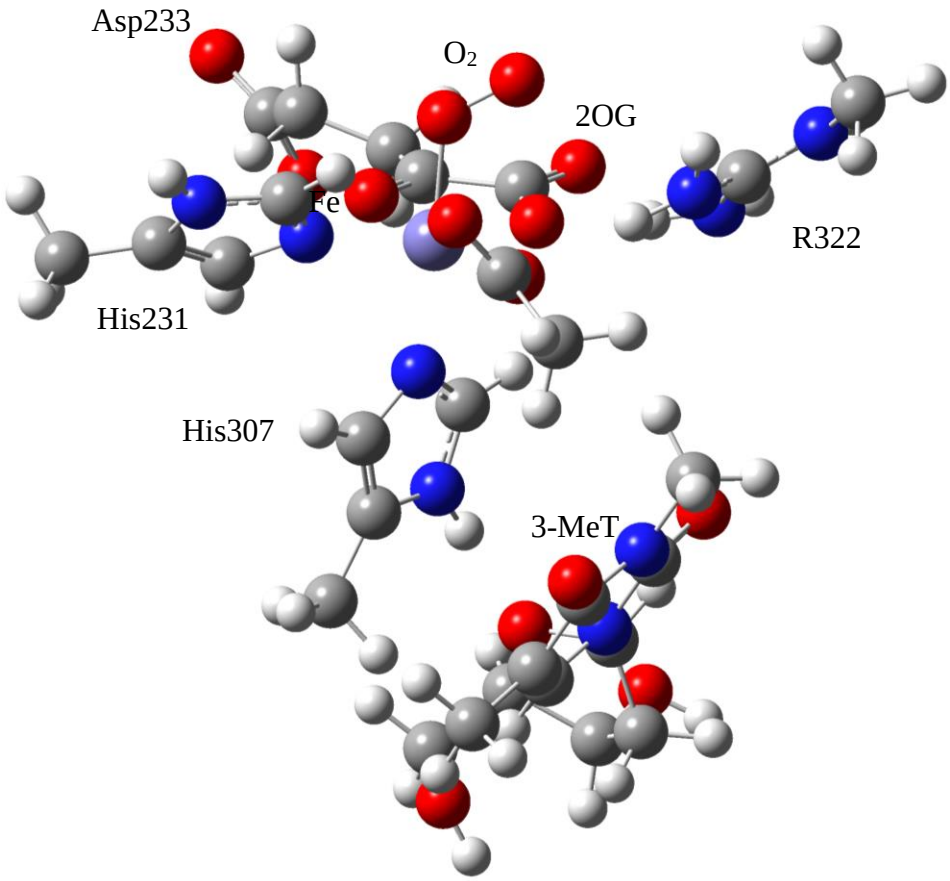

Figure 23: The extended QM cluster including the sidechain of R322. S=2, Charge=0

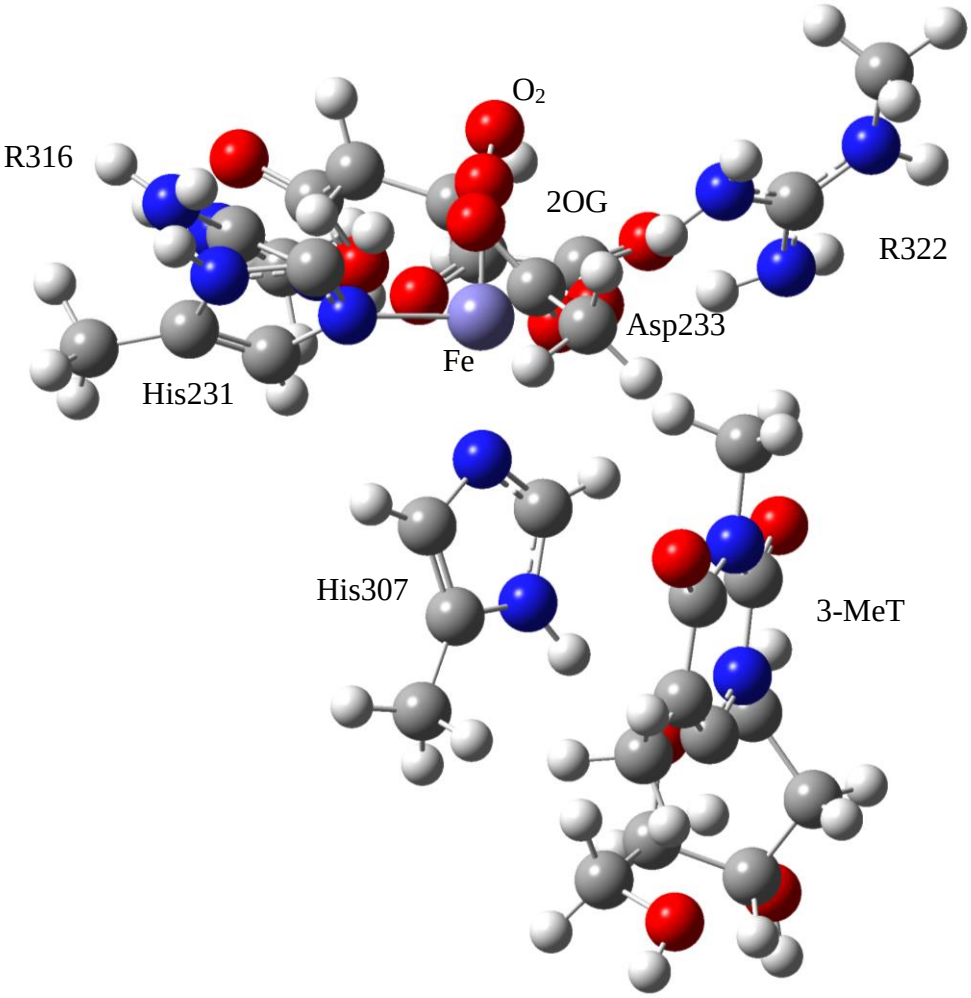

Figure 24: The extended QM cluster including the sidechains of R316 and R322. S=2, Charge=+1

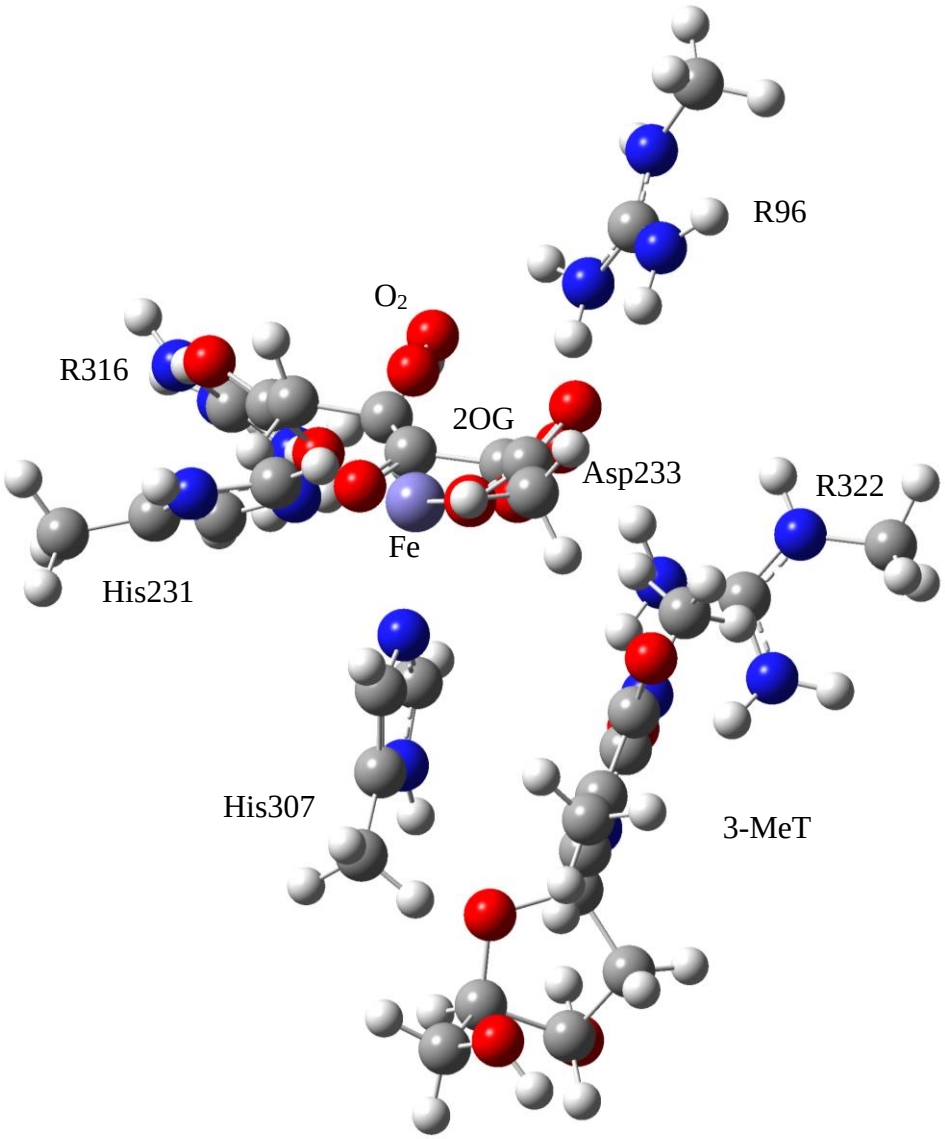

Figure 25: The extended QM cluster including the sidechains of R96, R316 and R322.

S=2, Charge=+2

	H307	H231	D233	2OG (C1-O)	2OG (C2-O)	2OG C1-C2	Fe-O (prox.)	O(prox.)-O(dist.)	2OG C2-O(dist.)	3Me-T(CH3)-Fe
Standard QM cluster	2.20	2.10	1.96	1.96	2.22	1.48	1.93	1.28	3.12	4.23
R316 extended region	2.14	2.09	1.93	2.02	2.27	1.56	1.91	1.28	2.86	4.43
R322 extended region	2.13	2.10	2.00	2.06	2.19	1.55	1.92	1.27	4.10	5.48
R322, R316 extended region	2.02	1.97	2.10	1.98	2.18	1.53	1.89	1.30	2.57	4.34
R96, R316, R322 extended region	2.11	2.09	1.92	2.02	2.13	1.54	2.09	1.29	2.32	4.63
QM/MM optimised structure	2.13	2.13	2.09	2.03	2.43	1.55	2.21	1.22	2.56	4.30

Table 6: The measured distances for the extended QM cluster models compared to the standard QM cluster, all distances are displayed in Å.

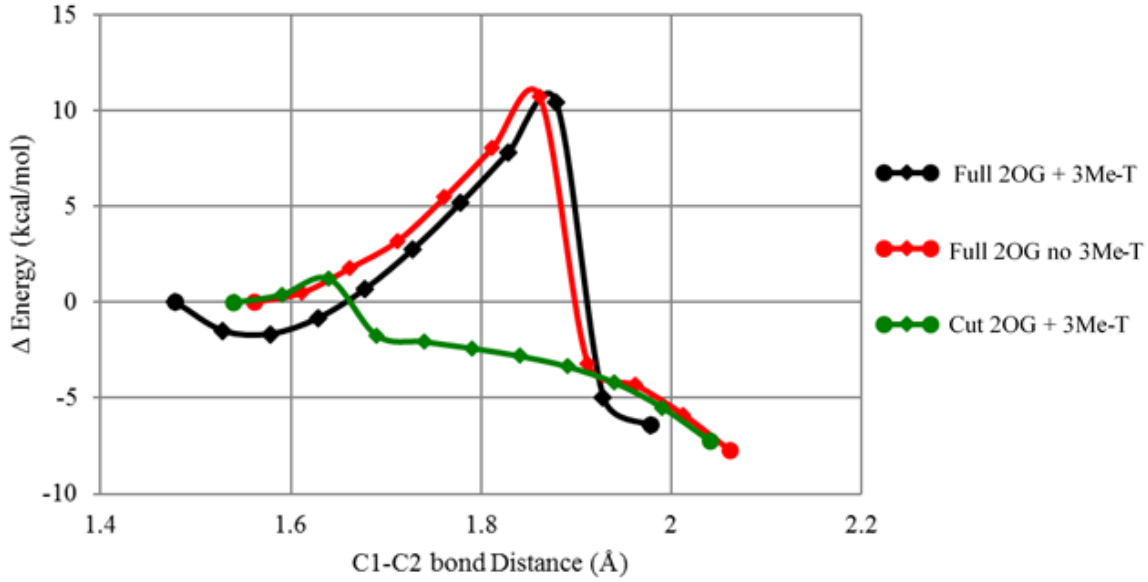

Figure 26: The 1-dimensional scans of the C1-C2 bond of 2OG reaction coordinate.

	H307	H231	D233	2OG (C1-O)	2OG (C2-O)	2OG C1-C2	Fe-O (prox.)	O(prox.)-O(dist.)	2OG C2-O(dist.)	3Me-T(CH3)-Fe
Full 2OG with 3Me-T optimised starting structure	2.20	2.10	1.96	1.96	2.22	1.48	1.93	1.28	3.12	4.23
Full 2OG with 3Me-T global maxima structure	2.18	2.10	1.95	2.02	2.21	1.89	1.91	1.28	2.58	4.25
Full 2OG with 3Me-T global minima structure	2.20	2.08	1.98	2.36	2.07	1.99	1.98	1.40	2.07	4.32
Full 2OG with 3Me-T optimised without CO2	2.14	2.10	2.01	n/a	2.24	n/a	1.91	1.48	1.33	4.10

Full 2OG without 3Me-T optimised starting structure	2.15	2.13	1.69	1.99	2.2	1.56	1.92	1.28	2.93	n/a
Full 2OG without 3Me-T global maxima structure	2.13	2.13	1.95	2.05	2.18	1.86	1.90	1.28	2.66	n/a
Full 2OG without 3Me-T global minima structure	2.15	2.10	1.99	2.51	2.08	2.06	1.99	1.41	1.37	n/a
Full 2OG without 3Me-T optimised without CO2	2.13	2.10	2.03	n/a	2.24	n/a	1.91	1.48	1.32	n/a
Cut 2OG with3Me-T optimised starting structure	2.04	1.99	1.95	1.93	2.32	1.54	1.88	1.29	2.43	4.2
Cut 2OG with 3Me-T global maxima structure	2.15	2.06	1.92	1.99	2.36	1.64	1.91	1.28	2.79	4.17
Cut 2OG with 3Me-T global minima structure	2.19	2.07	1.96	2.46	2.09	2.04	2.00	1.40	1.37	4.31

Table 7: Measured distances from the PES scans of the C1-C2 bond of 2OG, all distances are displayed in Å.

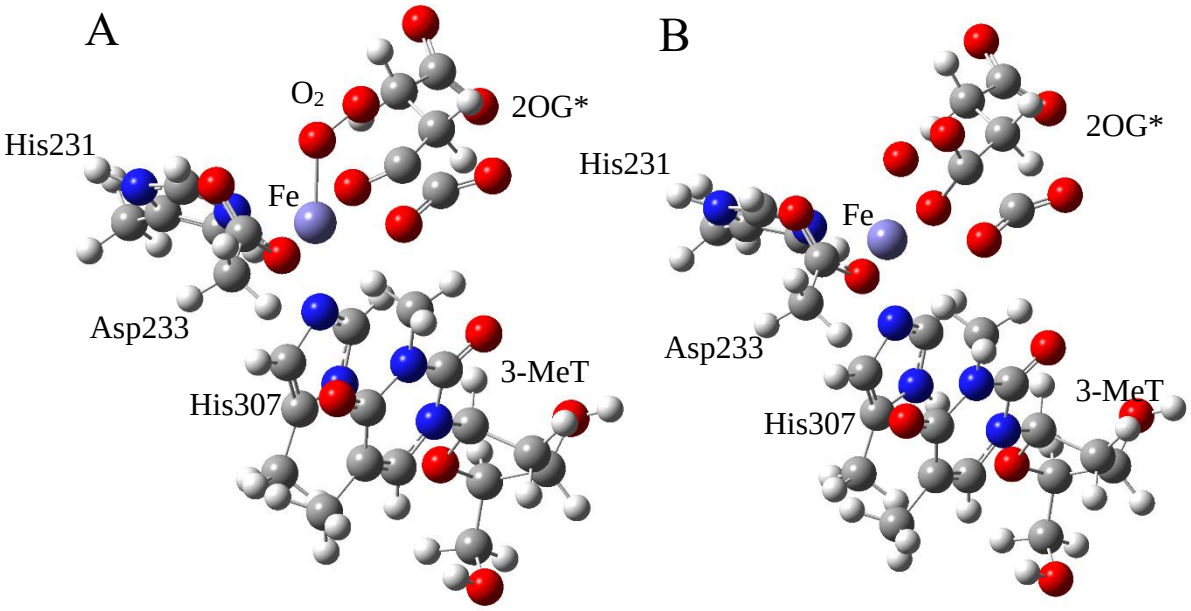

Figure 27: A) the global maximum structure, B) the global minimum structure for the standard QM cluster model from the PES scan of the C1-C2 2OG bond. (2OG* is labelled as such to represent the probable breakage of the C1-C2 bond and the formation of CO_2 and succinate)

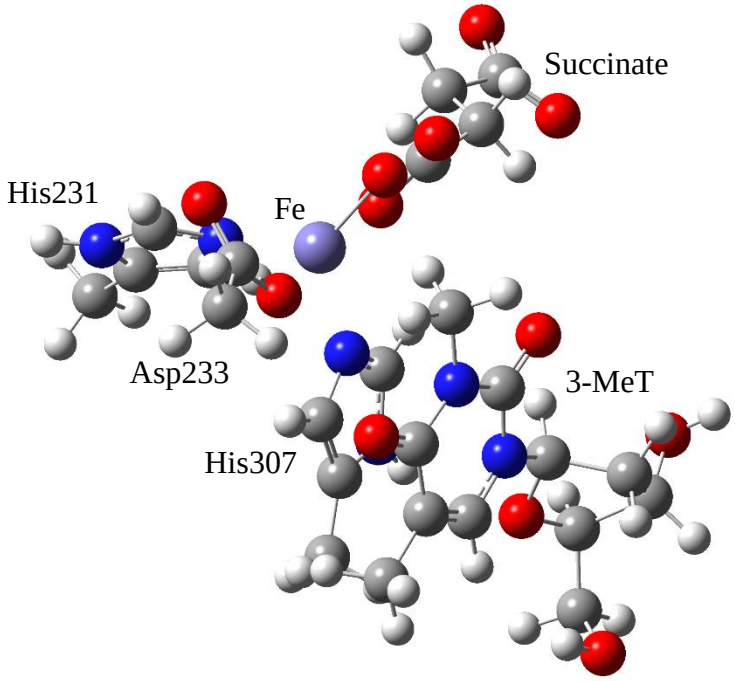

Figure 28: The freely optimised global minimum structure of the standard QM cluster model without CO_2. S=2

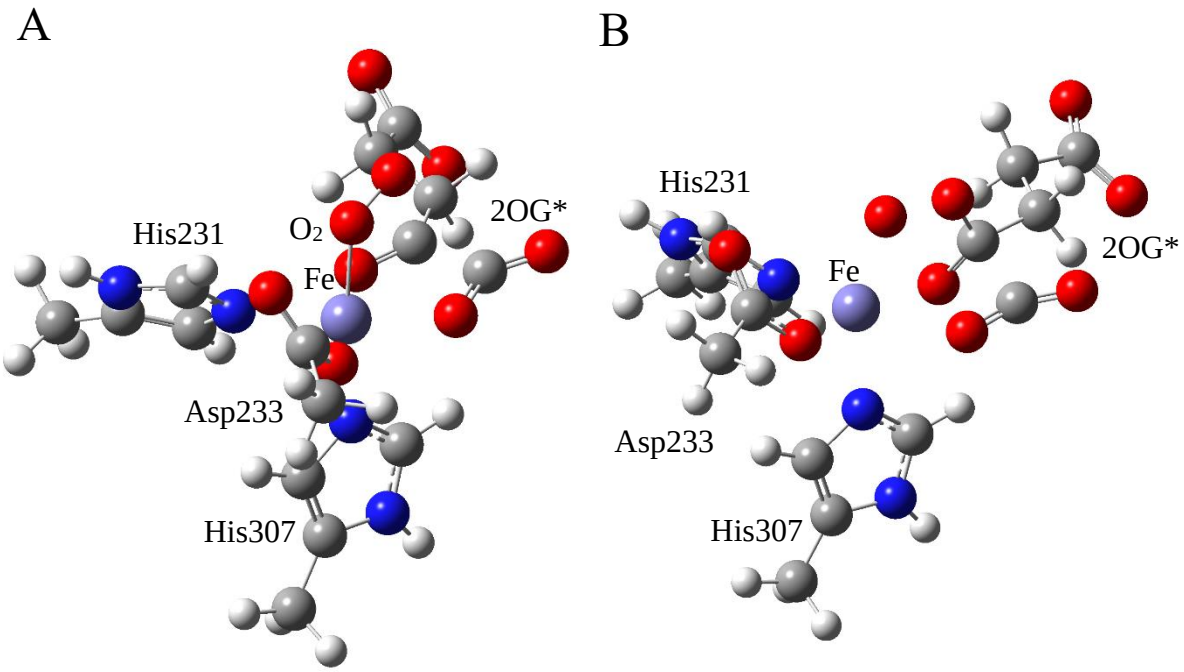

Figure 29: A) the global maximum structure, B) the global minimum structure for the full 2OG no 3Me-T QM cluster model from the PES scan of the C1-C2 2OG bond. . (2OG* is labelled as such to represent the probable breakage of the C1-C2 bond and the formation of CO_2 and succinate)

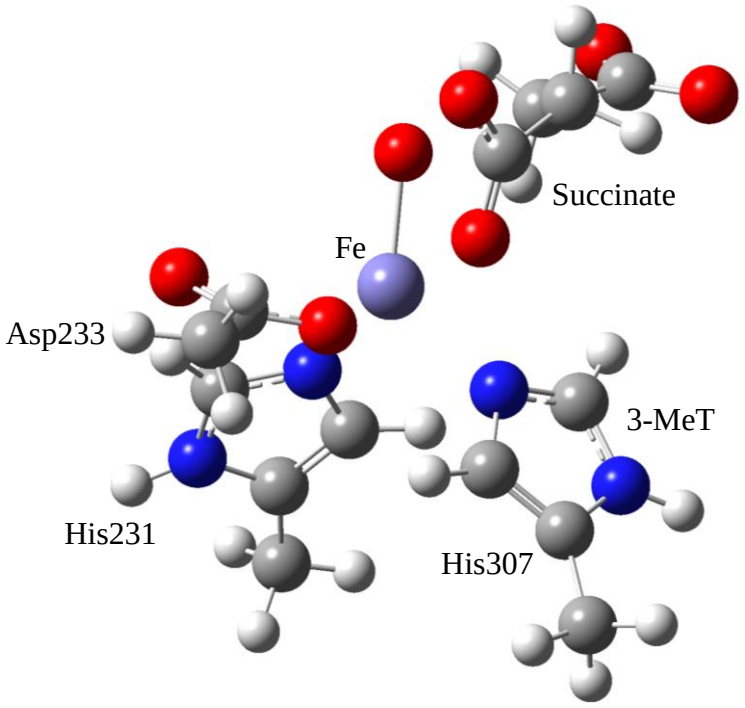

Figure 30: The freely optimised global minimum structure of the full 2OG no 3Me-T QM cluster model without CO_2. S=2

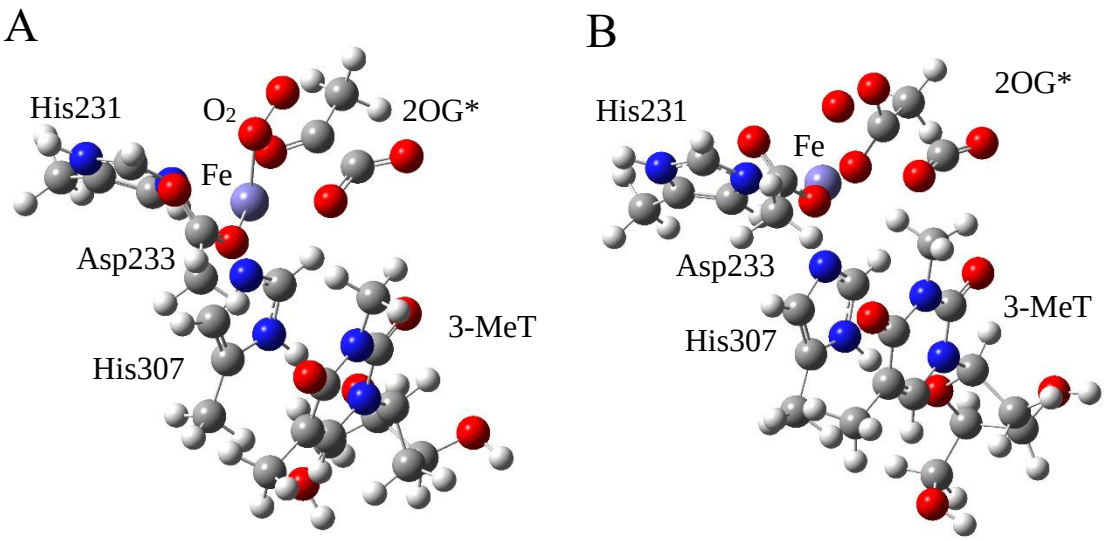

Figure 31: A) the global maximum structure, B) the global minimum structure for the Cut 2OG with 3Me-T QM cluster model from the PES scan of the C1-C2 2OG bond. (2OG* is labelled as such to represent the probable breakage of the C1-C2 bond and the formation of CO_2 and succinate)

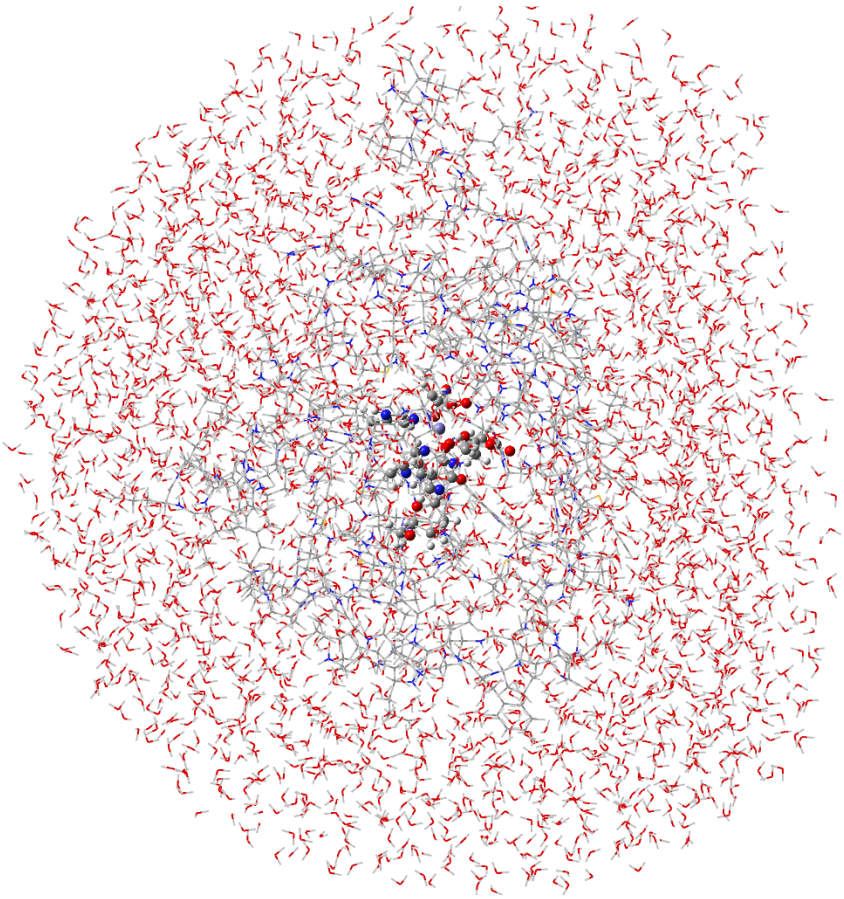

Figure 32: The whole QM/MM system, the atoms of the MM region are rendered as wireframes and the atoms of the QM system are rendered as balls and sticks.

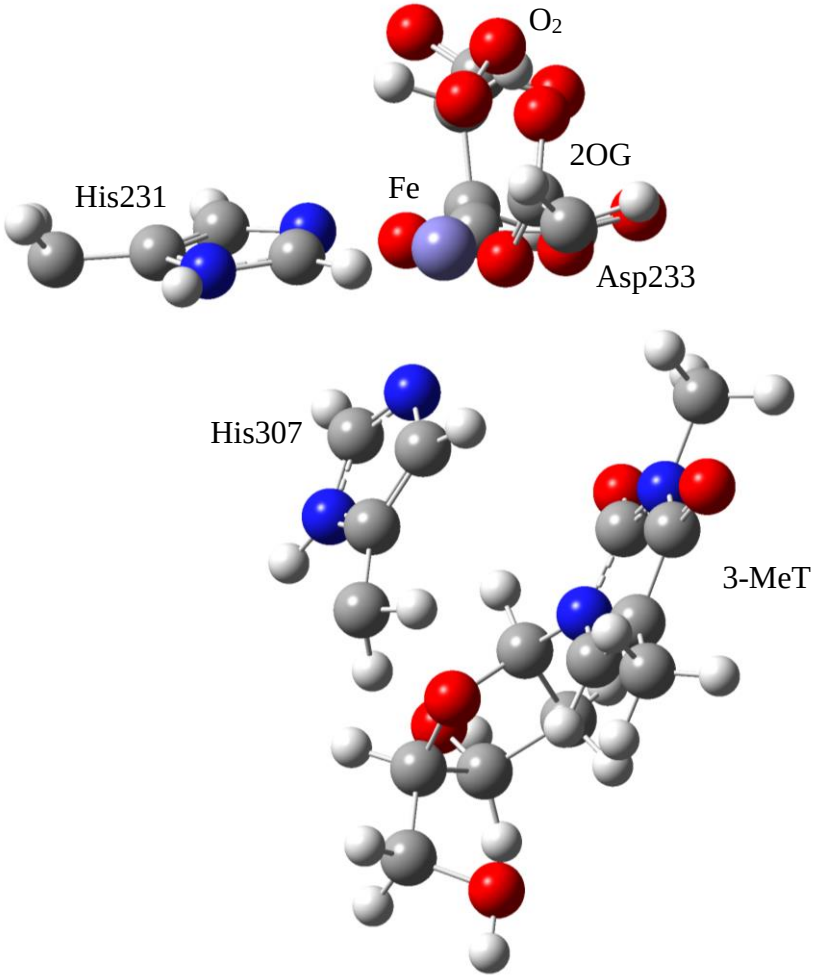

Figure 33: The QM atoms of the QM/MM model. S=2

3.5 DISCUSSION

3.5.1 Comparing the Effects of Different Cluster Size on Active Site Geometry

The distances measured were chosen to monitor the binding of iron to the coordinating amino acid residues and 2OG, the binding and geometry of oxygen in relation to iron and the viability of the target methyl group of 3Me-T to react with the formed Fe(IV)-oxo species. Of the three cluster models: the standard sized QM cluster and the cluster without 3Me-T show quite good agreement in the distance between H307, H231 and D233 to iron (Table 5). The cut 2OG with 3Me-T cluster shows good agreement with the other two clusters in the distance of D233 to iron but shows shorter distances of ~0.1Å between H307 and H231. The distance between the binding oxygen atoms of 2OG to iron are once again quite consistent between the standard sized QM cluster and the full 2OG without 3Me-T structure, whereas the cut 2OG with 3Me-T shows slightly longer binding distances for the C2-O atom of 2OG to iron by ~0.1Å (Table 5). All three QM cluster sizes show good agreement in Fe-$O_{proximal}$ and $O_{proximal}$-O_{distal} distances with only slight variation between them. The variation in distances between the three clusters is most pronounced in the distance between the 2OG-C2 to the distal oxygen atom. This distance in the standard sized QM cluster and the cluster without 3Me-T are somewhat similar, but is significantly shorter by ~0.5Å in the cut 2OG cluster (Table 5). The distance between the 3Me-T methyl carbon atom and iron are almost identical between the standard sized cluster and the cut 2OG with 3Me-T cluster, which is in turn similar to that of the QM/MM model and the pdb structure.

The distances measured in the initial pdb structure show reasonable agreement with the QM/MM model, with the only major exception being the 2OG C2-O to Iron distance that is 0.36Å longer in the QM/MM optimised structure (Table 5). However, comparisons between the crystal structure and the QM clusters and QM/MM model are difficult to judge as the

crystal structure was crystallised without 3Me-T or oxygen bound so represents a different stage of the full reaction mechanism.

The QM/MM optimised structure shows reasonable agreement with the bond length found in the standard sized QM cluster for the two coordinating histidine residues, however the bond length between aspartate and iron is somewhat longer ~0.15Å in the QM/MM model as compared to the other QM cluster sizes (Table 5). This lengthening of the bond between D233 and iron may be caused by interaction of D233 with the positively charged residue R322 weakening the D233 to iron interaction. A similar effect may contribute to the increased distances between iron and the 2OG binding oxygen atoms 2OG-C1-O and 2OG-C2-O in the QM/MM model, particularly the 2OG-C2-O atom is ~0.2Å further from iron in the QM/MM model compared to the standard sized QM model (Table 5). The sidechains of both R316 and R322 are found in close proximity to the negatively charged 2OG ligand, and their electrostatic interactions likely influence 2OG positioning and its binding to iron. The most striking difference between the QM/MM model and the other three QM cluster models is found in the geometry of the bound oxygen, in the QM/MM model the Fe-O bond distance is significantly longer by about ~0.2Å and the $O_{proximal}$-O_{distal} bond distance is shorter by about 0.05Å. The geometry of oxygen in the QM/MM model means the distal oxygen atom is found much closer to the C2 atom of 2OG by about 0.5Å, this is an important finding as this distance is along the proposed path of reaction that initiates the decarboxylation of 2OG and formation of the Fe(IV)-oxo species. This finding seems to suggest that the surrounding residues of the FTO active site have a facilitating effect on the formation of the Fe(IV)-oxo species.

Overall, the cut 2OG model seems to give the most inconsistent results, both in comparison to the other two QM cluster models and the QM/MM model. Based on our results this may

indicate that the reduced 2OG group may be an unsuitable replacement for the full 2OG ligand when modelling the FTO active site. In comparison 3Me-T seems to have only a subtle effects on the QM cluster model geometry of the iron binding site in FTO. In future the effect of cutting 2OG and not including 3Me-T in the QM cluster models will be investigated to better ascertain the influence that these two important components have on the overall geometry of the iron binding site.

3.5.2 Comparison of Extended Cluster Sizes

To investigate the effect of the surrounding arginine residues further, the extended cluster models were created and their measured geometries are displayed in Table 6. The R316 extended cluster was primarily created to investigate the influence that the positively charged sidechain had on the binding of the negatively charged 2OG. In the R316 extended cluster we see similar binding distances between the H307, H231 and D233 residues to iron as those found in the standard sized QM cluster (Table 6). This was the expected result as R316 is mainly thought to electrostatically interact with the C5 carboxylate moiety of 2OG so shouldn't have much of an effect on the distance between coordinating amino acid residues and iron. The distance between the 2OG binding oxygen atom 2OG-C1-O was longer in relation the standard sized QM cluster model and a lot more similar to the distance measured in the QM/MM model. Although the 2OG-C2-O distance was found to be longer in the R316 extended cluster relative to the standard QM cluster, it was still 0.16Å shorter than the same distance in the QM/MM cluster model. This result indicates that although R316 is an important interacting residue for 2OG, there must be other interacting residues in the proximity of 2OG which cause the 2OG-C2-O to iron distance to be ~0.2Å longer in the QM/MM model as compared to the standard sized QM model. Residues such as Y295 and S318 are also found to hydrogen bond the C5 carboxylate moiety of 2OG in the crystal structure and perhaps cause the observed differences in the geometry of 2OG and its binding

to iron. The addition of R316 to the standard QM cluster model seems to have shortened the important distance of 2OG-C2 to distal oxygen by about 0.25Å when compared to that of the standard QM cluster. This change in the position of 2OG causes the plane of the C1-C2 bond to sit above that of the two iron binding oxygen atoms of 2OG, in the standard QM cluster 2OG is more planar in respect to iron, this makes the 2OG-C2 to distal oxygen slightly longer. This observed change in 2OG positioning may be one of the ways in which R316 influences active site geometry to encourage the formation of the Fe(IV)-oxo species.

