Structural Glycobiology by NMR and Molecular Modelling: Ligand recognition by the carbohydrate active proteins LYVE-1, SseK1/2 and *Ps*LBP

Samuel Walpole

Thesis submitted in fulfilment of the requirement for the degree of

Doctor of Philosophy

University of East Anglia

School of Pharmacy

May 2019

Ad maiorem Dei gloriam

Abstract

The term carbohydrate-active protein (CAP) encompasses the group of proteins that either act as carbohydrate receptors (e.g. lectins) or enzymatically catalyse reactions involving carbohydrates as at least one of their substrates (e.g. glycosyltransferases). The fundamental importance of CAPs has only been realised over the past few decades, with carbohydrates playing a profound role in cellular signalling, adhesion and migration. Furthermore, carbohydrates are routinely used by pathogens in immune evasion and to modify host function, and changes in glycosylation patterns are associated with a range of disease states, including cancer and inflammation.

However, the study of CAP-carbohydrate interactions has been challenging, in part due to the inherently low affinity of many CAP-carbohydrate interactions that precluded detection by many techniques and make it difficult to obtained experimentally derived structures of their complexes. Fortunately, STD NMR spectroscopy is ideally suited to detecting weak interactions of this nature and provides structural information about the interaction through ligand epitope mapping. Furthermore, quantitative analysis of STD intensities allows three-dimensional models of the validated in the solution state, whether these models be derived from experiment or molecular modelling.

Within this thesis, a combination of STD NMR spectroscopy and molecular modelling has been used to unravel structural and dynamic detail of CAP-ligand interactions in three biologically or industrially relevant systems - (1) CD44/LYVE-1 which may play a role in cell trafficking across the lymphatics in cancer; (2) *Ps*LBP which may lead to new developments in the field of enzymatic carbohydrate synthesis; (3) SseK1/2 which exhibit a novel enzymatic mechanism involving glycosylation of arginine residues and may be involved in *Salmonella* virulence.

Table of Contents

Acknowledgements

I am incredibly grateful for the privilege I have had to spend these past four years studying for this PhD. It has been a great period of learning and discovery for me, not only academically but also in myself. Although I have ultimately decided that my future career lies elsewhere, I can say with certainty that I am a much different person now than I was when I started this degree (hopefully for the better!) and the experiences I have had over this time have been invaluable. Above all I can see how I have grown in confidence and self-assurance, and I am forever thankful to those who have been with me - to guide me and to give me that push when I've needed it.

I would like to start by thanking my supervisor, Dr. Jesus Angulo for his enthusiastic support and tutelage over the past four years - firstly for giving me this opportunity in the first place by selecting me as a PhD candidate, but also for all the time that you have invested in teaching and mentoring me during my studentship. I am thankful that you have been so supportive and understanding, particularly in allowing me to pursue activities outside of my direct research and in my decision to change the direction of my career. You have worked so hard to create a friendly research group that produces excellent science, making you thoroughly deserving of the nomination for PhD Supervisor of the Year.

I am also very grateful for the team behind the Norwich Research Park Doctoral Training Partnership, not only for supporting me financially, but also for the excellent training program that they have worked tirelessly to deliver to myself and my cohort. They have been particularly attentive to actioning our feedback to continuously improve our training to best meet our needs and the needs of future students to come.

Next, I would like to take time to thank everyone has lived with me in the NMR 'dungeon' - down in the deep recesses of the CAP basement. Thanks for being good colleagues and friends: for the fun we have had, for all the discussions - both academic and otherwise, and for putting up with me when my outlook hasn't been so great. I apologise if I

don't mention everyone personally - I am truly grateful for the time I have spent with all of you - but in particular I want to mention Serena for all of the science that we have done together and for teaching me a lot about NMR, and I want to mention Alex for all the conversations we've had, for keeping me sane and, of course, for all the memes.

Thank you to my parents for giving me a great upbringing in a supportive household, for always encouraging me to pursue my interests, and for believing in my success - even when I haven't. Thank you for all of the time that you have invested in me to shape who I am now.

Finally, I want to make a special mention to Danica, who arrived at probably the most difficult time in my PhD, who has seen me at my best and at my worst, and without whom, I can honestly say, I doubt I would have completed this thesis. Thank you for being a true partner. You have been a constant source of joy and fun that has kept me going during this time, but also you have offered firm encouragement when I've needed it. You always have my best interests at heart and in this relatively short time you have been responsible for so much of my development into a better person. I look forward to all that is in store for both of us in the future as we move forward together.

Chapter 1

Introduction

1.1 Carbohydrates

1.1.1 Origins of Glycobiology

In the field of molecular biology, it has long been believed that the flow of information from DNA to RNA to protein sufficiently explains the complex and diverse behaviour of life. In fact, this idea has become so prolific that it is now known as the central dogma.[1] However, it is now apparent that this is not the case and an accurate description requires a better understanding of other key players.

Carbohydrates are another major class of biomolecule and their role in metabolism and structural integrity has been understood for a long time[2]. However, further progress in the study of carbohydrates has been hampered, principally due to their inherent complexity. Unlike proteins, which are linear chains of amino acids, each carbohydrate monomer, known as a monosaccharide, can polymerise at multiple positions, leading to highly branched structures. A further distinction from proteins is that carbohydrates lack any template encoding. The sequence of a protein can be accurately predicted if the DNA sequence is known. Conversely, the fine balance of glycosyltransferases and glycosidases of the endoplasmic reticulum and Golgi apparatus determine the final carbohydrate structure, which is therefore dependant on physiological conditions and on cell type.[3,4]

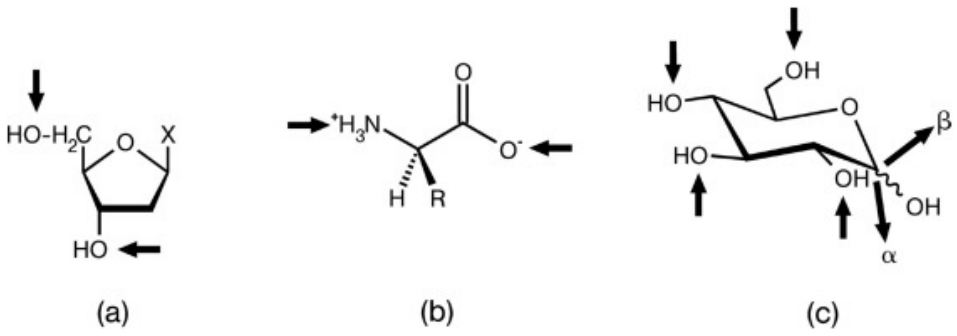

Figure 1.1: The possible linkage points for DNA (a), proteins (b) and carbohydrates (c), shown with arrows. Both DNA and proteins are template encoded and only possess two linkage points, leading to linear chains. Carbohydrates have many linkage points, so can be highly branched and diverse. Taken from [5]

It has now become evident that carbohydrates play an integral role in biological systems and as such the field of glycobiology, a term first coined by Raymond Dwek in 1988,[6] is now rapidly expanding. As an example of their importance, carbohydrates are often found covalently bound to extracellular proteins (glycoproteins), in which they have important roles in modulating stability and activity.[6,7] In the first case, the carbohydrate acts as a physical barrier to prevent recognition by proteases and antibodies. In the latter, carbohydrate residues can occupy the binding site or interact with key sidechains. Commonly, this is achieved via terminal sialic acid residues, which can be cleaved by sialidases to allow switching between low- and high-activity states.[8]

Carbohydrates also play a key role in cellular signalling.[7] Specific recognition of carbohydrates in the extracellular matrix (ECM) allows the cell to sense its environment and facilitates adhesion and migration. Carbohydrates are also key in inter-cell communication, by binding to specific receptors on adjacent cells that in turn facilitate a downstream signalling response. This also plays a key role in infection, in which many pathogens aim to mimic the carbohydrates of their hosts to both evade the immune response, and to facilitate host entry by binding to host cell receptors.[9]

Clearly then the precise structure of a particular carbohydrate has a profound impact on its function - addition or removal of single monosaccharide unit can determine whether or not a carbohydrate is recognised by its receptor, which may be responsible for a number

of downstream processes. Despite the lack of templating, it is still incredibly useful to understand the structural features of certain carbohydrates that give rise to their particular functionalities. These features have been termed the sugar code and much work is currently underway to elucidate these features.[5,10,11] Receptors that recognise these specific carbohydrates are termed 'readers' of sugar-encoded information, whilst other carbohydrate-active proteins that add and remove sugar units are termed 'writers' and 'erasers' of sugar-encoded information respectively. A good example of the specificity of the glycan code is the comparison between Blood Groups A and B - the former presenting a N-acetyl-galactosamine as the terminal residue, whereas the terminal residue is simply a galactose residue in the latter. This seemingly small change (removing the N-acetyl group) can be highly immunogenic where the wrong blood type is given to the host, showing high specificity in carbohydrate recognition.

Due to their intimate roles in many biological processes, carbohydrates are also implicated in a range of genetic and acquired diseases. The most common carbohydrate-related genetic defect is congenital disorder of glycosylation (CDG) Ia, which is characterised by a deficiency in phosphomannomutase, a key enzyme in the pathway that leads to incorporation of mannose into carbohydrates.[12,13] CDG Ia has high mortality rates due to susceptibility to infection and organ failure, and leads to reduced muscle tone and mental retardation. Many acquired diseases are characterised by a change in either glycosylation patterns or receptor expression. For example, inflammatory diseases are caused by changes in glycosylation in certain tissues, leading to leukocyte homing and immune activation.[14] Furthermore, many cancers exhibit increased levels of sialylation, which allows them to ignore extracellular signals, such as apoptosis death factors, and resist anti-cancer therapies.[14,15] Increased sialylation is also associated with increased aggressiveness and metastatic potential.

Much of the understanding of carbohydrate function we now have has been made possible with the advent of new technologies. For example, recombinant DNA technology and chemical mutagenesis allow protein glycosylation sites and carbohydrate biosynthesis to be modified, facilitating *in vivo* understanding of carbohydrate function.[16,17] In addition, isolated carbohydrates can be characterised by techniques such as antibody-lectin arrays,[18] mass spectrometry[19] and NMR

spectroscopy.[20] Techniques such as these allow the glycome, the total set of expressed carbohydrates under given conditions, to be characterised, facilitating a dynamic understanding of carbohydrate expression.[21] In addition, X-ray diffraction, NMR spectroscopy and molecular modelling can be used to create atomic resolution 3D models of carbohydrates and their complexes, allowing the details of the interaction to be understood.[22–26] In addition, the latter two techniques also present dynamic information, which is of particular importance in understanding the behaviour of carbohydrates.

1.1.2 Chemical Structure

In their simplest form, monosaccharides exist as polyhydroxyaldehydes (aldoses) or polyhydroxyketones (ketoses) with an empirical formula of $C_m(H_2O)_n$. Each hydroxymethylene unit is chiral, allowing for a huge amount of diversity with relatively few atoms; for a given molecule with n stereocentres, there are 2n stereoisomers. Nature tends only to use a subset of these however, the majority of natural monosaccharides existing in the D-configuration. In addition, only 9 monosaccharides are commonly observed in vertebrates (Fig. 1.2).[27] Despite this, a huge number of carbohydrate structures are possible. For example, considering only disaccharides of glucose, 20 unique carbohydrates can be made.