The R322 extended cluster model was created to assess the effect that R322 had on the active site geometry given its close proximity to the two negatively charged moieties of D233 and the C1 carboxylate of 2OG. The binding of H307 to iron in the R322 extended cluster model showed good agreement with the QM/MM model, being slightly shorter than the distance measured in the standard sized QM cluster. The binding of H231 was mostly unaffected and remained in good agreement with the standard sized QM cluster model and consequently also the QM/MM model. Although the binding distance between D233 and iron was longer in the R322 cluster as compared to the standard QM cluster it was still shorter than the same distance measured in the QM/MM model, this likely indicates that the longer D233 to iron distance observed in the QM/MM model is influenced by other factors than R322 alone. The length of the 2OG-C1-O to iron bond is found to be more consistent with that of the QM/MM model than the standard sized QM cluster, this is consistent with the idea that the binding between the C1 carboxylate moiety of 2OG and iron is influenced by the close proximity of R322. The Fe-O and $O_{proximal}$-O_{distal} bond distances are nearly unchanged in the R322 cluster as compared to the other QM cluster models, however the distance between the distal oxygen atom and the C2 atom of 2OG is much longer than in the other QM clusters and the QM/MM model. In the R322 extended cluster model we find the distal oxygen atom to be drawn more towards the sidechain of R322 (Figure 23), therefore it seems that R322 in isolation causes

oxygen to adopt a different, and probably non-productive, conformation as compared to the observed conformation of the other QM cluster models. This seems to indicate that R322 doesn't have a facilitating effect on the formation of the Fe(IV)-oxo species.

When both R322 and R316 are included in an extended QM cluster model we find that the iron to coordinating histidine bonds are shorter when compared to both the other QM clusters and the QM/MM model. The iron to D233 bond is slightly longer than in the other QM cluster models and is found to be in closer agreement to the distance in the QM/MM model. The binding of 2OG in this cluster more resembles that of the standard sized QM model and R322 extended QM cluster model. The inclusion of the two arginine residues seems to have only a subtle effect on the binding of oxygen to iron, the Fe-O bond distance in the R316 and R322 cluster model (Figure 24) is positioned similarly to that of the R316 extended cluster model (Figure 22) and the QM/MM model (Figure 33). The distal oxygen to 2OG-C2 distance is shorter than the standard QM model by about 0.5Å and is more similar to the distance found in the QM/MM model. These observations seem to indicate that the inclusion of the two residues R316 and R322 together cancel out the disrupting effect that R322 has on oxygen orientation relative to the 2OG-C2 atom that was observed in the R322 extended cluster model, this might indicate that the two residues have a synergistic effect on facilitating distal oxygen attack of 2OG. Further calculations will be required to determine the exact nature of the inter-play between these two important arginine residues.

The triple arginine extended cluster included R96, R316 and R322 and was created to investigate the effect that R96 had on the geometry of the iron binding site. The distances of the iron coordinating amino acid residues and 2OG showed good agreement with the other extended clusters, this seemed to agree with our previous observations of the proposed effects of R316 and R322 on the iron binding site geometry. The most profound effect on

the active site geometry was found in the Fe-O bond distance, in the triple arginine extended cluster the Fe-O bond is increased by ~0.1Å compared to both the standard sized QM cluster model and the other extended cluster models. This makes the Fe-O bond distance in the triple arginine extended cluster more similar if still slightly shorter than that of the QM/MM optimised structure, this possibly indicates to us that R96 is an important residue for oxygen binding. The triple arginine extended cluster also seems to possess the shortest 2OG C2 to distal oxygen distance of all of the QM cluster models. To test this hypothesis, more calculations on new QM cluster models will be required to ascertain whether R96 has this influence in isolation or if it is a synergistic effect with either one or two of the other arginine residues R316 and R322.

Overall, the extension of the QM cluster model to include arginine seemed to have profound effects on the 2OG C2 to distal oxygen distance, with every extended cluster apart from the R322 extended cluster showing significant shortening of this important measurement. However, comparisons between the QM/MM model and any cluster model no matter how large the cluster size are not direct comparisons due to the differences in solvation methodology. The explicit solvation of the QM/MM model as compared to the implicit solvation model of the QM cluster models will lead to different active site geometries and electronic structures.

3.5.3 PES Scans for the Formation of the Fe(IV)-oxo Complex

The profile of the PES scans of the standard sized QM cluster and the full 2OG without 3Me-T are incredibly similar (Figure 26), despite their initial C1-C2 bond distances being 1.48Å and 1.56Å respectively, the two clusters share a similar energy barrier height of ~10.5 kcal/mol. The two clusters also achieve their global maxima when the C1-C2 bond distance is 1.89Å for the standard sized QM model and 1.86Å for the full 2OG without 3Me-T cluster.

The measured distances from their global maxima structures nearly identical (Table 7). The structures of the global maxima of these two clusters are shown in Figure 27a and Figure 29a, both structures show the partial dissociation of the carboxylate moiety however at this point the distal oxygen is still positioned around 2.6Å from the C2 of 2OG. The end point structures of the scan differ but this may be due to their different starting 2OG C1-C2 bond distances, when carbon dioxide is removed and the two clusters are optimised freely their geometries once again become nearly identical (Figure 28 and Figure 30). All of this indicates that any role that 3Me-T plays on the calculated reaction path is subtle and not easily quantified by our QM cluster calculations, this is a reasonable assessment given the uncharged nature of 3Me-T and it being positioned >4Å from iron for the duration of the PES scan. However to more fully understand the influence of 3Me-T on the active site geometry during Fe(IV)-oxo formation further calculations using QM/MM models will need to be performed.

In contrast, the PES scan of the cut 2OG with 3Me-T cluster is very different from the other two clusters scans, despite sharing a similar initial 2OG C1-C2 bond distance as the full 2OG without 3Me-T. The barrier height of the cut 2OG with 3Me-T cluster is very low at 1.25 kcal/mol and the 2OG C1-C2 bond distance at this point is much shorter at only 1.64Å compared to the other two clusters PES scans. However, the distal oxygen to C2 of 2OG distance at this point isn't too dissimilar to the other two clusters at a distance of 2.79Å. Overall, the comparison of the three clusters that underwent 1-D PES scans of the 2OG C1-C2 distance reinforces our hypothesis that the cut 2OG model doesn't adequately replace the full 2OG ligand for the purposes of modelling the reaction mechanism of FTO. To better understand the electronic structural transitions that are occurring over the course of the PES scans more extensive calculations are required, with orbital and spin population analysis. The global maxima structures don't actually represent true transition states but provide good

starting points for transition state search algorithms to define actual transition state structures.

3.6 CONCLUSIONS

Despite the results presented being an incomplete set we can draw some initial conclusions from the work. With regards to the cutting of the 2OG molecule we have shown that this leads to results, in both the active site geometry and energetic results, that are inconsistent and probably erroneous when compared to the clusters with the full 2OG molecule and the QM/MM model. Including the 3Me-T molecule in the cluster calculations seems to have only a subtle effect on the results of the cluster calculations and for the sake of computational efficiency, i.e. reducing the number of atoms in the QM system, it could probably be excluded without drastic effect to the calculations.

The inclusion of the nearby arginine residues in the cluster calculations had a large effect on the geometry of the active site, with binding between iron and oxygen being most affected. In every cluster, excluding the R322 cluster, we observed a shortened distal oxygen to 2OG C2 distance. This distance is thought to be a key reaction coordinate in the attack of 2OG by oxygen and subsequent formation of the Fe(IV)-oxo intermediate complex. Comparison between the QM/MM model and extended cluster sizes was confounded by the use of implicit solvation in the QM clusters as opposed to explicit solvation in the QM/MM model making the influence of the surrounding amino acid residues difficult to distinguish from solvent model effects. More QM/MM modelling using alanine mutations or free energy calculations will be necessary to assess the individual contributions of neighbouring residues to the reaction path.

3.7 FUTURE WORK

Due to the computational complexity of the calculations and the lack of effective large scale parallelisation for QM calculations not all of the calculations could be finished in time for publication in this thesis. The first priority should be to create and optimise more QM cluster models that give a more direct comparison of the individual contributions of each component to the geometry and electronic structure of the QM model. Some examples of additional clusters would include: cut 2OG without 3Me-T, an extended cluster including the sidechain of R96, an extended cluster also including the sidechains of R96 and R316, an extended cluster including the sidechains of R96 and R322. The optimised QM and QM/MM models would then undergo extensive orbital and spin population analysis to allow the differences in their electronic structures to be assessed.

Investigations into the effect that functional and basis set have on the geometry and electronic structures of the different QM clusters and QM/MM model will also be investigated. Optimisations and scans with different functionals and basis sets would seek to replicate published results of similar calculations for different non-heme iron 2OG dependent oxygenases. These type of calculations would allow for comparisons between the electronic structure of FTO and other non-heme iron 2OG dependent enzymes.

Additionally, specific optimised structures will undergo different PES scans to test other reaction coordinates and calculate the energy barrier associated with the formation of the Fe(IV)-oxo complex. A reaction path for the QM/MM model would also give a better comparison to that calculated for the QM cluster models. Some of the structures of these calculated reaction paths could represent the structures of reaction intermediates and transition states, however these structures need to be refined with transition state search algorithms and more rigorous optimisations of their geometry until they possess only a single

negative frequency. The orbital and spin populations of theses optimised transition state structure would also be analysed to better understand the changes in the electronic structure occurring over the course of the calculated reaction path.

4 CONFORMATIONAL CHANGES OF THE GLUT5 RECEPTOR IN A MEMBRANE BOUND SYSTEM: A MOLECULAR SIMULATION STUDY

4.1 PREFACE

This chapter describes ongoing research done in collaboration with Dr. Marina Tanasova and her research group based at Michigan Technical University, with the aim of producing one or more joint publications featuring both experimental and computational research. The docking portion of this chapter describes the use of docking software to generate reasonable structures for several ligands that could then be verified against experimental results relating to their affinity to the GLUT5 receptor. The MD portion of this chapter describes the MD simulations created to study the conformational dynamics and substrate interactions of the GLUT5 receptor embedded in a lipid bilayer. The chapter will contain an overview of the study so far as well as a detailed description of the contributed computational research.

4.2 INTRODUCTION

GLUT proteins are part of a wide group of proteins classified by their role in facilitating the movement of glucose and other sugars through the cell membrane, forms of glucose transporter proteins are found throughout nature as glucose is a near universal energy source for most organisms [174]. Mammalian GLUT proteins can be roughly split into three different classes, within each class the GLUT proteins share certain sequence and structural similarities [175-177]. Class I GLUTs account for GLUT1-GLUT4 and are mainly involved in the uptake of glucose, galactose and other sugars into the cell. Class II GLUTs (GLUT5, 7, 9 and 11) instead facilitate the transport of fructose and glucose across the cell membrane. Class III GLUT proteins are the class of more recently identified and more poorly understood glucose transporter proteins found to be distinct from Class I and II types [175, 177, 178]. GLUT5 is mainly found to be expressed in the cells of the small intestine where it transports fructose from the lumen, where fructose is in relatively higher concentration, into the enterocytes (the cells that make up the inner lining of the small intestine); however it is also found to expressed in lower levels all over the body in the kidneys, brain, testes, skeletal muscles and adipose tissues [179]. High levels of GLUT5 expression have been observed in type 2 diabetes [180, 181] and other metabolic diseases [182, 183] as well as breast [184, 185] and other forms of cancer [186]. These diseases all have the common factor that they are indicative of having states of altered metabolism, cancer cells in particular rely on anaerobic glycolytic respiration pathways that bypass the mitochondrial stage of metabolism (aka as the Warburg effect) to fuel their rapid proliferation [187]. For this type of metabolism they need a ready source of sugars which they recruit from outside the cell by over expressing glucose transporter protein such as GLUT5 [186-188]. GLUT5 is a particularly interesting transporter protein because it is the only know one that solely transports fructose, all other GLUT proteins are concerned with the transport of glucose and additional sugars [178].

GLUT5 is a facilitator transporter, this means it's a membrane bound protein responsible for the transport of a substrate across the mostly impermeable hydrophobic cell membrane [174]. Although some small molecules can pass through the cell membrane by diffusion alone, this process is slow and unregulated, facilitated transportation is needed to reliably move amino acids, peptides, sugars and ions across the cell membrane [189]. Specialised proteins fulfil the role of transporting substrates across the cell membrane, different types of these proteins utilise a range of strategies to achieve this. GLUT5 as a transporter protein is defined as having two gates, one for the intracellular and one for the extracellular side. These gates can be either open or closed, if one is open the other by definition is closed in a mechanism known as alternating access [174-177, 189-193]. This defining characteristic separates them from the similar channel family of proteins, which are by definition either simultaneously open or simultaneously closed to both the intra and extracellular sides at any given time [191]. This distinction allows transporter proteins to move their substrates both with and against a diffusion gradient, whereas channel type transporters can only move substrates from high to low concentrations [175]. The mechanism of alternating access is achieved through large conformational motions in the protein that act in response to substrate binding to open and close the gates in a concerted way to move the substrate from one side of the membrane to the other [190]. Although the process is a cycle where each step is reversible, for the sake of clarity the mechanism will be described as follows. The protein favours and adopts the outward open conformation when no substrate is bound [194], upon substrate binding a conformational change occurs to close the outward facing gate restricting access to the binding site from the extracellular side of the protein. The substrate bound outward occluded structure undergoes a conformational shift that sees the two domains rotate ~15° to open the inward facing gate of the protein, this allows the substrate to unbind and enter the intracellular space. In the absence of the sugar substrate the transporter defaults

to the outward open conformation and is again ready to accept a substrate [175, 192]. A stylised diagram of the proposed alternating access mechanism is shown in Figure 34.

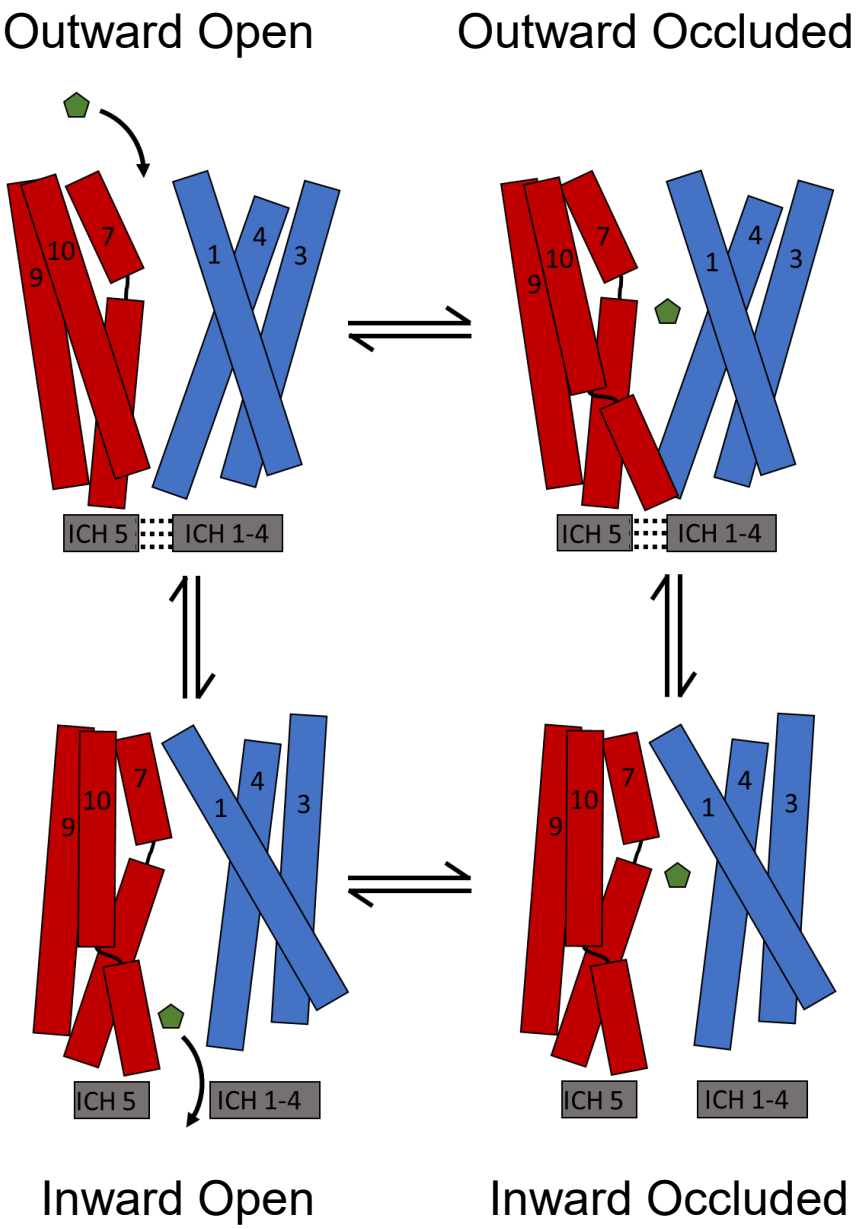

Figure 34: A stylised diagram of the proposed alternating access mechanism of GLUT5. The C-terminal transmembrane bundle helices are represented by red boxes, the N-terminal transmembrane bundle helices are represented by blue boxes, and the intracellular helices are represented by grey boxes. Dashed lines between the intracellular helices represents the proposed hydrogen bonding between ICH5 and the other intraceullar helices. The fructose substrate is represented by a green pentagon.

GLUT5 is composed of twelve Trans-Membrane α-helices (TM) and five Intra-Cellular α-Helices (ICH). The TM helices are arranged into two bundles, six in the C-terminal domain and six in the N-terminal domain [194]. Throughout the cycle of alternate access the N-terminal domain remains rigid, whereas the C-terminal domain undergoes internal structural rearrangement (particularly at transmembrane helices 7 and 10) in order to facilitate the conformational motion necessary for opening and closing the inner and outer gates of the transporter, this motion is sometimes referred to be a rocker-switch-like movement [191, 194]. The ICH are thought to be responsible for stabilising the outward open conformation of the protein, and are probably the reason the protein defaults to the outward open conformation when no ligand is bound. Salt bridges form between the IC helices which link the N-terminal and C-terminal domains, experimental mutational studies of the key ICH salt bridge forming residues E336, E400 and R407 to neutral alanine have been shown to freeze the transporter into an inward facing conformation [193]. The crystal structure of GLUT5 (PDBID 4YB9) in the inward open conformation is shown in Figure 35, a stylised diagram of the proposed alternating access mechanism of GLUT is shown in Figure 34. GLUT5 only transports fructose and no other similar hexose sugars, to achieve this the protein must have a binding site that is specific only to fructose, and produce no conformational motion to transport the sugar through the membrane in the presence of any other sugar. Structural comparisons to understand the specificity of GLUT 5 towards fructose in contrast to the other GLUT1-4 transporters are hampered by the lack of experimental structures of the receptors in complex with different sugar substrates [175]. Further frustrations to the structural study of the GLUT transporters are caused by the unknown contributions of the lipids in the membrane, in the same way that solvated protein structures differ from their crystallised counterparts, so do the structure of membrane bound and crystallised transporter proteins. The lipids in the membrane clearly support the receptor and maintain its structure,

however their contribution to the structure function relationships and the conformational dynamics of the receptor at the atomistic level is poorly understood [195-197].

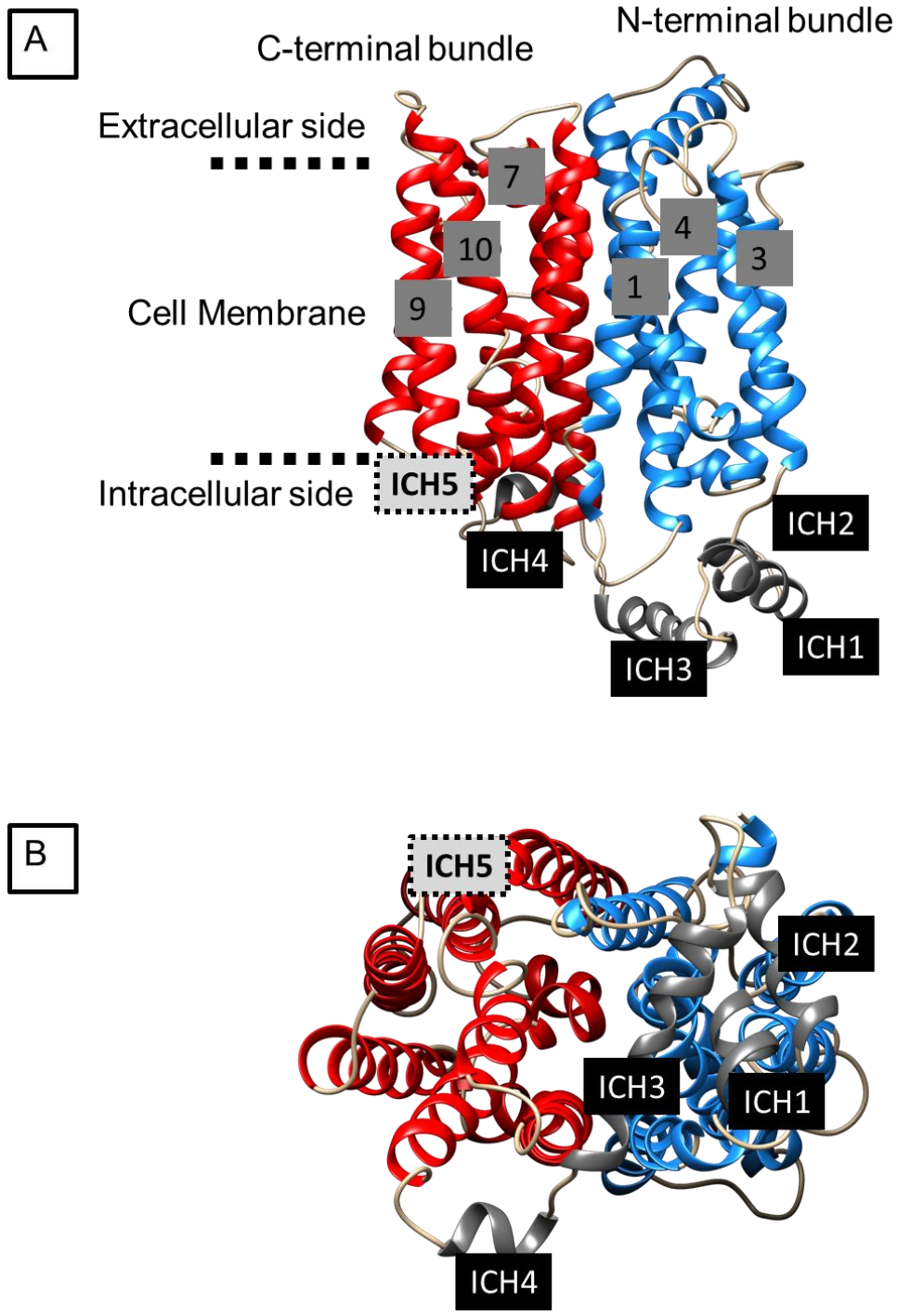

Figure 35: The crystal structure of GLUT5 (PDBID 4YB9) is in the open inward facing conformation. The C-terminal transmembrane bundle helices are rendered as red ribbons, the N-terminal transmembrane bundle helices are rendered as blue ribbons, the intracellular helices are

rendered as grey ribbons and the loops joining the helices are coloured beige. A) Shows a cross membrane view of the receptor with the important transmembrane helices labelled with grey boxes and black text. The intracellular helices are labelled with black boxes and white text, ICH5 is not present in the crystal structure but its position is still labelled with a light grey box with a dashed line border. B) Shows the same representation of the GLUT5 structure but rotated 90 degrees to give a view from the intracellular side of the membrane. In this panel the intracellular helices are labelled as above.

Given their link to a range of diseases and conditions [180-186], inhibitors of GLUT5 are extremely sought after as no potent and selective drug targeting GLUT5 is currently available [198]. Development of ligands to inhibit GLUT5 has been frustrated by the lack of a reliable experimental high throughput-assay with which to test potential compounds with, the binding affinity between GLUTs and their sugar ligands cannot be measured by a range of commonly used techniques such as isothermal calorimetry titration, surface plasmon resonance or microscale thermophoresis [175]. As well as being a drug target, GLUT5 is an interesting marker for cancer and metabolic disease. Development of selective fluorescent probes that bind to GLUT5 could be used to quickly diagnose unusual overexpression of GLUT5 that might be linked to tumour formation or metabolic disease. This strategy is already utilised with the use of glucose radio-labelled with the isotope ^{18}F (Fludeoxyglucose), cancer cells take in the radio-labelled glucose compound at a rate that is significantly higher than surrounding cells; the accumulated radiotracer can then be imaged with a positron emission tomography scan [199]. This strategy however isn't perfect, not all cancer cells have increased glucose uptake and therefore won't show up on the PET scan [200, 201]. Being able to detect fructose transport into cells through GLUT5 could provide an alternative strategy for imaging cancer cells as well as measuring how aggressive their growth rate is, this could allow therapies to be tailored towards how quickly a particular tumour will grow and potentially metastasize [178]. Radiotracer dyes are suited towards detection inside the body as the emitted radiation can be detected through tissue, however they are considered expensive due to the on-site cyclotrons necessary to produce them, and their short half-lives mean they cannot be stockpiled in significant amounts [202]. Fluorophore conjugated sugars offer a cheaper alternative for use in laboratory settings, cells that take up the dye can be easily detected using spectrofluorometers to image the transport of the dye into cells. Although not as suited to *in-vivo* imaging as radiotracer dyes due to the

need for photon penetration, fluorescent dyes have been gaining traction as a diagnostic tool for early tumour detection [203-205].

A range of Fluorescent probes were designed and synthesised by our experimental collaborators in Dr. Tanasova's group, the ManCou probes were created by conjugating the blue fluorescent coumarin group to the aminosugar 1-amino-2,5-anhydro-d-mannitol. Conjugation to mannitol was chosen as it was shown to only rely on fructose transport so was likely to pass only through GLUT5 as opposed to fructose conjugation which produced conjugates that were reliant on both glucose and fructose transport and likely to pass through both GLUT1 and GLUT5 [206]. Variation of the C4 functional group on the coumarin moiety was designed to give a range of ManCou probes with different GLUT5 binding properties as well as different fluorescent colours. To aid in the design of new ManCou probes, molecular docking can provide an atomistic insight into the binding of the probes to the GLUT5 transporter. For the reasons described above, the crystal structure of the GLUT5 transporter may not be representative of the protein structure in its physiological, membrane bound, and solvated environment. The receptor will also flex and change over time in response to its own internal dynamic motions as well as those of the lipid membrane and bound substrate molecule, these movements may influence the active site interactions. The study of the interactions between the substrate and receptor over the course of an MD simulation will give insights into how the probes bind within the transporter that may be helpful in the design of new fluorescent probes.

4.3 COMPUTATIONAL METHODOLOGY

The structure of the receptor in an inward open form from *Bos Taurus* (PDBID 4YB9 [194]) is missing the coordinates of a loop in the C-terminal region, these coordinates were modelled using MODELLER [207, 208] for Chimera [131]. The structure was superimposed with a structure of human GLUT3 (PDBID 4ZWB) with maltose bound in the receptor site.

NDBM and the 3 ManCou probes (1,3 and 4) were modelled using GaussView [167] and their structures were optimised using the B3LYP/6-31G functional [209, 210] and basis set in Gaussian09 [94]. Maltose was docked to a rigid model of the GLUT5 receptor using AutoDock4[42] to verify the proposed location for fructose and NDBM binding. Having verified their similar binding sites NDBM and fructose were then also docked to the rigid GLUT5 receptor.

The 3 ManCou probes as well as being rigidly docked also underwent flexible receptor docking with GLUT5. The residues identified as key binding residues from the crystal structure as well as those observed to be in the rigid docking results, were selected to have rotatable sidechain bonds. The poses from the flexible docking results were found to agree well with the poses from the rigid docking results, based on this it was deemed that the flexible receptor docking offered no clear advantage over the rigid receptor docking method in this case.

The lowest energy binding poses from the docking of fructose, ManCou1 and ManCou3 were selected to undergo extended MD simulation. The protein was prepared with PROPKA web server [166] to define the protonation states of the titratable residues. To build the membrane bound system of GLUT5, the membrane builder module from the CHARMM-GUI website [211] was used. Based on the methodology of similar previous MD simulations of GLUT1[212, 213], 246 POPC lipids were chosen to construct the membrane with the receptor bound in the centre of the membrane. The system was solvated with TIP3P [21] water molecules with K+ and Cl- ions added to create a neutral system with an ion concentration of 0.15M with tleap in Amber14 [25] to give total system dimensions of ~100Å X ~100Å X ~110Å. The AmberFF14SB [214] forcefield was used with the lipid14 forcefield [215] for the membrane. Fructose and ManCou1 and ManCou3 were parametrised

with Antechamber in Amber14 [25] using the GAFF forcefield [26]. The system was minimised in Amber14 using pmemd with 5000 steps of steepest descent followed by 5000 steps of conjugate gradient minimisation. Following this, the lipids in the system were heated in two stages, first to 100K whilst the lipids in the system were restrained with a 10kcal/mol force constant. The second stage of heating warmed the system to 300K with the lipids restrained as in the first heating step, in addition at this stage the pressure in the system was equilibrated to one atmosphere. A short MD run of 5ns was first run to equilibrate the systems PBC box dimensions before the productive MD run. The productive MD simulations used the pmemd.cuda code on a GPU workstation for one microsecond. Analysis of the simulation trajectories was performed with CPPTRAJ for Amber14 [25] and the MD movie module of Chimera [131]. All images of molecular models for this chapter were created using Chimera [131].

4.4 RESULTS

4.4.1 Docking of NDBM, Fructose and Maltose

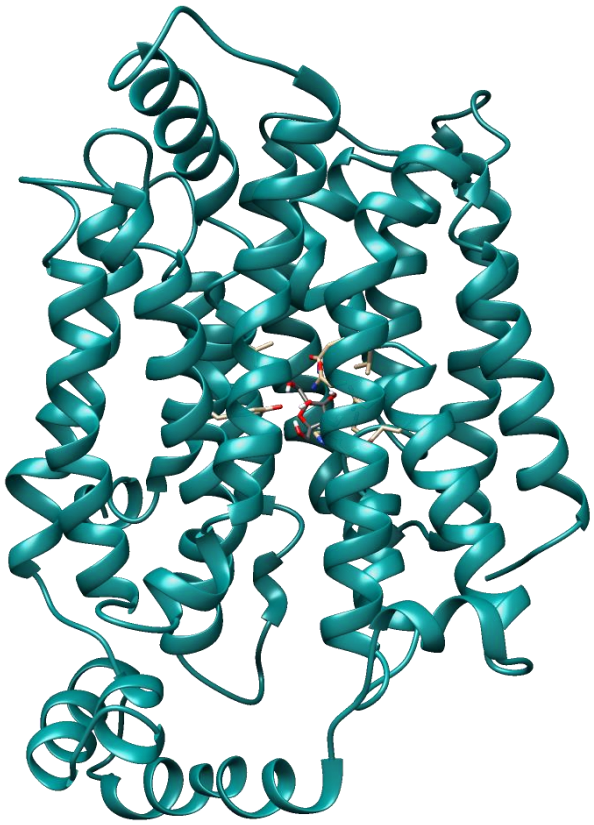

Figure 36: The ribbon view structure of GLUT5 from the PDBID 4BY9 structure. The protein's ribbons are shown in cyan, the atoms of fructose and its interacting residues are shown in stick rendering with grey carbons for fructose and beige carbons for those of the protein. The rest of the atoms are coloured as follows: hydrogen-white, nitrogen-blue, oxygen-red.

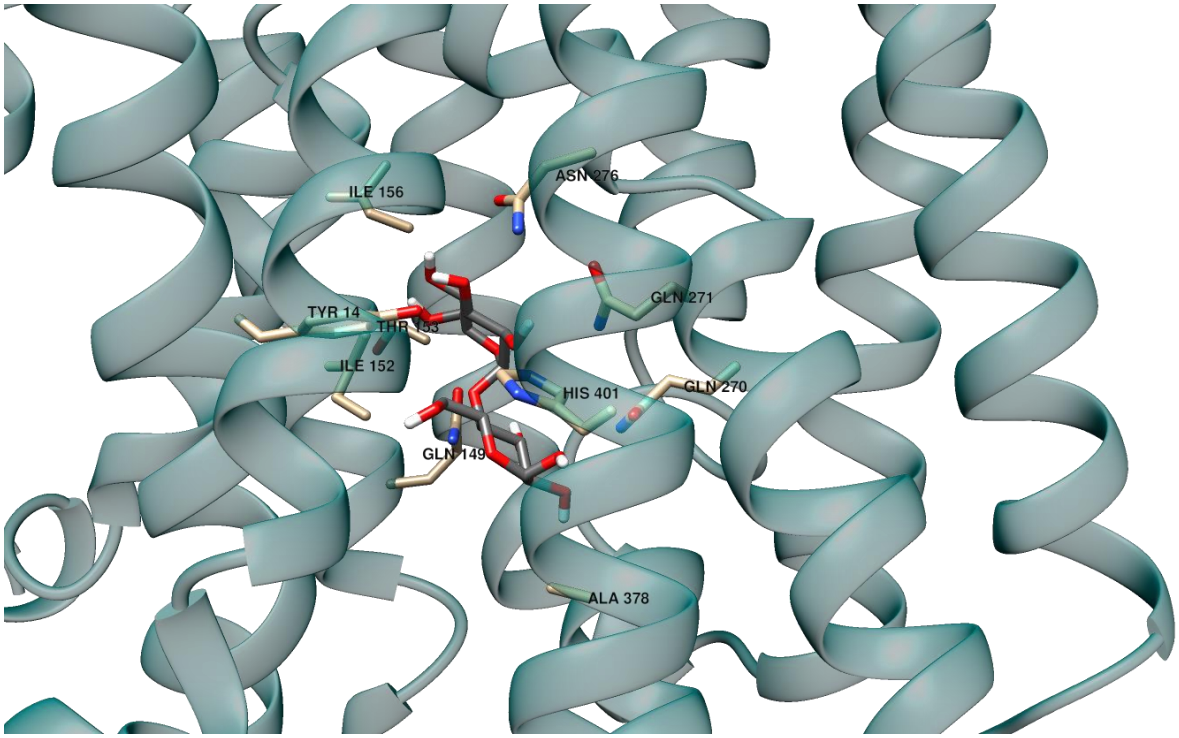

Figure 37: The docked structure of maltose with the interacting residues of the GLUT5 receptor labelled. The ribbons of the protein are partially transparent to improve clarity. The atom colours follow the same scheme as Figure 36.

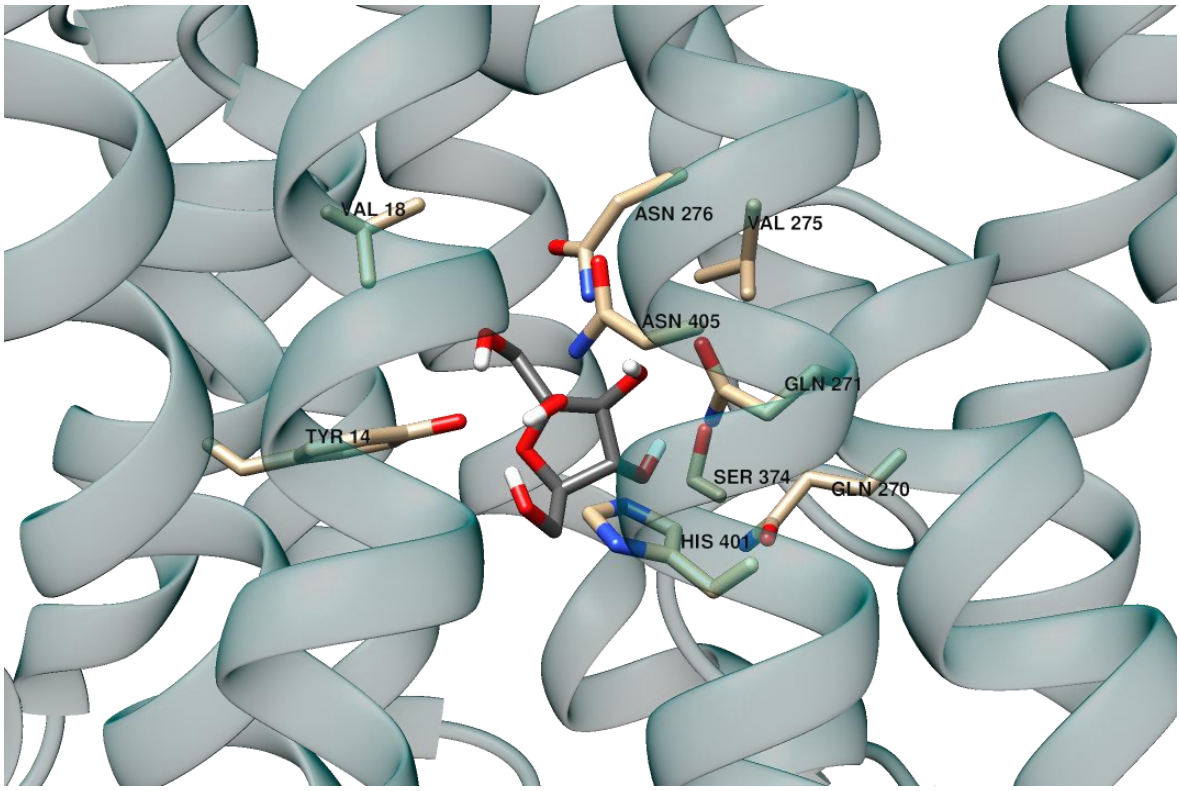

Figure 38: The docked structure of fructose with the interacting residues of the GLUT5 receptor labelled. The atom colours follow the same scheme as Figure 36.

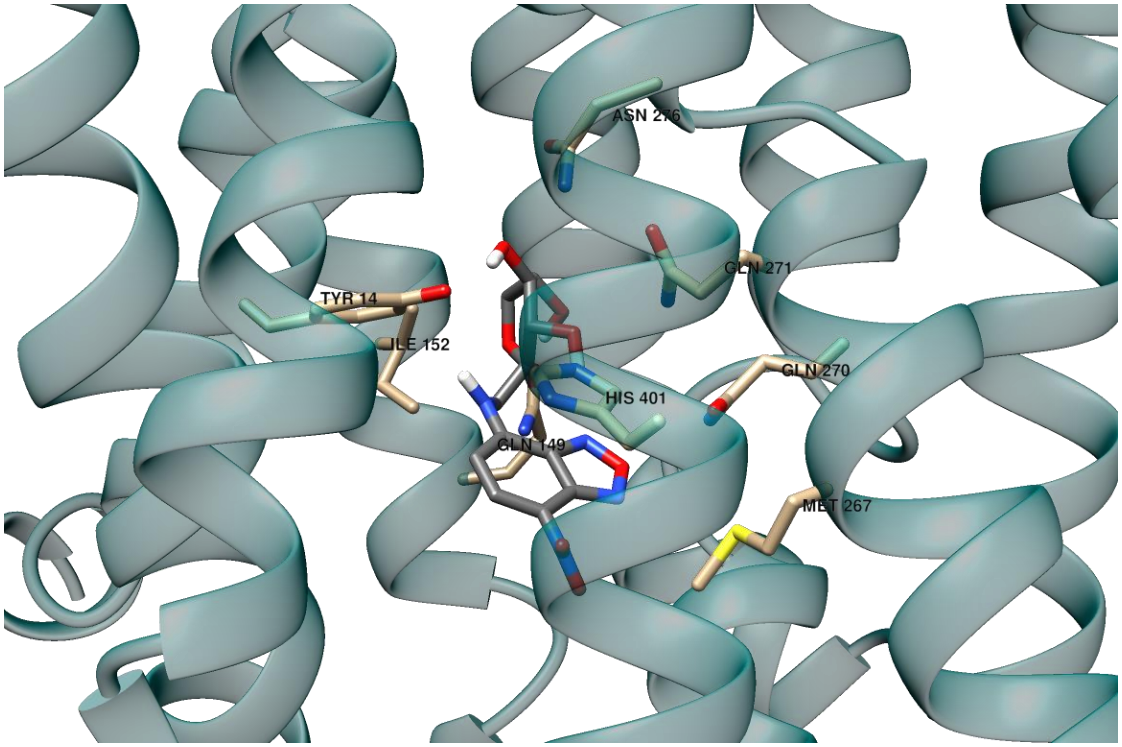

Figure 39: The docked structure of NDBM with the interacting residues of the GLUT5 receptor labelled. The atom colours follow the same scheme as Figure 36.

		[1]		[2]	[3]	[4]
Docked Compound	Estimated Free Energy of Binding* (kcal/mol)	VDW + H-bond +desolvated Energy (kcal/mol)	Electrostatic Energy (kcal/mol)	Final Total Internal Energy (kcal/mol)	Torsional Free Energy (kcal/mol)	Unbound System's Energy (kcal/mol)
Maltose	-7.28	-8.46	-0.01	-0.92	1.19	-0.92
Fructose	-3.37	-5.33	-0.12	-3.80	2.09	-3.80
NDBM	-6.11	-8.23	-0.27	-2.13	2.39	-2.13

Table 8: The different energy components of the lowest energy docking pose of each sugar ligand as calculated by Autodock.

4.4.2 Docking of the ManCou ligands

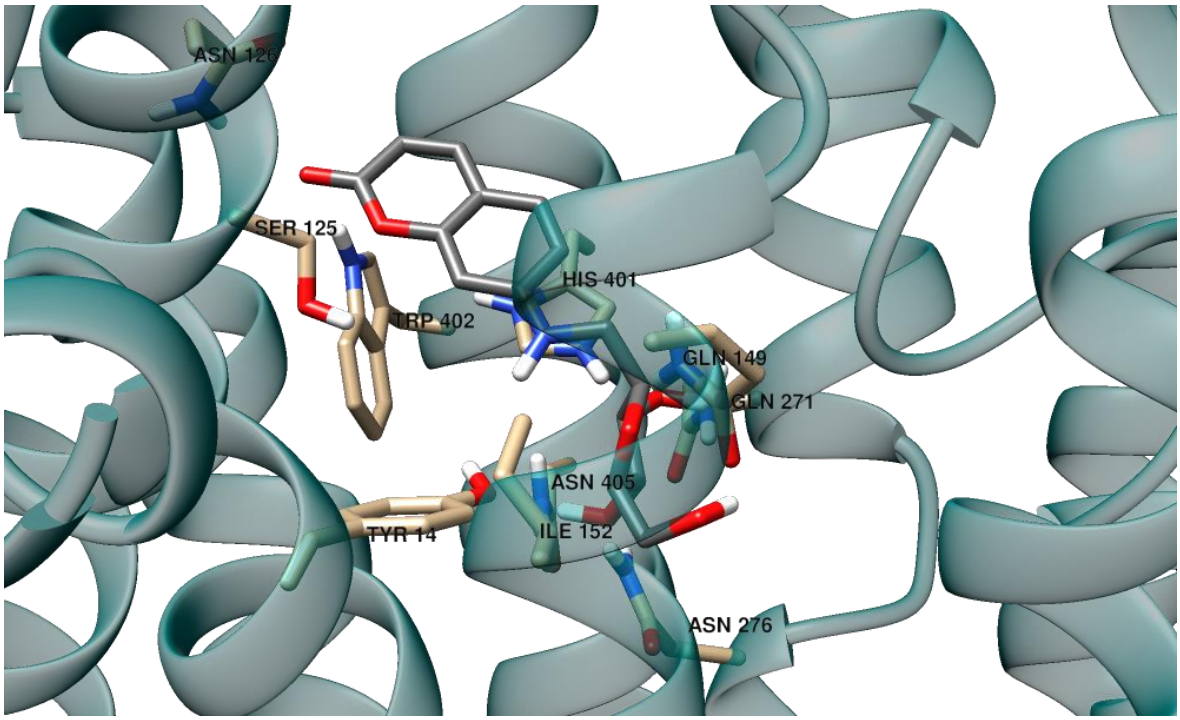

Figure 40: The docked structure of ManCou1 with the interacting residues of the GLUT5 receptor labelled. The atom colours follow the same scheme as Figure 36.

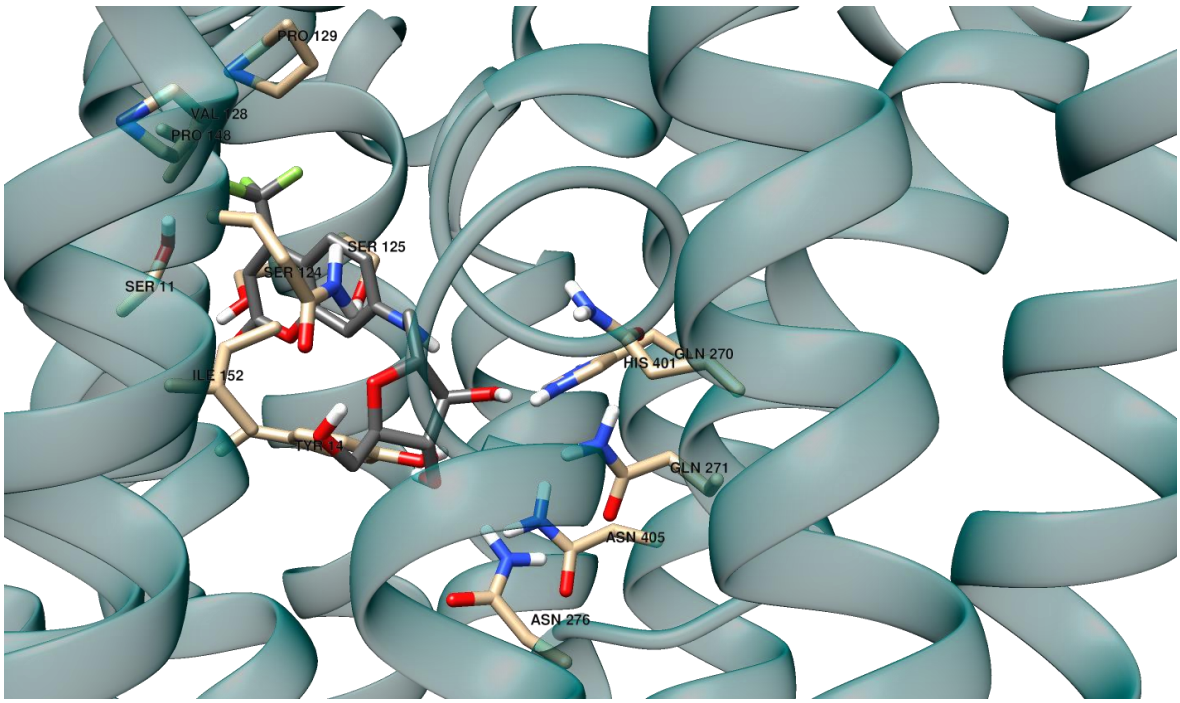

Figure 41: The docked structure of ManCou3 with the interacting residues of the GLUT5 receptor labelled. The atom colours follow the same scheme as Figure 36.

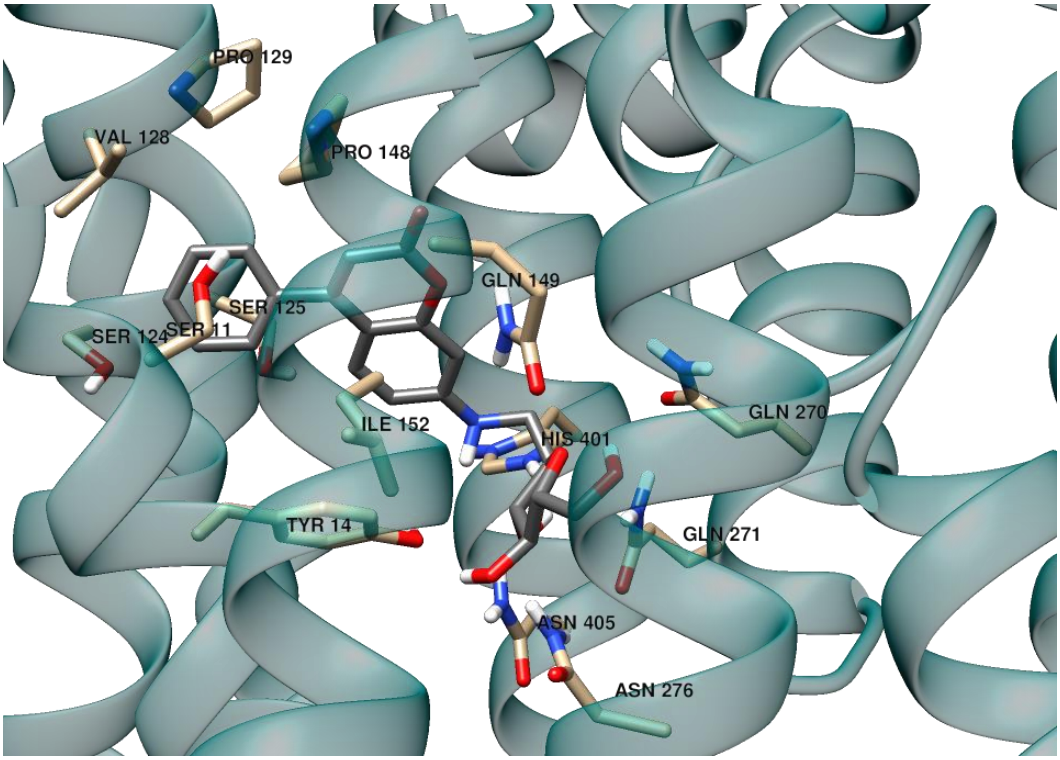

Figure 42: The docked structure of ManCou4 with the interacting residues of the GLUT5 receptor labelled. The atom colours follow the same scheme as Figure 36.

		[1]		[2]	[3]	[4]
Docked Compound	Estimated Free Energy of Binding* (kcal/mol)	VDW + H-bond +desolvated Energy (kcal/mol)	Electrostatic Energy (kcal/mol)	Final Total Internal Energy (kcal/mol)	Torsional Free Energy (kcal/mol)	Unbound System's Energy (kcal/mol)
ManCou1	-6.93	-8.79	-0.23	-1.33	2.09	-1.33
ManCou2	-7.71	-9.60	-0.20	-1.45	2.09	-1.45
ManCou3	-7.39	-9.61	-0.17	-1.44	2.39	-1.44
ManCou4	-7.85	-10.08	-0.15	-1.87	2.39	-1.87

Table 9: The different energy components of the lowest energy docking pose of each ManCou ligand as calculated by Autodock.

4.5 DISCUSSION

4.5.1 Docking of NDBM, Fructose and Maltose

The lowest energy binding poses of NDBM and maltose were found to form hydrogen bonds with Y14, Q149, Q270, Q271 and N276 (

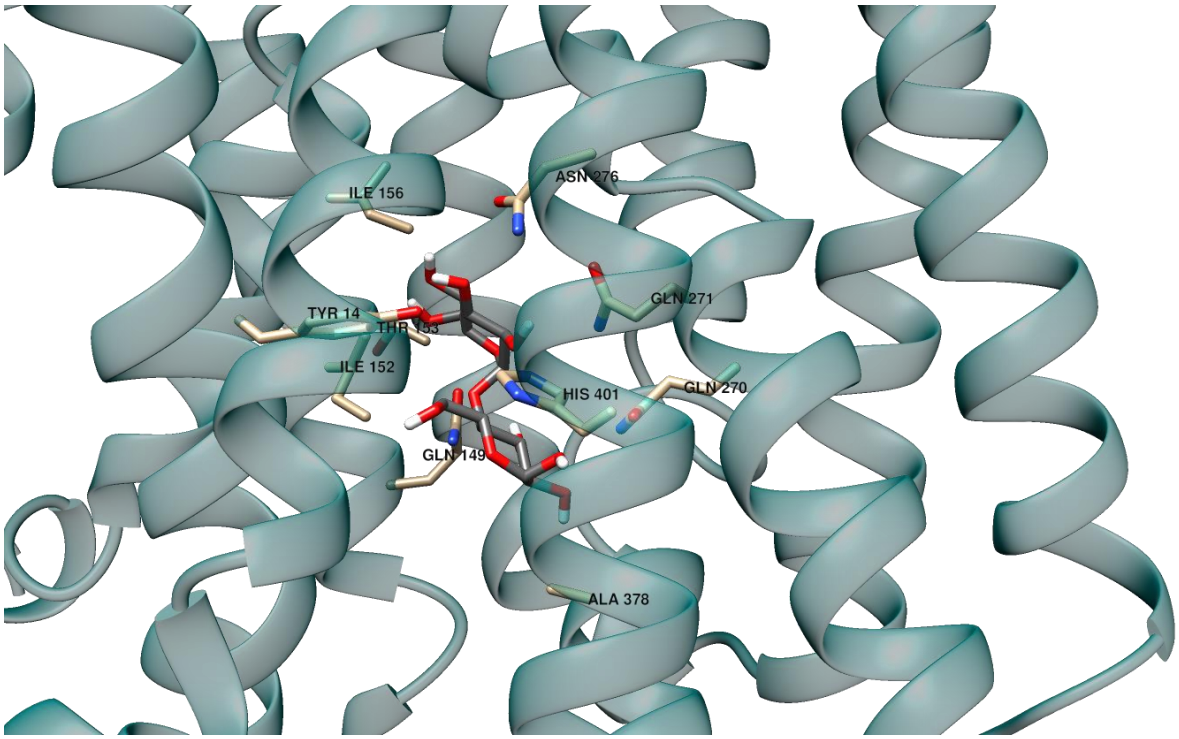

Figure 37 and Figure 38). The large number of hydrogen bond donors and acceptors in the structure of maltose and NDBM allow the two ligands to form several hydrogen bonds with the interior of the GLUT5 receptor. The only major difference in the binding of NDBM as opposed to maltose is the VDW interaction between the aromatic ring of NDBM and H401. The similarities between the binding modes of NDBM and maltose are reflected in their similar values for the estimated free energy of binding energy from Table 8.

Fructose makes a lot of the same hydrogen bonding interactions as NDBM and maltose, however it does not interact with Q149 due to its smaller size or H401 due to its lack of an aromatic ring moiety. The predicted binding site of fructose calculated by our docking is in good agreement with the binding site identified in the literature based on homology with similar GLUT transporter proteins [194]. The higher free energy of binding for fructose, may

be related to the function of GLUT5 as a fructose transporter, because of this we wouldn't expect it to strongly bind fructose. This allows fructose to easily dissociate from the transporter and move into the intracellular space.

4.5.2 Docking of the ManCou ligands

The lowest energy docked conformation of ManCou1 to GLUT5 (

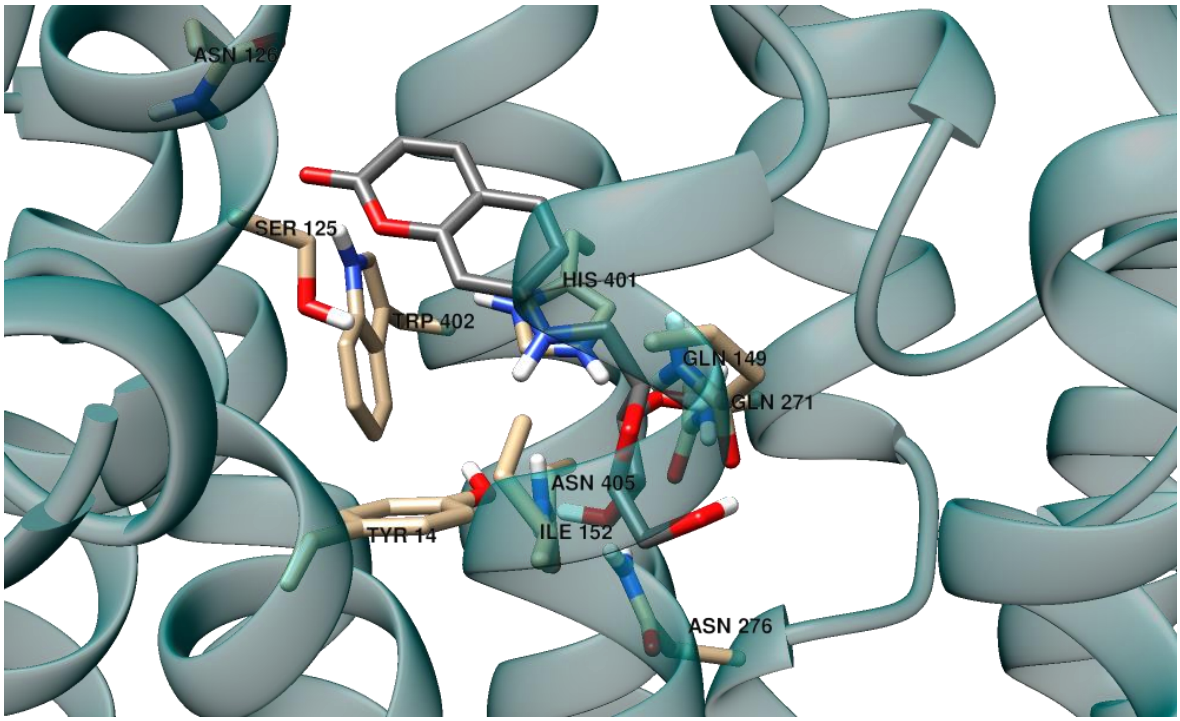

Figure 40) has the sugar moiety of forming hydrogen bonds with Q149, Q271, Q276, N405 and H401. This pose is distinct from the conformation found in the previous docking of fructose, this suggest to us that the addition of the coumarin moiety has influenced the binding position of ManCou1. The coumarin moiety of ManCou1 forms hydrogen bonding interactions with Y14, S125 and N126 and makes non-polar interactions with the sidechain of W402 and H401. The binding pose of ManCou3 is quite distinct from the binding pose of ManCou1, both the sugar and substituted coumarin moiety are found in subtly different places than the docked conformation of ManCou1 (

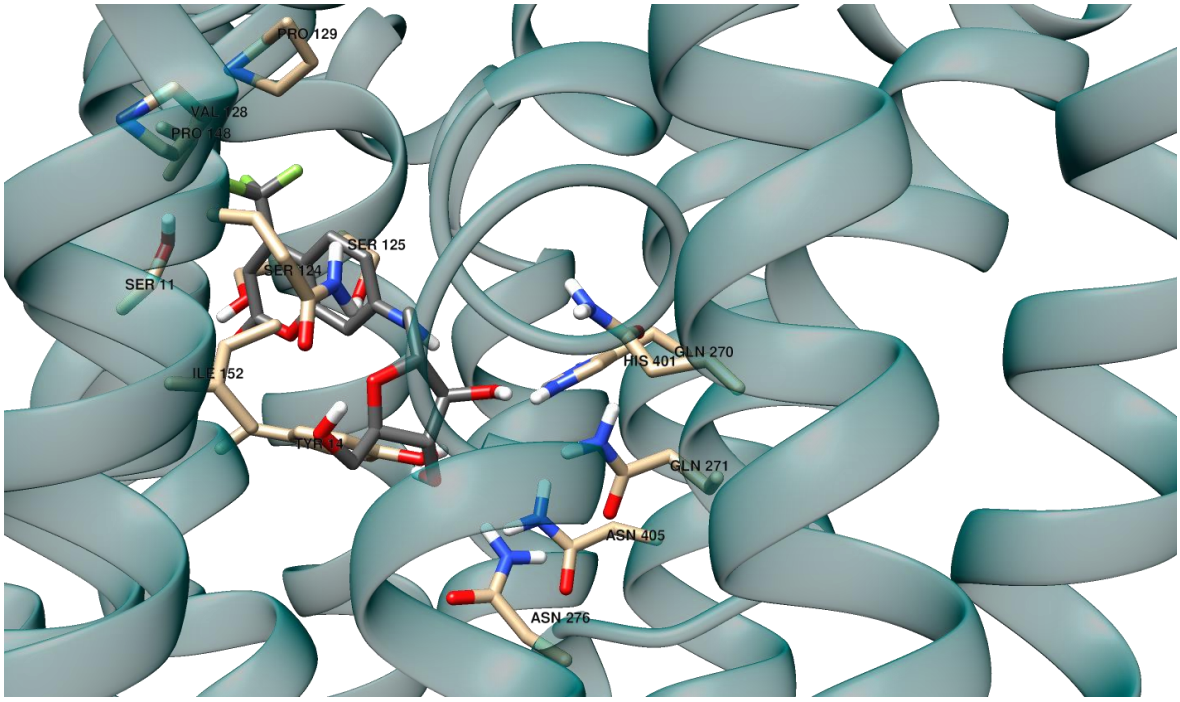

Figure 41). The sugar moiety makes hydrogen bonding interactions with Y14, Q270, Q271, N276, and N405. The substituted coumarin moiety likely forms hydrogen bonding interactions with S11, S124 and S125. The majority of the residues surrounding the substituted coumarin moiety of ManCou3 are hydrophobic residues such as Y14, P118, V128, P129 and I152, however no interaction with W402 or H101 are observed. The docked conformation of ManCou4 is more similar to that of ManCou3 than ManCou1, the substituted coumarin moiety seems to occupy a similar hydrophobic region as that of ManCou3 and the sugar moiety makes a lot of similar hydrogen bonding interactions (Figure 42). The sugar moiety of ManCou4 makes hydrogen bonding interactions with Y14, N149, Q270, Q271, N276, H401 and N405. The substituted coumarin moiety of ManCou4 doesn't appear to make any hydrogen bonding interaction and instead makes non-polar contacts with V128, P129, P148 and I152. The observed similarity of the interactions between ManCou3 and ManCou4 are reflected in their very similar estimated free enegy of binding in Table 9.

4.5.3 Glut5 Membrane MD Simulations

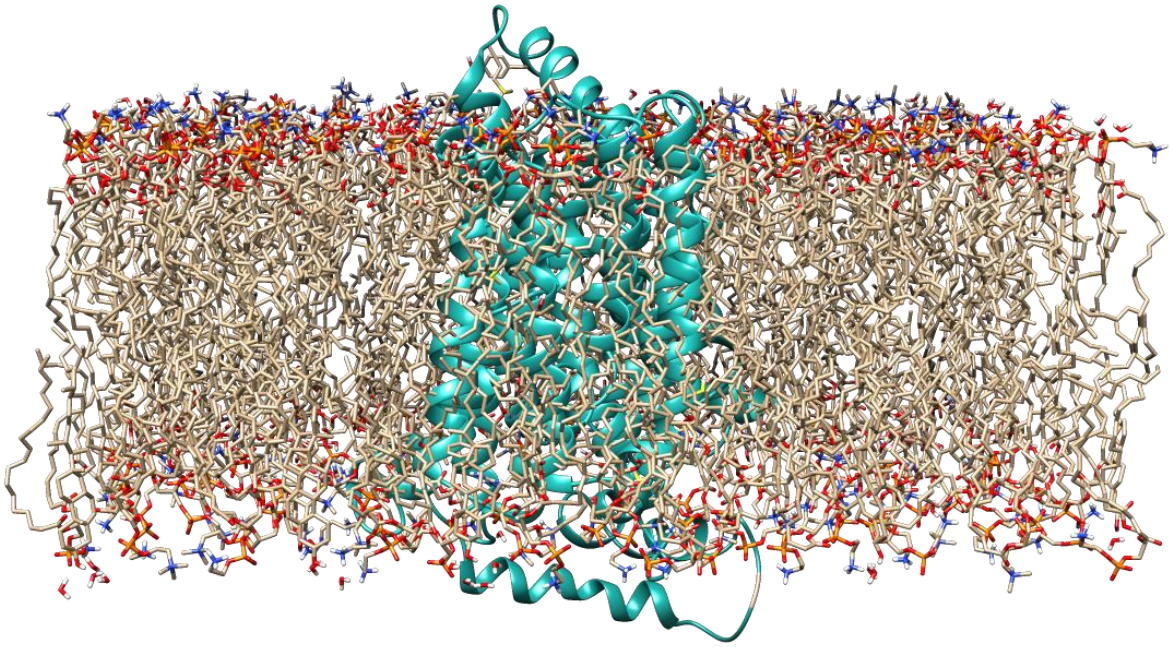

Figure 43: GLUT5 in a POPC lipid membrane model built using the CHARMM-GUI membrane builder tool. The atom colours follow the same scheme as Figure 36, with the addition of phosphorus which is coloured orange.

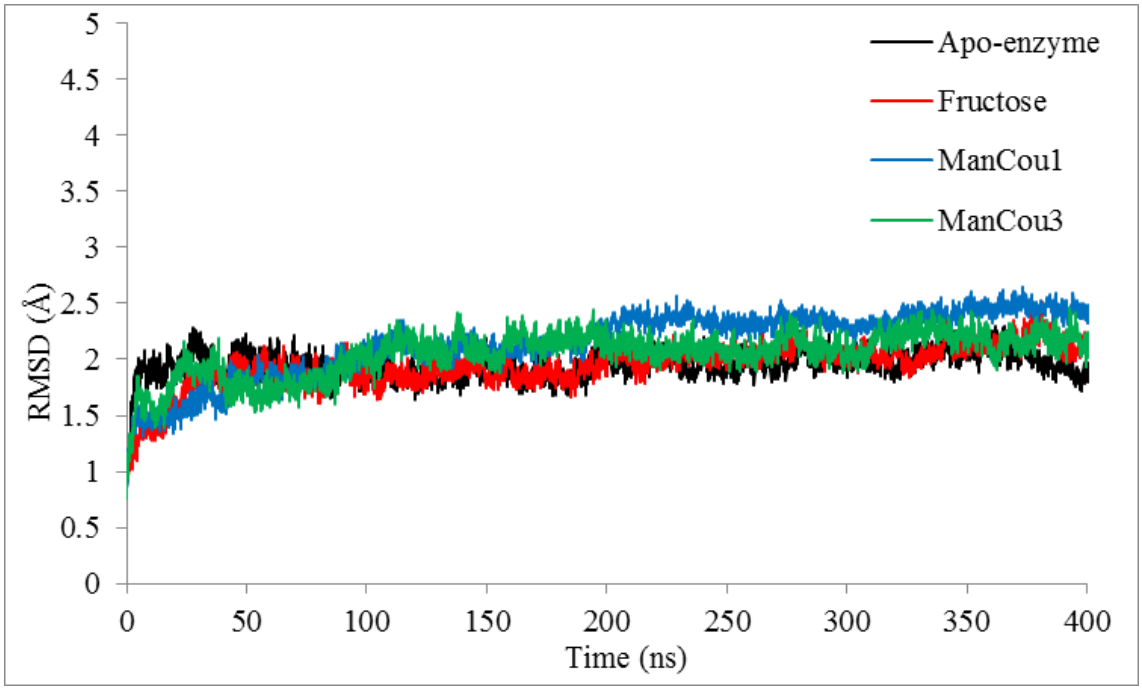

Figure 44: RMSD graph of the carbon-α alpha atoms of the four GLUT5 membrane simulations up to 400ns.

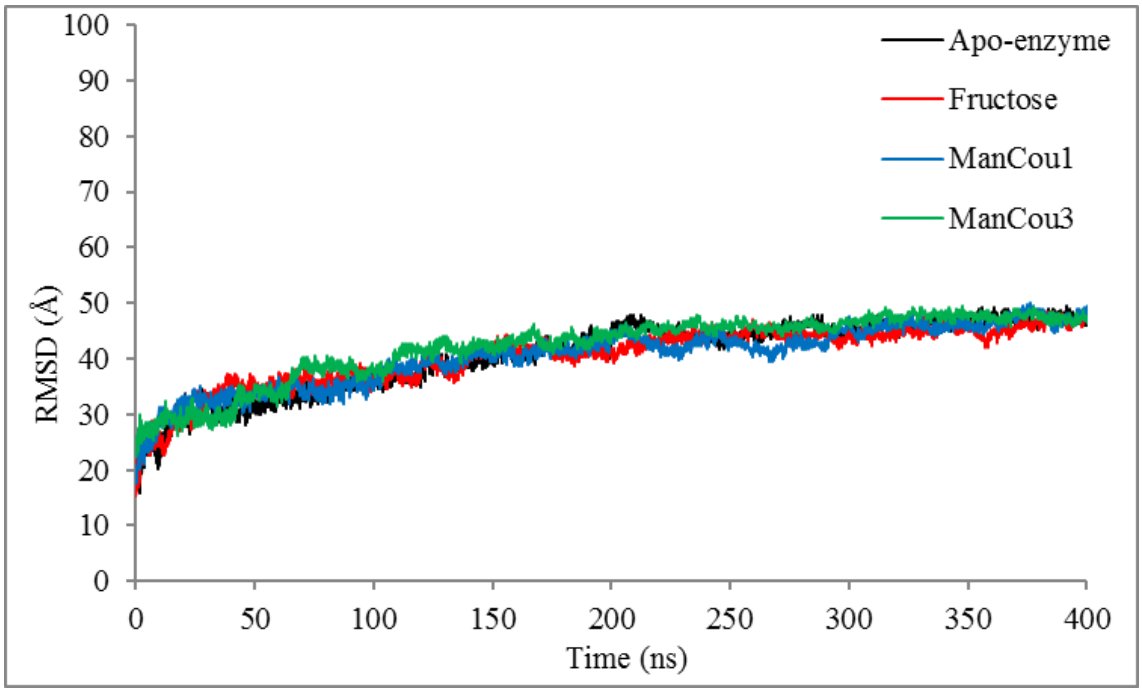

Figure 45: RMSD graph of the heavy atoms of the POPC membrane of the four GLUT5 membrane simulations up to 400ns.