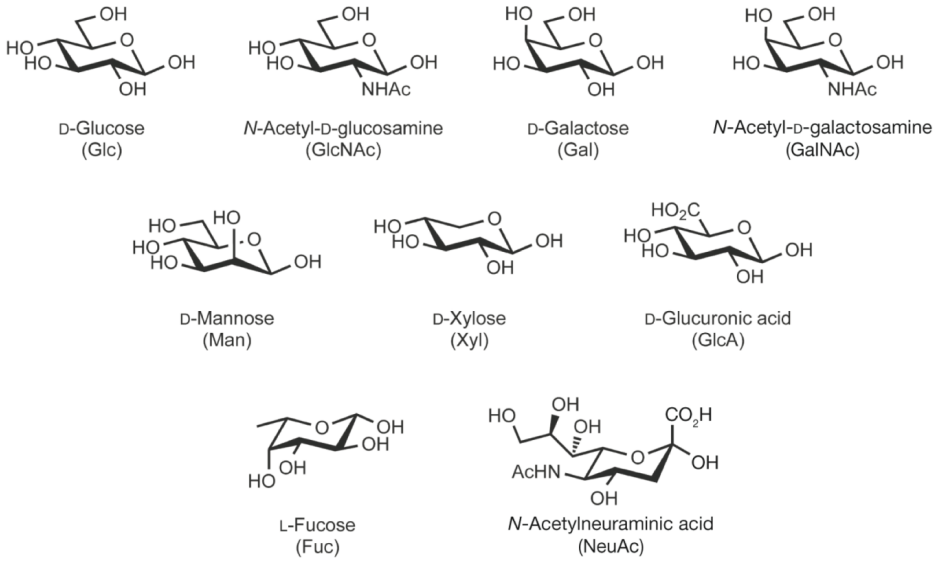

Figure 1.2: The nine monosaccharides commonly found in vertebrates, depicted here as β-anomers. All exist as pyranoses, while only Fuc is in the L-configuration, and NeuAc, the most common form of sialic acid, is the only ketose. Taken from [27]

In solution, monosaccharides predominantly exist in either a cyclic furanose (5-membered ring) or pyranose (six-membered ring) form generated through reaction of the carbonyl with one of the hydroxyl groups to form a hemiacetal. On cyclisation, a new stereocentre is generated at the so-called anomeric carbon (C1 in aldoses), giving rise to two new stereoisomers. These are known as the α- and β-anomers and are defined by the absolute configuration of the anomeric carbon compared to that of the highest-numbered stereocentre.[27] If both carbons have the same configuration then it is said to be the α-anomer, otherwise it is the β-anomer. In solution, these anomers can readily interconvert through a ring-opening and recyclisation process called mutarotation. From sterics alone, it would be expected that the α-anomer would be strongly disfavoured due to diaxial interactions. However, the α-anomer is significantly populated in many monosaccharides, and is even the major anomer in mannose.[28] This is attributed to favourable overlap between a non-bonding orbital of the endocyclic oxygen and the σ^* orbital of the carbon to exocyclic oxygen bond, known as the endo-anomeric effect (Fig. 1.3).[29]

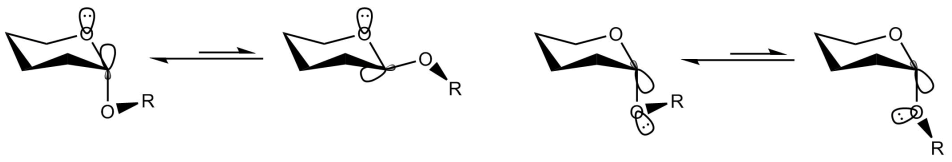

Figure 1.3: Graphical representation of the endo-anomeric (left) and exo-anomeric (right) effects. Both the axial configuration and gauche conformation are stabilised by hyperconjugation between heteroatom non-bonding and carbon-heteroatom anti-bonding orbitals.

As for cyclohexanes, the most favourable ring-conformation of pyranoses is usually the chair (C) conformation. This can further be defined as $^4C_1$ or $^1C_4$, which refers to the carbon atoms that fall above or below the plane of the ring when drawn as shown (Fig. 1.4). In most cases, the $^4C_1$ conformation is favoured as axial interactions are minimised. Other possible conformations include half-chair (H), boat (B), skew-boat (S) and envelope (E) (Fig. 1.4) but these are only accessible in more exotic monosaccharides[30] or by distortion in an enzyme active site.[31,32] Ring conformations can readily be quantified by using the spherical coordinate system developed by Cremer and Pople to describe ring puckering (Fig. 1.4).[33] These coordinates can also be applied to furanoses, which are inherently more flexible than pyranoses and can readily sample multiple ring conformations.[34,35]

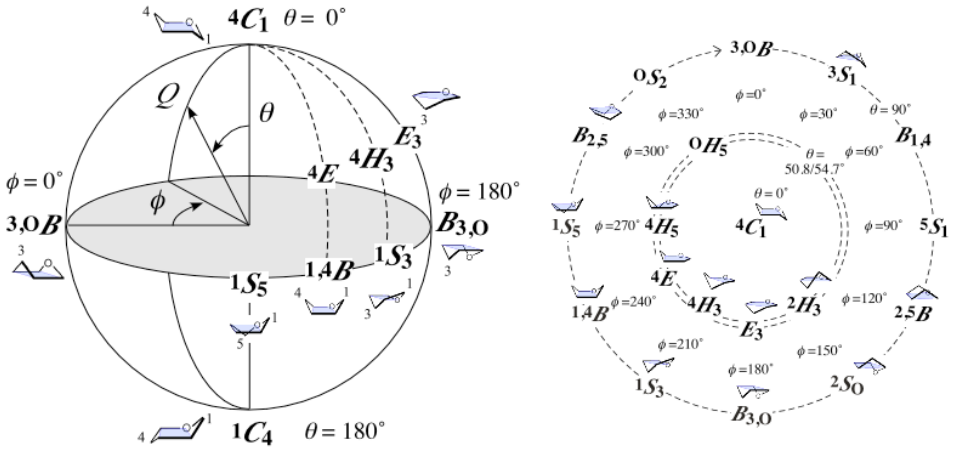

Figure 1.4: Stereographic representation of the Cremer-Pople ring-puckering coordinates proposed by Jeffery and Yates (1979). The polar (θ) and azimuthal (ϕ) coordinates quantitatively describe each ring conformation. For example, $\theta = 0°, 180°$ (ϕ is undefined) describe the $^4C_1$ and $^1C_4$ conformations respectively.

Polymerisation of monosaccharides occurs by condensation of the C1 hemiacetal of one monosaccharide with a hydroxyl group of another monosaccharide, forming a glycosidic bond. The, now, acetal is no

longer susceptible to reduction whereas the remaining hemiacetal in the added monosaccharide is, giving the polymer a distinct polarity. These are therefore termed the non-reducing and reducing ends respectively and are equivalent to the N- and C-terminal terminology in proteins. In addition, mutarotation is no longer possible at the acetal, so the anomericity of each monosaccharide within the polymer is preserved. In some carbohydrates, such as sucrose, the glycosidic bond is formed between the anomeric positions of both monosaccharides. These are termed non-reducing carbohydrates as no hemiacetal is present at either terminus. The nomenclature of glycosidic linkages is described in terms of the numbering of the hydroxyl atoms involved and the anomeric form of the non-reducing residue. For example, the disaccharide formed between β-glucose (Glc) and O4 of galactose (Gal) would be named Glc-β-1,4-Gal. Polymers of monosaccharides are known collectively as glycans but may be termed oligosaccharides for short polymers (< 20 residues) or polysaccharides for longer chains. As each monosaccharide contains multiple hydroxyl groups, condensation can occur multiple times on the same residue forming branched structures, which confers glycans an immense capacity for bearing structurally diverse information.

While the pyranose rings of each monosaccharide can be considered essentially rigid, flexibility can be achieved by rotation about the glycosidic bonds. The angles, ϕ and ψ, describe the conformation of the glycosidic linkages, and are defined here as $O5_n$-$C1_n$-Ox_{n+1}-Cx_{n+1} and $C1_n$-Ox_{n+1}-$C(x)_{n+1}$-$C(x\text{-}1)_{n+1}$ respectively (Fig. 1.5).[2] A gauche conformation is usually favoured for ϕ due to the exo-anomeric effect which, similar to the endo-anomeric effect, involves favourable overlap of a non-bonding orbital of the exocyclic oxygen with the σ^* orbital of the carbon to endocyclic oxygen bond (Fig. 1.3).[29] The key relevance of the stereoelectronic component of the exo-anomeric effect has been experimentally demonstrated very recently by the elegant use of fluorinated glycomimetics.[36] Glycosidic bonds involving the oxygen of the exocyclic hydroxymethyl group at C5 have an extra degree of freedom by rotation around the C5-C6 bond. This allows a further torsion angle, ω, to be defined as O6-C6-C5-O5 (Fig. 1.5). By considering ω and the torsion angle defined by O6-C6-C5-C4, three rotamers can be defined. With respect to the above torsions, these are gauche-gauche (gg), gauche-trans (gt) and trans-gauche (tg) (Fig. 1.6). In many cases, the gg and gt conformations are largely populated, while the

tg conformation is virtually non-existent.[37,38] For simple molecules, the preference for two adjacent electronegative groups to be gauche to one another, known as the gauche effect, is often rationalised as a stereoelectronic effect.[29] While this may be a contributing factor in carbohydrates, studies have shown that solvation and sterics play a much more important role.[39]

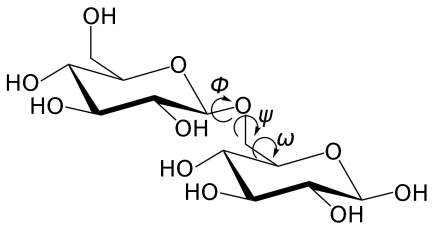

Figure 1.5: The torsions important for carbohydrate flexibility using Glc-β-1,6-Glc as an example. The angles ϕ and ψ are common to all glycosidic linkages, whereas ω is unique to the 1,6 linkage.

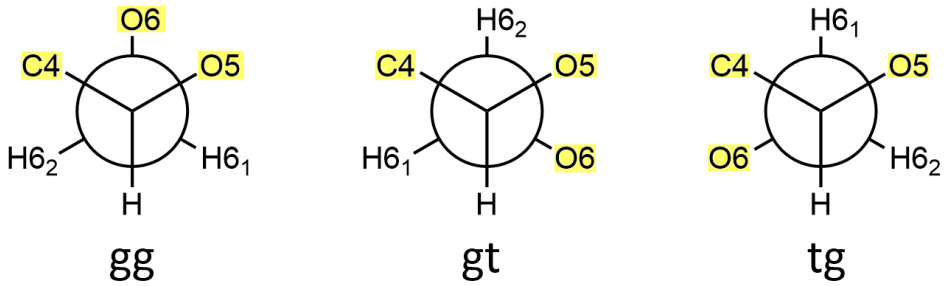

Figure 1.6: Newman projections viewed along the C5-C6 bond showing the gauche-gauche (gg), gauche-trans (gt) and trans-gauche (tg) rotamers. These rotamers describe the conformation of O6 relative to C4 and O5 respectively (highlighted in yellow).

Carbohydrate structure is further complicated by chemical modification following synthesis, which is known as post-glycosylational modification.[40] Common modifications include sulfation, methylation and acetylation, all of which provide unique structural motifs that add a further level of diversity to the carbohydrate 'toolbox'.