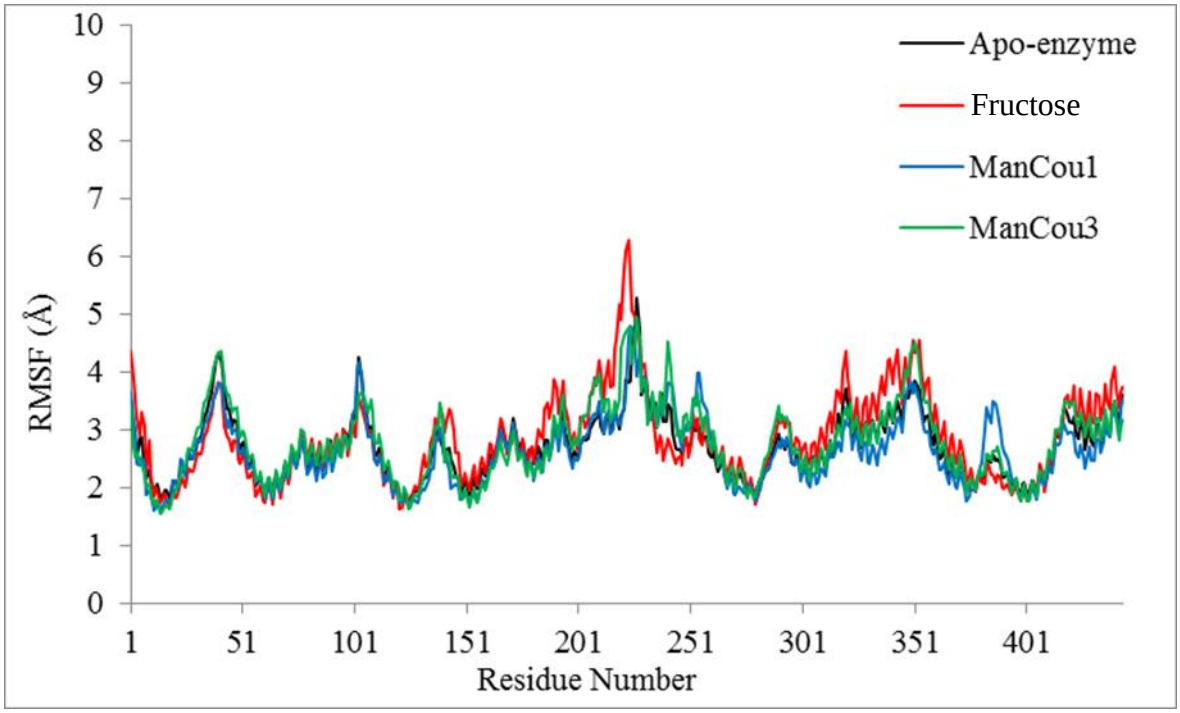

Figure 46: RMSF graph of the carbon-α alpha atoms of the four GLUT5 membrane simulations up to 400ns.

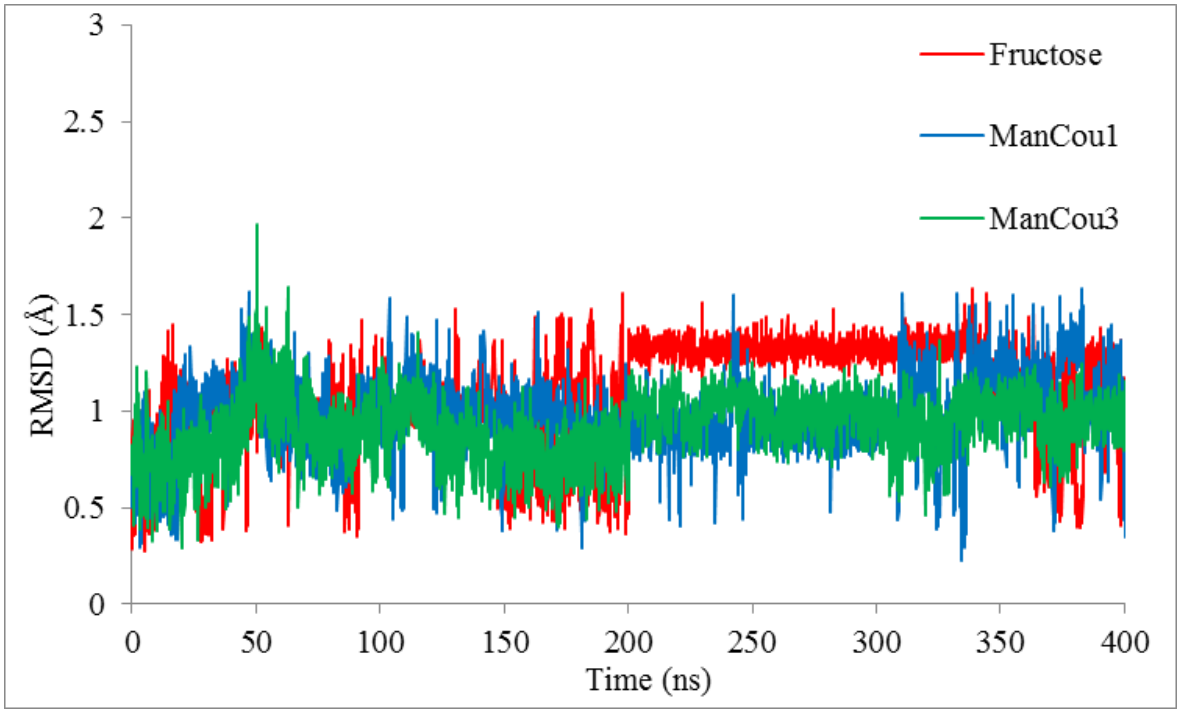

Figure 47: RMSD graph of the heavy atoms of fructose, ManCou1 and ManCou3 ligands up to 400ns.

4.5.4 Glut5 Membrane MD Simulations

The RMSD of the Cα atoms of the four MD simulations: the apo-enzyme, fructose bound, ManCou1 bound and ManCou3 bound is shown in Figure 44. The four MD simulations seem to equilibrate by about 100ns and with the exception of ManCou1 to have RMSD values of slightly below 2Å. The ManCou1 simulation seems to adopt a new distinct conformation after about 100ns, this conformation seems to be stable with relatively little fluctuation from the 2.5Å level. The RMSD profile of fructose and the apo-enzyme are the most similar while the RMSD profile of ManCou3 is alike it is found to be slightly more flexible than the other two. In general though the RMSD of all four simulations are quite low compared to other protein structures, this may be due to the stabilising effect that the membrane has on the protein structure lowering the receptors overall flexibility. Throughout all of the MD simulations the protein remains embedded in the lipid bilayer, this indicates that the system is stable. The RMSD plot of the heavy atoms of the lipid bilayer is shown in Figure 45, the lipid membrane seems to fluctuate even up to 250ns into the simulation. After 250ns, the fluctuations in the lipid RMSD do appear to level off but more simulation time is necessary to know if the lipid membrane has equilibrated fully. The RMSF plot of the Cα atoms of the four GLUT 5 membrane simulations are shown in Figure 46, overall the profiles of the RMSF plots are quite similar between the four MD simulations, this is likely correspond with much of the structure being embedded in the lipid membrane not allowing much structural flexibility. Fructose, despite having overall low RMSD values has some regions that have higher relative flexibility than the other three MD simulations. The areas of the protein with low RMSF values tend to part of the transmembrane helices, these areas are entirely enclosed and supported by the lipid membrane and have little conformational flexibility. The regions of high RMSF tend to correspond to the solvent exposed loop regions that connect the helices of the receptor, these regions have no defined secondary structure and are much more conformationally flexible. The particularly flexible area of residues

spanning 213-226 is a flexible region that was not present in the crystal structures, possibly due to its many possible conformations. This region (shown in Figure 48) is located between two intracellular helices, these helices are usually joined by salt bridges that function to hold the receptor in its outward open conformation, our MD simulation is based on the inward open conformation so these salt bridges are not present and the lack of inter helix electrostatic interactions possibly making the region more flexible. The RMSD of the heavy atoms of the ligands for the MD simulations that contain fructose, ManCou1 and ManCou3 are found in

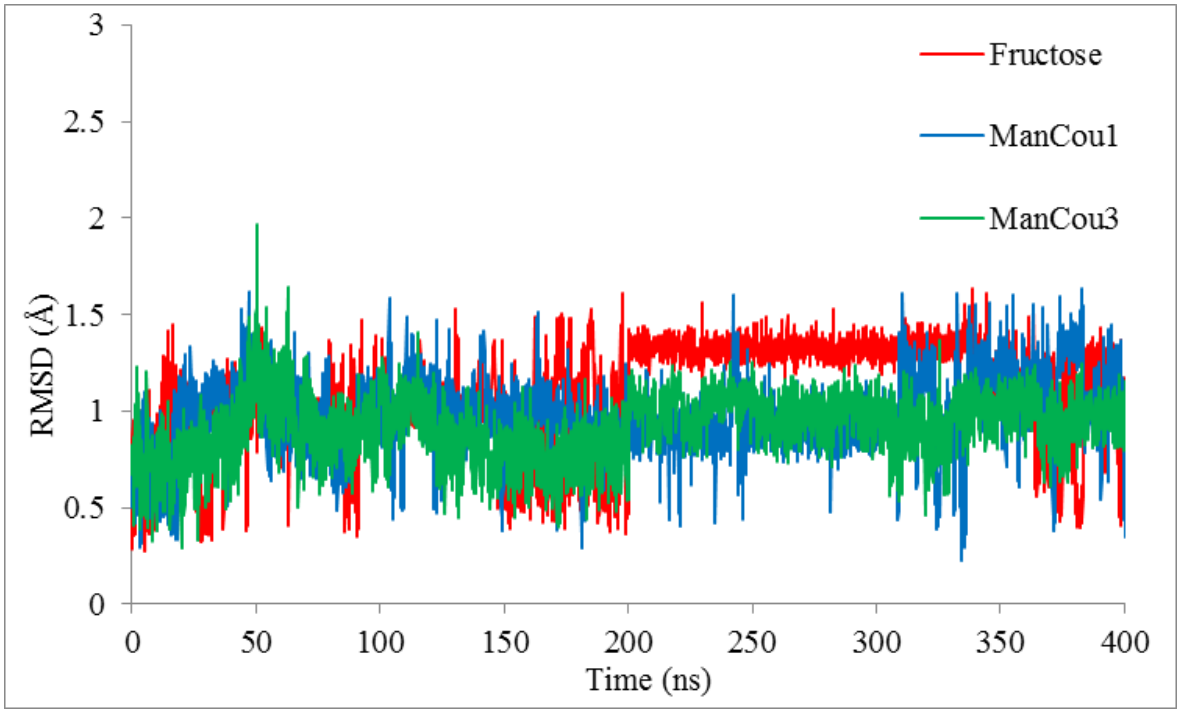

Figure 47. The positions of the ligands seem be highly variable within the receptor, fructose in particular seems to adopt a new conformation around 200ns, it then stay in this conformation for around 150ns before returning to its previous conformation again. This kind of fluctuation in ligand position may be indicative of a conformational change in the binding site causing fructose to adopt this new conformation, these kind of conformational changes have already been investigated and found to play a role in the function of GLUT1 transporter proteins [212].

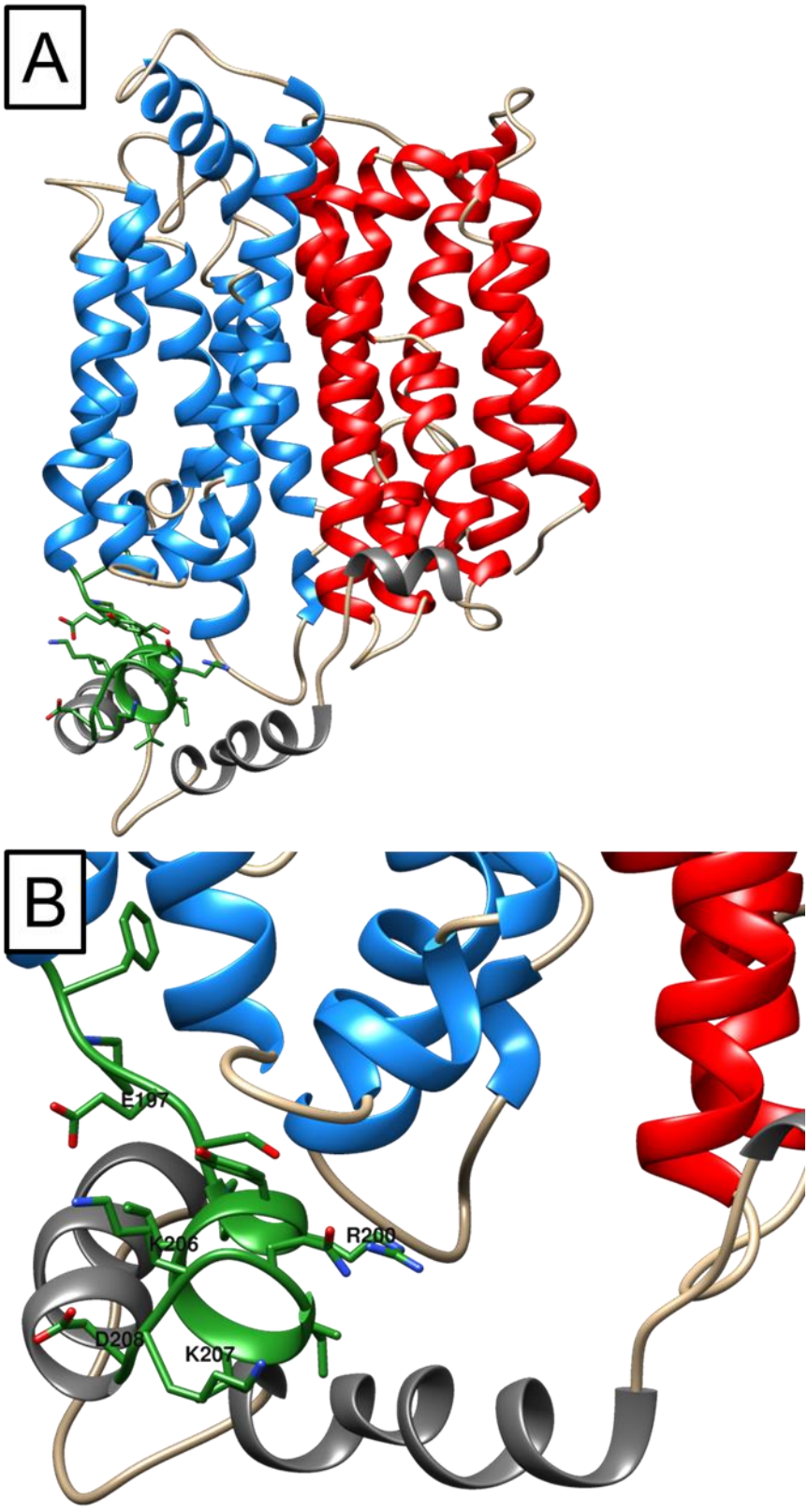

Figure 48: A) A rendering of the GLUT5 receptor with the C-terminal transmembrane bundle helices rendered as red ribbons, the N-terminal transmembrane bundle helices rendered as blue ribbons, the intracellular helices are rendered as grey ribbons, the flexible region of residues 213-226 is rendered a green ribbon and the loops joining the helices are coloured beige. B) Shows an

enlarged close up image entered on the flexible loop region 213-226, the sidechain atoms of the are also displayed and the important charged residues are labelled. Atom element colours are as follows; carbon is dark green, nitrogen is dark blue, oxygen is red and hydrogen is white.

4.6 CONCLUSIONS

Overall the docking of fructose, maltose and NDBM was found to align well with the experimental observations of GLUT5's role as a fructose transporter, where the low binding energy of fructose allows for it to pass through the receptor as intended. Whereas maltose, which is not biologically transported by the receptor has a high binding affinity indicating it would probably not be able to pass through the receptor. NDBM acting as a probe for GLUT5 would also bind to the receptor with a high affinity, this is reflected in its low docking energy indicating a relatively strong binding to the GLUT5 receptor. Additionally the binding site for fructose predicted by our docking, is similar to the binding site predicted by the X-ray crystallographic structure paper.

The docking of the ManCou ligands shows good agreement with the docking of fructose, in that the sugar moieties of the ManCou ligands occupy a similar spatial positioning in the GLUT5 receptor site as that of fructose. This positioning of the ManCou ligands is important to their role as GLUT5 probes, so that they will selectively bind to GLUT5 due to its preference for only fructose transportation. The initial results from the MD of the GLUT5 look promising. The ligands remain bound to the GLUT5 receptor and their positional movements inside the receptor may give insights into the structural changes occurring in the receptor in response to their binding. The RMSD of the C-alpha remains relatively stable and appears to have equilibrated. The RMSD of the POPC membrane also appears to have equilibrated or be in the process of equilibrating.

4.7 FUTURE WORK

Although the current docking phase of the project is completed, new ManCou probes are being researched for which new docked structures will be generated and possibly lead to additional MD simulation studies as well.

The MD simulations of GLUT5 are still incomplete and will eventually reach the 1μs timescale, thorough atomistic analysis of the resultant trajectories will be performed as well as conformational analyses. The free energy of binding between the ligands and the GLUT5 receptor structure will be calculated using the MMGBSA method [216].

Given the role of the GLUT5 receptor in sugar transport, it would be of interest to model the transit of different sugars and ManCou probes through the transporter. For structural changes occurring at long timescales non-conventional MD methods can be used such as accelerated MD and steered MD could be utilised, these methods have recently been utilised in modelling the mechanism of sugar transport in GLUT1 receptors [212, 213].

In-silico mutagenesis studies could allow us to study the contributions of individual residues on the protein's structure/function, ligand binding and conformational dynamics. Several experimental mutant structures have been created, some of these mutants have altered the selectivity of the receptor from fructose to glucose [194]. Using MD simulations to study these mutant structures would allow us to gain insight into why the selectivity of the receptor is changed at an atomistic level in specific mutant forms of the receptor.

5 MODELLING STUDY OF THE BINDING OF MONOSACCHARIDES TO ODORRANALECTIN AND TWO SYNTHETIC LECTINOMIMICS

5.1 PREFACE

This chapter describes work done in collaboration with experimental researchers that led to the publication of a paper entitled: “Targeting cancer-specifc glycans by cyclic peptide lectinomimics”, published in August 2017, the full paper is included in the appendix of the thesis, and gives full credit to all of the authors. The chapter contains an overview of the whole study, a summary of the full methodology and results of the work, and a more detailed description of the computational work contributed to the research. The docking study was undertaken to study the interactions of odorranalectin and two artificial lectinomimics with a range of sugar molecules, these structural insights could then be used to explain experimental results relating to the affinity of the three lectins to the studied compounds.

5.2 INTRODUCTION

Lectins are a sub-set of proteins and peptides that are dedicated to the binding of carbohydrates, found ubiquitously throughout nature, they are produced by nearly all forms of unicellular and multicellular life as well as some viruses [217, 218]. They recognise and bind free sugar molecules as well as the sugar moieties found in glycolipids and glycoproteins, usually in a highly specific manner [219]. Lectins fulfil important roles in cell adhesion, cell signalling and recognition [220]. In certain cases the binding of lectins to sugar moieties can act as a defence mechanism in the immune systems of animals and plants as well as their widespread utilization as toxins in both eukaryotic and prokaryotic organisms [218, 221]. Given the enormous variety of lectins that exist, only a fraction of the total amount have had their structures resolved [222, 223]. Newly characterised lectins are interesting lead compounds for the development of new drugs, their ability to recognise specific sugars, glycolipids or glycoproteins means they have the potential to be targeted agonists/antagonists for use in new treatment and detection strategies [223, 224]. In healthy cells glycosylation, the process of adding sugar moieties to lipids and proteins, serves a variety of functions. The process of glycosylation often occurs as a post-translational modification to proteins and enzymes. Glycan based modifications can help to fold proteins, provide them with structural support, add new functionality, or allow the protein to be recognised by specific lectins or antibodies [217].

The role of glycosylation in cancer has become increasingly studied in recent years, with abnormal glycosylation being recognised as not only an indicator of cancer but playing a key role in tumour development [225, 226]. Tumour cells produce unique glycosylated structures not found in healthy cells, some of these structures are not only indicative of the aberrant metabolism of cancer cells but are directly linked with the cancer's survival, proliferation and metastasis [227]. These unique tumour-associated glycans can be used as

markers to predict the development of cancer before other symptoms have developed, earlier diagnoses generally lead to better outcomes due to the increased effectiveness of treatments acting on early stage tumours [228]. An example of glycans as tumour markers is the detection of alpha fetoprotein in high levels in adults, this glycosylated protein is usually only found in high amounts in the developing foetus. High amounts of glycosylated alpha fetoprotein in adults is associated with liver malfunction [228, 229]. However, high serum levels of alpha fetoprotein alone are not always indicative of cancer and can be caused by liver cirrhosis or hepatitis C infection. Specifically though detection of the fucose glycolsylated form of alpha fetoprotein (AFP-L3) using the lectin *Lens agglutinin* is a proven diagnostic method specifically for the detection of hepatocellular carcinoma [230, 231]. Another developing therapy based on targeting cancer specific glycans is based on interfering with the ability of metastasising tumour cells to adhere to surrounding tissues [217, 232]. Anti-metastasising agents can be used along-side conventional treatments to reduce the likelihood that a tumour will invade other tissues to form secondary tumours. Some of these agents work by binding to the sugar moieties of the target cancer specific glycans responsible for cell adhesion and migration preventing them from attaching to the leucocytes and endothelial cells that are known to be responsible for facilitating the development of metastases [233]. Some plant based lectins have shown promise as anti-tumour agents, by binding to certain cell surface receptors the lectins can directly trigger the programmed cell death pathways of the tumour cells [234]. A lectin isolated from the seeds of a specific variety *Phaseolus coccineus* (a leguminous plant) was found to selectively bind to sialic acid conjugates, increased amounts of sialic acid is a known hallmark of cancer specific glycans not abundantly found in healthy cells [235]. The lectin was shown to bind in response to the elevated levels of sialic acid and induce apoptosis in L929 fibrosarcoma cells [236]. Antibodies and lectins are ideally suited to selectively binding these adhesion and migration related cancer specific glycans, examples of lectins and antibodies utilizing

this strategy are currently under development and can be found throughout the literature [234, 237-241].

Development of other glycan targeting strategies for cancer detection and treatment is developing as more and more cancer specific glycans are discovered [227], approaches either focus on creating antibodies or lectins that are specific to tumour-associated glycans. The antibody biased approach is hampered by problems in activating an immune response to the glycan molecules needed to produce new antibodies, the glycans seem to make poor immunogens and antibodies derived in this way exhibit poor binding affinities [242]. Of the available antibodies known to bind to glycans it can be difficult to determine their exact specificities, as well as binding to the target glycan the antibody often binds to other off-target glycans potentially will undesirable effects [243]. Lectins suffer from similar problems with a lack of known specificity producing off-target effects, in addition they are often inherently unstable and can denature or be broken down by enzymes in the body reducing their effectiveness as drugs [244]. Artificially modified lectins, lectinomimics offer an alternative solution, by modifying the structure of natural lectins it is hoped that the stability and selectivity of the lectinmimics can be improved to the point where they might be considered as effective agents for binding cancer specific glycans, for use as diagnostic probes and potential anti-cancer therapies [245, 246].

A small 17 amino acid cyclic peptide (Odorranalectin) isolated from the skin of a frog found throughout South Asia (*Odorrana grahami*), was found to possess lectin-like properties with a binding preference to L-fucose [247]. The peptides structure was solved by NMR spectroscopy (PDBID2JQW) and found to adopt a β-turn-like conformation with a disulphide bridge linking residues C6 and C16. Odorranalectin is the smallest peptide found to have selective sugar binding affinities, this makes it of particular interest to drug development [246, 247]. Large peptides and proteins are undesirable as lead compounds in drug development due to their potential to evoke an immune response, their inability to be

administered orally, poor tissue penetration, and their rapid breakdown by endogenous enzymes [248]. Odorranalectin is a small peptide known not to provoke an immune response, and is being investigated for use as a component in drug delivery systems designed to cross the blood brain barrier to carry drug payloads to specific targets [247, 249]. Odorranalectin could also be used as a starting point for the development of lectinomimics for the purposes of binding to cancer specific glycans to create tumour sensing probes as well as new potential anti-cancer agents through the defined mechanisms described above. With this aim in mind our experimental collaborators designed and synthesised two novel Odorranalectin lectinomimics, replacing the disulphide bridge in Odorranalectin with a lactam bridge to improve stability whilst maintaining a similar geometry to that of the native lectin.

The experimental portion of the study aimed to assess: (1) if the modifications made to the Odorranalectin structure affected its conformational structure and therefore its ability to bind to specific monosaccharides; (2) how the modifications to the Odorranalectin structure would affect its in vitro toxicity to healthy cells; (3) how Odorranalectin and the two lectinomimics would be broken down by human serum and what products would be formed; (4) if the ability of Odorranalectin and the two lectinomimics to bind to specific monosaccharides could potentially be used as a probe to distinguish different types of cancer cells from healthy cells based on their cell surface glycosylated structures; (5) if the binding to cancer specific glycans by Odorranalectin and the two lectinomimics could have an anti-metastasising effect on certain human cancer cell lines in-vitro. Our molecular docking experiments would attempt to provide insight into the atomistic details of the binding between Odorranalectin and different monosaccharides known to be found as conjugates in the glycosylated cell surface structures. Additionally, docking of the modified lectins to a known natural substrate of the natural lectin in order to investigate how the lactam bridge moiety affected substrate binding.

5.3 SUMMARY OF EXPERIMENTAL METHODS

Full description of the experimental methods can be found in the paper attached in the appendix, a summary of the experimental methods to give context to the aim and findings of the overall publication is given below.

The two synthesised lectinomimics were distinguished from one another by the orientation of the lactam bridge, in Synthetic Peptide (SP2) the modified amino acid diaminopropionic acid (Dap) replaces cysteine on residue 6 and residue 16 is replaced by aspartate. The sidechains of residue 6 and 16 can then be cross-linked to produce the lactam bridge. In Synthetic Peptide (SP3) Dap replaces cysteine on residue 16 and aspartate on residue 6, the two substituted residues are then linked to produce a lactam bridge between residues 6 and 16. A linear control peptide with the same sequence as Odorranalectin but no disulphide linking was also produced. Fluorescently labelled analogues of Odorranalectin, SP2 and SP3 were also synthesised for use in fluorescence based binding assays, these were created by linking the peptides to fluorescein. A polyethylene glycol linker molecule was added to link the amino end of the peptides to the fluorescein moiety, this was done to increase the solubility of the compounds and minimise any potential steric effects caused by the bulky fluorescein moiety on the binding of the fluorescently labelled analogues to other molecules. The conformations of the synthesised peptides were studied using Circular Dichroism (CD) spectroscopy. Additional assessment of the conformations adopted by the synthesised peptides was performed using Macromodel 9.9 [250] for Maestro [251].

The fluorescently labelled analogues of Odorranalectin, SP2 and SP3 were assayed with the Bovine Serum Albumin (BSA) conjugated monosaccharides: l-fucose, d-galactose, d-glucose, N-acetyl-d-neuranimic acid, N-acetyl-d-galactosamine and N-acetyl-d-glucosamine. The binding affinities of each of the lectins to each of the BSA-conjugated sugars could then be measured with the fluorescence intensity. A similar fluorescently

tagged control lectin (*Aleuria aurantia* lectin) with a known fucose specificity was assayed in a similar fashion to act as a comparison against Odorranalectin, SP2 and SP3.

The binding of Odorranalectin, SP2 and SP3 against the model glycoproteins fetuin and asialofetuin was assessed with Isothermal titration calorimetry. The structures of these two glycoproteins are highly similar, fetuin is distinguished from asialofetuin by the terminal sialic acid moieties found on its glycan chains. This assay allowed for measurement of the binding affinities of the three peptides with glycoproteins which have well-defined glycan structures.

The stability of Odorranalectin, SP2 and SP3 in human serum was tested by incubating the three peptides at 37°C in a human serum solution, samples were taken at different times (0, 45min, 2h, 4h, 8h and 24h) these samples were analysed with high performance liquid chromatography and mass spectrometry.

The binding specificity of the fluorescently labelled analogues of Odorranalectin, SP2 and SP3 to bind cell surface glycosylated structures was measured using a fluorescence staining assay with a several varieties of human cancer cell cultures as well as a control of healthy human skin fibroblasts. Comparison fluorescence assays of other fluorescently labelled lectins, *Aleuria aurantia* lectin, *Sambucus nigra* lectin and *Ulex europaeus I* with a variety of human cancer cell lines were also performed. A competitive binding assay to compare the binding affinity of fluorescently labelled Odorranalectin vs unlabelled Odorranalectin to the MCF-7 (human breast epithelial Adenocarcinoma) was also performed to assess if the fluorescent labelling of the lectin structure affected its ability to bind to the cell surface glycosylated structures. The *in vitro*toxicity of Odorranalectin, SP2 and SP3 towards healthy human skin fibroblasts at a range of concentrations was tested with a luminescence based cell viability assay.

A model to test the ability of Odorranalectin, SP2 and SP3 to bind to cell surface cancer specific glycans and act as cell migration inhibitors was performed with a transwell

migration assay of three metastatic human breast cancer cell lines and an additional mouse breast tumour model cell line.

5.4 COMPUTATIONAL METHODS

Twenty odorranalectin NMR structures were downloaded from the PDB (PDBID 2JQW) and ensemble clustered using UCSF Chimera [131]. Eight of these structures were grouped in the first cluster, and model number 10 was selected as the most representative for the average odorranalectin structure. Structures of the four sugar ligands were created using Gaussview [167] and optimized using Gaussain09 [94] with the Hartree–Fock method and 3-21G basis set. The optimized sugar structures were rigidly docked to the lectin peptide using Autodock 4 [42]. The defined site for the docking search was based on the binding site from the NMR structure. The initial positions and orientations of the ligands were randomized. Of the ten docked structures, only the lowest binding energy values for each ligand are given (Table 10 and Table 11) and the corresponding structures were visualized in Maestro [251] (Figure 50 and Figure 51). The lactam bridge of the modified lectin structures SP2 and SP3 was created by modifying the disulfide bridge between C6 and C16 to become a lactam bridge using Gaussview [167], the structures were then minimised. The two lactam bridged peptides and the original lectin also underwent docking with L-fucose using Autodock 4 [42] to produce 50 structures of lectin-ligand complexes. All images of molecular models in this chapter were created using Maestro [251].

5.5 SUMMARY OF EXPERIMENTAL RESULTS AND CONCLUSIONS

The CD spectra showed that SP3 mainly adopts a similar conformation to that of Odorranalectin in solution, the signal of SP3 however did indicate the presence of other

conformations not similar to that of Odorranalectin. SP2 gave CD spectra results that were indicative of a disordered structure and was found to be similar to that of the linear control peptide. The computational conformational prediction of Odorranalectin agreed with the findings of the CD spectra, only one conformational cluster was found, with no alternative conformations being predicted. The computational structure prediction indicated that SP3 was more likely to adopt a conformation that was similar to Odorranalectin than SP2.

The assay to test the binding affinity of the BSA conjugated monosaccharides to the fluorescently labelled Odorranalectin showed a strong preference to l-fucose with lesser affinity to both galactose and N-acetyl-d-galactosamine. As expected the fluorescently tagged control lectin AAL only showed a binding interaction to l-fucose. The ITC measured assay of Odorranalectin, SP2 and SP3 with the model glycoproteins fetuin and asialofetuin showed similar binding affinities of both Odorranalectin and SP3 with a much greater affinity for asialofetuin over fetuin. The lower affinity of fetuin to Odorralectin and SP3 may be caused by the terminal sialic acid moiety of the glycan chains blocking the binding of the lectins. SP2 was found to have no detectable binding to the two model glycoproteins.

The results of in vitro the cell toxicity tests showed that Odorranalectin was considered somewhat toxic to health human skin fibroblasts at relatively high concentrations (2mg/L). Surprisingly the two lectinomimics SP2 and SP3 showed no appreciable in-vitro toxicity at even the highest tested concentrations.

The human serum stability tests showed that all three peptides experienced some level of proteolytic degradation, Odorranalectin and SP3 were however more to this resistant than SP2. It was also found that SP2 was more likely to suffer from rapid cleavage of linear residues 1-5 whereas Odorranalectin and SP3 only suffered cleavage of residues A1 and Y2. The resulting metabolites of A1 and Y2 cleavage in Odorranalectin and SP3 were also found to degrade much more slowly than any of cyclic metabolites of SP2.

The staining assays of the fluorescently labelled analogues of Odorranalectin SP2 and SP3 with human cancer cell lines showed that Odorranalectin, SP3 exhibited similar binding profiles to one another. However, the fluorescence intensity of the SP3 results were less intense than those of Odorranalectin. SP2 was found not to bind to any of the tested cancer cell lines. Odorranalectin was shown to produce a positive binding fluorescent reaction to T-47D, MCF-7, HEP-G2 and MDA-MB-231 cell lines but no binding reaction to the control healthy human skin fibroblasts. This is an encouraging result for the use of Odorranalectin to detect cancer associated cell surface glycan in vitro with a similar profile to AAL with the only notable exception being the MDA-MB-231 cell line for which AAL produced only a weakly fluorescent result. The competitive binding assay of the MCF-7 cancer cell line with fluorescently labelled Odorranlectin and native Odorranalectin showed that, the native form displaces the fluorescently labelled form, this means the native Odorranalectin has a greater binding affinity for cancer specific cell surface glycans than the fluorescent labelled form.

The results of the transwell cell migration assay showed both Odorranalectin and SP3 do possess similar antimetastatic abilities against MDA-MB-231 and 231mfp human cancer cell lines, they achieve max inhibition at concentration levels of 50μg/L, increase of concentration to 100μg/L showed no appreciable improvement. The two lectins showed a greater inhibitory effect on the human 231mfp cell line, but only a minor effect on the migration of the mouse 4T1 cell line. These results indicate that the three metastatic cancer cell lines are having their cell mobility inhibited by the binding of Odorranalectin and SP3 to their cell surface glycan, however all three cell lines to lesser degrees have alternative non-inhibited mechanisms for cell migration not affected by the lectins.

5.6 DOCKING RESULTS AND DISCUSSION

Figure 49: The 20 NMR structures of Odorranalectin taken from the pdb structure of 2JQW. The peptides are displayed in ribbon view with each structure being randomly assigned a different colour to show the structural variance between them.