1.1.3 Carbohydrates in Biology

The majority of carbohydrates are found covalently attached to other biomolecules (glycoconjugates), either as glycolipids or glycoproteins. Monosaccharides have also been reported bound to DNA, but this appears only to be associated with damage and ageing.[41–43]

Glycolipids are found both in the outer and internal (ER, Golgi etc.) membranes of the cell, where they function as biosurfactants and have roles in cell-cell binding and signalling (Fig. 1.7).[44] For example, *Vibrio cholera* initiate host cell invasion through binding to the GM_1 ganglioside.[45] In signalling, glycolipids come together, along with cholesterol, in concentrated regions of the membrane known as lipid rafts.[46,47] This facilitates signalling by bringing specific signalling proteins into close association. Gram-negative bacteria produce unique glycolipids known as lipopolysaccharides (LPS). These are complex structures consisting of an anchoring lipid (endotoxin) covalently bound to a core oligosaccharide.[48] This is then attached to a repeating polysaccharide, known as the O-antigen, which extends into the extracellular space. LPS molecules cover most of the surface of the outer membrane and act as a physical barrier to protect the bacterium from antibiotics, the immune system and environmental stress. The endotoxin can be released from the membrane as a result of cell division or death, and is highly immunogenic.[49] While such a response helps the host fight the invading organism, Gram-negative infections can result in septicaemia due to prolonged inflammation.[50]

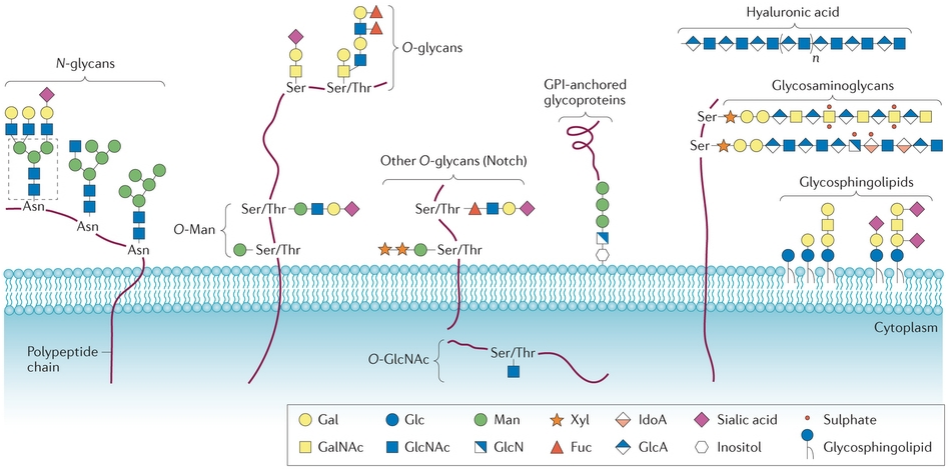

Figure 1.7: Schematic representation of common classes of glycans depicted as glycoconjugates to protein or lipid molecules. The exception is hyaluronan, which is never conjugated to other biomolecules. Taken from [51].

Glycans are covalently attached to protein sidechains via either a nitrogen (N-glycan) or oxygen (O-glycan) atom. N-glycosidic bonds are formed between the reducing terminal, almost always β-GlcNAc, of the glycan and an asparagine sidechain (Fig. 1.7)[3,52]. Sterics permitting, this occurs at a consensus sequence of N-X-S/T (where X is any amino acid except proline), which is known as the N-glycan sequon. While all N-glycans start out as $Glc_3Man_9GlcNAc_2Asn$, var-

ious trimming and further glycosylation steps allow N-glycans to be separated into 3 distinct classes, all with a common biantennary core of $Man_3GlcNAc_2$ (Fig. 1.8). These are: oligomannose, in which the core is attached only to mannose residues, complex, in which oligosaccharides are linked to the core via GlcNAc, and hybrid, in which one antenna is high mannose and the other is complex. As well as being essential for mediating glycoprotein activity and stability, N-glycans are also crucial in the steps before the mature glycoprotein is secreted, in roles such as quality control[53] and trafficking within the cell.[54]

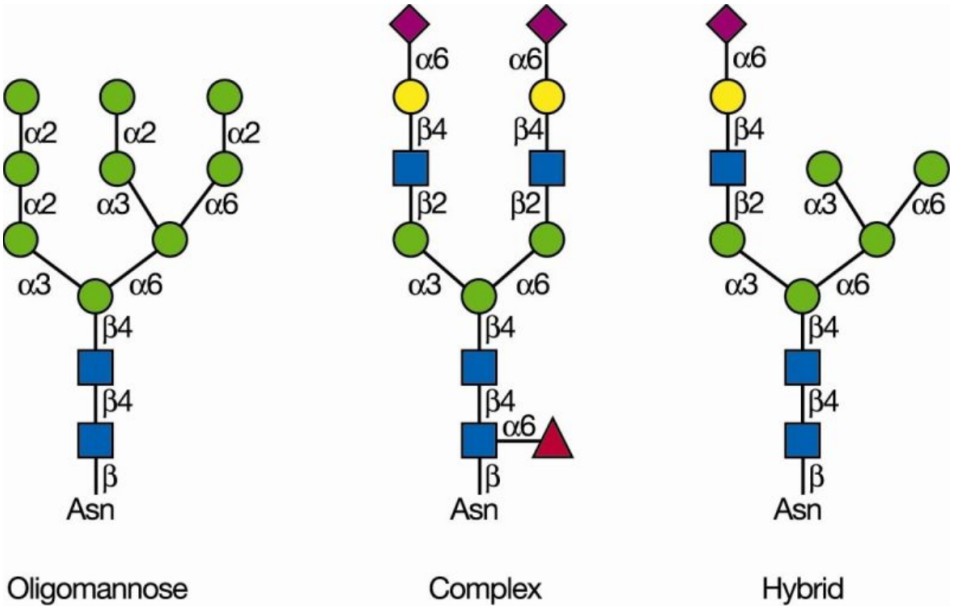

Figure 1.8: Classification of N-glycans as oligomannose, complex, or hybrid based on oligosaccharides attached to a common core glycan. Monosaccharides are represented according to the same scheme as Fig. 1.7. Taken from [3].

O-glycosylation differs somewhat from N-glycosylation. Firstly, O-glycosylation occurs by stepwise addition of monosaccharides to the existing glycan chain, which is initiated via an O-glycosidic bond to a serine or threonine residue (Fig. 1.7).[55–57] An O-glycosylation sequon is yet to be identified. A number of different types of monosaccharide can initiate the O-glycan. Glycans initiated by O-GalNAc are common and found on mucin glycoproteins, where they play mostly a protective role in the extracellular environment.[58] Inside the cell a unique form of O-glycosylation occurs involving GlcNAc, in which glycosylation and deglycosylation occurs readily, providing a regulatory switching mechanism.[59]

Proteoglycans are a specific class of glycoprotein which consist of a protein core bound to at least one, usually many, glycosaminogly-

cans (GAGs), linear polysaccharides consisting of repeating disaccharide units (Fig. 1.7).[60,61] They are usually found in the extracellular matrix, where they are known to contribute to tissue architecture. More recently however, they have also been shown to play an important role in signalling by recruiting leukocytes[62] and interacting with cytokines.[63]

1.1.3.1 Carbohydrate Binding Proteins

Many biological roles of carbohydrates are mediated through binding to carbohydrate-specific proteins. Despite the immensity of the glycome, most carbohydrate-binding proteins (CBPs) are highly specific, often recognising only a single glycan.[64] This interaction tends to be very weak however, with dissociation constants typically in the millimolar range. Nevertheless, these interactions usually display high avidity through receptor clustering and multivalent ligands. Specificity is achieved by providing a contact surface complementary to the unique geometry of the particular carbohydrate.[65] The amphipathicity of carbohydrates means that both hydrogen bonding and hydrophobic interactions are important, as well as electrostatics.[66] Two major classes of CBP exist, lectins and GAG-binding proteins (GBPs).

Lectins are CBPs involved in cell-cell adhesion, known as agglutination.[67] They typically bind with only the terminal carbohydrate residue contacting the protein. Lectins are categorised based on highly conserved carbohydrate recognition domains (CRDs) (Fig. 1.9). The most common of these are the Ca^{2+}-dependant C-type lectins, which are cell-surface receptors that have a dual role in both activation and inhibition of the immune system.[68] The C-type lectin fold is characterised by two antiparallel β-sheets, two α-helices and two loops - as an example, see the structure of DC-SIGN in Fig. 1.10A. Stabilisation of these loops by Ca^{2+} binding facilitates glycan binding. In addition, two disulphide bonds are highly conserved, as well as the glycan-binding E-P-N and W-N-D motifs.

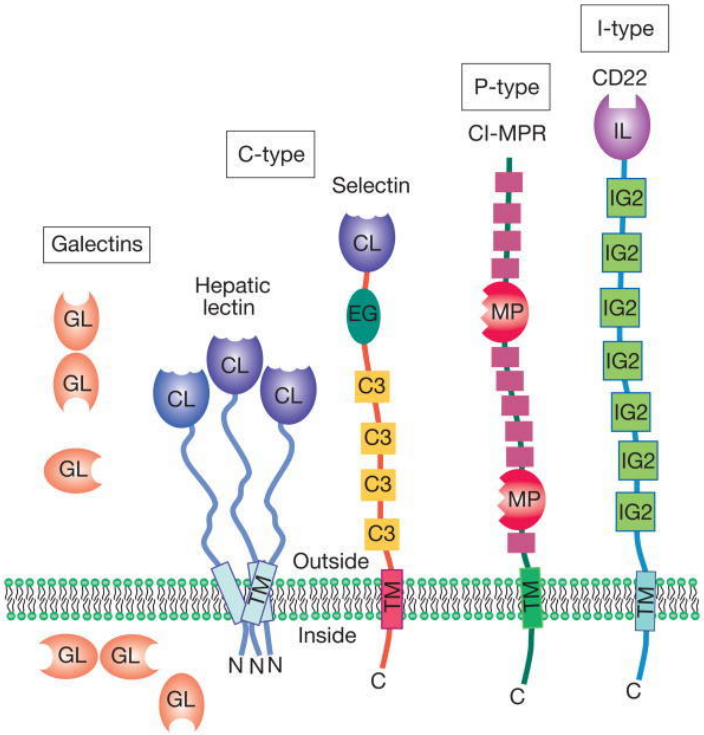

Figure 1.9: Schematic representation of examples of common lectin classes. Each class is characterised by conserved CRDs (GL, CL, MP, IL). Other domains shown are EGF-like (EG), immunoglobulin C2 (IG2), transmembrane (TM), and complement regulatory repeat (C3). Taken from [69]

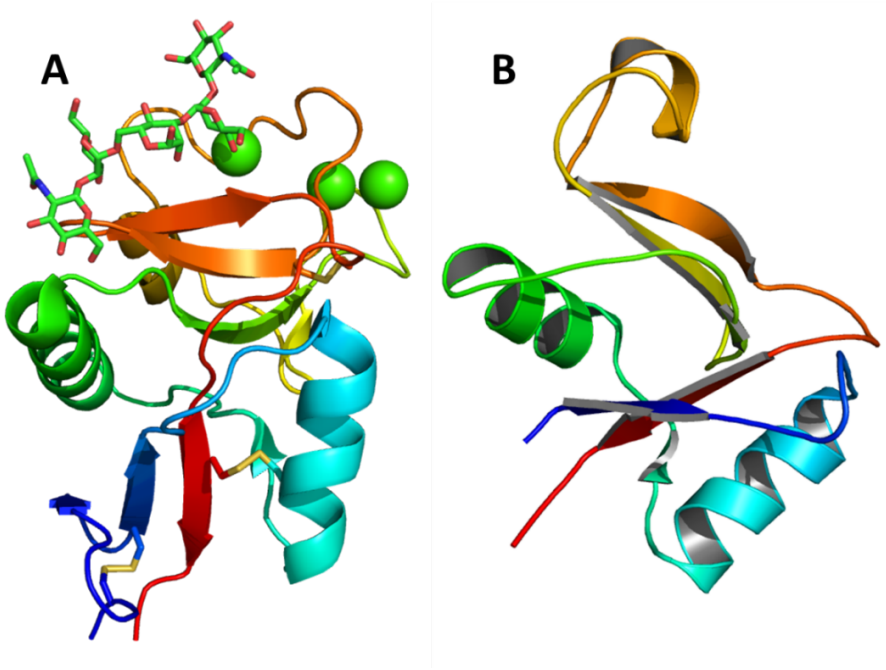

Figure 1.10: (A) Cartoon representation of DC-SIGN, a C-type lectin, bound to $GlcNAc_2Man_3$ used to demonstrate the C-type lectin fold. Ca^{2+} ions are represented as green spheres. PDB accession code: 1K9I [70] (B) The Link domain of TSG-6, showing structural homology to the C-type fold minus the Ca^{2+}-binding loop. No ligand-bound structure was available at time of writing, but NMR and modelling studies suggest binding along a shallow groove formed between the hook-like loop and β-sheet (orange) [71]. PDB accession code 2PF5 [72].

Another interesting class of lectin are the galectins, which are unique in that they are soluble, and are found both inside the cell and in the extracellular matrix. They are specific for galactosides (galactose-containing glycans) and are characterised by an essential conserved H-N-R motif within the CRD that directly interacts with bound galatose.[73] An important function of galectins, which contain multiple CRDs, is intra-cell crosslinking of transmembrane glycoproteins. Clustering of such proteins facilitates downstream signalling, and galectins are again implicated mostly in inflammation and the immune response.[74]

Unlike lectins, GBPs contain a binding groove that contacts several GAG residues, allowing the repeating disaccharide pattern to be recognised.[75] The majority of reported GBPs bind to heparan sulphate (HS), which consists of variably sulphated disaccharides unit of β/α-1,4-GlcNAc-α-1,4-GlcA/IdoA. Understanding HS specificity is complex as different proteins have been shown to bind preferentially to different HS oligosaccharides.[76] However, it is clear that

complementarity between the sulphate and basic amino acid residues is essential. Specificity is also achieved through IdoA, which is able to access the $^2S_0$ conformation.[77–79] This reorients the sulphate group, creating a unique contact surface. This explains why dermatan sulphate, which also contains IdoA, can bind to many HS-binding proteins.[75] GBPs are implicated in a wide range of functions, including cell adhesion, migration and inflammation. A particularly well-studied system is the ternary complex formed by HS, thrombin and anti-thrombin. Here HS acts as an anti-coagulant by increasing anti-thrombin activity via both allostery and proximity.[80]

Specific binding of proteins to hyaluronan (HA), which is neither conjugated nor sulphated, has also been extensively studied. The HA-binding Link domain is common to most of these proteins and is structurally similar to the C-type lectin domain (Fig. 1.10 B).[81] It is stabilised by two conserved disulphide bonds, and interacts with HA mostly through specific recognition of the carboxylate and N-acetyl groups. The Link module is found in many proteoglycans, which facilitates the formation of supramolecular aggregates that are essential for tissue integrity.[82] Several cell-surface receptors containing the Link domain have also been identified. Of these, CD44 is the primary HA receptor in many cell types.[83] This allows HA to act as a signalling molecule, and is implicated in cell adhesion and migration, as well as survival and differentiation.

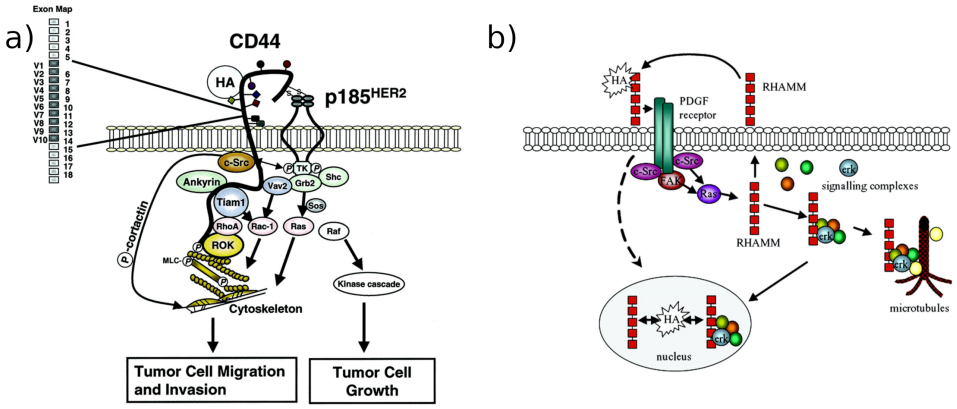

Figure 1.11: Examples of hyaluronan acting as a signalling molecule by interacting with the cell-surface receptors CD44 (a) and RHAMM (b), both of which induce downstream intracellular signalling. Taken from [84]

1.1.3.2 Glycan Processing Enzymes

Glycan processing enzymes can be broadly separated into two categories: glycosyltransferases, which catalyse the formation of new glycosidic bonds between a monosaccharide donor and a mono-, oligo- or polymeric glycan chain, and glycosidases, which catalyse the removal of monosaccharides from the glycan chain, either for trimming as part of the glycan synthesis process, or for degradation.[85]

1.1.3.2.1 Glycosyltransferases Glycosyltransferases (GTs) are known to be extremely specific for both their donor and acceptor substrates, to the extent that, in most cases only the exact donor and acceptor will be tolerated. For example, human blood group B α-1,3 galactosyltransferase transfers a galactose residue to a galactose-α-1,2-fucose disaccharide, but the reaction will not occur, for example, for a galactose-α-2,6-sialic acid moiety.[85] Interestingly, this same enzyme is capable of binding both UDP-galactose and UDP-glucose in the donor substrate binding site, although only UDP-galactose is active, highlighting that binding isn't sufficient for enzymatic activity.[86] The high specificity of GTs lead to the belief that each GT was uniquely responsible for a single reaction in a particular biosynthetic pathway. However, this is now known to not be absolutely true, with a rare few GTs tolerating multiple substrates, and in some cases there is redundancy in that multiple GTs can form the same type of glycosidic linkage. For example, EXTL2 can use either N-acteylglucosamine or N-acetylgalactosamine as a donor,[87] whereas there is redundancy in fucosyltransferases (FUTs), with α-1,3-linkages with fucose being catalysed by FUTs 3-7, 9 and 11.[88]

Regardless of these few exceptions, the rule holds true for the most part, as exemplified by more than 500,000 known GT sequences, spread over 106 sequence-distinct families at time of writing.[89] Despite the large sequence variability between GT families, of all the GTs whose 3-dimensional structures have been determined experimentally, almost all fall into one of two structurally distinct folds, GT-A or GT-B (Fig. 1.12).[90]

The GT-A type fold consists of two β-α-β motifs, each described as Rossman-like, that are closely associated to form a single continuous

β-sheet core. The GT-A fold is usually described as a single protein domain, although some argue that this is not strictly true due to each Rossman-like motif being separately responsible for either acceptor or donor substrate binding. The majority of GT-A type GTs are metal-ion dependant and coordinate a divalent metal cation through an Asp-X-Asp (termed DXD) motif.

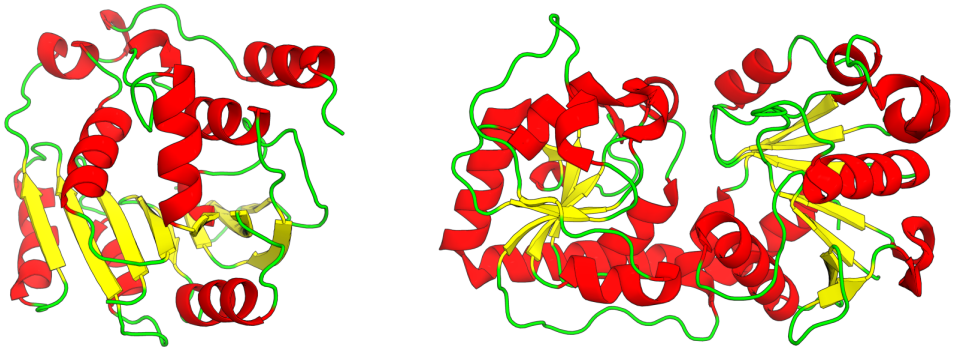

Figure 1.12: Cartoon representations of the the GT-A and GT-B folds. Left: The crystal structure of *Bacillus subtilis* SpsA (PDB 1QGQ), a GT-A type enzyme. Right: The crystal structure of T4 bacteriophage β-glucosyltransferase (PDB 1JG7). In both cases, α-helices (red), β-strands (yellow) and loops (green) are highlighted.

The structure of GT-B type GTs also consists of two Rossman-like folds, although here they are usually described as two separate domains since they are normally separated by a longer, more flexible linker. Each fold is again responsible for either acceptor or donor substrate binding and the active site of GT-B GTs falls within the cleft between the two Rossman-like domains.

GTs accept a nucleotide-sugar as a donor substrate and form a new glycosidic bond between the donor sugar and the acceptor glycan through cleavage of the nucleotide diphosphate, which is stabilised either by the divalent metal cation in metal-dependant GTs or a positively charged side chain. The mechanism is considered to be either inverting, in which the anomeric configuration of the donor sugar is inverted as a result of the reaction, or retaining, in which the anomeric configuration of the donor sugar is persevered from reactant to product.[85,90]

The inverting mechanism follows an S_N2 reaction in which the acceptor initiates a backside attack of anomeric carbon of the sugar-nucleotide donor (Fig. 1.14). A new glycosidic bond is formed between the attacking hydroxyl of the acceptor and the donor anomeric carbon on the pyranose face opposite the original glycosidic bond, and the

nucleotide diphosphate is ejected, resulting in inversion of stereochemistry around the anomeric centre. A nearby aspartate or glutamate residue usually acts as a base to deprotonate the attacking hydroxyl during the reaction.

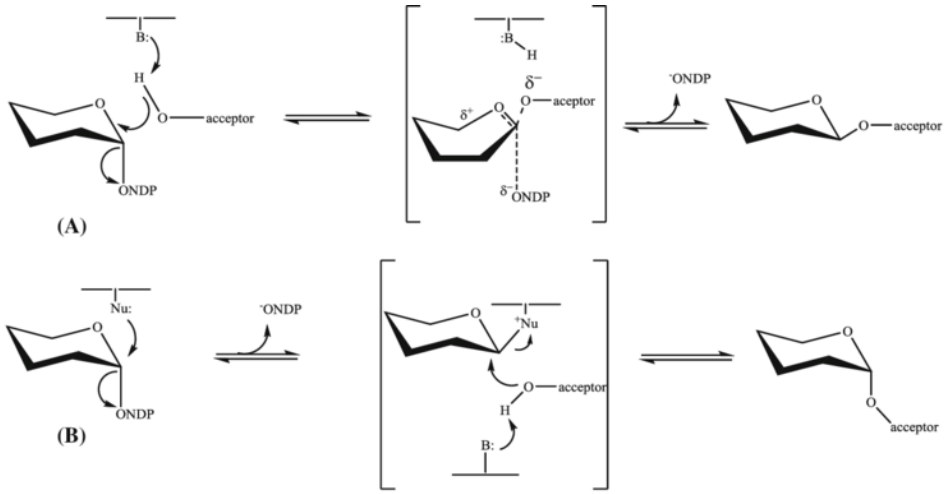

Figure 1.13: General mechanisms for inverting and retaining glycosyltransferases. Top: The inverting mechanism showing the backside attack of the acceptor substrate leading to inversion of anomeric configuration. Bottom: The retaining mechanism showing the formation of a donor-protein intermediate before attack by the acceptor substrate, leading to overall retention of anomeric configuration. Used with permission of Dr.Brock.Schuman under the Creative Commons license.