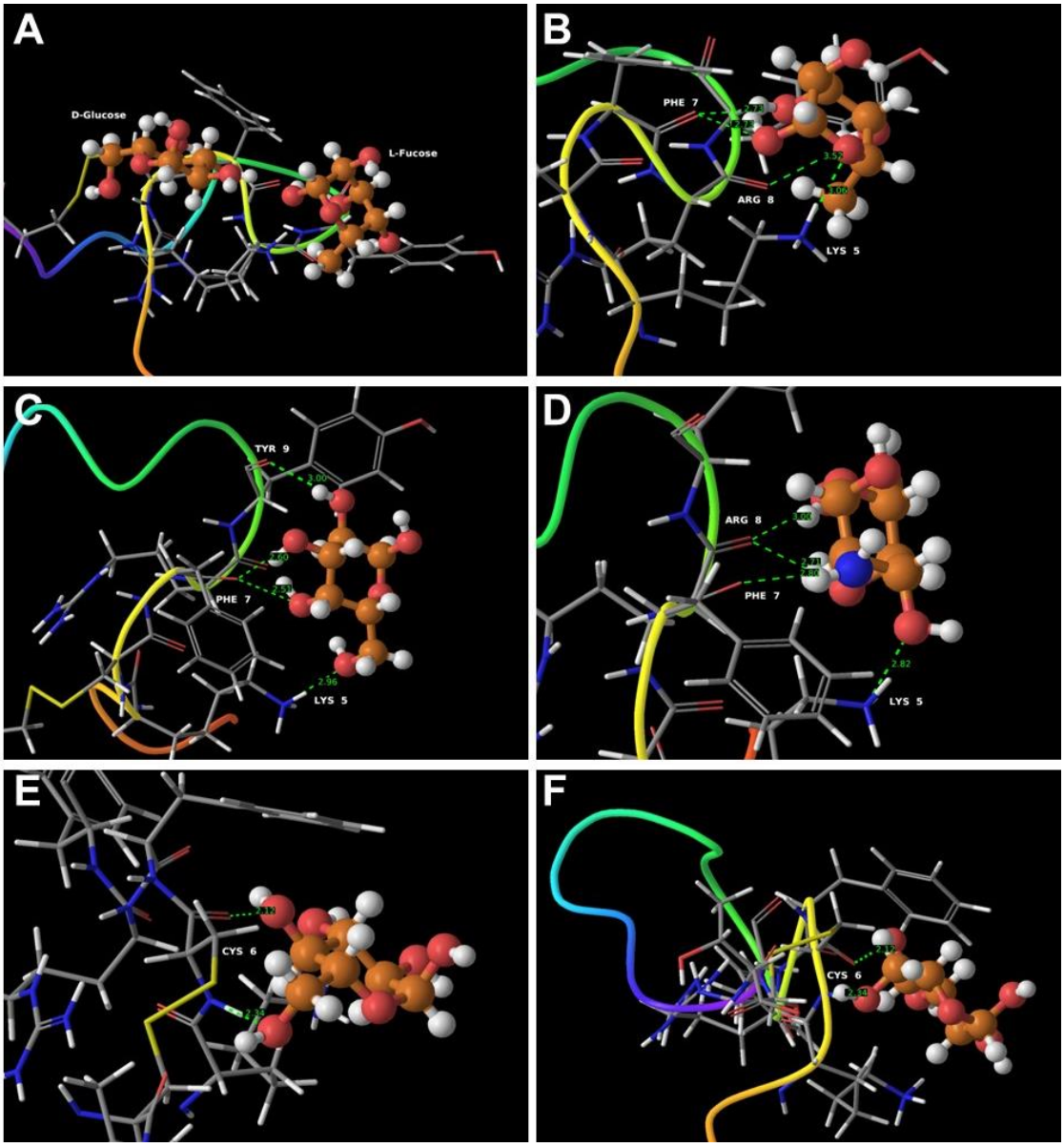

Figure 50: The backbone ribbon of the peptide is shown as a rainbow tube, the binding residues of the peptide are shown in a sticks rendering with C, O, N, H, S coloured grey, red, blue, white and yellow. The sugar molecules are shown in CPK rendering with the C atoms coloured in orange, the rest of the atoms are coloured the same as the peptide binding residues. Potential hydrogen bonds and the distance between donor and acceptor are shown in a light green. A) Highlights the difference in the binding sites for L-Fucose and D-Glucose. B) The binding site for L-Fucose. C) The binding site for D-Galactose. D) The binding site for D-Galactosamine. E) The binding site for D-Glucose. F) An Alternative view of the binding site for D-Glucose.

		[1]		[2]	[3]	[4]
Sugar Compound	Estimated Free Energy of Binding* (kcal/mol)	VDW + H-bond +desolvated Energy (kcal/mol)	Electrostatic Energy (kcal/mol)	Final Total Internal Energy (kcal/mol)	Torsional Free Energy (kcal/mol)	Unbound System's Energy (kcal/mol)
D-Galactosamine	-2.32	-4.10	-0.35	-2.93	+1.79	-2.93
D-Galactose	-2.52	-3.89	-0.41	-3.62	+1.79	-3.62
D-Glucose	-1.50	-3.26	-0.02	0.00	+1.79	0.00
L-Fucose	-2.60	-3.43	-0.36	-2.53	+1.19	-2.53

Table 10: Table of the energy components for the docking of the different sugar compounds to Odorranalectin where the *The estimated free energy of binding is defined as: [1]+ [2] + [3] - [4].

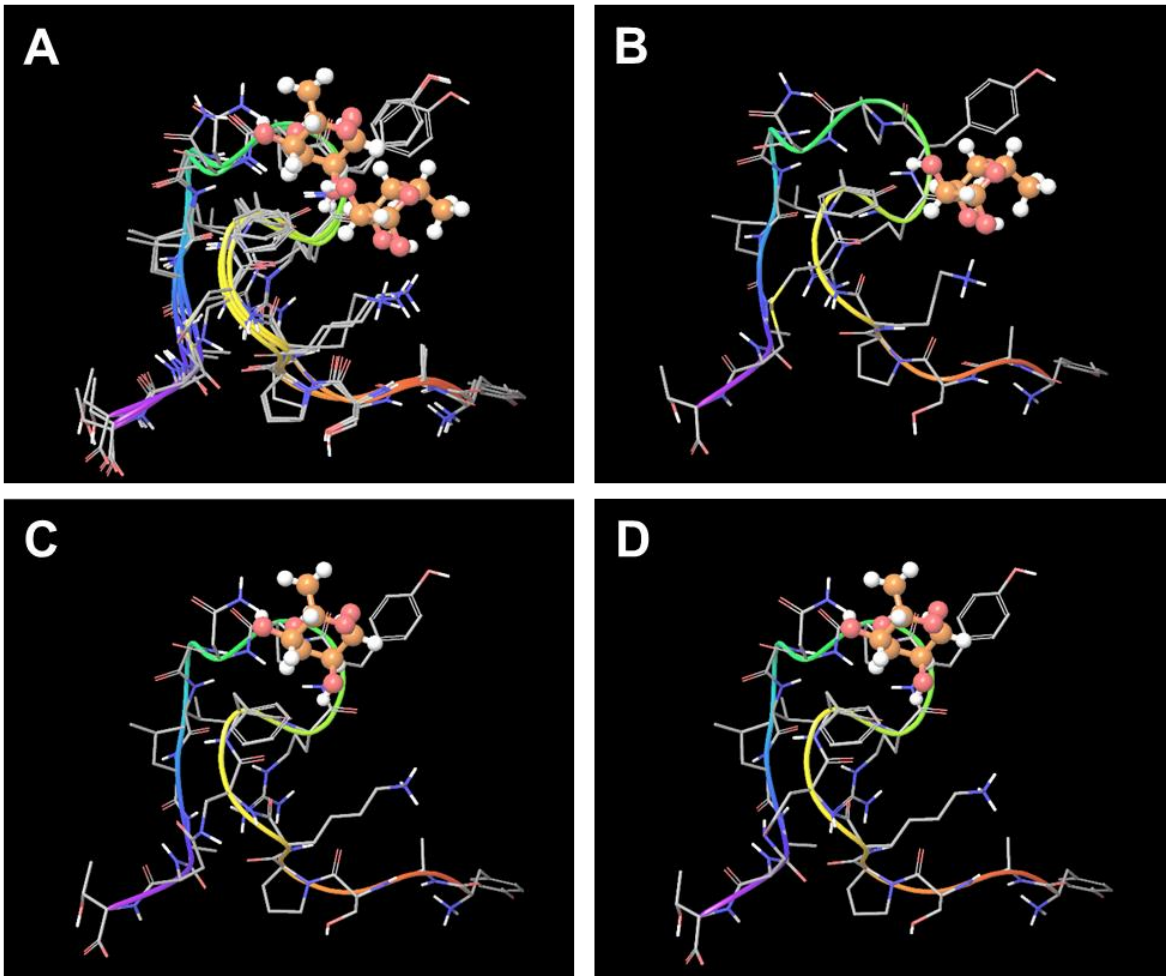

Figure 51: The backbone ribbon of the peptides is shown as a rainbow tube, the binding residues of the peptide are shown in a sticks rendering with C, O, N, H, S coloured grey, red, blue, white

and yellow. l-fucose is shown in CPK rendering with the C atoms coloured in orange, the rest of the atoms are coloured the same as the peptide binding residues. A) Odorranalectin, SP2 and SP3 with the docked conformations of l-fucose all overlaid to show the similarity between their structures. B) Odorannalectin with docked l-fucose. C) SP2 with docked l-fucose. D) SP3 with docked l-fucose.

		[1]		[2]	[3]	[4]
Lectin Compound	Estimated Free Energy of Binding* (kcal/mol)	VDW + H-bond +desolvated Energy (kcal/mol)	Electrostatic Energy (kcal/mol)	Final Total Internal Energy (kcal/mol)	Torsional Free Energy (kcal/mol)	Unbound System's Energy (kcal/mol)
Odorranalectin	-2.60	-3.43	-0.36	-2.53	+1.19	-2.53
SP2	-2.84	-3.87	-0.16	-2.53	1.19	-2.53
SP3	-2.83	-3.86	-0.17	-2.52	1.19	-2.52

Table 11: Table of the energy components for the docking of the l-fucose to Odorranalectin, SP2 and SP3 where the *The estimated free energy of binding is defined as: [1]+ [2] + [3] - [4].

Molecular docking of the four monosaccharides (l-fucose, d-galactose, N-acetyl-d-galactosamine, and d-glucose) was conducted to insight into their binding modes with Odorranalectin. Energetically favorable interactions were observed for l-fucose, d-galactose and N-acetyl-d-galactosamine. The binding pose of d-glucose is distinct in position and higher in energy to that of the other docked sugars causing its binding energy on average to be higher than the other three sugars by ~1 kcal/mol (Table 10). These results are in good agreement with the observed preferential binding of Odorranalectin to l-fucose in comparison to the undetectable binding to d-glucose in the fluorescence based binding assay. The hydrophobic side chains of the peptide enclose the binding sites of all four of the sugars. In the case of d-glucose, these are the side chain of F7 and the aliphatic moiety of K5; and for the other three sugars, in addition to F7 and K5, we have also the side chain of Y9. The similar binding orientations of l-fucose, d-galactose and d-galactosamine make it hard to assess the individual van der Waals' contributions of the neighbouring non-polar residues, to the binding energy. l-fucose, d-galactose and d-galactosamine were found to have the same four hydrogen bonds with K5, F7, R8 and Y9; whereas d-glucose, was found to form only two hydrogen bonds with the backbone of C6 (Figure 50). This may contribute to its relatively higher binding affinity. In addition to the four hydrogen bonds observed in the

complexes of d-galactose and d-galactosamine, l-fucose can also form an additional weaker hydrogen bond with the backbone carbonyl of R8. The charge of the ring oxygen of l-fucose is −0.68. Thus l-fucose can form an electrostatic dipole interaction with the positively charged side chain of K5, this may contribute to its relatively strong binding affinity [252]. The docking of l-fucose to SP2 and SP3 showed similar interacting residues to that of the native peptide Odorranalectin, however the poses differed in their orientations (Figure 51), as well as similar estimated free energy of binding (Table 11). The main difference observed in the position of l-fucose in SP2 and SP3 as opposed to Odorranalectin is the loss of the hydrogen bond with the side chain on K5. Instead the sugar is found to hydrogen bond the side chain of N11, this interaction is not found in the docking of l-fucose to Odorranalectin. The difference in binding is probably due to conformational changes induced by replacing the disulphide bridge with a lactam bridge. Importantly, the structures of SP2 and SP3 were minimized structures only, while the structure of Odorranalectin is a representative structure from NMR spectroscopy. To more accurately compare the binding affinities of Odorranalectin to SP2 and SP3, it would be necessary to account for the conformational flexibility of SP2 and SP3 with a conformational sampling method such as MD. The proposed alternative binding site for l-fucose interacting with C6, C16 and T17, suggested by the NMR titration experiment [253] was not observed in any of the binding poses of any of the dockings, even after significant enlargement of the docking area. However, future studies involving the consideration of lectin and glycan flexibility, as well as explicit solvation, are necessary to confirm this observation.

5.7 CONCLUSIONS

Our molecular docking results for the docking of Odorranalectin to the monosaccharides l-fucose, d-galactose, N-acetyl-d-galactosamine, and d-glucose showed good agreement with the experimental fluorescence based binding assay. The docking of SP3 showed that the modified lectin bound l-fucose in a similar yet distinct way than Odorranlectin, this confirmed that the lactam bridge modification probably does affect the sugar binding site of the lectin. Docking of SP3 to l-fucose showed that the experimentally observed lower binding affinity of SP3 to l-fucose was probably not based on the position of the lactam bridge in the sugar binding site; instead the lower binding affinity of SP3 was probably caused by a larger change in the conformation that was suggested by the CD spectra. The observed breakdown of SP2 and Odorranalectin in human serum to their respective cyclic metabolites, likely doesn't affect their ability to bind the sugar molecules as the cleaved residues (A1 and Y2) aren't directly involved in sugar binding according to our predicted docking poses. Overall the study showed promising findings for modifying lectin disulphide linkages to lactam bridges for the purposed of creating lectinomimics.

6 ELECTRONIC STRUCTURE AND ABSORPTION SPECTRA OF NEW SYNTHETIC FLUORESCENT COMPOUNDS

6.1 PREFACE

This chapter describes work done in collaboration with experimental researchers that has so far lead to the publication of a paper entitled “Luminescent Probes for Sensitive Detection of pH Changes in Live Cells through Two Near-Infrared Luminescence Channels” being published in June 2017, the full paper included in the appendix of this thesis gives full credit to all of the authors. There is also another paper that has recently been accepted “Fluorescent Probes for Sensitive and Selective Detection of pH Changes in Live Cells in Visible and Near-infrared Channels” which is also attached in the appendix. A third manuscript is in the final stages of preparation and will also soon be submitted. In each of these studies, QM calculations were used to determine the 3D structures, absorption spectra, excited states and HOMO/LUMO gap energies. This information was used to design and improve new pH sensitive luminescent probes.

The chapter contains an overview of each study and a summary of the full methodology and results of the works, as well as a more detailed description of the computational work contributed to the research.

6.2 INTRODUCTION

Fluorescence is a property defined as the ability to absorb a photon, and almost instantaneously re-emit another photon at a different usually shorter wavelength. A fluorophore is a chemical compound or group that possesses fluorescent properties, fluorophores can be incorporated into other molecules to act as fluorescent labels [254]. These fluorescent probes have become an integral tool for biologists, as they can be used to tag molecules of interest in order to track their movements in biological systems, and their incorporation into fluorescence microscopy has provided scientists with an important technique for the non-invasive real-time imaging of live cells [255]. Of particular interest are fluorescent probes that can respond to a change in conditions in the extracellular and intracellular environments, these allow chemical information to be gathered from the interior of small biological structures in a non-invasive manner [256, 257]. Fluorescent probes can be designed for a range of uses: to sense macromolecules (specific enzymes, proteins, DNA and RNA), specific moieties of larger molecules, and concentrations of specific small molecules or ions [258-261]. A fluorescent probe that is sensitive to the concentration of H^+ ions can be used to detect the pH of the solute in which it is dissolved, and allow for the measurement of pH based on the resultant fluorescence spectra [261]. The majority of pH sensitive probes work by interconverting between a protonated and deprotonated forms, where the two forms have distinct fluorescent properties. In response to the concentration of H^+ ions the equilibrium between the two forms changes leading to a shift in the fluorescence spectra, this allows for the measurement of pH [261, 262]. Studying pH at a cellular level is an important technique for understanding the internal architecture of the subcellular environment, as well as for understanding the pH differences that arise as a result of altered metabolism and disease [263]. Changes in cellular pH are associated with altered metabolic pathways than can develop as a result of inflammatory responses, cancer, neurodegenerative

diseases, and certain metabolic diseases [264, 265]. In particular, tumour formation is associated with increased acidity of the extracellular fluid, and increased basicity of the tumour cell internal environment [266, 267]. The acidification of the surrounding extracellular environment around tumour cells is thought to be caused by their altered and unusually anaerobic metabolic pathways [268]. In normal cells low pH regions can be found inside lysosomes these organelles harness their acidic internal conditions to provide the optimal conditions for hydrolysing enzymes responsible for the breakdown of proteins, peptides, lipids and carbohydrates. Lysosomes fulfil important roles such as the autophagy of unnecessary cellular structures, metabolism of waste products and the destruction of pathogens [269]. Visualising pH in the intracellular environment with pH sensitive fluorescent probes in live cells can give insights into important cellular functions relating to lysosome activity, distribution and formation [270].

However developing fluorescent probes for biological imaging has many challenges (in no particular order): (1) probes should not have cytotoxic properties that can harm cells and disrupt the very processes they are trying to study; (2) many natural substances in cells such as proteins can fluoresce at similar wavelengths to those of the probe causing a phenomena called autofluorescence, this can obscure images with background fluorescence and give false positive results; (3) probes used for intracellular imaging must be able to cross the cell membrane, this usually means they should be water soluble and rely on cellular active transport mechanisms; (4) high energy shorter wavelength light can damage cells, so ideally probes must have absorption and emission wavelengths higher than 600nm; (5) for in vivo imaging, the absorbed and emitted wavelengths of light must be able to penetrate tissue for the detection of fluorescence; (6) the stability of the probe, the probe must be chemically stable both to intracellular conditions as well as being reasonably stable with regards to photobleaching effects (the breakdown of the fluorophore by light); (7) probes should have reasonable quantum yields, this is the efficiency of photons that are emitted by the probe vs

the number of photons they absorb, poor quantum yields result in probes that don't give strong fluorescent signals and can't be detected; (8) the probe should be sensitive only to the desired analyte and not show false positive results in the presence of other chemicals [256-259, 262].

Most of the aforementioned problems concerning autofluorescence, the photodamaging effect on cells and the tissue penetrating power of the light can be avoided by ensuring that the absorbed and emitted light wavelengths of the probe are in the Near-Infrared (NIR) region (650-900nm) [258, 259, 271] . The development of NIR probes is an emerging field and new NIR probes are being designed for a range of biologically important uses such as detecting reactive oxygen species, inorganic ions, thiol containing species, hydrogen sulphide, specific enzymes and pH [258, 259]. It is hoped that these strategies could be used to selectively image cancer cells, leading to simpler and non-invasive fluorescence based screening of cancerous cells in patients [259, 272].

Most fluorophores exhibit a Stokes shift, this is the increase in wavelength between the absorbed and emitted photon. Fluorophores with this property require excitation with higher energy shorter wavelengths of light, and this is problematic due to the tendency of higher frequency photons to damage cells and cause autofluorescence effects [271, 273]. Certain fluorophores exhibit anti-Stokes shift properties, this means they absorb lower energy photons, longer wavelengths of light, and upconvert them to emit shorter wavelength photons [274]. These anti-Stokes shifting fluorophores are particularly useful for the imaging of biological systems because the light needed to activate the fluorescent properties can be very low energy, this means there is no cell damaging effect and less photobleaching of the dye while still providing high levels of detection sensitivity [273, 275]. Most Anti-stokes shift dyes are lanthanide based nano-particles, however these compounds generally have problems with cytotoxicity making them unsuitable for use with living cells [276, 277].

In the two papers described in this chapter our experimental collaborators have successfully developed and synthesised a pair of novel pH sensitive probes that display anti-stokes shift fluorescent properties, and another three novel probes that have fluorescent properties in both acidic and neutral basic conditions and for use in imaging intracellular pH in living cells [273].

6.3 "LUMINESCENT PROBES FOR SENSITIVE DETECTION OF PH CHANGES IN LIVE CELLS THROUGH TWO NEAR-INFRARED LUMINESCENCE CHANNELS" – SUMMARY

These two novel probes were designed around anti-Stokes shifting fluorophores known to fluoresce in the NIR range in response to pH changes that were developed previously, however the spirolactam based fluorophores alone were found to be quite cytotoxic and had poor water solubility making them unsuitable for intracellular imaging [261, 278, 279]. The fluorophore uses a reversible spirolactam ring opening/closing reaction to "sense" pH changes, in acidic conditions (pH <7.4) the protonated form of the fluorophore's spirolactam ring is opened and possesses both conventional and anti-stokes shifting fluorescent properties. In neutral and basic pH conditions (pH >7.4) the spirolactam ring is deprotonated and closed, in this state the fluorophore is colourless and shows no fluorescent properties. The opening of the spirolactam ring in response to pH causes increased π-conjugation in the fluorophore and the emergence of their fluorescent properties. This range of pH sensitivity means that the fluorophore is very weakly fluorescent at extracellular pH levels and become very fluorescent in the acidic conditions of lysosomes.

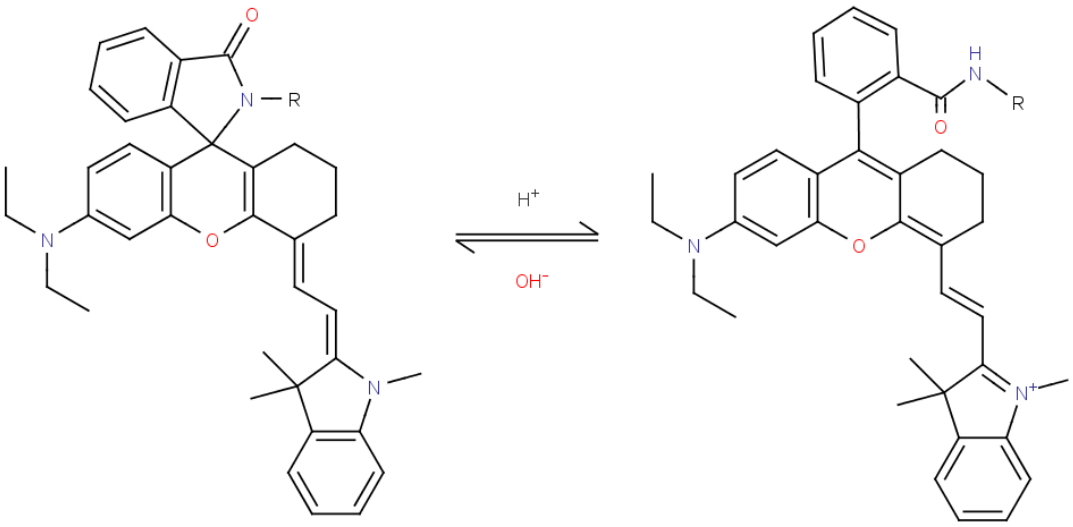

Figure 52: The generic structure of the pH sensitive rhodamine fluorophore in both its protonated and deprotonated forms.

In “Luminescent Probes for Sensitive Detection of pH Changes in Live Cells through Two Near-Infrared Luminescence Channels” the aforementioned fluorophore was conjugated to mannose sugar molecules by a linker composed of a 2,2`-(ethylenedioxy)diethylamine to produce two distinct probes. Probe A containing a single mannose residue and Probe B containing two mannose residues. In the study different properties of the probes were investigated: (1) the response of the absorption and fluorescence spectra of the two probes to varying pH levels; (2) the selectivity of the probes towards H^+ ions as opposed to other positively charged metal ions; (3) the photostability of the two probes in response to constant excitation; (4) The cytotoxicity of the two probes to live human cells; (5) the permeability and resultant fluorescence of the probes in live human cells.

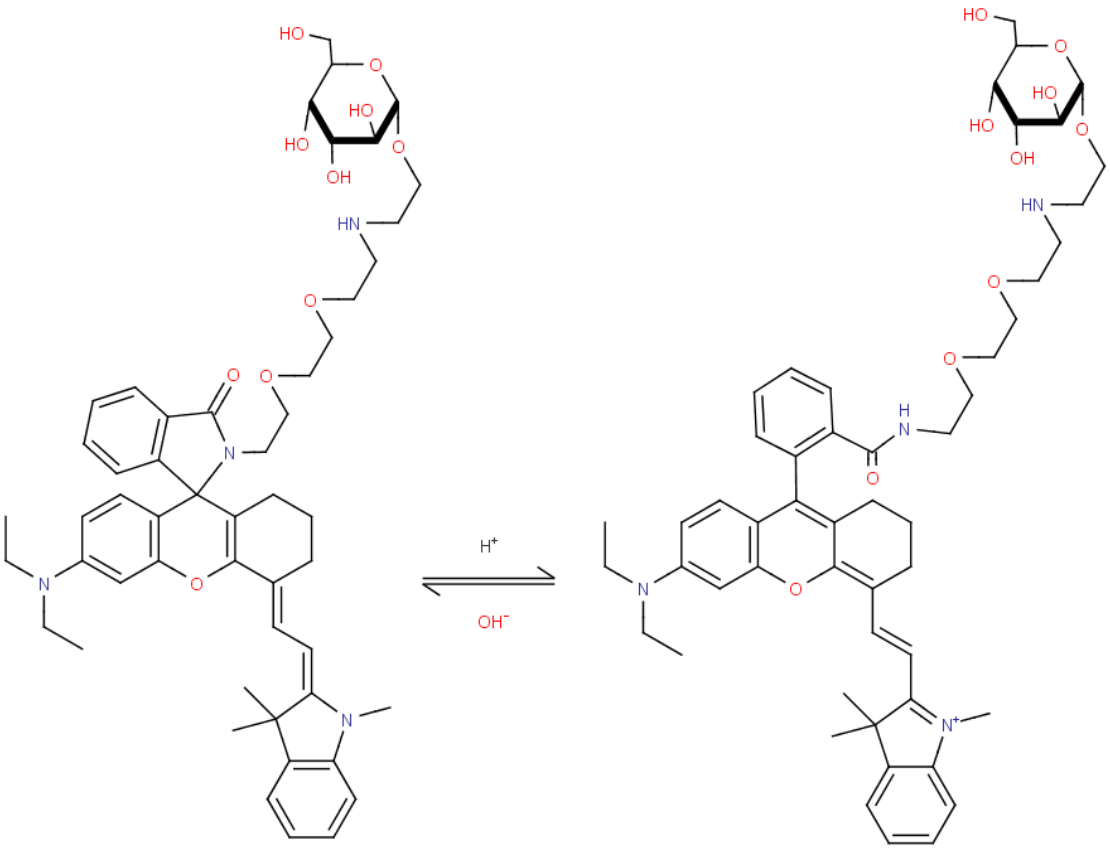

Figure 53: The structure of probe A from “Luminescent Probes for Sensitive Detection of pH Changes in Live Cells through Two Near-Infrared Luminescence Channels” in its protonated and deprotonated forms.

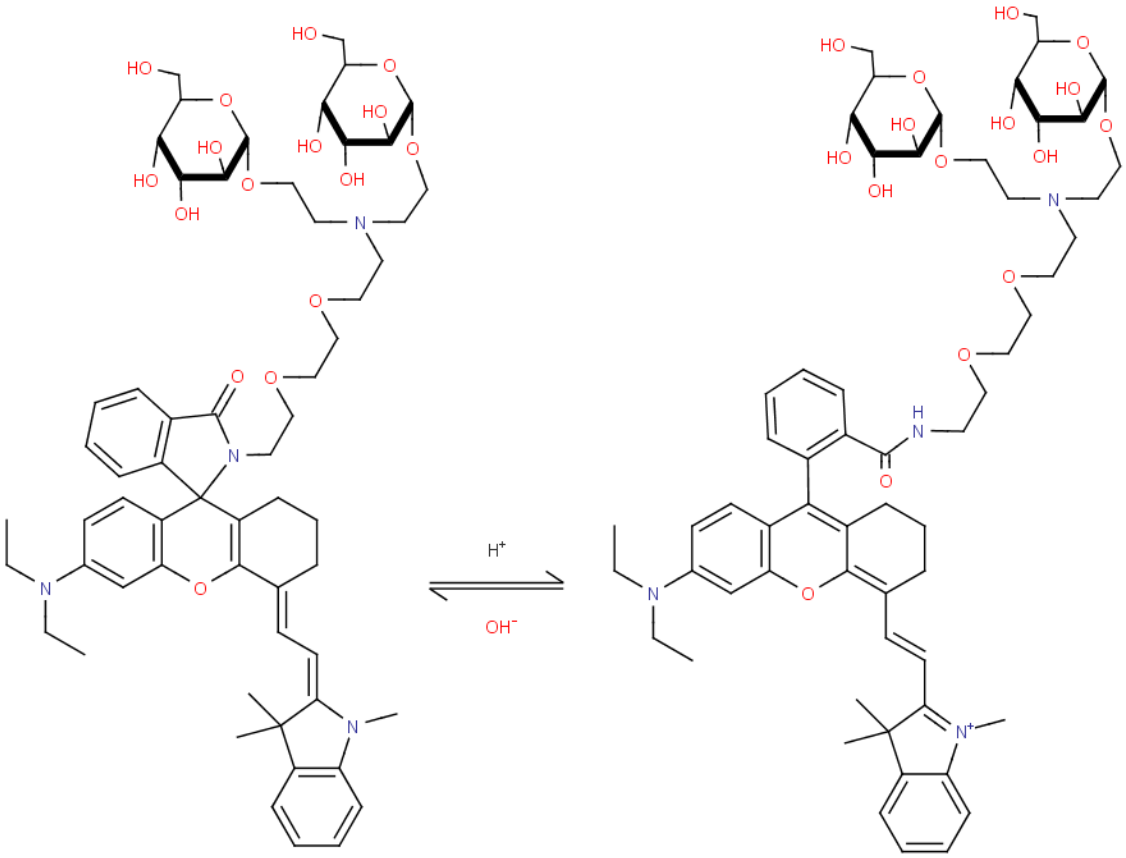

Figure 54: The structure of probe B from “Luminescent Probes for Sensitive Detection of pH Changes in Live Cells through Two Near-Infrared Luminescence Channels” in its protonated and deprotonated forms.

6.3.1 Computational Methods

The structures of the two probes A and B in both acidic and neutral-basic pH conditions were built using GaussView [167]. The four structures were then optimised with Gaussian09 [94] in a Polarizable Continuum Model (PCM) of solvent water with a 6-31G(d) basis set using the DFT B3LYP functional. The HOMO/LUMO orbitals of the two probes in acidic and basic conditions were generated from their optimised structures and visualised in GaussView [167].

The protonated structures of probe A and B underwent excited state Time Dependent - DFT (TD-DFT) optimisations of the first six singlet (S=0) excited states in PCM solvent water

with the B3LYP functional and a 6-31G* basis set in Gaussian09 [94]. From the excited state optimisations, the absorption spectra data was extracted using GaussView [167]. All images of molecular models for this chapter were created using GaussView [167].

6.3.2 Summary of Experimental Results

The two probes had their absorption and fluorescence spectra measured in a range of pH conditions between pH 2.2-8.0. As pH decreases the absorbance peak observed at ~375nm decreases. Whereas the absorbance peak at ~715nm and its accompanying shoulder peak at ~650nm is found to reach maximum absorbance at pH 4.4, decreasing the pH below 4.4 results in less observed absorbance in this region. It is thought the tertiary amine may become protonated at particularly low pH levels affecting the π-conjugation of the fluorophore system and therefore its optical absorbance. The fluorescence spectra of the two probes showed them to be not fluorescent when pH >7.4, as the pH is decreased peaks are observed to form and increase in intensity at ~740nm in both the conventional and anti-stokes shift fluorescence spectra. Peak levels of fluorescence are observed at pH 4.4, in more acidic conditions the intensity of the fluorescence is decreased, this is again thought to be due to the protonation of the tertiary amine group. In acidic conditions where the spirolactam ring opened form of the probes is most prevalent the strongest conventional fluorescence is observed with an excitation wavelength of 690nm, the strongest anti-stokes shift fluorescence is observed with an excitation wavelength of 808nm. In both probes this pH response is reversible in the measured pH range of 2.2-8.0. The fluorescence quantum yields of probes A and B at pH 4.4 were calculated to be 8.4% and 8.1% respectively.

Both probes showed very little change in their fluorescent intensity at both neutral-basic and acidic conditions in solutions of a range of metal ions (Cu^{2+}, Na^{+}, Fe^{2+}, Zn^{2+}, Al^{3+}, Ca^{2+}, Mg^{2+}, Mn^{2+}, Ag^{+}, K^{+}, Ni^{2+} and Co^{2+}), indicating that they make good selective probes of H^{+} and therefore pH even in the presence of metal ions.

The probes were excited continuously by 690nm light and in intervals of time the conventional fluorescence intensity was observed to assess the photostability of the probes, the conventional fluorescence intensity of both probe A and B showed a reduction in intensity of about ~11% after 3 hours of interval excitation. The anti-stokes shift fluorescence intensity of the two probes was measured in a similar manner with an excitation wavelength of 808nm, after 3 hours the anti-stokes shift fluorescence intensity of both probes had barely decreased. This demonstrated the excellent photostability of the two probes anti-stokes shifting fluorescent properties.

The cytotoxic properties of the two probes were assessed with an XTT assay with live human fibroblasts, cell viability was reduced somewhat by both probes, however even at a concentration of 50μM the viability greater than 70%. At the concentration levels used for cell for the cell staining portion of the experiment (5-20 μM) cell viability would not be too negatively affected by the probes.

To test the ability of the probes to stain live cells, they were incubated at varying concentrations (5-20 μM), with HeLa and KB cell lines. The distribution and intensity of fluorescent regions in cells treated with the two probes was compared against lysotracker, an established dye for staining acidic compartments in live cells. Probe A and B show fluorescent spots that are shown to match well to the same cellular areas of low pH as those areas stained by lysotracker. Both conventional and anti-stokes shift fluorescence intensity was found to increase with probe concentration. The ability of the probes to respond to extra and intracellular pH levels was tested by incubating the cells in buffer solutions at a range of pH values (4.5, 5.5, 6.5 and 7.5) with nigericin, a compound used to equilibrate intracellular pH levels to those of the extracellular environment. At the normal physiological pH 7.5 both probes showed only weakly fluorescent results, a decrease in pH level saw both the conventional and anti-stokes fluorescence intensity of the stained cells increase. The observed pH sensitive change in fluorescence was in good agreement with the optical

responses of the probes in different pH solutions. The two probes also both showed a greater sensitivity to pH in the range tested than the commercial lysotracker.

6.3.3 Computational Results and Discussion

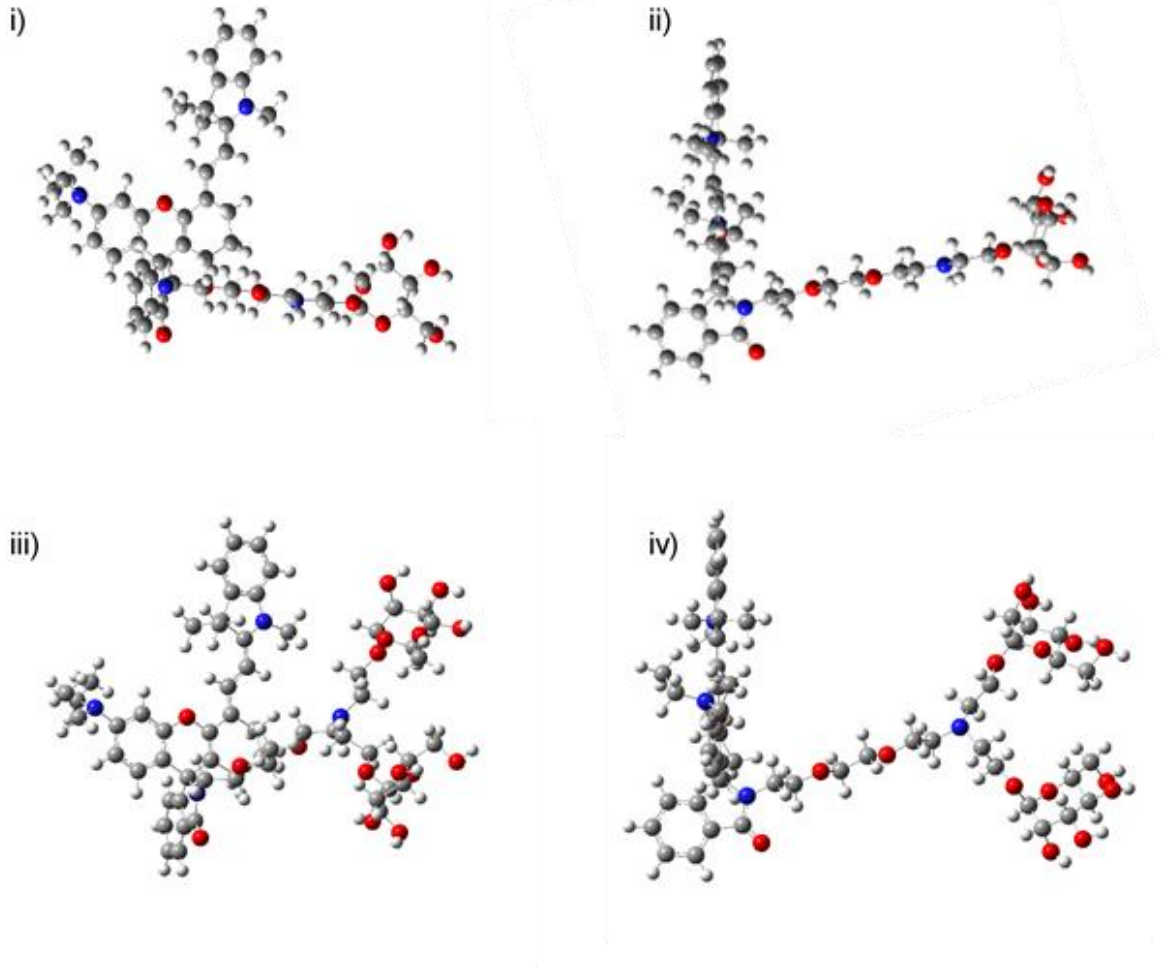

Figure 55: i) and ii) show probe A in the deprotonated form, ii) is a rotated view of i) emphasising the angle between the spacer region and the plane of the fluorophore. iii) and iv) show probe B in the blue form, iv) is a rotated form of iii).