The retaining mechanism is not well understood, but is believed to take place *via* a so-called double displacement mechanism. In such mechanisms, the sugar of the donor substrate is first covalently linked to the GT *via* nucleophilic attack by an aspartate or glutamate residue (Fig. 1.14). This results in a protein-sugar intermediate in which the stereochemistry is inverted. The acceptor substrate can then attack the anomeric carbon of this intermediate, inverting the configuration of the anomeric centre again, resulting in overall retention of anomeric stereochemical configuration.

1.1.3.2.2 Glycosidases Glycosidases are responsible for the cleavage of glycosidic bonds, either by forming the free sugar product (glycoside hydrolases), a sugar-1-phosphate (glycoside phosphorylases) or by β-elimination of a uronic acid residue (glycoside lyases). In nature, glycosidases are used for a range of functions, including degradation/-turnover of glycans and glycoconjugates, metabolism of sugars and for processing of glycans during their biosynthesis. For example, it

is common for the $Glc_3Man_9GlcNAc_2$ glycan, which is conjugated to an acceptor protein as a single unit, to be subsequently trimmed by glycosidases in order to produce a wider range of complex glycans.[85]

Far more glycosidase structures have been solved compared to GTs, due predominantly to their greater stability.[91] These structures have revealed that there is far greater diversity in glycosidase structure compared to GTs, with many different glycosidase folds being observed. However, the majority of glycosidases follow a Koshland type enzymatic mechanism, which is comparable to the reverse mechanism described above for GTs. This mechanism can be either retaining or inverting (Fig. 1.14), with the retaining mechanism forming a glycosyl-protein intermediate, which is why GTs are expected to follow this same double-displacement mechanism. For glycoside hydrolases the attacking nucleophile is water, while inorganic phosphate is used in glycoside phosphorylases.

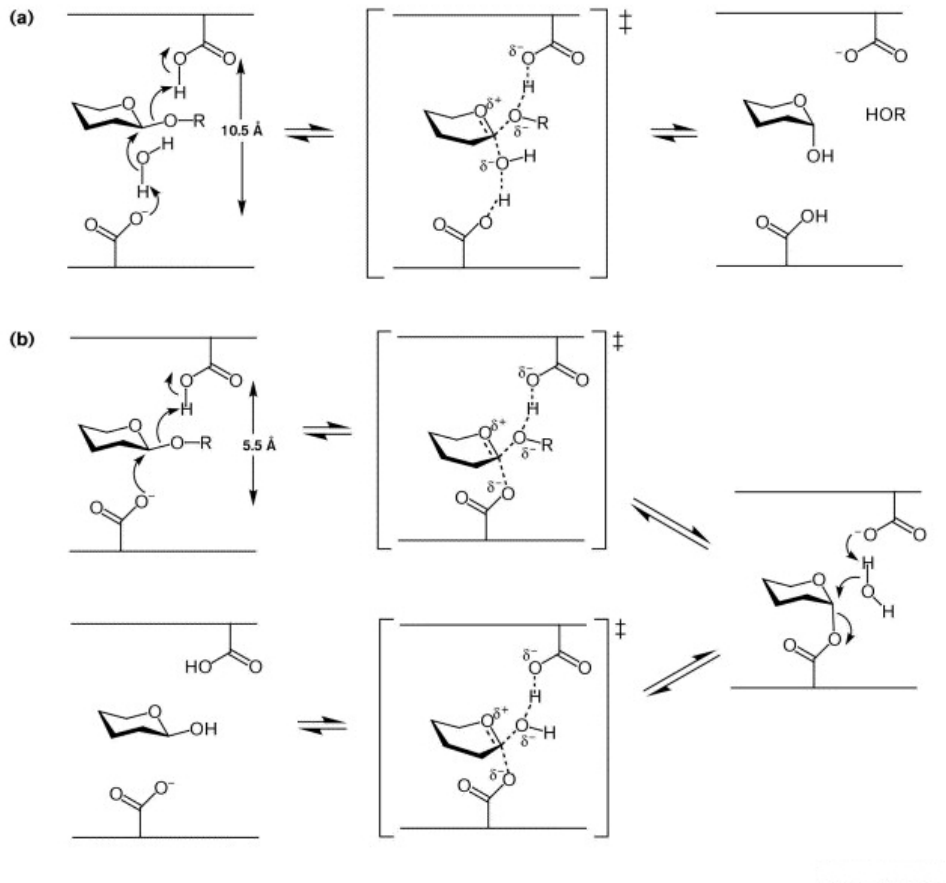

Figure 1.14: General mechanisms for inverting and retaining glycosidases. Top: The inverting mechanism showing the backside attack of the nucleophile, in this case water (glycoside hydrolase), leading to cleavage of the glycan and inversion of anomeric configuration. Bottom: The retaining mechanism showing the formation of a sugar-protein intermediate before attack by the nucleophile, again water in this case, leading to cleavage of the glycan and overall retention of anomeric configuration. Taken from [92]

Glycosidases are also of great interest for carbohydrate synthesis, both in research and for industrial applications.[93,94] The glycosidase enzymatic mechanism is reversible and so can be used to form glycosidic bonds, if an excess of the reaction products are present, allowing for the facile synthesis of glycans in a specific and stereocontrolled manner. Glycosidases are more attractive than GTs for this role for a number of reasons. Firstly, glycosidases are usually more stable and easier to isolate than GTs,[91] making production of large quantities needed for industry more accessible. Furthermore, the donor substrates of their reverse reaction, monosaccharides or sugar-1-phosphates, are cheaper to produce than the donor substrates for GTs (sugar-nucleotides).[95] Finally, glycosidases typically have a far broader substrate specificity than GTs, allowing them to be useful in a variety of reactions, including those involving unnatural con-

jugates. For example, β-glycosidases have been used to synthesise alkyl glycosides.[96] Furthermore, there are numerous accounts of directed evolution and mutagenesis in order to modify the specificity of glycosidases.[97–99]

1.2 Nuclear Magnetic Resonance

1.2.1 Spin and Energy Levels

Spin is an intrinsic angular momentum possessed by fundamental particles. This spin angular momentum is a vector ($\boldsymbol{S}$) with magnitude:

(1) $|\boldsymbol{S}| = [S(S+1)]^{\frac{1}{2}}\hbar$

where S is the spin quantum number (which may have integer or half-integer values) and $\hbar$ is Planck's constant (h) over 2π. The direction of S is quantised such that the z-component (S_z) value satisfies the equation:

(2) $S_z = m_s\hbar$

where m_s is the spin magnetic quantum number, and may adopt $2S + 1$ values between S and $-S$ as shown below (Fig. 1.15):

(3) $m_s = S, S-1, ..., -S+1, -S$

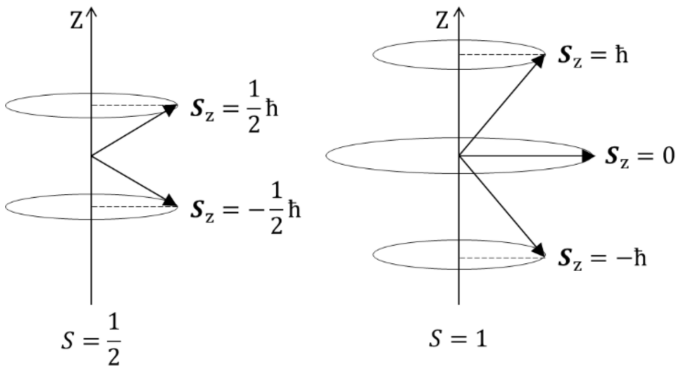

Figure 1.15: Quantisation of the spin angular momentum vector. As only the z-component and the magnitude are known, the vector may lie anywhere along the cone. If $S = \frac{1}{2}$ (left) then $m_s = \pm\frac{1}{2}$, whereas if $S = 1$ (right) then $m_s = 1, 0, -1$.

Nucleons are spin-half particles ($S = \frac{1}{2}$) and therefore their spin angular momentum may have m_s values of $\pm\frac{1}{2}$. In a nucleus the spins of the nucleons combine to give a total nuclear spin angular momentum ($\boldsymbol{I}$) described by the nuclear spin quantum number, I. Using the ^{2}H nucleus as an example, the two nucleons may both have the same m_s value (the spins are parallel) giving $I = 1$, or they may have opposite m_s values (the spins are antiparallel) giving $I = 0$ (Fig. 1.16). This can be generalised with the following expression:

(4) $I = |S_1 - S_2|, |S_1 - S_2| + 1, ..., |S_1 + S_2|$

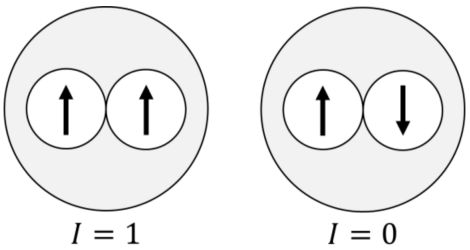

Figure 1.16: Possible nuclear spin states of the ^{2}H nucleus. A parallel arrangement of spins ($m_{S1} = m_{S2} = \frac{1}{2}$) gives $I = 1$, while an anti-parallel arrangement ($m_{S1} = \frac{1}{2}, m_{S2} = \frac{1}{2}$) gives $I = 0$.

A similar approach may be taken for larger nuclei. In the case of ^{2}H, the $I = 1$ state is the ground state and the $I = 0$ is the excited state. The energy difference between nuclear spin states is so large that, ignoring exceptional circumstances, for any nucleus it may always be assumed that it is in the ground nuclear spin state.[100] There is no way to predict exactly what the ground nuclear spin state will be; this must simply be determined experimentally. However the following may be said: nuclei with an odd number of nucleons have a ground state with half-integer spin ($I = \frac{1}{2}, \frac{3}{2}, ...$), while nuclei with an even number of nucleons have a ground state with zero- or integer-spin ($I = 0, 1, ...$).[101] The following discussion will concern only spin-half nuclei ($I = \frac{1}{2}$) as the techniques employed in this thesis employ only such nuclei and quadrupolar nuclei ($I > \frac{1}{2}$) suffer from additional complexities such as quadrupolar coupling and relaxation.

The direction of the nuclear spin angular momentum is quantised as for fundamental particles, giving a nuclear magnetic quantum number (m_I) with $2I + 1$ possible values. As nuclei possess both charge and an angular momentum they generate a magnetic moment ($\boldsymbol{\mu}$) defined as:

(5) $\boldsymbol{\mu} = \gamma \boldsymbol{I}$

where γ is the gyromagnetic ratio, an intrinsic property of a particular nucleus type defined as the ratio of the nucleus' magnetic moment to its angular momentum. Most nuclei, such as ^{1}H and ^{13}C, possess a positive gyromagnetic ratio. In such cases the nuclear spin and the magnetic moment are parallel (Fig 1.17).[100] However a minority of nuclei, including ^{15}N, have a negative gyromagnetic ratio resulting in antiparallel alignment of nuclear spin and magnetic moment.