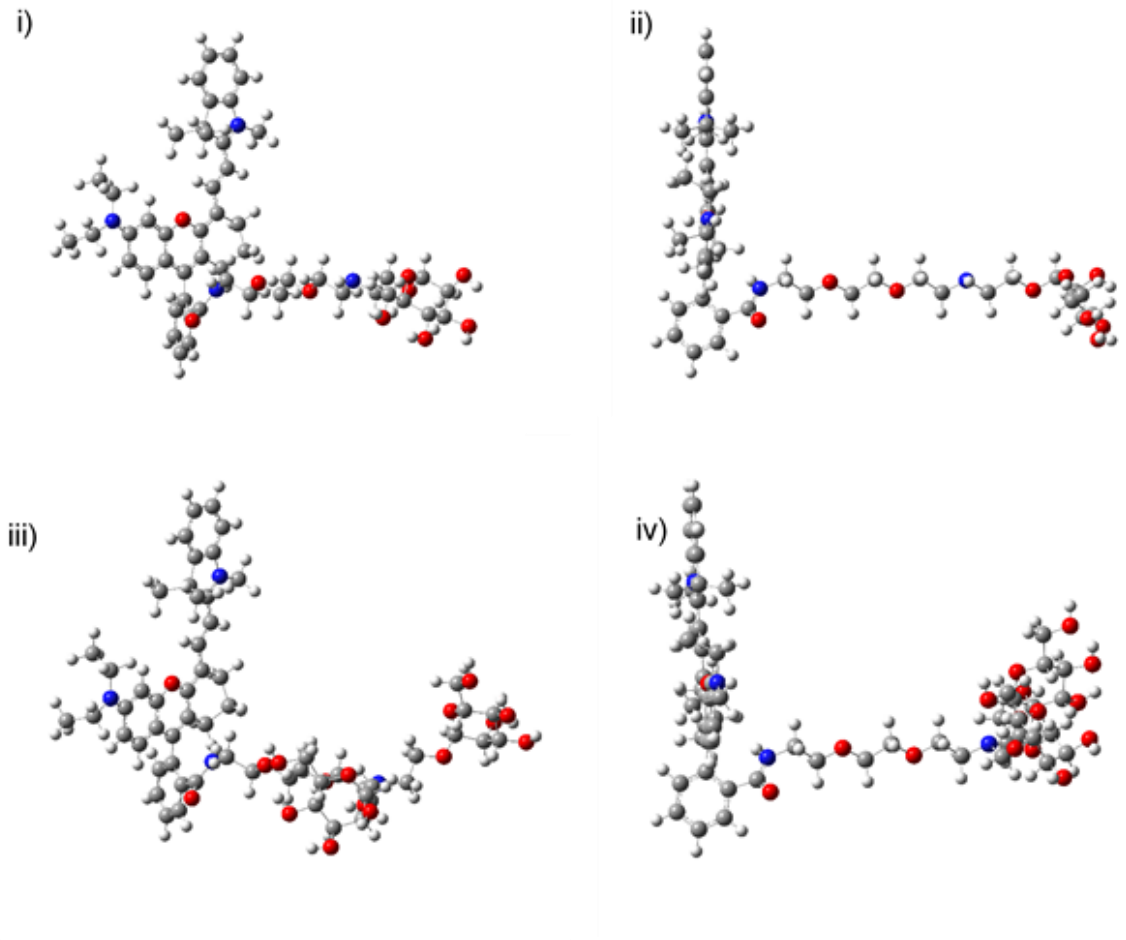
i)
ii)
iii)
iv)

Figure 56: i) and ii) show probe A in the protonated fluorescent form, ii) is a rotated view of i) emphasising the angle between the spacer region and the plane of the fluorophore. iii) and iv) show probe B in the pink form, iv) is a rotated form of iii).

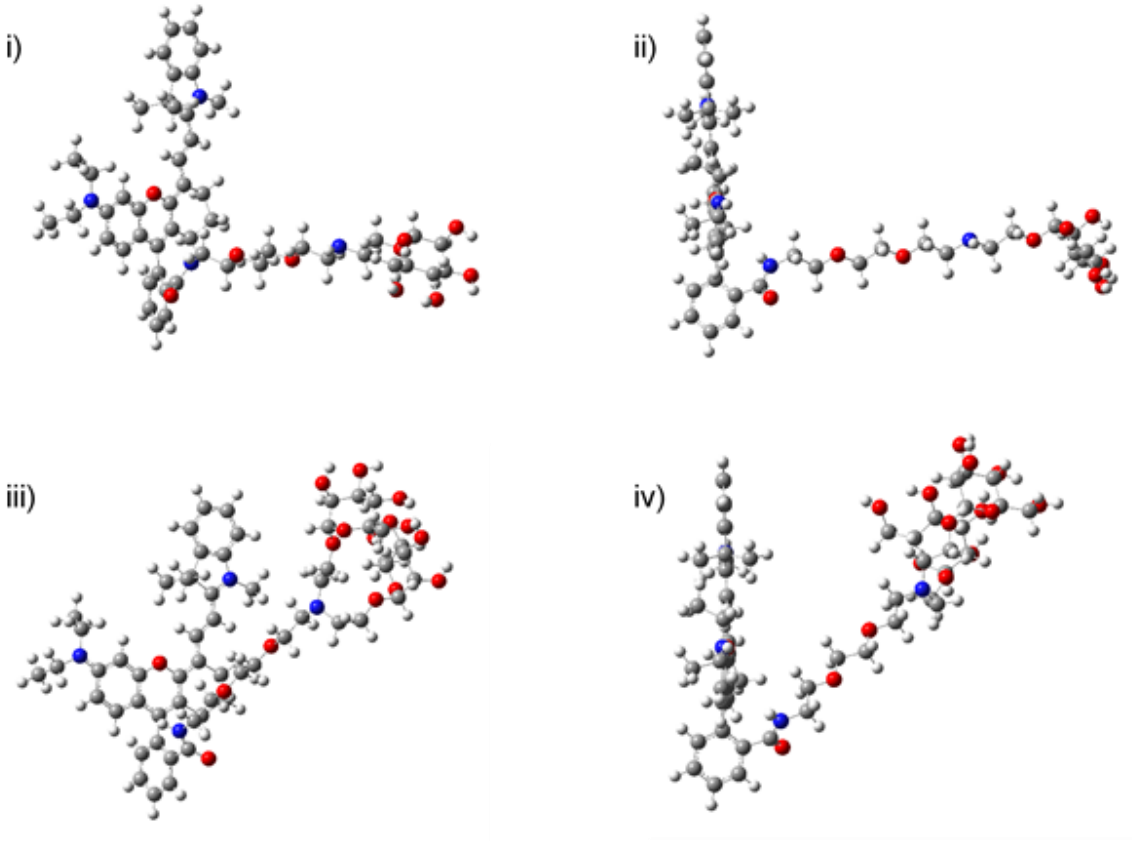

Figure 57: i) and ii) show the excited state geometry of probe A in its protonated form, ii) is a rotated view of i) emphasising the angle between the spacer region and the plane of the fluorophore. iii) and iv) show the excited state geometry of probe B in its protonated form, iv) is a rotated form of iii).

Structure	HOMO/LUMO gap (eV)
Probe A deprotonated	0.128
Probe B deprotonated	0.12813
Probe A protonated	0.08492

Probe B protonated	0.08483
Excited state Probe A protonated	0.08042
Excited state probe B protonated	0.08039

Table 12: The calculated HOMO-LUMO gap energies for the two fluorescent probes A and B in their optimised deprotonated and protonated forms as well as the excited state optimisations of the protonated form.

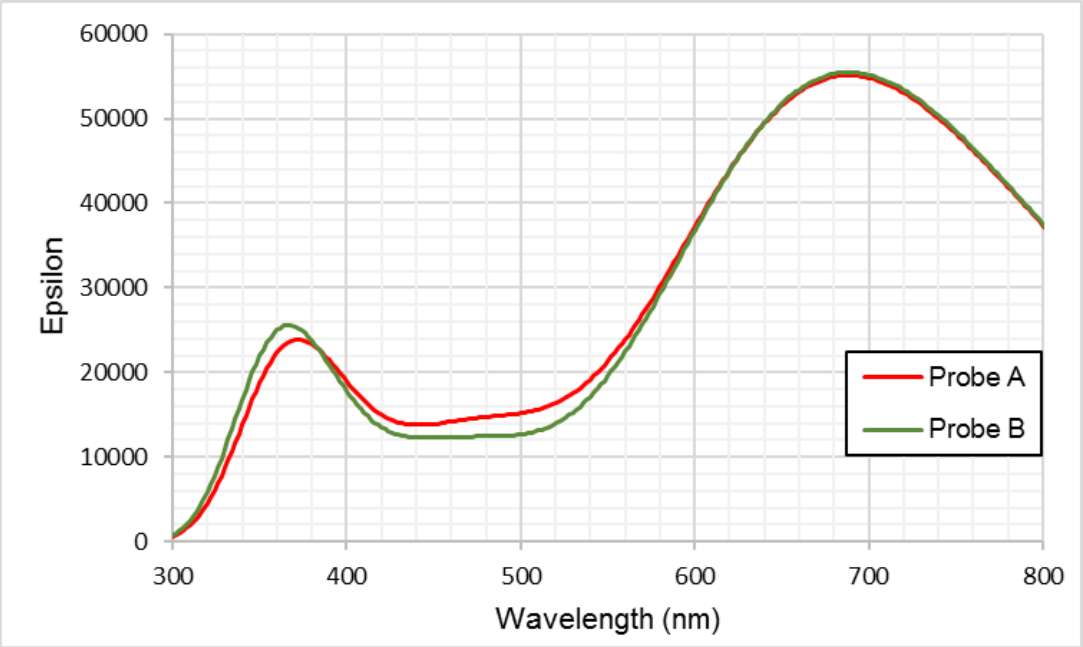

Figure 58: The absorption spectra of the excited state optimisations of the protonated forms of probes A and B.

Probe	Singlet first excited state energy (eV)	Wavelength (nm)	Oscillatory strength
A	1.797	689.94	1.3534
B	1.7982	689.48	1.3661

Table 13: Results from the excited state calculations of Probe A and B.

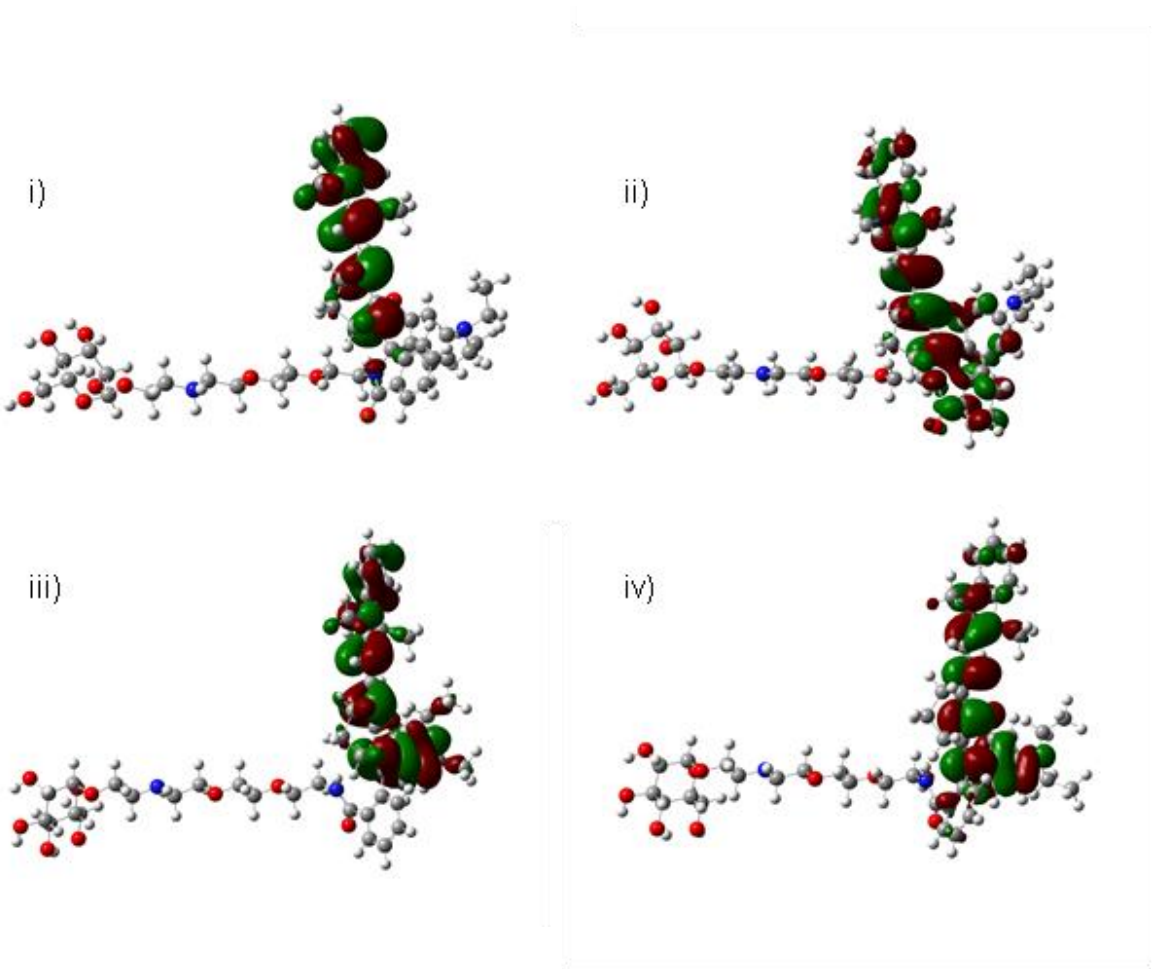

Figure 59: i) HOMO orbital of Probe A in its deprotonated form, ii) LUMO orbital of Probe A in its deprotonated form, iii) HOMO orbital of Probe A in its protonated form, iv) LUMO orbital of Probe A in its protonated form.

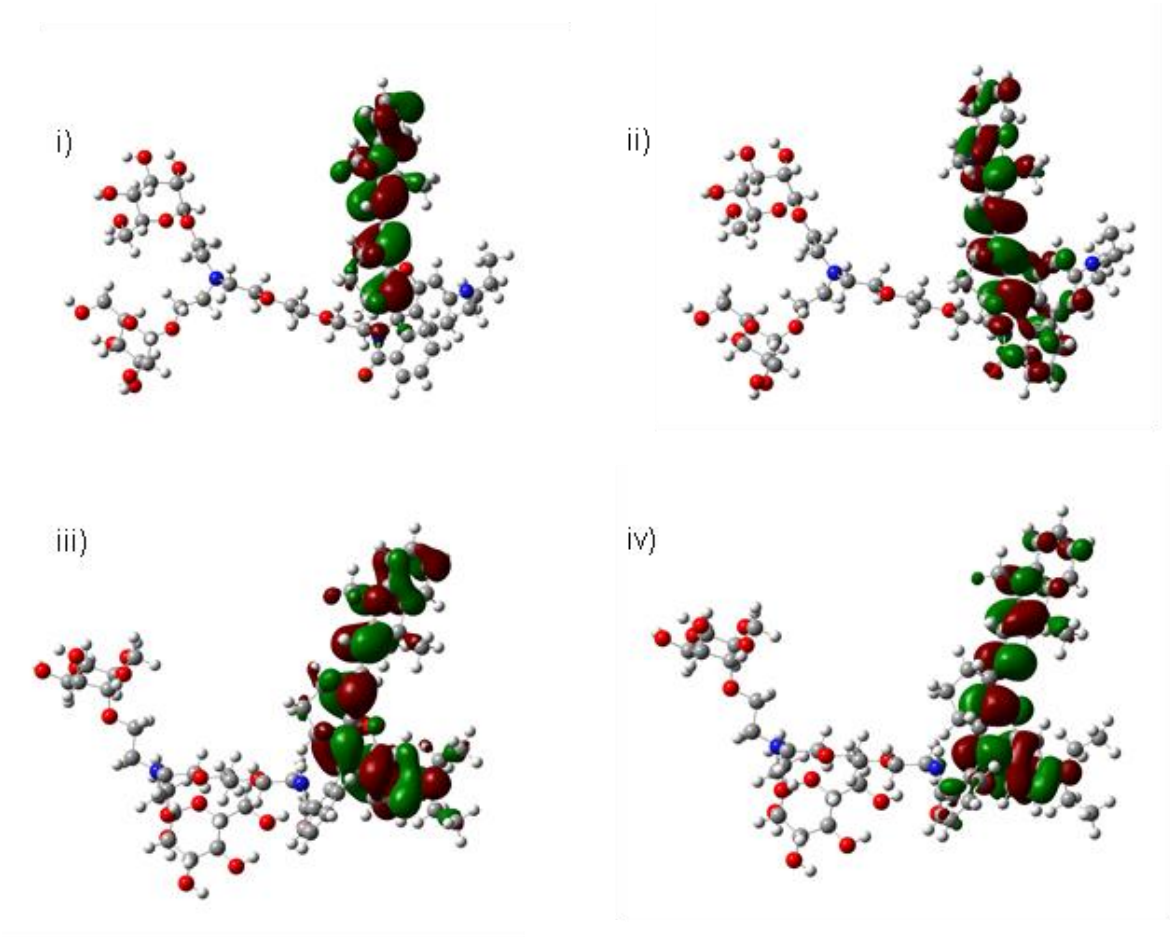

Figure 60: i) HOMO orbital of Probe A in its deprotonated form, ii) LUMO orbital of Probe A in its deprotonated form, iii) HOMO orbital of Probe A in its protonated form, iv) LUMO orbital of Probe A in its protonated form.

Despite the only difference between probes A and B being the addition of another mannose group to the spacer moiety, the two probes show observable differences in fluorescence at acidic pH. The optimised spirolactam ring closed structures of probes A and B at basic pH are structurally quite similar, their optimised geometries can be seen in Figure 55. The optimised structures of the ring opened forms of probes A and B in acidic conditions are also alike, a minor difference is that the mannose in probe B is roughly parallel to the plane of the fluorophore region (Figure S2). The main noticeable structural difference between the opened and closed spirocyclic ring forms of the two probes is the angle between the mannose tail moiety and the plane of the fluorophoric ring moiety. The spirocyclic ring closed form

of the probes in neutral-basic conditions shows a more acute angle between the spacer moiety and the fluorophore plain whereas in the acidic form the angle is closer to being perpendicular, this means in the closed spirocyclic form the tail moiety lies closer to the fluorophoric ring moiety (Figure 56). The energy differences between the LUMO and HOMO orbitals is very similar in both probes A and B. In the acidic forms the HOMO-LUMO energy gap is smaller than in the basic forms (Table 12). The greatest structural difference between probes A and B is seen in the structure of the excited state optimised structures of the spirocyclic ring opened forms in acidic conditions (Figure 57). In this probe B is found to adopt a more compact and distorted in geometry compared to the other structures of probe A, the angle between the spacer and the fluorophore is very acute and the mannose residues lie directly over the fluorophore, this could possibly cause fluorescence quenching by consuming energy via molecular rotation of the mannose residue of probe B. These differences in the optimised geometry may explain the difference in the experimentally observed fluorescence. The absorption spectra for the two excited state optimisations was shown in Figure 58 and Table 13, although similar, some variation is seen in the region 300-550 nm. The HOMO and LUMO orbitals are presented in Figure 59 and Figure 60 for probes A and B, respectively. The opening of the spirocylic ring under acidic conditions causes the HOMO orbital of both probe A and B to shift from the substituted indole ring and its conjugated alkene chain region to become more spread out over the rhodamine-like fluorophore moiety. The LUMO orbitals of probe B seem to show a similar if less pronounced effect where the orbital becomes more spread out over the rhoadmine-like fluorophore moiety upon ring opening.

6.4 "FLUORESCENT PROBES FOR SENSITIVE AND SELECTIVE DETECTION OF PH CHANGES IN LIVE CELLS IN VISIBLE AND NEAR-INFRARED CHANNELS" - SUMMARY

The three probes A, B and C described in this paper improved upon the fluorescent functionality of the probes developed in the previous paper. Most pH sensitive probes like those from “Luminescent Probes for Sensitive Detection of pH Changes in Live Cells through Two Near-Infrared Luminescence Channels” don’t have fluorescent properties in neutral-basic conditions, this can lead to “blind spots” at these pH levels. Probe A, B and C were created by conjugating the spirolactam rhodamine based fluorophore to different coumarin moieties, the coumarin fluorophore moiety is known to have fluorescent properties in the visible light spectral range, and the spirolactam rhodaminefluorophore known to have pH sensitive NIR fluorescent properties. By combining these two fluorophores it was hoped that the probes would have the desired distinctive fluorescent properties in both acidic and neutral-basic conditions allowing for the probes detection in non-acidic conditions to remove imaging “blind spots”. To increase the water solubility of the probes an azide moiety joined to the fluorophore by an oligo(ethylene glycol) spacer was attached to the fluorophore moiety. The terminal azide group is a commonly used moiety in click chemistry and allows for functionalisation through the addition of diverse moieties such as sugars and peptides for the future development of derivative compounds. As well as testing the spectral responses of the probes to pH levels both in solution and in vivo, the probes would be tested for their sensitivity to H^+ over commonly found metal ions, photostability over time, cell membrane permeability, and cytotoxicity.

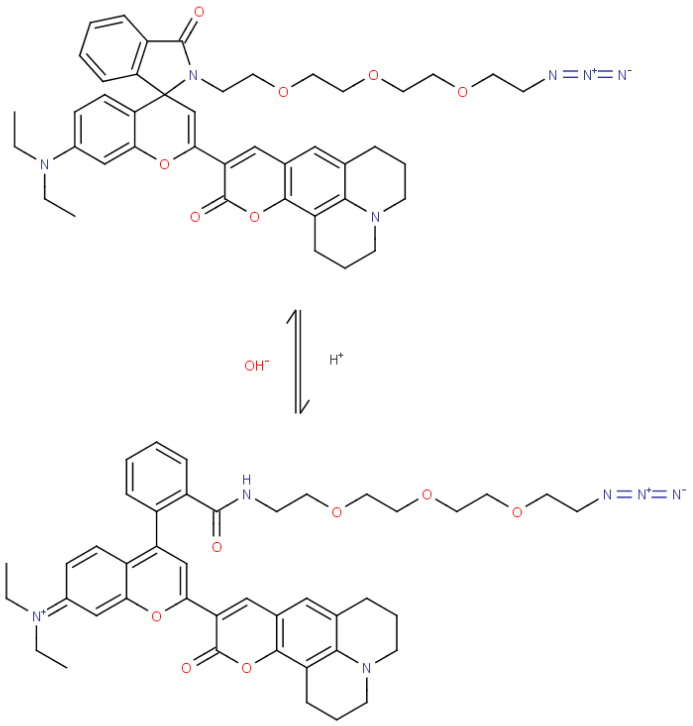

Figure 61: The structure of probe A from “Fluorescent Probes for Sensitive and Selective Detection of pH Changes in Live Cells in Visible and Near-infrared Channels” in both its protonated and deprotonated form.

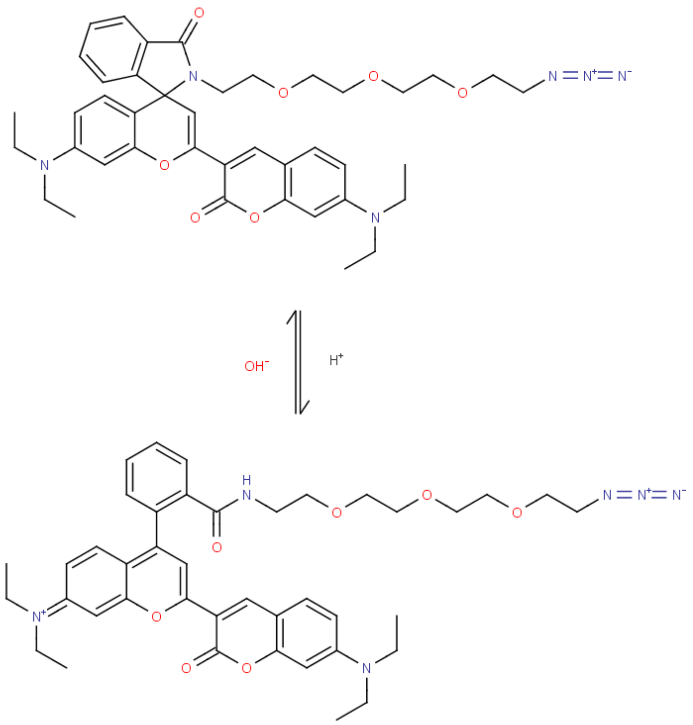

Figure 62: The structure of probe B from “Fluorescent Probes for Sensitive and Selective Detection of pH Changes in Live Cells in Visible and Near-infrared Channels” in both its protonated and deprotonated form.

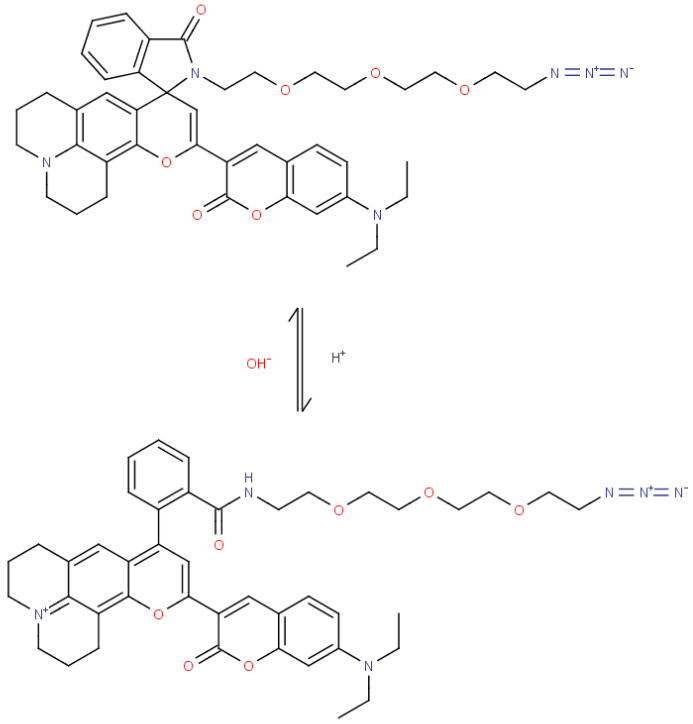

Figure 63: The structure of probe C from “Fluorescent Probes for Sensitive and Selective Detection of pH Changes in Live Cells in Visible and Near-infrared Channels” in both its protonated and deprotonated form.

6.4.1 Summary of Experimental Results

The fluorescence spectra of the 3 probes dissolved in buffer solutions were recorded as the pH of the solutions was decreased from pH 7.6 to 1.6. At neutral-basic pH the three probes A, B and C showed absorption peaks at 420nm, 445nm and 415nm respectively with some additional very weak absorbance at ~650nm also being recorded. As pH is decreased new absorption peaks at 655nm 670nm and 653nm are observed for probes A, B and C respectively, this absorption was found to increase in intensity as the pH is lowered from 7.4 to 4.0, below 4.0 the absorption remains high but doesn’t increase further. The measured

fluorescence spectra of the probes A, B and C show two sets of new NIR peaks emerging at 711nm, 696nm and 707nm respectively, in response to the pH being lowered from 7.4 to 1.6. Additional new non-NIR fluorescent peaks were observed at 497nm, 482nm and 498nm for probes A, B and C respectively were also observed with a similar pH sensitive response as the NIR peaks. All three probes were shown to possess these optical response properties in a completely reversible manner in the range of pH 2.0-10.0. The quantum yields of probes A, B and C were calculated to be 2.77%, 1.85% and 0.65% at pH 2.0 with an excitation wavelength of 625nm, with a shorter excitation wavelength of 425nm the quantum yields of the three probes were 2.73%, 2.64% and 4.86% respectively.

All three probes showed little response in their fluorescence spectra intensity in both neutral-basic (pH 7.6) and acidic (pH 2.0) conditions in solutions with different metal ions (Fe^{2+}, Cr^{2+}, Mn^{2+}, Fe^{3+}, Zn^{2+}, Mg^{2+}, Ag^{+}, Ca^{2+}, Al^{3+}, Cu^{2+}, Ni^{2+}, Co^{2+}, Hg^{2+} and Cd^{2+}) indicating their selectivity for H^{+} and therefore their ability to be pH sensitive probes even in the presence of metal ions.

To test the photostability of the three probes they were repeatedly excited by 690nm wavelength light in 5min time intervals while their fluorescence spectra was recorded in 10min intervals for a total of 3 hours. After 3 hours all three probes had lost <10% of their fluorescence intensity indicating their excellent photostability.

The cytotoxic properties of the three probes were assessed by incubating them at a range of concentrations (5-50μM) with live cells from the MDA-MB231 cell line. Probe A had a moderate effect on cell viability with concentrations 5-20μM having >70% cell viability. Probes B and C were less tolerated and had a more severe effect on cell viability, and only retained >70% cell viability at 5μM and 10μM levels of concentration. All three probes had profoundly negative effects on cell viability at the 50μM level of concentration.

Live cell staining was performed with the MDA-MB231 and HUVEC-C cell lines for all three probes at 10μM concentrations, the commercially available lysosensor blue probe was

used to compare the ability of the probes to detect areas of low pH at the intracellular level. Based on the initial observations it was found that probes A and C showed no NIR fluorescent signals in the live cell staining tests, probe B however performed well and underwent further live cell staining. Probe B is less hydrophobic than the other two probes, this may be the reason it was more able to permeate into the live cells and produce fluorescence results in the NIR range; as opposed to probe A and C which performed equally well in solution but weren't able to fluorescently stain live cells. Probe B underwent additional live cell testing and was incubated with MDA-MB231 and HUVEC-C cells in buffer solutions at a range of pH levels 4.5, 5.5, 6.5, 7.5 and 8.5, nigericin was used to equilibrate the intracellular pH levels to those of the buffer solution. Increasing NIR fluorescence in response to decreasing pH was observed at the pH levels tested, with maximum fluorescence being observed in the pH 4.5 buffer solution. Comparing probe B stain locations to those of the lysosensor blue probe showed that the two probes were found in the same intracellular regions of low pH.

6.4.2 Computational Methods

GaussView [167] was used for building the structures of probes A, B and C both in protonated and deprotonated state. Geometry optimisation of the six structures was performed using DFT with the B3LYP functional [209, 210] and 6-31G* basis set in Gaussian09 [94]. The Polarizable Continuum Model (PCM) [280, 281] with a dielectric constant of 78.39 was used to represent the water solvent. Computationally predicted absorption spectra were calculated using Time Dependent - DFT (TD-DFT) with B3LYP functional [209, 210] and 6-31G* basis set in PCM [280, 281] in Gaussian09 [94]. From the optimisations of the first six singlet (S=0) excited states in PCM solvent water with the B3LYP functional and a 6-31G* basis set in Gaussian09 [94] the absorption spectra data was extracted using GaussView [167]. From these calculations, the HOMO and the LUMO

orbitals and absorption spectra were visualised using GaussView [167]. All images of molecular models for this chapter were created using GaussView [167].

6.4.3 Computational Results and Discussion

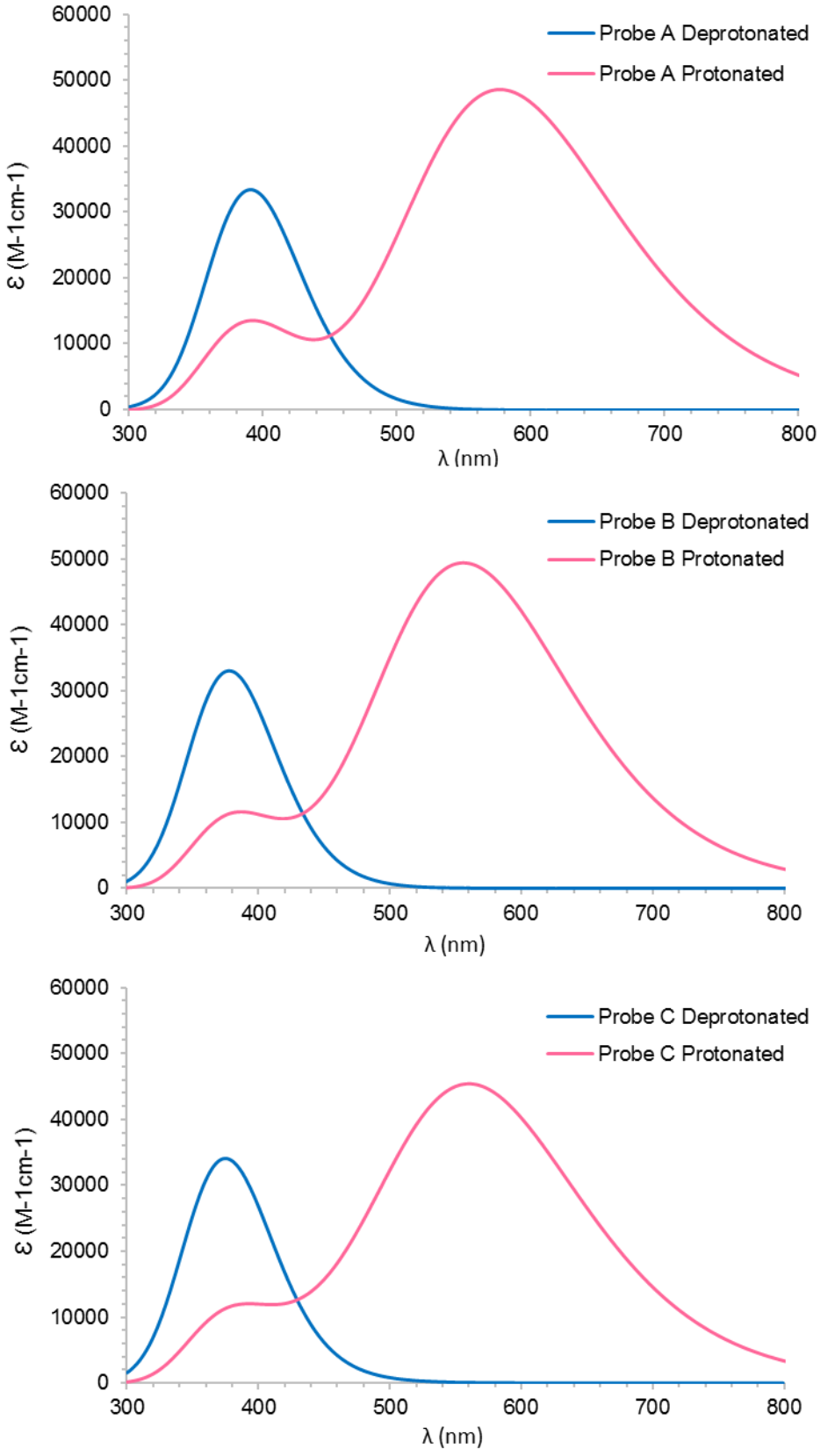

Figure 64: The calculated absorption spectra of the three fluorescent probes A, B and C in both their protonated (pink) and deprotonated (blue) forms.