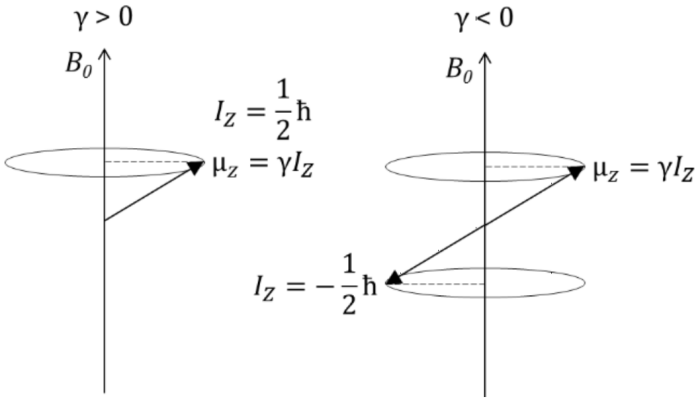

Figure 1.17: Alignment of the nuclear magnetic moment and the nuclear spin angular momentum in a magnetic field. The magnetic moment aligns with an external magnetic field such that the α-state is the ground state for nuclei with positive gyromagnetic ratios (left) and the β-state is the ground state for nuclei with negative gyromagnetic ratios (right).

In the absence of an external magnetic field, where the z-axis is arbitrary, there is no energetic preference for the nuclear spin angular momentum to be orientated in a particular direction. Therefore the $m_I = \frac{1}{2}$ and $m_I = \frac{-1}{2}$ states are equally populated. However when an external magnetic field is applied, now defined as the z-axis, the degeneracy is lifted because the nuclear magnetic moment interacts differently with the external magnetic field depending on its orientation. For nuclei, this phenomenon is known as the as the Nuclear Zeeman Effect (Fig. 1.18).

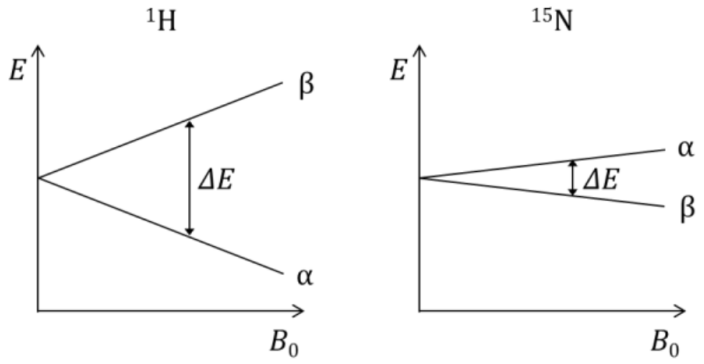

Figure 1.18: The Nuclear Zeeman Effect. The energy of α and β states as a function of external magnetic field strength. These states are split to a greater extent in nuclei with large gyromagnetic ratios such as ^{1}H (left) compared to nuclei with small gyromagnetic ratios such as ^{15}N (right).

A population difference is therefore generated with $m_I = \frac{1}{2}$ (or α-state) as the ground state and $m_I = \frac{-1}{2}$ (or β-state) as the excited state, for nuclei with positive gyromagnetic ratios. It is important here to emphasise the difference between nuclear spin states, which involve a change in the value of m_I, and nuclear magnetic states which involve a change in the value of m_I.[100] Nuclear magnetic resonance spectroscopy concerns transitions between m_I states, not changes in I. The difference in energy between m_I states is such that there is only a very slight preference for the α-state. For example, for ^{1}H nuclei in a 500 MHz field (298 K) the population difference is only 1 in every 10,000, giving an inherently weak NMR signal. This is made worse by nuclei with small gyromagnetic ratios, the states of which are split to a lesser extent by the external magnetic field than nuclei with larger gyromagnetic ratios.[102]

The energy of a particular nuclear magnetic state may be defined as:

$$(6)\quad E_{m_I} = -m_I \hbar \gamma \boldsymbol{B_0}$$

where $\boldsymbol{B_0}$ is the magnitude of the external magnetic field in Tesla (T). Therefore the difference in energy (when $I = \frac{1}{2}$) between m_I states may be defined as:

$$(7)\quad \Delta E = \gamma \hbar \boldsymbol{B_0}$$

If energy matching this difference is supplied to the system, a transition between the m_I states may be induced. As practically this energy

is delivered using a linearly oscillating magnetic field, it is often more convenient to describe the difference in energy in terms of a frequency, redefining the above equation as:

$$(8)\ \ \nu_0 = -\frac{\gamma \boldsymbol{B_0}}{2\pi} \text{ (in Hz) or } \omega_0 = -\gamma \boldsymbol{B_0} \text{ (in rad s}^{-1}\text{)}$$

where ν_0 or ω_0 are known as the Larmor frequency. The negative sign is a consequence of Larmor precession, which is discussed later.

1.2.2 Chemical Shift

The above description predicts that every nucleus of the same type would have the same Larmor frequency, whereas real NMR spectra contain a multitude of peaks across a range of frequencies. For example, a typical ^{1}H spectrum is expected to contain a peak for each ^{1}H nucleus in the molecule (ignoring chemical equivalence). To explain this, it is said that each nucleus has a chemical shift (δ in ppm), which modifies the Larmor frequency of each individual nucleus with the expression:[103]

$$(9)\ \ \nu_0 = -\frac{-\gamma \boldsymbol{B_0}[1 + (\delta \times 10^{-6})]}{2\pi} \text{ or } -\gamma[1 + (\delta \times 10^{-6})]\boldsymbol{B_0}$$

The chemical shift arises because each individual nucleus within a molecule actually experiences a slightly different magnetic field.[102] In a molecule the nucleus is surrounded by electrons in molecular orbitals. These electrons which also possess an angular momentum and charge (but of opposite sign to the nucleus) generate a local magnetic field opposed to the external field. The nucleus therefore experiences a field weaker than the actual external field and is said to be shielded. This effective field will be different for each nucleus depending on the exact electronic structure surrounding it. The current generated from electrons within orbitals, which in turn generates the opposing local magnetic field described above, is known as a diamagnetic current. It is also possible for electrons to generate a paramagnetic current, which generates a magnetic field aligned with the external magnetic field. In this case the nucleus experiences a magnetic field greater than the external magnetic field and is deshielded. This paramagnetic current is generated by movement of electrons between orbitals aligned in the xy-plane, for example between p_x and p_y orbitals. This can occur when

distortion causes orbital mixing; in the above example the p_x orbital will have some p_y character and vice versa. The net chemical shift of an individual nucleus is therefore the sum of local diamagnetic and paramagnetic currents, magnetic fields generated around neighbouring atoms and other sources that cause perturbation of the local electronic environment, such as hydrogen bonding and charged moieties

1.2.3 The Vector Model

In the absence of an external magnetic field, the magnetic moments of individual nuclei are oriented randomly such that the sample as a whole has no net magnetic moment. When the sample is subjected to an external magnetic field there is a slight preference for the magnetic moments of the nuclei to align parallel to the magnetic field. Therefore the vector sum of the magnetic moments of the whole sample gives a bulk magnetisation parallel to the external magnetic field (conventionally the z-axis) known as the equilibrium magnetisation.[103] It should be noted that formation of the equilibrium magnetisation is not instantaneous; it develops over a finite period of time (Fig. 1.19). The bulk magnetisation can be manipulated using weak, linearly oscillating magnetic fields. The interaction between this magnetic field and the nuclei in the sample is greatest at the Larmor frequency. For the ^{1}H nucleus within the magnetic field of a typical spectrometer, the Larmor frequency is of the order of hundreds of MHz, which corresponds to the radiofrequency (RF) region of the electromagnetic spectrum. For this reason, these weak magnetic fields are commonly termed RF fields. The most common manipulation of the bulk magnetisation is the $90_x°$ pulse, which applies a RF field along the x-axis for a given amount of time such that the bulk magnetisation is rotated 90° from the z-axis. Due to chemical shift, the RF field cannot be precisely on resonance with every nucleus in the sample. Therefore in a 90° pulse the magnetisation of nuclei slightly off-resonance with the pulse will not be fully on the xy-plane (transverse), but the effect of this is usually negligible. In some cases this effect may be utilised to selectively excite certain regions of the spectrum by intentionally choosing the strength and frequency of the RF field such that only specific resonances are effectively manipulated by the field.[103]

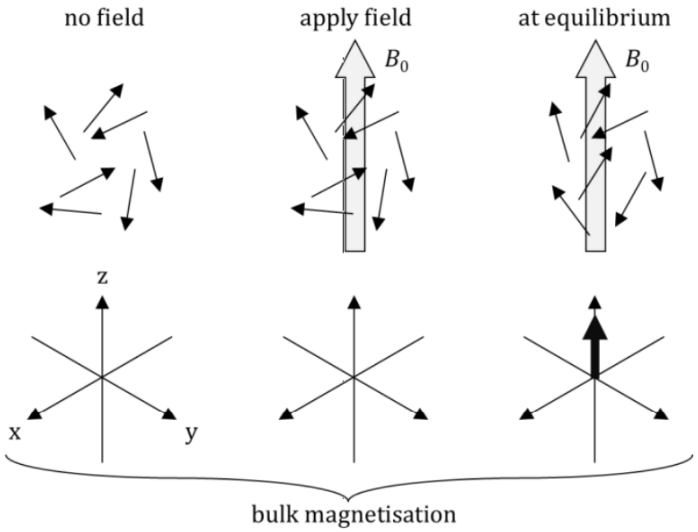

Figure 1.19: The bulk magnetisation. In a sample each nuclear magnetic moment is aligned randomly such that the bulk magnetisation is zero. Applying an external magnetic field gives a preference for the magnetic moments to be aligned with the magnetic field such that, after a period of time, an equilibrium magnetisation develops aligned parallel to the external magnetic field.

Immediately following the $90_x°$ pulse the bulk magnetisation is aligned along the y-axis but then begins to precess at the Larmor frequency as a consequence of a torque generated as a product of the nuclear magnetic moments and the external magnetic field. As mentioned earlier, the Larmor frequency is defined as being negative with respect to the gyromagnetic ratio. This is because nuclei with a positive gyromagnetic ratio precess in a clockwise direction about the z-axis, which in the axis system used corresponds to a negative rotation.[103] The precession of the bulk magnetisation about the z-axis is able to induce a current in a coil aligned along the xy-plane. The signal generated in this manner is known as the free induction decay (FID), due to the decay back to equilibrium magnetisation caused by relaxation. This is the essence of the most basic of NMR experiments, known as pulse acquire. A sample is subjected to an external magnetic field and after a finite period of time reaches an equilibrium magnetisation. The FID resulting from a 90° pulse is then recorded, generating a spectrum with peaks at the Larmor frequency of each nucleus in the sample (Fig. 1.20).