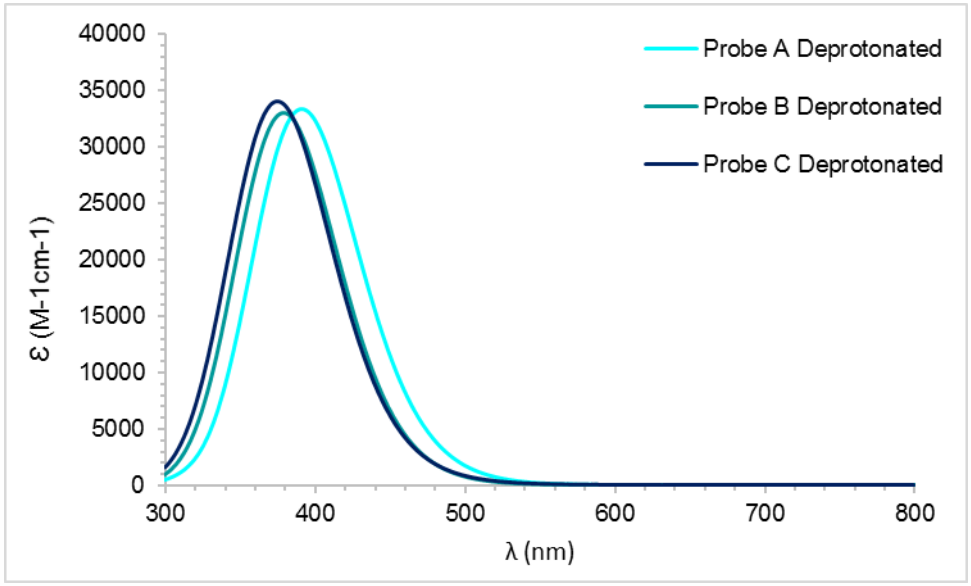

Figure 65: The combined calculated absorption spectra of the three fluorescent probes in their deprotonated forms.

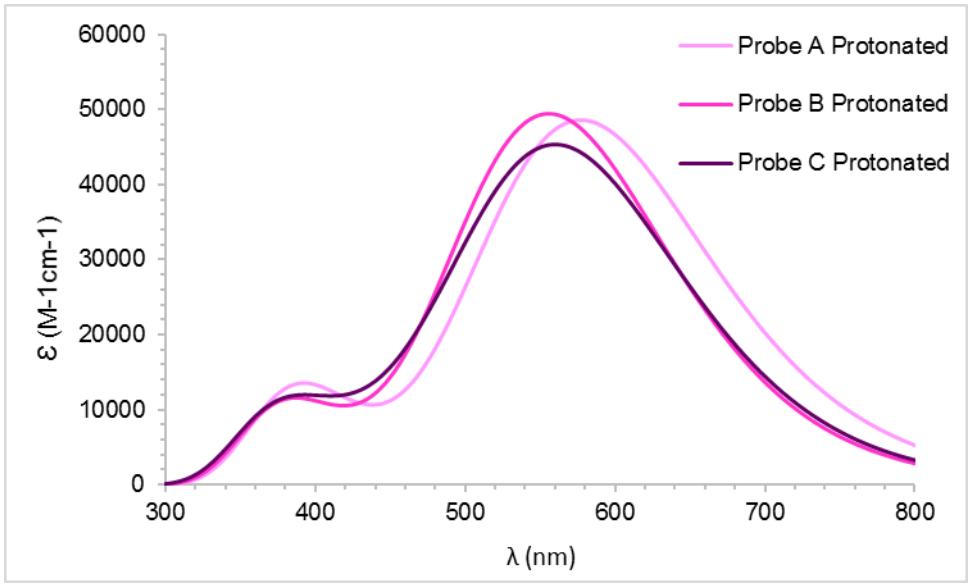

Figure 66: The combined calculated absorption spectra of the three fluorescent probes in their protonated forms.

Compound	HOMO Energy (eV)	LUMO Energy (eV)	HOMO-LUMO Gap Energy (eV)
A Deprotonated	-0.1882	-0.0608	0.1274
A Protonated	**-0.2021**	**-0.1141**	**0.0880**
B Deprotonated	-0.1941	-0.0645	0.1296
B Protonated	**-0.2081**	**-0.1160**	**0.0922**
C Deprotonated	-0.1810	-0.0634	0.1176
C Protonated	**-0.2036**	**-0.1119**	**0.0917**

Table 14: The calculated HOMO and LUMO energy as well as HOMO-LUMO gap energies of the three fluorescent probes in both their protonated and deprotonated forms.

Compound	Excitation Energy Wavelength (nm)	Oscillator Strength
A Deprotonated	392.19	0.6798
A Protonated	**578.34**	**1.1863**
B Deprotonated	397.81	0.0415
B Protonated	**558.32**	**1.1952**
C Deprotonated	444.82	0.0102
C Protonated	**566.28**	**1.0661**

Table 15: The calculated excitation energy wavelengths and oscillator strengths three fluorescent probes in both their protonated and deprotonated forms.

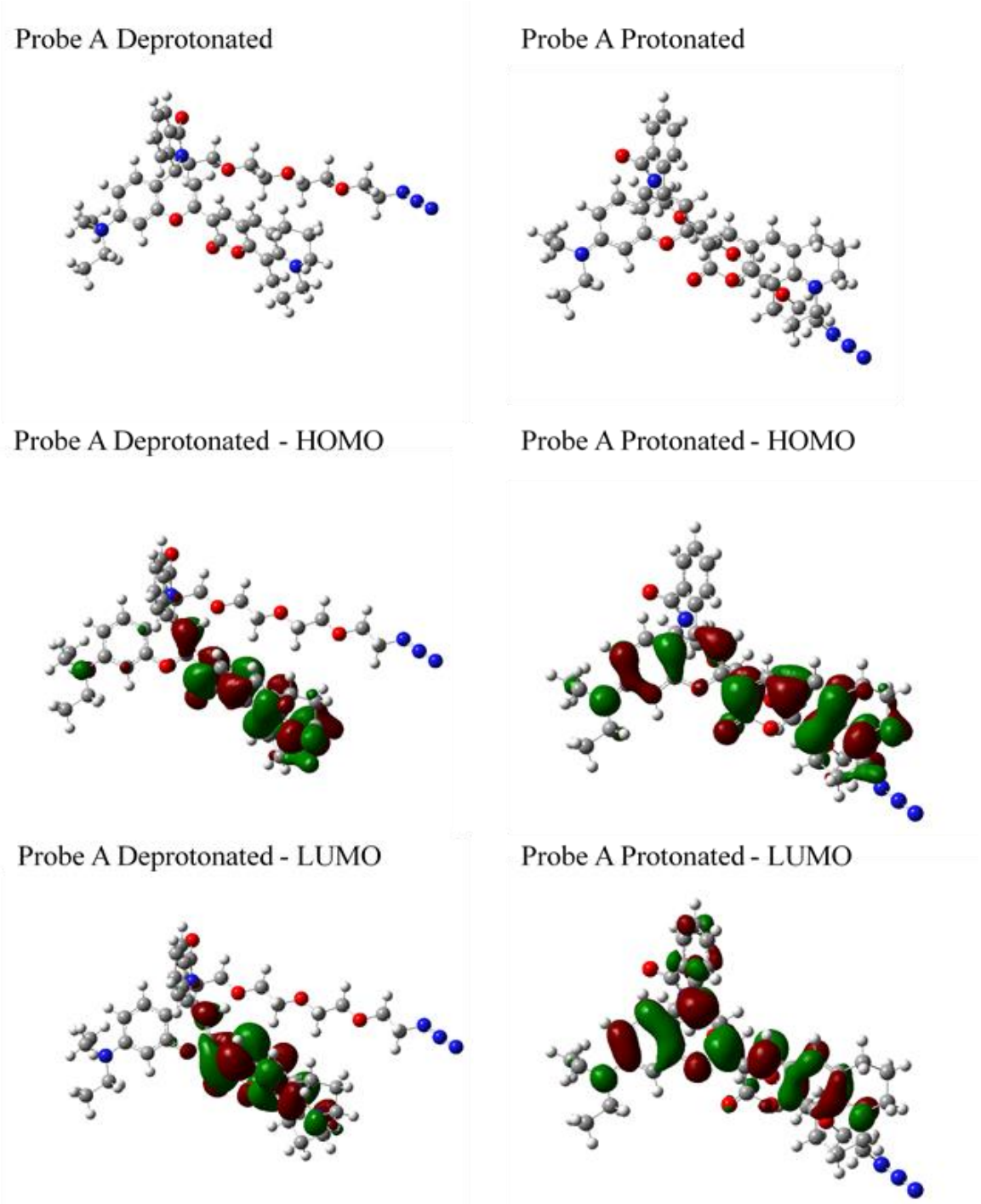

Figure 67: The optimised structures of the protonated and deprotonated forms of fluorescent probe A , with atoms coloured as follows carbon is grey, hydrogen is white, nitrogen is blue and oxygen is red. Visualisations of their HOMO and LUMO orbitals are coloured by wavefunction in either red or green.

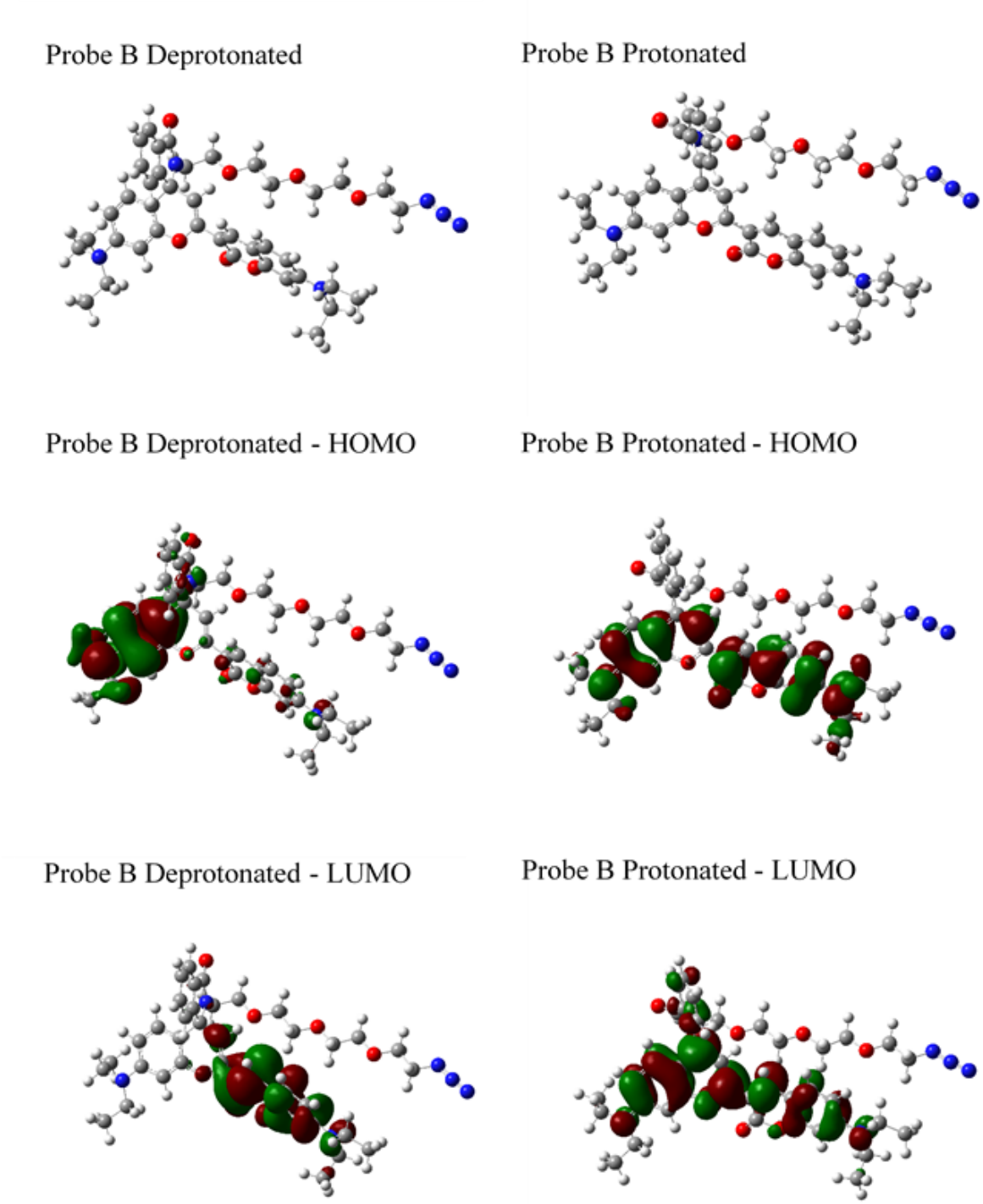

Figure 68: The optimised structures of the protonated and deprotonated forms of fluorescent probe B, with atoms coloured as follows carbon is grey, hydrogen is white, nitrogen is blue and oxygen is red. Visualisations of their HOMO and LUMO orbitals are coloured by wavefunction in either red or green.

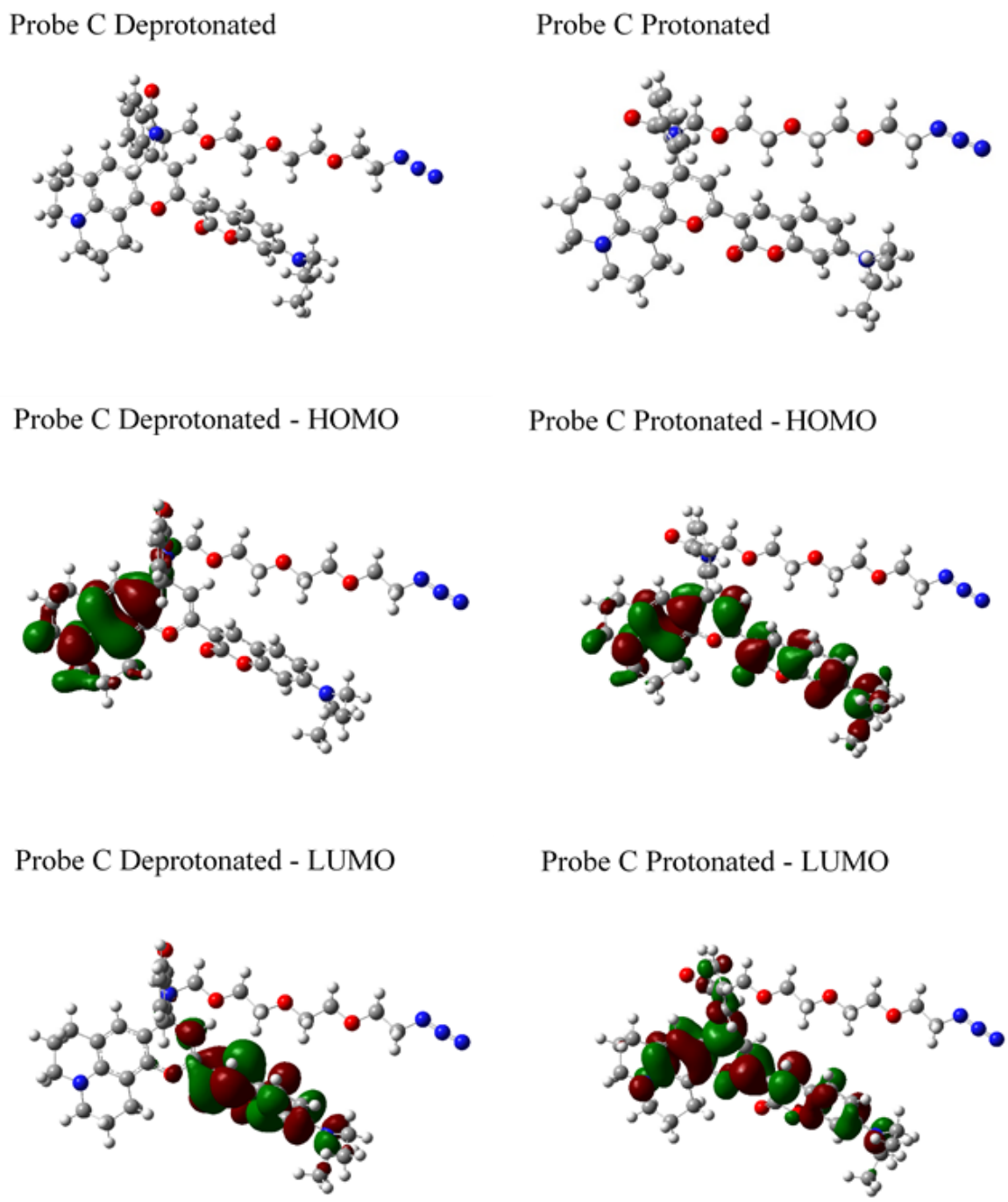

Figure 69: The optimised structures of the protonated and deprotonated forms of fluorescent probe C, with atoms coloured as follows carbon is grey, hydrogen is white, nitrogen is blue and oxygen is red. Visualisations of their HOMO and LUMO orbitals are coloured by wavefunction in either red or green.

In general, the structures of the deprotonated forms of the probes showed a conformation in which the tail moiety is found to become parallel to the fluorophore region (Figure 67, Figure 68 and Figure 69). Upon protonation and breakage of the spirolactam bridge the structures did not show much structural change, the tail moiety remained roughly parallel to the fluorophore moiety. The calculated absorption spectra of probes A, B and C in their deprotonated forms show maxima between 360nm and 400nm (Figure 65). The absorption spectra of the protonated forms of the probes show a broad peak around 570nm with a shoulder found at ~380nm (Figure 66). The calculations show similar differences in the positions of the maxima between the deprotonated and protonated forms as in the experimental spectra. Probe B in its protonated form also shows the highest oscillator strength as the experimental absorption spectra. The maximum in the spectra of protonated form of B is position at lowest wavelength in respect to protonated forms of A and C. The inspection of the energy difference between HOMO and LUMO shows a similar trend with Probe B having the largest energy gap in both protonated and deprotonated forms (Table 14). The HOMO orbital of the deprotonated probes B and C mainly shows contributions from the aromatic nitrogen atom of fluorophore moiety and some contribution from the phenyl ring adjacent to the spirolactam bond (Figure 67, Figure 68 and Figure 69). In the HOMO orbital of the deprotonated form of probe A, there is a major contribution from the coumarin moiety. The LUMO orbitals of all 3 probes in their deprotonated forms are found to be centred on the coumarin moiety. In the protonated form the HOMO orbitals of all 3 probes are very similar and exhibit contributions from the coumarin and fluorophore moiety. The LUMO orbitals of the protonated forms of the 3 probes are also very similar between the 3 probes, as well as showing distribution across the coumarin and fluorophore moeity we see some of the orbital also on the phenyl ring adjacent to the amide bond.

7 CONCLUSIONS

The molecular dynamics simulations of tryptophan halogenase gave us insight into the structural similarities of the two halogenase enzymes, PrnA and PyrH. From our MD simulations we found that the selectivity of the two enzymes probably arises from their distinct tryptophan binding sites which orient the tryptophan in the binding site so only one halogenated product is produced. Our comparisons of the MD simulations to crystal structures revealed the differences brought about when the proteins are crystallised. An FAD strap region that had previously been identified in PyrH was also shown to probably exist in PrnA and may hold the key to the two observed differences in the kinetics of the two enzymes. Analysis of the mutant structures of PrnA revealed that the likely reason for their inactivity is their inability to bind hypochlorous acid in the active site.

Our QM cluster models and QM/MM model of the FTO active site showed geometric parameters that were consistent with experimental results. Some of these QM clusters were used to calculate the formation of the Fe(IV)-oxo species successfully. Our differently sized QM clusters models seem to indicate that the cut form of 2OG is not an adequate replacement for the realistically sized 2OG ligand for the purposes of studying the FTO active site.

Our docking studies successfully predicted the structure of the bound complex of fructose to GLUT5 with results that were consistent with those predicted by experimental methodology. We successfully created a solvated membrane bound simulated model of GLUT5 and have performed some preliminary analysis of its conformational dynamics and shown the model structure to be stable at the timescale of our MD simulations.

The docked structures of the different sugar ligands to Odorranalectin were successfully created and their relative free energy of binding results reflected their actual relative binding

affinities measured by experimental methods. From the docked structures we were able to suggest why some sugars had a greater affinity to the lectin structure than others.

Our computational structures of the various fluorescent probes revealed possible structural reasons for their different fluorescent results. Our calculated absorption spectra of the probes showed some agreement with those derived by experimental means. The computed orbital structures were used to suggest electronic structural differences that may have led to the observed differences in fluorescent results between the various probes.

8 REFERENCES

1. Blake, C., et al., *Structure of hen egg-white lysozyme: a three-dimensional fourier synthesis at 2 Å resolution.* Nature, 1965. **206**(4986): p. 757-761.
2. Garcia-Viloca, M., et al., *How enzymes work: analysis by modern rate theory and computer simulations.* Science, 2004. **303**(5655): p. 186-195.
3. Friesner, R.A. and V. Guallar, *Ab initio quantum chemical and mixed quantum mechanics/molecular mechanics (QM/MM) methods for studying enzymatic catalysis.* Annu. Rev. Phys. Chem., 2005. **56**: p. 389-427.
4. Antoniou, D., et al., *Computational and theoretical methods to explore the relation between enzyme dynamics and catalysis.* Chemical reviews, 2006. **106**(8): p. 3170-3187.
5. Henzler-Wildman, K.A., et al., *A hierarchy of timescales in protein dynamics is linked to enzyme catalysis.* Nature, 2007. **450**(7171): p. 913.
6. van der Kamp, M.W. and A.J. Mulholland, *Combined Quantum Mechanics/Molecular Mechanics (QM/MM) Methods in Computational Enzymology.* Biochemistry, 2013. **52**(16): p. 2708-2728.
7. Warshel, A. and M. Karplus, *Calculation of ground and excited state potential surfaces of conjugated molecules. I. Formulation and parametrization.* Journal of the American Chemical Society, 1972. **94**(16): p. 5612-5625.
8. Kitchen, D.B., et al., *Docking and scoring in virtual screening for drug discovery: methods and applications.* Nat Rev Drug Discov, 2004. **3**(11): p. 935-949.
9. Ferreira, L.G., et al., *Molecular docking and structure-based drug design strategies.* Molecules, 2015. **20**(7): p. 13384-421.
10. Foresman, J.B. and Æ. Frisch, *Exploring chemistry with electronic structure methods: a guide to using Gaussian.* 1996.
11. Adcock, S.A. and J.A. McCammon, *Molecular dynamics: survey of methods for simulating the activity of proteins.* Chemical reviews, 2006. **106**(5): p. 1589-1615.
12. Hansson, T., C. Oostenbrink, and W. van Gunsteren, *Molecular dynamics simulations.* Current opinion in structural biology, 2002. **12**(2): p. 190-196.
13. Martín-García, F., et al., *Comparing Molecular Dynamics Force Fields in the Essential Subspace.* PLOS ONE, 2015. **10**(3): p. e0121114.
14. Hinchliffe, A., *Molecular modelling for beginners*. 2005: John Wiley & Sons.
15. Jones, J.E., *On the Determination of Molecular Fields. II. From the Equation of State of a Gas.* Proceedings of the Royal Society of London. Series A, 1924. **106**(738): p. 463-477.
16. Baker, C.M., *Polarizable force fields for molecular dynamics simulations of biomolecules.* Wiley Interdisciplinary Reviews: Computational Molecular Science, 2015. **5**(2): p. 241-254.
17. Frenkel, D. and B. Smit, *Understanding molecular simulation: from algorithms to applications*. Vol. 1. 2001: Academic press.
18. Rupley, J.A. and G. Careri, *Protein hydration and function.* Advances in protein chemistry, 1991. **41**: p. 37-172.
19. Berendsen, H.J., et al., *Interaction models for water in relation to protein hydration*, in *Intermolecular forces*. 1981, Springer. p. 331-342.
20. Berendsen, H., J. Grigera, and T. Straatsma, *The missing term in effective pair potentials.* Journal of Physical Chemistry, 1987. **91**(24): p. 6269-6271.
21. Jorgensen, W.L., et al., *Comparison of simple potential functions for simulating liquid water.* The Journal of chemical physics, 1983. **79**(2): p. 926-935.
22. Teeter, M.M., *Water-protein interactions: theory and experiment.* Annual review of biophysics and biophysical chemistry, 1991. **20**(1): p. 577-600.
23. Weiner, S.J., et al., *A new force field for molecular mechanical simulation of nucleic acids and proteins.* Journal of the American Chemical Society, 1984. **106**(3): p. 765-784.
24. Pearlman, D.A., et al., *AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to*

simulate the structural and energetic properties of molecules. Computer Physics Communications, 1995. **91**(1-3): p. 1-41.
25. Case, D.A., et al., *Amber 14.* 2014.
26. Wang, J., et al., *Development and testing of a general amber force field.* Journal of computational chemistry, 2004. **25**(9): p. 1157-1174.
27. Van Gunsteren, W. and H. Berendsen, *Groningen molecular simulation (GROMOS).* Library manual, Biomos, Groningen, The Netherlands, 1987: p. 1-221.
28. Brooks, B.R., et al., *CHARMM: A program for macromolecular energy, minimization, and dynamics calculations.* Journal of computational chemistry, 1983. **4**(2): p. 187-217.
29. Jorgensen, W.L. and J. Tirado-Rives, *The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin.* Journal of the American Chemical Society, 1988. **110**(6): p. 1657-1666.
30. Van Aalten, D.M., et al., *PRODRG, a program for generating molecular topologies and unique molecular descriptors from coordinates of small molecules.* Journal of computer-aided molecular design, 1996. **10**(3): p. 255-262.
31. Zoete, V., et al., *SwissParam: a fast force field generation tool for small organic molecules.* Journal of computational chemistry, 2011. **32**(11): p. 2359-2368.
32. Wang, J., et al., *Antechamber: an accessory software package for molecular mechanical calculations.* J. Am. Chem. Soc, 2001. **222**: p. U403.
33. Li, P. and K.M. Merz Jr, *MCPB. py: A python based metal center parameter builder*. 2016, ACS Publications.
34. Case, D., et al., *AMBER16 Package.* 2016.
35. Chothia, C. and A.M. Lesk, *The relation between the divergence of sequence and structure in proteins.* Embo j, 1986. **5**(4): p. 823-6.
36. Drews, J., *Drug discovery: a historical perspective.* Science, 2000. **287**(5460): p. 1960-1964.
37. Lengauer, T. and M. Rarey, *Computational methods for biomolecular docking.* Current Opinion in Structural Biology, 1996. **6**(3): p. 402-406.
38. Morris, G.M. and M. Lim-Wilby, *Molecular docking.* Methods Mol Biol, 2008. **443**: p. 365-82.
39. Xie, Z.R. and M.J. Hwang, *Methods for predicting protein-ligand binding sites.* Methods Mol Biol, 2015. **1215**: p. 383-98.
40. Davis, I.W., et al., *Blind docking of pharmaceutically relevant compounds using RosettaLigand.* Protein science, 2009. **18**(9): p. 1998-2002.
41. Ghersi, D. and R. Sanchez, *Improving accuracy and efficiency of blind protein-ligand docking by focusing on predicted binding sites.* Proteins: Structure, Function, and Bioinformatics, 2009. **74**(2): p. 417-424.
42. Morris, G.M., et al., *AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility.* Journal of computational chemistry, 2009. **30**(16): p. 2785-2791.
43. Jones, G., et al., *Development and validation of a genetic algorithm for flexible docking.* Journal of molecular biology, 1997. **267**(3): p. 727-748.
44. Yuriev, E. and P.A. Ramsland, *Latest developments in molecular docking: 2010–2011 in review.* Journal of Molecular Recognition, 2013. **26**(5): p. 215-239.
45. Taylor, R.D., P.J. Jewsbury, and J.W. Essex, *A review of protein-small molecule docking methods.* Journal of computer-aided molecular design, 2002. **16**(3): p. 151-166.
46. Halperin, I., et al., *Principles of docking: An overview of search algorithms and a guide to scoring functions.* Proteins: Structure, Function, and Bioinformatics, 2002. **47**(4): p. 409-443.
47. Whitley, D., *A genetic algorithm tutorial.* Statistics and computing, 1994. **4**(2): p. 65-85.
48. Ördög, R. and V. Grolmusz, *Evaluating genetic algorithms in protein-ligand docking.* Bioinformatics Research and Applications, 2008: p. 402-413.
49. Atkins, P.W., *Quanta: a handbook of concepts*. 1991: Clarendon Press.
50. Born, M., *The statistical interpretation of quantum mechanics.* Nobel Lecture, 1954. **11**: p. 1942-1962.

51. Atkins, P.W., *Quanta*. 1991: Clarendon Press.
52. Szabo, A. and N.S. Ostlund, *Modern quantum chemistry: introduction to advanced electronic structure theory*. 2012: Courier Corporation.
53. Born, M., *Born-Oppenheimer Approximation.* Quantum, 1927. **2**(2014/12): p. 4.
54. Ma, Z., *Quantum three-body problems.* Science in China Series A: Mathematics, 2000. **43**(10): p. 1093-1107.
55. Atkins, P.W., *Molecular quantum mechanics*. 1970: Oxford university press.
56. Schuster, P. and P. Wolschann, *Computational chemistry.* Monatshefte für Chemie - Chemical Monthly, 2008. **139**(4): p. III-IV.
57. Hartree, D.R. *The wave mechanics of an atom with a non-Coulomb central field. Part I. Theory and methods*. in *Mathematical Proceedings of the Cambridge Philosophical Society*. 1928. Cambridge Univ Press.
58. Jensen, F., *Introduction to computational chemistry*. 2013: John Wiley & Sons.
59. Hayward, D.O., *Quantum mechanics for chemists*. Vol. 14. 2002: Royal Society of Chemistry.
60. Dykstra, C., et al., *Theory and applications of computational chemistry: the first forty years*. 2011: Elsevier.
61. Bultinck, P., et al., *Computational medicinal chemistry for drug discovery*. 2003: CRC Press.
62. Leach, A.R., *Molecular modelling: principles and applications*. 2001: Pearson Education.
63. Thiel, W., *Semiempirical quantum–chemical methods.* Wiley Interdisciplinary Reviews: Computational Molecular Science, 2014. **4**(2): p. 145-157.
64. Zerner, M.C., *Semiempirical Molecular Orbital Methods*, in *Reviews in Computational Chemistry*. 2007, John Wiley & Sons, Inc. p. 313-365.
65. Eschrig, H., *The fundamentals of density functional theory*. Vol. 32. 1996: Springer.
66. Senn, H.M. and W. Thiel, *QM/MM methods for biomolecular systems.* Angewandte Chemie International Edition, 2009. **48**(7): p. 1198-1229.
67. Zhao, Y. and D.G. Truhlar, *Benchmark Databases for Nonbonded Interactions and Their Use To Test Density Functional Theory.* Journal of Chemical Theory and Computation, 2005. **1**(3): p. 415-432.
68. Goerigk, L. and S. Grimme, *A thorough benchmark of density functional methods for general main group thermochemistry, kinetics, and noncovalent interactions.* Physical Chemistry Chemical Physics, 2011. **13**(14): p. 6670-6688.
69. Becke, A.D., *A new mixing of Hartree–Fock and local density-functional theories.* The Journal of Chemical Physics, 1993. **98**(2): p. 1372-1377.
70. Zhang, Y., X. Xu, and W.A. Goddard, *Doubly hybrid density functional for accurate descriptions of nonbond interactions, thermochemistry, and thermochemical kinetics.* Proceedings of the National Academy of Sciences, 2009. **106**(13): p. 4963-4968.
71. Graham, D.C., et al., *Optimization and basis-set dependence of a restricted-open-shell form of B2-PLYP double-hybrid density functional theory.* The Journal of Physical Chemistry A, 2009. **113**(36): p. 9861-9873.
72. Huenerbein, R., et al., *Effects of London dispersion on the isomerization reactions of large organic molecules: a density functional benchmark study.* Physical Chemistry Chemical Physics, 2010. **12**(26): p. 6940-6948.
73. Goerigk, L. and S. Grimme, *Double-hybrid density functionals.* Wiley Interdisciplinary Reviews: Computational Molecular Science, 2014. **4**(6): p. 576-600.
74. Meier, K., et al., *Multi-Resolution Simulation of Biomolecular Systems: A Review of Methodological Issues.* Angewandte Chemie-International Edition, 2013. **52**(10): p. 2820-2834.
75. Mulholland, A.J., *Modelling enzyme reaction mechanisms, specificity and catalysis.* Drug Discovery Today, 2005. **10**(20): p. 1393-1402.
76. Ranaghan, K.E., et al., *Transition state stabilization and substrate strain in enzyme catalysis: ab initio QM/MM modelling of the chorismate mutase reaction.* Organic & biomolecular chemistry, 2004. **2**(7): p. 968-980.

77. Hu, H. and W.T. Yang, *Free energies of chemical reactions in solution and in enzymes with ab initio quantum mechanics/molecular mechanics methods*, in *Annual Review of Physical Chemistry*. 2008. p. 573-601.
78. Blumberger, J., *Free energies for biological electron transfer from QM/MM calculation: method, application and critical assessment.* Physical Chemistry Chemical Physics, 2008. **10**(37): p. 5651-5667.
79. Senn, H.M. and W. Thiel, *QM/MM studies of enzymes.* Current opinion in chemical biology, 2007. **11**(2): p. 182-187.
80. Sousa, S.F., P.A. Fernandes, and M.J. Ramos, *Computational enzymatic catalysis - clarifying enzymatic mechanisms with the help of computers.* Physical Chemistry Chemical Physics, 2012. **14**(36): p. 12431-12441.
81. Kang, J. and M. Tateno, *Recent Applications of Hybrid Ab Initio Quantum Mechanics–Molecular Mechanics Simulations to Biological Macromolecules*. Some Applications of Quantum Mechanics. 2012: InTech.
82. Svensson, M., et al., *ONIOM: A multilayered integrated MO+ MM method for geometry optimizations and single point energy predictions. A test for Diels-Alder reactions and Pt (P (t-Bu) 3) 2+ H2 oxidative addition.* The Journal of Physical Chemistry, 1996. **100**(50): p. 19357-19363.
83. Lonsdale, R., J.N. Harvey, and A.J. Mulholland, *A practical guide to modelling enzyme-catalysed reactions.* Chemical Society Reviews, 2012. **41**(8): p. 3025-3038.
84. Singh, U.C. and P.A. Kollman, *A combined ab initio quantum mechanical and molecular mechanical method for carrying out simulations on complex molecular systems: applications to the CH3Cl+ Cl− exchange reaction and gas phase protonation of polyethers.* Journal of Computational Chemistry, 1986. **7**(6): p. 718-730.
85. Hu, H. and W.T. Yang, *Development and application of ab initio QM/MM methods for mechanistic simulation of reactions in solution and in enzymes.* Journal of Molecular Structure-Theochem, 2009. **898**(1-3): p. 17-30.
86. Lin, H. and D.G. Truhlar, *QM/MM: what have we learned, where are we, and where do we go from here?* Theoretical Chemistry Accounts, 2007. **117**(2): p. 185-199.
87. Zhang, Y., T.-S. Lee, and W. Yang, *A pseudobond approach to combining quantum mechanical and molecular mechanical methods.* The Journal of chemical physics, 1999. **110**(1): p. 46-54.
88. Philipp, D.M. and R.A. Friesner, *Mixed ab initio QM/MM modeling using frozen orbitals and tests with alanine dipeptide and tetrapeptide.* Journal of computational chemistry, 1999. **20**(14): p. 1468-1494.
89. Amara, P., et al., *The generalized hybrid orbital method for combined quantum mechanical/molecular mechanical calculations: Formulation and tests of the analytical derivatives.* Theoretical Chemistry Accounts, 2000. **104**(5): p. 336-343.
90. Gao, J., et al., *A generalized hybrid orbital (GHO) method for the treatment of boundary atoms in combined QM/MM calculations.* The Journal of Physical Chemistry A, 1998. **102**(24): p. 4714-4721.
91. Vreven, T., et al., *Combining quantum mechanics methods with molecular mechanics methods in ONIOM.* Journal of Chemical Theory and Computation, 2006. **2**(3): p. 815-826.
92. Vreven, T., et al., *Geometry optimization with QM/MM, ONIOM, and other combined methods. I. Microiterations and constraints.* Journal of computational chemistry, 2003. **24**(6): p. 760-769.
93. Pronk, S., et al., *GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit.* Bioinformatics, 2013: p. btt055.
94. Frisch, M., et al., *Gaussian 09, revision D. 01*. 2009, Gaussian, Inc., Wallingford CT.
95. Jukes, T.H., *Some historical notes on chlortetracycline.* Reviews of infectious diseases, 1985. **7**(5): p. 702-707.
96. Moellering, R.C., *Vancomycin: a 50-year reassessment.* Clinical Infectious Diseases, 2006. **42**(Supplement 1): p. S3-S4.