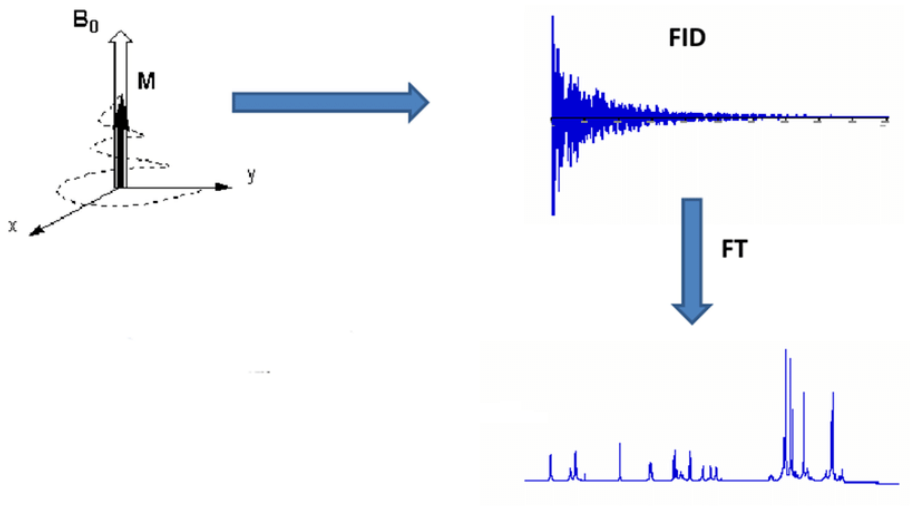

Figure 1.20: Diagram showing the origin of the NMR spectrum. Precession of the bulk magnetisation in the transverse plane induces a signal that is recorded by the spectrometer in the time domain (FID). The Fourier transform (FT) of the FID leads to the spectrum in the frequency domain, in which each peak corresponds to a frequency deconvoluted from the FID. Taken from [104]

Another common pulse is the 180° pulse which is named as such because it will flip the equilibrium magnetisation through an angle of 180° to be aligned along the -z-axis. However, when the bulk magnetisation is aligned in the xy-plane, the $180°_x$ pulse has the effect of a reflection in the xz-plane. The 180° pulse operates in the same manner as the 90° pulse except that the RF field is left on for twice as long. In fact, the 90° pulse is calibrated by finding the pulse length that gives no FID. This corresponds to a 180° pulse, as complete inversion gives no transverse magnetisation, so the 90° pulse must be half this time.

1.2.4 The Rotating Frame

The above description uses a frame of reference known as the laboratory frame; the external observer 'sees' the spins precessing as they do in reality. However it simplifies the mathematics of more complex experiments if a rotating frame is considered, as the time dependence of the RF fields is removed.[102] Here the frame of reference rotates about the z-axis at the frequency of the RF field (ω_{rf}). The offset (Ω) is defined as:

(10) $\Omega = \omega_0 - \omega_{rf}$

and gives the apparent frequency of precession for a given nucleus. Nuclei that are completely on resonance with the RF field will therefore appear static. When the gyromagnetic ratio is positive, nuclei that precess faster than ω_{rf} appear to rotate clockwise, while nuclei that precess slower appear to rotate anticlockwise.

1.2.5 Relaxation

Relaxation is the process in which the bulk magnetisation returns to its equilibrium position, either following manipulation or on first introduction of the sample into the external magnetic field. This can be separated into two separate processes: longitudinal relaxation, which describes the return of the z-component of the bulk magnetisation to its equilibrium position, and transverse relaxation, which describes the loss of coherence in the xy-plane such that, at equilibrium, it sums to zero. Both of these processes are dependent on fluctuations in the local magnetic field. For spin-half nuclei, the changes in the local magnetic field are dominated by dipolar interactions and chemical shift anisotropy.[101]

The dipole-dipole mechanism is the interaction of the magnetic field of one nucleus with the magnetic field of another. The interaction has an inverse-cube distance dependence and is dependent on the gyromagnetic ratios of the interacting nuclei. As a result, dipolar interactions rarely take place over more than a few Ångstroms. The strength of the interaction is also dependent on the angle between the two field vectors, meaning that this dipolar coupling changes as the molecule tumbles in solution. The chemical shift was previously described as the nucleus experiencing a field different from the external magnetic field due to shielding from the surrounding electrons. However, as alluded to above, the magnitude of this shielding is dependent on the relative orientations of the nucleus and the surrounding electrons. Therefore, the chemical shift of a nucleus changes as the orientation of the molecule within the sample changes. This is known as chemical shift anisotropy. The effect of this is not observed in solution-state spectra due to averaging by fast molecular tumbling. However, it still provides a mechanism for relaxation. From the above description, it is clear that the molecular motions present in a sample have a direct impact on relaxation. This motion is typically described by the cor-

relation time (τ_c), which is defined as the average amount of time taken for a molecule in a sample to rotate through an angle of one radian. The correlation time is affected by factors such as temperature and viscosity, but is principally dependent on molecular weight; the larger the molecule, the longer the correlation time. The issue with describing motion using the correlation time is that it only describes the average amount of motion present. It would be much more useful to be able to describe how this motion is distributed across different timescales.[103] This can be achieved using the spectral density function ($J(\omega)$), which describes the relative intensity of motion as a function of frequency:

$$(11)\quad J(\omega) = \bar{B}_{loc}^2 \frac{2\tau_c}{1+\omega^2\tau_c^2}$$

A plot of $J(\omega)$ against ω takes the form of a Lorentzian curve with a maximum at $J(0)$ (Fig. 1.21). An important feature of $J(\omega)$ is that its integral is independent of the correlation time. In other words, the area under the plot remains constant for all values of τ_c. Therefore, long correlation times correspond to a spectral density function with a large value of $J(0)$ that quickly decays with increasing frequency. However, short correlation times will give a plot with a smaller value of $J(0)$ but that extends into higher frequencies.

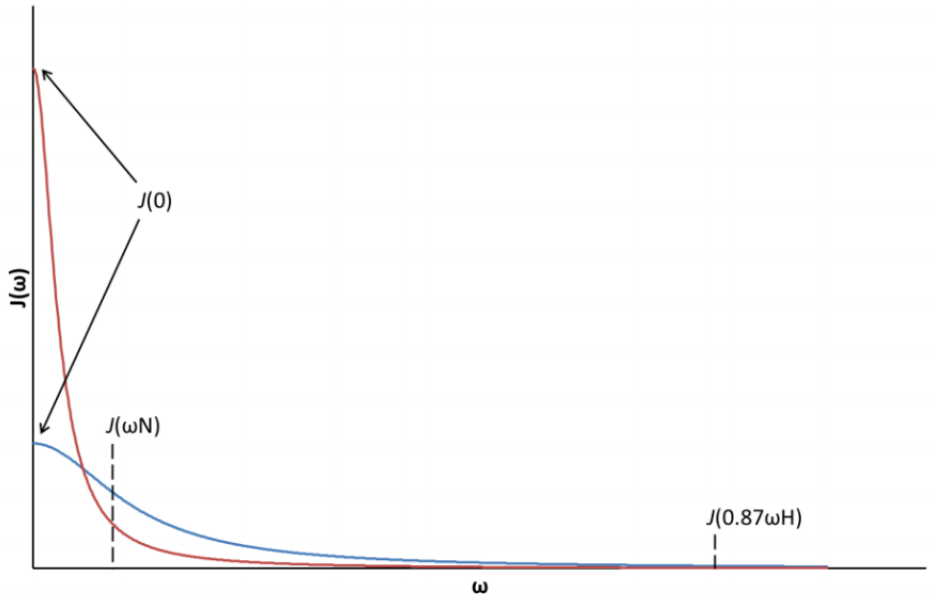

Figure 1.21: The spectral density function plotted for long (red) and short (blue) values of τ_c. Frequencies analysed by reduced spectral density analysis are marked.

1.2.5.1 Longitudinal Relaxation

Longitudinal relaxation is the process that returns the z-component of the bulk magnetisation to its equilibrium position, which requires nuclei to lose energy to their surroundings. For longitudinal relaxation to occur, the spectral density function must therefore have a component of motion at the Larmor frequency, in order to induce transitions between the two nuclear magnetic states. The longitudinal relaxation time constant (T_1) is therefore related to the spectral density function as follows:[102]

$$(12)\quad \frac{1}{T_1} = \gamma^2 \langle B^2 \rangle J(\omega_0)$$

Analysis of the spectral density equation shows that $J(\omega_0)$ is maximised, making T_1 relaxation most rapid, when $\tau_c = \frac{1}{\omega_0}$ (Fig. 1.22). Rearranging this to $\omega_0 \tau_c = 1$ allows two motional regimes to be defined. Fast motion (or the extreme narrowing limit) is defined as $\omega_0 \tau_c << 1$. In this regime analysis of the spectral density function equation shows that $J(\omega_0) = 2\tau_c$, and therefore $J(\omega_0)$ increases with longer correlation times up to a maximum at $\tau_c = \frac{1}{\omega_0}$. The slow motion regime (or spin diffusion limit) is defined as $\omega_0 \tau_c >> 1$ and analysis of the spectral density function equation gives $J(\omega_0) = \frac{J(0)}{\omega_0^2 \tau_c^2}$, meaning at the spin diffusion limit $J(\omega_0)$ increases with shorter values of τ_c.

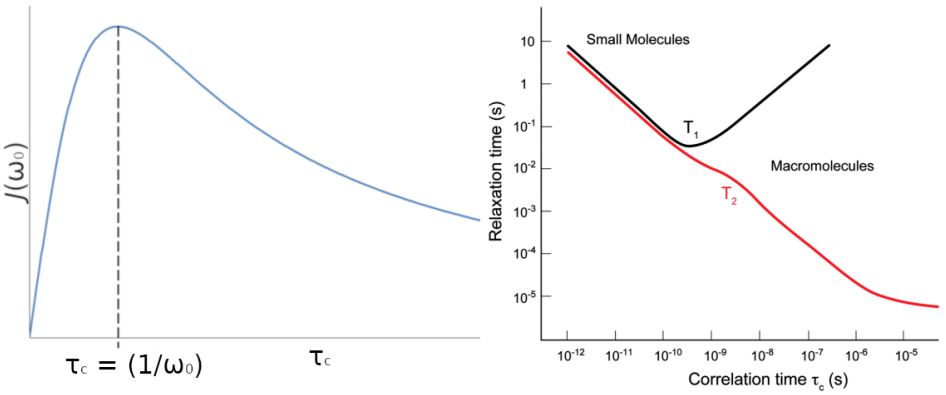

Figure 1.22: A plot of $J(\omega_0)$ as a function of correlation time (left) shows a maximum at $\tau_c = \frac{1}{\omega_0}$. As T_1 relaxation is dependent on the amount of motion at the Larmor frequency, it is also shortest at $\tau_c = \frac{1}{\omega_0}$ (right, taken from [105]). The plot also shows how T_2 becomes more rapid with longer correlation times.

A common method for measuring the longitudinal relaxation time is inversion recovery (Fig. 1.23). First the equilibrium magnetisation is inverted using a 180° pulse such that it points along the –z-axis. During the following delay period (τ) the magnetisation begins to relax back to equilibrium. A 90° pulse is then applied and the resulting FID is recorded. The experiment is performed multiple times with varying values of τ.

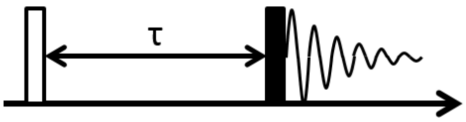

Figure 1.23: The inversion recovery pulse sequence. The magnetisation is first inverted by a 180° pulse (blank). After a delay (τ), a 90° pulse is applied (filled) and the resulting FID is recorded.

If τ is short then the bulk magnetisation will still point along the –z-axis immediately before the 90° pulse. In this case, the magnetisation is rotated towards the -y-axis, and the peaks in the resulting spectrum will be negative (Fig. 1.24). As τ increases, the magnitude of the bulk-magnetisation decreases to zero, giving negative spectra with decreasing intensities. If τ is extended further, the bulk magnetisation will begin to increase along the +z-axis up to the equilibrium magnetisation. In these cases the 90° pulse rotates the magnetisation along the y-axis, giving positive spectra with increasing peak intensities.