97. Hammer, P.E., et al., *Four genes from Pseudomonas fluorescens that encode the biosynthesis of pyrrolnitrin.* Applied and Environmental Microbiology, 1997. **63**(6): p. 2147-2154.
98. Feling, R.H., et al., *Salinosporamide A: a highly cytotoxic proteasome inhibitor from a novel microbial source, a marine bacterium of the new genus Salinospora.* Angewandte Chemie International Edition, 2003. **42**(3): p. 355-357.
99. Neumann, C.S., D.G. Fujimori, and C.T. Walsh, *Halogenation strategies in natural product biosynthesis.* Chemistry & biology, 2008. **15**(2): p. 99-109.
100. Long, B.H., et al., *Discovery of antitumor indolocarbazoles: rebeccamycin, NSC 655649, and fluoroindolocarbazoles.* Current Medicinal Chemistry-Anti-Cancer Agents, 2002. **2**(2): p. 255-266.
101. Eguchi, S., *Bioactive heterocycles*. 2006: Springer.
102. Young, I.S. and P.S. Baran, *Protecting-group-free synthesis as an opportunity for invention.* Nature chemistry, 2009. **1**(3): p. 193-205.
103. Wang, X., B.S. Lane, and D. Sames, *Direct C-arylation of free (NH)-indoles and pyrroles catalyzed by Ar-Rh (III) complexes assembled in situ.* Journal of the American Chemical Society, 2005. **127**(14): p. 4996-4997.
104. Anderson, J.R. and S.K. Chapman, *Molecular mechanisms of enzyme-catalysed halogenation.* Molecular Biosystems, 2006. **2**(8): p. 350-357.
105. Shepherd, S.A., et al., *Extending the biocatalytic scope of regiocomplementary flavin-dependent halogenase enzymes.* Chemical Science, 2015. **6**(6): p. 3454-3460.
106. Zehner, S., et al., *A regioselective tryptophan 5-halogenase is involved in pyrroindomycin biosynthesis in Streptomyces rugosporus LL-42D005.* Chemistry & biology, 2005. **12**(4): p. 445-452.
107. Seibold, C., et al., *A flavin-dependent tryptophan 6-halogenase and its use in modification of pyrrolnitrin biosynthesis.* Biocatalysis and Biotransformation, 2006. **24**(6): p. 401-408.
108. Shepherd, S.A., et al., *A Structure-Guided Switch in the Regioselectivity of a Tryptophan Halogenase.* ChemBioChem, 2016. **17**(9): p. 821-824.
109. Bitto, E., et al., *The structure of flavin-dependent tryptophan 7-halogenase RebH.* Proteins: Structure, Function, and Bioinformatics, 2008. **70**(1): p. 289-293.
110. Dong, C., et al., *Tryptophan 7-halogenase (PrnA) structure suggests a mechanism for regioselective chlorination.* Science, 2005. **309**(5744): p. 2216-2219.
111. Zhu, X., et al., *Structural insights into regioselectivity in the enzymatic chlorination of tryptophan.* Journal of molecular biology, 2009. **391**(1): p. 74-85.
112. Dong, C., et al., *Crystallization and X-ray diffraction of a halogenating enzyme, tryptophan 7-halogenase, from Pseudomonas fluorescens.* Acta Crystallographica Section D: Biological Crystallography, 2004. **60**(8): p. 1438-1440.
113. Yeh, E., et al., *Chlorination by a Long-Lived Intermediate in the Mechanism of Flavin-Dependent Halogenases†,‖.* Biochemistry, 2007. **46**(5): p. 1284-1292.
114. van Pée, K.-H., *Biosynthesis of halogenated metabolites by bacteria.* Annual Reviews in Microbiology, 1996. **50**(1): p. 375-399.
115. Yeh, E., S. Garneau, and C.T. Walsh, *Robust in vitro activity of RebF and RebH, a two-component reductase/halogenase, generating 7-chlorotryptophan during rebeccamycin biosynthesis.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(11): p. 3960-3965.
116. Flecks, S., et al., *New insights into the mechanism of enzymatic chlorination of tryptophan.* Angewandte Chemie International Edition, 2008. **47**(49): p. 9533-9536.
117. Prlić, A., et al., *Pre-calculated protein structure alignments at the RCSB PDB website.* Bioinformatics, 2010. **26**(23): p. 2983-2985.
118. Smith, T.F. and M.S. Waterman, *Identification of common molecular subsequences.* Journal of molecular biology, 1981. **147**(1): p. 195-197.
119. Karplus, M. and J.A. McCammon, *Molecular dynamics simulations of biomolecules.* Nature Structural & Molecular Biology, 2002. **9**(9): p. 646-652.

120. Karplus, M. and J. Kuriyan, *Molecular dynamics and protein function.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(19): p. 6679-6685.
121. Kruschel, D. and B. Zagrovic, *Conformational averaging in structural biology: issues, challenges and computational solutions.* Molecular Biosystems, 2009. **5**(12): p. 1606-1616.
122. Henzler-Wildman, K.A., et al., *Intrinsic motions along an enzymatic reaction trajectory.* Nature, 2007. **450**(7171): p. 838-844.
123. Csermely, P., R. Palotai, and R. Nussinov, *Induced fit, conformational selection and independent dynamic segments: an extended view of binding events.* Trends in biochemical sciences, 2010. **35**(10): p. 539-546.
124. Orozco, M. and F.J. Luque, *Theoretical methods for the description of the solvent effect in biomolecular systems.* Chemical Reviews, 2000. **100**(11): p. 4187-4226.
125. Grant, B.J., A.A. Gorfe, and J.A. McCammon, *Large conformational changes in proteins: signaling and other functions.* Current opinion in structural biology, 2010. **20**(2): p. 142-147.
126. Release, S.d., *3: Maestro;, version 9.9 Schrödinger.* LLC: New York, NY, 2014.
127. SchuÈttelkopf, A.W. and D.M. Van Aalten, *PRODRG: a tool for high-throughput crystallography of protein–ligand complexes.* Acta Crystallographica Section D: Biological Crystallography, 2004. **60**(8): p. 1355-1363.
128. Schuler, L.D., X. Daura, and W.F. Van Gunsteren, *An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase.* Journal of Computational Chemistry, 2001. **22**(11): p. 1205-1218.
129. Malde, A.K., et al., *An automated force field topology builder (ATB) and repository: version 1.0.* Journal of chemical theory and computation, 2011. **7**(12): p. 4026-4037.
130. Fiser, A. and A. Šali, *Modeller: generation and refinement of homology-based protein structure models.* Methods in enzymology, 2003. **374**: p. 461-491.
131. Pettersen, E.F., et al., *UCSF Chimera—a visualization system for exploratory research and analysis.* Journal of computational chemistry, 2004. **25**(13): p. 1605-1612.
132. Humphrey, W., A. Dalke, and K. Schulten, *VMD: visual molecular dynamics.* Journal of molecular graphics, 1996. **14**(1): p. 33-38.
133. Grant, B.J., et al., *Bio3d: an R package for the comparative analysis of protein structures.* Bioinformatics, 2006. **22**(21): p. 2695-2696.
134. Studio, R., *RStudio: integrated development environment for R.* RStudio Inc, Boston, Massachusetts, 2012.
135. Singh, W., et al., *Effects of Mutations on Structure–Function Relationships of Matrix Metalloproteinase-1.* International Journal of Molecular Sciences, 2016. **17**(10): p. 1727.
136. Grossfield, A. and D.M. Zuckerman, *Quantifying uncertainty and sampling quality in biomolecular simulations.* Annual reports in computational chemistry, 2009. **5**: p. 23-48.
137. Andorfer, M.C., et al., *Directed evolution of RebH for catalyst-controlled halogenation of indole C–H bonds.* Chemical Science, 2016. **7**(6): p. 3720-3729.
138. Frayling, T.M., et al., *A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity.* Science, 2007. **316**(5826): p. 889-94.
139. Loos, R.J.F. and G.S.H. Yeo, *The bigger picture of FTO – the first GWAS-identified obesity gene.* Nature reviews. Endocrinology, 2014. **10**(1): p. 51-61.
140. Gerken, T., et al., *The obesity-associated FTO gene encodes a 2-oxoglutarate-dependent nucleic acid demethylase.* Science, 2007. **318**(5855): p. 1469-1472.
141. Goode, E.L., C.M. Ulrich, and J.D. Potter, *Polymorphisms in DNA repair genes and associations with cancer risk.* Cancer Epidemiology Biomarkers & Prevention, 2002. **11**(12): p. 1513-1530.
142. Lurie, G., et al., *The obesity-associated polymorphisms FTO rs9939609 and MC4R rs17782313 and endometrial cancer risk in non-Hispanic white women.* PLoS One, 2011. **6**(2): p. e16756.
143. Lewis, S.J., et al., *Associations between an obesity related genetic variant (FTO rs9939609) and prostate cancer risk.* PLoS One, 2010. **5**(10): p. e13485.

144. Kaklamani, V., et al., *The role of the fat mass and obesity associated gene (FTO) in breast cancer risk.* BMC medical genetics, 2011. **12**(1): p. 52.
145. Hess, M.E. and J.C. Brüning, *The fat mass and obesity-associated (FTO) gene: obesity and beyond?* Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 2014. **1842**(10): p. 2039-2047.
146. Keller, L., et al., *The obesity related gene, FTO, interacts with APOE, and is associated with Alzheimer's disease risk: a prospective cohort study.* Journal of Alzheimer's Disease, 2011. **23**(3): p. 461-469.
147. Ng, M.C., et al., *Implication of genetic variants near TCF7L2, SLC30A8, HHEX, CDKAL1, CDKN2A/B, IGF2BP2, and FTO in type 2 diabetes and obesity in 6,719 Asians.* Diabetes, 2008. **57**(8): p. 2226-2233.
148. Abu-Omar, M.M., A. Loaiza, and N. Hontzeas, *Reaction mechanisms of mononuclear non-heme iron oxygenases.* Chemical reviews, 2005. **105**(6): p. 2227-2252.
149. Almén, M.S., et al., *Genome wide analysis reveals association of a FTO gene variant with epigenetic changes.* Genomics, 2012. **99**(3): p. 132-137.
150. Quesne, M.G., et al., *Quantum mechanics/molecular mechanics study on the oxygen binding and substrate hydroxylation step in AlkB repair enzymes.* Chemistry-A European Journal, 2014. **20**(2): p. 435-446.
151. Schofield, C.J. and R.P. Hausinger, *2-oxoglutarate-dependent Oxygenases*. 2015: Royal Society of Chemistry.
152. Han, Z., et al., *Crystal structure of the FTO protein reveals basis for its substrate specificity.* Nature, 2010. **464**(7292): p. 1205.
153. Ozer, A. and R.K. Bruick, *Non-heme dioxygenases: cellular sensors and regulators jelly rolled into one?* Nature chemical biology, 2007. **3**(3): p. 144.
154. Markolovic, S., S.E. Wilkins, and C.J. Schofield, *Protein Hydroxylation Catalyzed by 2-Oxoglutarate-dependent Oxygenases.* The Journal of Biological Chemistry, 2015. **290**(34): p. 20712-20722.
155. Jia, G., et al., *Oxidative demethylation of 3-methylthymine and 3-methyluracil in single-stranded DNA and RNA by mouse and human FTO.* FEBS letters, 2008. **582**(23-24): p. 3313-3319.
156. Jia, G., et al., *N6-Methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO.* Nat Chem Biol, 2011. **7**(12): p. 885-887.
157. Aik, W., et al., *Structural basis for inhibition of the fat mass and obesity associated protein (FTO).* Journal of medicinal chemistry, 2013. **56**(9): p. 3680-3688.
158. McDonough, M.A., et al., *Structural studies on human 2-oxoglutarate dependent oxygenases.* Current opinion in structural biology, 2010. **20**(6): p. 659-672.
159. Rose, N.R., et al., *Inhibition of 2-oxoglutarate dependent oxygenases.* Chemical Society reviews, 2011. **40**(8): p. 4364-4397.
160. Schramm, V.L., *Enzymatic Transition States, Transition-State Analogs, Dynamics, Thermodynamics, and Lifetimes.* Annual Review of Biochemistry, 2011. **80**(1): p. 703-732.
161. Kollman, P.A., B. Kuhn, and M. Peräkylä, *Computational Studies of Enzyme-Catalyzed Reactions: Where Are We in Predicting Mechanisms and in Understanding the Nature of Enzyme Catalysis?* The Journal of Physical Chemistry B, 2002. **106**(7): p. 1537-1542.
162. Saen-oon, S., et al., *Atomic detail of chemical transformation at the transition state of an enzymatic reaction.* Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(43): p. 16543-16548.
163. Wang, B., et al., *Computations Reveal a Rich Mechanistic Variation of Demethylation of N-Methylated DNA/RNA Nucleotides by FTO.* ACS Catalysis, 2015. **5**(12): p. 7077-7090.
164. Blomberg, M.R., et al., *Quantum chemical studies of mechanisms for metalloenzymes.* Chemical reviews, 2014. **114**(7): p. 3601-3658.
165. Diebold, A.R., et al., *Activation of α-keto acid-dependent dioxygenases: application of an {FeNO} 7/{FeO2} 8 methodology for characterizing the initial steps of O2 activation.* Journal of the American Chemical Society, 2011. **133**(45): p. 18148-18160.

166. Dolinsky, T.J., et al., *PDB2PQR: an automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations.* Nucleic acids research, 2004. **32**(suppl_2): p. W665-W667.
167. Dennington, R., et al., *GaussView*. 2009, Version.
168. Thinnes, C.C., et al., *Targeting histone lysine demethylases—progress, challenges, and the future.* Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms, 2014. **1839**(12): p. 1416-1432.
169. Chowdhury, R., et al., *The oncometabolite 2-hydroxyglutarate inhibits histone lysine demethylases.* EMBO reports, 2011. **12**(5): p. 463-469.
170. Becke, A.D., *Density-functional exchange-energy approximation with correct asymptotic behavior.* Physical Review A, 1988. **38**(6): p. 3098-3100.
171. Perdew, J.P., *Density-functional approximation for the correlation energy of the inhomogeneous electron gas.* Physical Review B, 1986. **33**(12): p. 8822-8824.
172. Barone, V. and M. Cossi, *Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model.* The Journal of Physical Chemistry A, 1998. **102**(11): p. 1995-2001.
173. Svensson, M., et al., *ONIOM: a multilayered integrated MO+ MM method for geometry optimizations and single point energy predictions. A test for Diels− Alder reactions and Pt (P (t-Bu) 3) 2+ H2 oxidative addition.* The Journal of Physical Chemistry, 1996. **100**(50): p. 19357-19363.
174. Chen, L.Q., et al., *Transport of Sugars*, in *Annual Review of Biochemistry, Vol 84*, R.D. Kornberg, Editor. 2015, Annual Reviews: Palo Alto. p. 865-894.
175. Yan, N., *A Glimpse of Membrane Transport through Structures—Advances in the Structural Biology of the GLUT Glucose Transporters.* Journal of Molecular Biology, 2017. **429**(17): p. 2710-2725.
176. Deng, D. and N. Yan, *GLUT, SGLT, and SWEET: Structural and mechanistic investigations of the glucose transporters.* Protein Science, 2016. **25**(3): p. 546-558.
177. Long, W.T. and C.I. Cheeseman, *Structure of, and functional insight into the GLUT family of membrane transporters.* Cell Health and Cytoskeleton, 2015. **7**: p. 167-183.
178. Fedie, J.R., *Fluorescent Probe Development for Fructose Specific Transporters in Cancer*. 2017, Michigan Technological University.
179. Uldry, M. and B. Thorens, *The SLC2 family of facilitated hexose and polyol transporters.* Pflügers Archiv, 2004. **447**(5): p. 480-489.
180. Corpe, C.P., et al., *The regulation of GLUT5 and GLUT2 activity in the adaptation of intestinal brush-border fructose transport in diabetes.* Pflügers Archiv European Journal of Physiology, 1996. **432**(2): p. 192-201.
181. Castello, A., et al., *Regulation of GLUT5 gene expression in rat intestinal mucosa: regional distribution, circadian rhythm, perinatal development and effect of diabetes.* Biochemical Journal, 1995. **309**(1): p. 271-277.
182. Douard, V. and R.P. Ferraris, *Regulation of the fructose transporter GLUT5 in health and disease.* American Journal of Physiology-Endocrinology and Metabolism, 2008. **295**(2): p. E227-E237.
183. Lim, J.S., et al., *The role of fructose in the pathogenesis of NAFLD and the metabolic syndrome.* Nature reviews gastroenterology and hepatology, 2010. **7**(5): p. 251-264.
184. Zamora-León, S.P., et al., *Expression of the fructose transporter GLUT5 in human breast cancer.* Proceedings of the National Academy of Sciences, 1996. **93**(5): p. 1847-1852.
185. Godoy, A., et al., *Differential subcellular distribution of glucose transporters GLUT1–6 and GLUT9 in human cancer: ultrastructural localization of GLUT1 and GLUT5 in breast tumor tissues.* Journal of cellular physiology, 2006. **207**(3): p. 614-627.
186. Yamamoto, T., et al., *Over-expression of facilitative glucose transporter genes in human cancer.* Biochemical and biophysical research communications, 1990. **170**(1): p. 223-230.
187. Medina, R.A. and G.I. Owen, *Glucose transporters: expression, regulation and cancer.* Biological research, 2002. **35**(1): p. 9-26.

188. Szablewski, L., *Expression of glucose transporters in cancers.* Biochimica et Biophysica Acta (BBA)-Reviews on Cancer, 2013. **1835**(2): p. 164-169.
189. Deng, D., et al., *Crystal structure of the human glucose transporter GLUT1.* Nature, 2014. **510**(7503): p. 121.
190. Abramson, J., et al., *The lactose permease of Escherichia coli: overall structure, the sugar-binding site and the alternating access model for transport.* FEBS letters, 2003. **555**(1): p. 96-101.
191. Shi, Y., *Common folds and transport mechanisms of secondary active transporters.* Annual review of biophysics, 2013. **42**: p. 51-72.
192. Deng, D., et al., *Molecular basis of ligand recognition and transport by glucose transporters.* Nature, 2015. **526**(7573): p. 391.
193. Schürmann, A., et al., *Role of conserved arginine and glutamate residues on the cytosolic surface of glucose transporters for transporter function.* Biochemistry, 1997. **36**(42): p. 12897-12902.
194. Nomura, N., et al., *Structure and mechanism of the mammalian fructose transporter GLUT5.* Nature, 2015. **526**(7573): p. 397.
195. Lee, A.G., *How lipids affect the activities of integral membrane proteins.* Biochimica et Biophysica Acta (BBA)-Biomembranes, 2004. **1666**(1): p. 62-87.
196. Phillips, R., et al., *Emerging roles for lipids in shaping membrane-protein function.* Nature, 2009. **459**(7245): p. 379.
197. Laganowsky, A., et al., *Membrane proteins bind lipids selectively to modulate their structure and function.* Nature, 2014. **510**(7503): p. 172.
198. Thompson, A.M.G., et al., *Discovery of a specific inhibitor of human GLUT5 by virtual screening and in vitro transport evaluation.* Scientific reports, 2016. **6**.
199. Som, P., et al., *A fluorinated glucose analog, 2-fluoro-2-deoxy-D-glucose (F-18): nontoxic tracer for rapid tumor detection.* J Nucl Med, 1980. **21**(7): p. 670-5.
200. Chandra, R.V. and J.A.J. King, *10 - Advanced imaging of brain tumors A2 - Kaye, Andrew H*, in *Brain Tumors (Third Edition)*, E.R. Laws, Editor. 2012, W.B. Saunders: Edinburgh. p. 188-213.
201. McLean, I.D. and J. Martensen, *Chapter 2 - Specialized Imaging A2 - Marchiori, Dennis M*, in *Clinical Imaging (Third Edition)*. 2014, Mosby: Saint Louis. p. 44-78.
202. Krug, B., et al., *Activity-based costing evaluation of [18F]-fludeoxyglucose production.* European journal of nuclear medicine and molecular imaging, 2008. **35**(1): p. 80-88.
203. Frangioni, J.V., *In vivo near-infrared fluorescence imaging.* Current opinion in chemical biology, 2003. **7**(5): p. 626-634.
204. Sokolova, I.A., et al., *The development of a multitarget, multicolor fluorescence in situ hybridization assay for the detection of urothelial carcinoma in urine.* The Journal of Molecular Diagnostics, 2000. **2**(3): p. 116-123.
205. Weissleder, R., *Scaling down imaging: molecular mapping of cancer in mice.* Nature reviews. Cancer, 2002. **2**(1): p. 11.
206. Levi, J., et al., *Fluorescent fructose derivatives for imaging breast cancer cells.* Bioconjugate chemistry, 2007. **18**(3): p. 628-634.
207. Sánchez, R. and A. Šali, *Comparative protein structure modeling: introduction and practical examples with modeller.* Protein structure prediction: Methods and protocols, 2000: p. 97-129.
208. Webb, B. and A. Sali, *Protein structure modeling with MODELLER.* Protein Structure Prediction, 2014: p. 1-15.
209. Kim, K. and K.D. Jordan, *Comparison of Density Functional and MP2 Calculations on the Water Monomer and Dimer.* The Journal of Physical Chemistry, 1994. **98**(40): p. 10089-10094.
210. Stephens, P.J., et al., *Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields.* The Journal of Physical Chemistry, 1994. **98**(45): p. 11623-11627.

211. Jo, S., et al., *CHARMM-GUI Membrane Builder for mixed bilayers and its application to yeast membranes.* Biophysical journal, 2009. **97**(1): p. 50-58.
212. Fu, X., et al., *Mechanistic study of human glucose transport mediated by GLUT1.* Journal of chemical information and modeling, 2016. **56**(3): p. 517-526.
213. Park, M.-S., *Molecular dynamics simulations of the human glucose transporter GLUT1.* PloS one, 2015. **10**(4): p. e0125361.
214. Case, D., et al., *The FF14SB force field.* Amber, 2014. **14**: p. 29-31.
215. Dickson, C.J., et al., *Lipid14: the amber lipid force field.* Journal of chemical theory and computation, 2014. **10**(2): p. 865-879.
216. Kollman, P.A., et al., *Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models.* Acc Chem Res, 2000. **33**(12): p. 889-97.
217. Stanley, P., H. Schachter, and N. Taniguchi, *Essentials of glycobiology.* Varki, A, 2009.
218. Goldstein, I.J., et al., *What should be called a lectin?* Nature, 1980. **285**(5760): p. 66-66.
219. Ambrosi, M., N.R. Cameron, and B.G. Davis, *Lectins: tools for the molecular understanding of the glycocode.* Organic & biomolecular chemistry, 2005. **3**(9): p. 1593-1608.
220. Ernst, B. and J.L. Magnani, *From carbohydrate leads to glycomimetic drugs.* Nature reviews. Drug discovery, 2009. **8**(8): p. 661.
221. Sharon, N., *Lectins: past, present and future1*. 2008, Portland Press Limited.
222. Sharon, N. and H. Lis, *History of lectins: from hemagglutinins to biological recognition molecules.* Glycobiology, 2004. **14**(11): p. 53R-62R.
223. Liener, I., *The lectins: properties, functions, and applications in biology and medicine*. 2012: Elsevier.
224. Armstrong, G.D., *Carbohydrate receptor mimicry as a basis for antibacterial therapy.* Curr Opin Drug Discov Devel, 2000. **3**(2): p. 191-202.
225. Kannagi, R., et al., *Carbohydrate-mediated cell adhesion in cancer metastasis and angiogenesis.* Cancer science, 2004. **95**(5): p. 377-384.
226. Reis, C.A., et al., *Alterations in glycosylation as biomarkers for cancer detection.* Journal of clinical pathology, 2010. **63**(4): p. 322-329.
227. Dalziel, M., et al., *Emerging principles for the therapeutic exploitation of glycosylation.* Science, 2014. **343**(6166): p. 1235681.
228. Peracaula, R., et al., *Altered glycosylation in tumours focused to cancer diagnosis.* Disease markers, 2008. **25**(4-5): p. 207-218.
229. Aoyagi, Y., et al., *The fucosylation index of alpha-fetoprotein and its usefulness in the early diagnosis of hepatocellular carcinoma.* Cancer, 1988. **61**(4): p. 769-774.
230. Zhou, L., J. Liu, and F. Luo, *Serum tumor markers for detection of hepatocellular carcinoma.* World J Gastroenterol, 2006. **12**(8): p. 1175-81.
231. Sato, Y., et al., *Early recognition of hepatocellular carcinoma based on altered profiles of alpha-fetoprotein.* New England Journal of Medicine, 1993. **328**(25): p. 1802-1806.
232. Stock, A.M., et al., *Targets for anti-metastatic drug development.* Curr Pharm Des, 2013. **19**(28): p. 5127-34.
233. Köhler, S., et al., *E-/P-selectins and colon carcinoma metastasis: first in vivo evidence for their crucial role in a clinically relevant model of spontaneous metastasis formation in the lung.* British Journal of Cancer, 2010. **102**(3): p. 602-609.
234. Liu, B., H.-j. Bian, and J.-k. Bao, *Plant lectins: Potential antineoplastic drugs from bench to clinic.* Cancer Letters, 2010. **287**(1): p. 1-12.
235. Yin, J., et al., *Hypoxic Culture Induces Expression of Sialin, a Sialic Acid Transporter, and Cancer-Associated Gangliosides Containing Non–Human Sialic Acid on Human Cancer Cells.* Cancer Research, 2006. **66**(6): p. 2937.
236. Chen, J., et al., *A novel sialic acid-specific lectin from Phaseolus coccineus seeds with potent antineoplastic and antifungal activities.* Phytomedicine, 2009. **16**(4): p. 352-360.
237. Irie, A., et al., *Galectin-9 as a prognostic factor with antimetastatic potential in breast cancer.* Clinical Cancer Research, 2005. **11**(8): p. 2962-2968.

238. Park, W.-B., et al., *Inhibition of tumor growth and metastasis by Korean mistletoe lectin is associated with apoptosis and antiangiogenesis.* Cancer Biotherapy and Radiopharmaceuticals, 2001. **16**(5): p. 439-447.
239. Tung, T.C., *Orally administrable anti-metastatic lectin compositions and methods*. 1991, Google Patents.
240. Testa, J.E., et al., *Eukaryotic Expression Cloning with an Antimetastatic Monoclonal Antibody Identifies a Tetraspanin (PETA-3/CD151) as an Effector of Human TumorCell Migration and Metastasis.* Cancer Research, 1999. **59**(15): p. 3812-3820.
241. Zhang, C.C., et al., *PF-03732010: a fully human monoclonal antibody against P-cadherin with antitumor and antimetastatic activity.* Clinical Cancer Research, 2010: p. clincanres. 1343.2010.
242. Pinho, S.S. and C.A. Reis, *Glycosylation in cancer: mechanisms and clinical implications.* Nature reviews. Cancer, 2015. **15**(9): p. 540.
243. Sterner, E., N. Flanagan, and J.C. Gildersleeve, *Perspectives on anti-glycan antibodies gleaned from development of a community resource database.* ACS chemical biology, 2016. **11**(7): p. 1773-1783.
244. Hamman, J.H., G.M. Enslin, and A.F. Kotzé, *Oral delivery of peptide drugs.* BioDrugs, 2005. **19**(3): p. 165-177.
245. Rodriguez, M.C., et al., *Targeting cancer-specific glycans by cyclic peptide lectinomimics.* Amino Acids, 2017: p. 1-17.
246. Loffet, A., *Peptides as drugs: is there a market?* Journal of Peptide Science, 2002. **8**(1): p. 1-7.
247. Li, J., et al., *Odorranalectin Is a Small Peptide Lectin with Potential for Drug Delivery and Targeting.* PLOS ONE, 2008. **3**(6): p. e2381.
248. Edwards, C., M. Cohen, and S. Bloom, *Peptides as drugs*. 1999, Oxford University Press.
249. Wu, H., et al., *A novel small Odorranalectin-bearing cubosomes: Preparation, brain delivery and pharmacodynamic study on amyloid-β 25–35-treated rats following intranasal administration.* European Journal of Pharmaceutics and Biopharmaceutics, 2012. **80**(2): p. 368-378.
250. LLC, S., *MacroModel version 9.9.* New York, NY, 2012.
251. Release, S., *1: Maestro*. 2013, Schrodinger.
252. Samsonov, S.A., J. Teyra, and M.T. Pisabarro, *Docking glycosaminoglycans to proteins: analysis of solvent inclusion.* Journal of computer-aided molecular design, 2011. **25**(5): p. 477-489.
253. Li, J., et al., *Odorranalectin is a small peptide lectin with potential for drug delivery and targeting.* PLoS One, 2008. **3**(6): p. e2381.
254. Lichtman, J.W. and J.-A. Conchello, *Fluorescence microscopy.* Nature methods, 2005. **2**(12): p. 910.
255. Fernández-Suárez, M. and A.Y. Ting, *Fluorescent probes for super-resolution imaging in living cells.* Nature reviews. Molecular cell biology, 2008. **9**(12): p. 929.
256. Ueno, T. and T. Nagano, *Fluorescent probes for sensing and imaging.* Nature methods, 2011. **8**(8): p. 642.
257. Zhang, J., et al., *Creating new fluorescent probes for cell biology.* Nature reviews. Molecular cell biology, 2002. **3**(12): p. 906.
258. Guo, Z., et al., *Recent progress in the development of near-infrared fluorescent probes for bioimaging applications.* Chemical Society Reviews, 2014. **43**(1): p. 16-29.
259. Hilderbrand, S.A. and R. Weissleder, *Near-infrared fluorescence: application to in vivo molecular imaging.* Current Opinion in Chemical Biology, 2010. **14**(1): p. 71-79.
260. Chan, J., S.C. Dodani, and C.J. Chang, *Reaction-based small-molecule fluorescent probes for chemoselective bioimaging.* Nature chemistry, 2012. **4**(12): p. 973-984.
261. Yin, J., Y. Hu, and J. Yoon, *Fluorescent probes and bioimaging: alkali metals, alkaline earth metals and pH.* Chemical Society Reviews, 2015. **44**(14): p. 4619-4644.

262. Berezin, Mikhail Y., et al., *Near-Infrared Fluorescence Lifetime pH-Sensitive Probes.* Biophysical Journal, 2011. **100**(8): p. 2063-2072.
263. Casey, J.R., S. Grinstein, and J. Orlowski, *Sensors and regulators of intracellular pH.* Nature reviews. Molecular cell biology, 2010. **11**(1): p. 50.
264. Welch, W.J., *Mammalian stress response: cell physiology, structure/function of stress proteins, and implications for medicine and disease.* Physiological reviews, 1992. **72**(4): p. 1063-1081.
265. Smith, C., et al., *Excess brain protein oxidation and enzyme dysfunction in normal aging and in Alzheimer disease.* Proceedings of the National Academy of Sciences, 1991. **88**(23): p. 10540-10543.
266. Gerweck, L.E. and K. Seetharaman, *Cellular pH gradient in tumor versus normal tissue: potential exploitation for the treatment of cancer.* Cancer research, 1996. **56**(6): p. 1194-1198.
267. Urano, Y., et al., *Selective molecular imaging of viable cancer cells with pH-activatable fluorescence probes.* Nature medicine, 2009. **15**(1): p. 104-109.
268. Švastová, E., et al., *Hypoxia activates the capacity of tumor-associated carbonic anhydrase IX to acidify extracellular pH.* FEBS letters, 2004. **577**(3): p. 439-445.
269. Castro-Obregon, S., *The discovery of lysosomes and autophagy.* Nature Education, 2010. **3**(9): p. 49.
270. Wang, X., et al., *High-Fidelity Hydrophilic Probe for Two-Photon Fluorescence Lysosomal Imaging.* Journal of the American Chemical Society, 2010. **132**(35): p. 12237-12239.
271. Kiyose, K., H. Kojima, and T. Nagano, *Functional Near-Infrared Fluorescent Probes.* Chemistry–An Asian Journal, 2008. **3**(3): p. 506-515.
272. Luo, S., et al., *A review of NIR dyes in cancer targeting and imaging.* Biomaterials, 2011. **32**(29): p. 7127-7138.
273. Zhang, S., et al., *Luminescent Probes for Sensitive Detection of pH Changes in Live Cells through Two Near-Infrared Luminescence Channels.* ACS sensors, 2017. **2**(7): p. 924-931.
274. Zhu, X., et al., *Anti-Stokes shift luminescent materials for bio-applications.* Chemical Society Reviews, 2017. **46**(4): p. 1025-1039.
275. Yang, H., et al., *Upconversion luminescent chemodosimeter based on NIR organic dye for monitoring methylmercury in vivo.* Advanced Functional Materials, 2016. **26**(12): p. 1945-1953.
276. Tian, J., et al., *Intracellular adenosine triphosphate deprivation through lanthanide-doped nanoparticles.* Journal of the American Chemical Society, 2015. **137**(20): p. 6550-6558.
277. Park, Y.I., et al., *Theranostic Probe Based on Lanthanide-Doped Nanoparticles for Simultaneous In Vivo Dual-Modal Imaging and Photodynamic Therapy.* Advanced materials, 2012. **24**(42): p. 5755-5761.
278. Liu, Y., et al., *Near-infrared in vivo bioimaging using a molecular upconversion probe.* Chemical Communications, 2016. **52**(47): p. 7466-7469.
279. Vegesna, G.K., et al., *pH-activatable near-infrared fluorescent probes for detection of lysosomal pH inside living cells.* Journal of Materials Chemistry B, 2014. **2**(28): p. 4500-4508.
280. Miertuš, S., E. Scrocco, and J. Tomasi, *Electrostatic interaction of a solute with a continuum. A direct utilizaion of AB initio molecular potentials for the prevision of solvent effects.* Chemical Physics, 1981. **55**(1): p. 117-129.
281. Miertuš, S. and J. Tomasi, *Approximate evaluations of the electrostatic free energy and internal energy changes in solution processes.* Chemical Physics, 1982. **65**(2): p. 239-245.

9 SUPPORTING INFORMATION AND APPENDICES