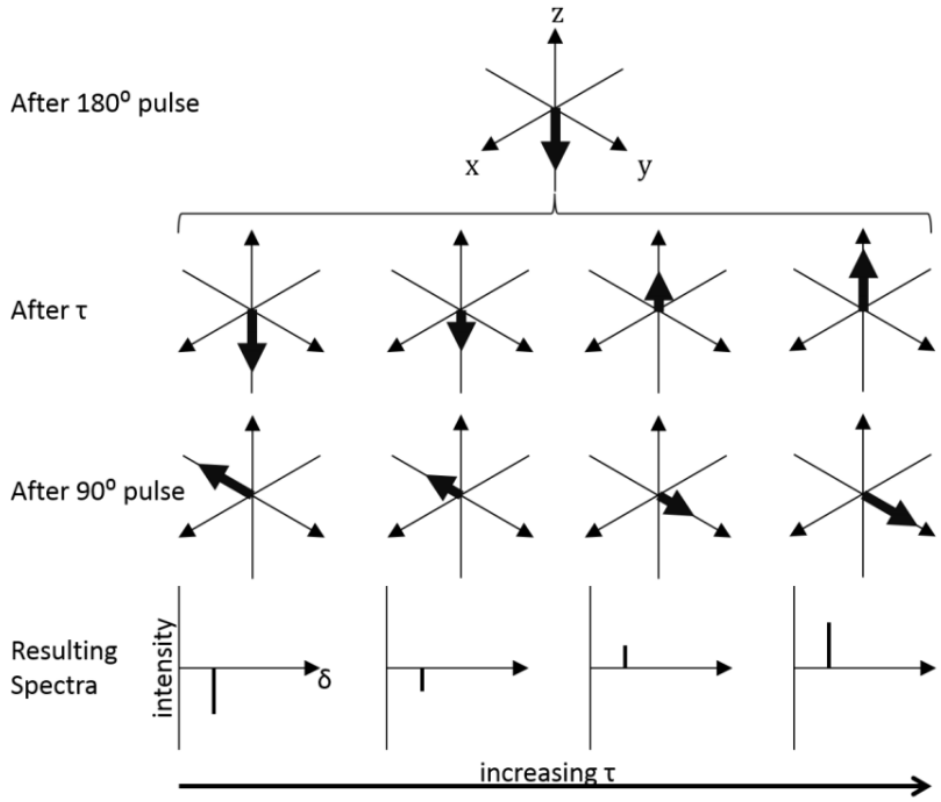

Figure 1.24: Relaxation recovery at increasing delay times (t). As t increases, the negative peak intensities decrease to zero before becoming positive.

Assuming that relaxation is exponential, the magnetisation at a given time ($M_z(\tau)$) can be expressed as:

$$(13)\quad M_z(\tau) = M_0(1 - 2e^{\frac{-\tau}{T_1}})$$

where M_0 is the equilibrium magnetisation, τ is the time between the 180° and 90° pulses, and T_1 is the longitudinal relaxation constant.

1.2.5.2 Transverse Relaxation

Transverse (T_2) relaxation is caused by loss of coherence in the transverse plane, meaning the individual spins precess at slightly different rates such that eventually the magnetisation in the xy-plane sums to zero. T_2 relaxation is responsible for line broadening in NMR spectra and is related to spectral linewidth as shown:

$$(14)\quad \Delta\nu_{\frac{1}{2}} = \frac{1}{\pi T_2}$$

where $\Delta\nu_{\frac{1}{2}}$ is linewidth at half height.

The above description for longitudinal relaxation also contributes to the transverse relaxation, as fields oscillating at the Larmor frequency can manipulate the orientation of the magnetisation.[103] The second contribution comes from the fact that each nucleus will precess at a slightly different rate due to experiencing different local magnetic fields. This contribution is maximised with slower molecular motion as each spin will experience a different magnetic field for longer, causing a rapid loss of coherence.[101] When a molecule tumbles quickly, the local magnetic field changes more rapidly and its effect is averaged out, reducing the effectiveness of this contribution (Fig. 1.22). The transverse relaxation time constant can therefore be described as:

$$(15)\quad \frac{1}{T_2} = \frac{1}{2}\gamma^2\langle B^2\rangle J(\omega_0) + \frac{1}{2}\gamma^2\langle B^2\rangle J(0)$$

As $J(0)$ is much greater than $J(\omega_0)$ (Fig. 1.21), the contribution from slow motion dominates T_2. Therefore it is expected that T_2 will always become shorter with longer correlation times (Fig. 1.22).

The observed transverse relaxation (T_2^*) is due to both external field inhomogeneities and molecular motion. However, the true transverse relaxation is only due to molecular motion. The two processes can be separated using a spin echo pulse sequence (Fig. 1.25). This pulse sequence begins with a $90°_x$ pulse followed by a delay of time, τ. During the delay some coherence will be lost due to spins precessing at slightly different rates (Fig. 1.26). The use of a $180°_x$ pulse causes the magnetisation to be reflected in the xz-plane. After a second delay, any coherence lost due to constant inhomogeneities (external magnetic field) will be refocused. However, random loss of coherence (molecular motion) will not be, so the observed loss in intensity of the FID must be from T_2.

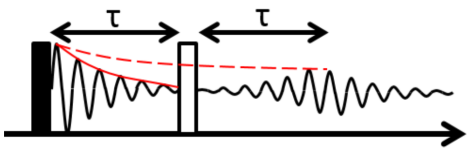

Figure 1.25: The spin-echo pulse sequence. After the 90° pulse (filled), the FID decays with a time constant, T_2^* (red, solid). After a delay (τ), a 180° pulse is applied. After a total time of 2τ, some coherence is refocused. The lost intensity is from T_2 (red, dashed).

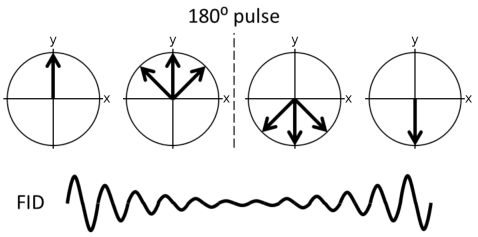

Figure 1.26: Refocusing of magnetisation in the xy-plane using a spin-echo pulse. Over a period of time (2τ) the magnetisation fans out due to loss of coherence. The 180° pulse flips the magnetisation in the xz-plane, and after 2τ the magnetisation is refocused.

The intensity of a peak in the spectrum following the spin echo ($I(2\tau)$) may be expressed as:

$$(16)\ \ I(2\tau) = I(0)exp\frac{-2\tau}{T_2}$$

where $I(0)$ is the intensity when $\tau = 0$.[102]

1.2.5.3 The Nuclear Overhauser Effect (NOE)

In a homonuclear system containing two spins (I and S), there are four energy levels with a total of six possible relaxation pathways (Fig. 1.27). Four of these pathways are single quantum transitions (W_1) that correspond to self relaxation as described above for longitudinal relaxation. These transitions are therefore dependent on the spectral density at the Larmor frequency of that spin:

$$(17)\ \ W_1^S = \frac{3}{40}b^2J(\omega_{0,I})$$

$$(18)\ \ W_1^I = \frac{3}{40}b^2J(\omega_{0,S})$$

where $b = \dfrac{\mu_0\gamma_I\gamma_S\hbar}{4\pi r^3}$

In addition to these self-relaxation mechanisms, it is also possible for cross relaxation to occur, in which the relaxation of one spin perturbs the energy state of the spin it is coupled to. The double quantum transition ($W_2^{I,S}$) between $\alpha_I\alpha_S$ and $\beta_I\beta_S$ corresponds to an energy difference of $\omega_{0,I} + \omega_{0,S}$. Therefore, according to the spectral density

this rate of this transition is dependent on the spectral density at $\omega_{0,I} + \omega_{0,S}$ (Eqn. 19). The zero quantum transition ($W_0^{I,S}$) between $\alpha_I\beta_S$ and $\beta_I\alpha_S$ corresponds to an energy difference of $\omega_{0,I} - \omega_{0,S}$. Therefore, according to the spectral density this rate of this transition is dependent on the spectral density at $\omega_{0,I} - \omega_{0,S}$ (Eqn. 20).

$$(19)\ W_2 = \frac{3}{10} b^2 J(\omega_{0,I} + \omega_{0,S})$$

$$(20)\ W_0 = \frac{1}{20} b^2 J(\omega_{0,I} - \omega_{0,S})$$

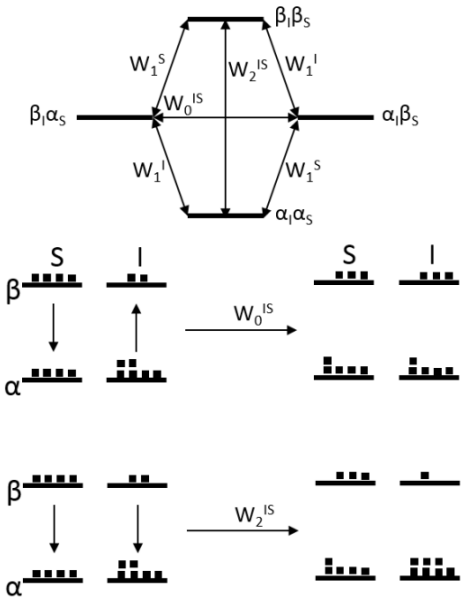

Figure 1.27: The energy level diagram for two dipole-dipole coupled spins, I and S, showing allowed transitions (top). Zero-quantum transitions (W_0) result in a reduction in the population difference of I (middle), leading to a negative enhancement. Double quantum transitions (W_2) result in an increase in the population difference of I (bottom), leading to a positive enhancement.

In the W_2 transition, relaxation of the S-spin also causes relaxation of the I-spin whereas in the W_0 transition, relaxation of the S-spin ($\beta \rightarrow \alpha$) causes excitation of the I-spin ($\alpha \rightarrow \beta$). To understand which transition will dominate in a given system, it is necessary to define the rate constants for self relaxation(R_z^I and R_z^S) and for cross relaxation ($\sigma_{I,S}$):

$$(21)\ R_z^I = 2W_1^I + W_2 + W_0$$

$$(22)\ R_z^S = 2W_1^S + W_2 + W_0$$

(23) $\sigma_{I,S} = W_2 - W_0$

Substituting in the definitions of the single- (Eqns. 17 - 18), double- (Eqn. 19) and zero-quantum (Eqn. 20) rates constants gives:

(24) $R_z^I = b^2[\frac{3}{20}J(\omega_{0,I}) + \frac{3}{10}J(\omega_{0,I} + \omega_{0,S}) + \frac{1}{20}J(\omega_{0,I} - \omega_{0,S})]$

(25) $R_z^S = b^2[\frac{3}{20}J(\omega_{0,S}) + \frac{3}{10}J(\omega_{0,I} + \omega_{0,S}) + \frac{1}{20}J(\omega_{0,I} - \omega_{0,S})]$

(26) $\sigma_{I,S} = b^2[\frac{3}{10}J(\omega_{0,I} + \omega_{0,S}) - \frac{1}{20}J(\omega_{0,I} - \omega_{0,S})]$

From here, it is clear that self-relaxation will always dominate since all the single, double and zero-quantum pathways contribute in an additive manner to the rate constant. Therefore the effect of cross relaxation is not usually directly observable unless conditions are set up to do so. Such experiments will be discussed later.

The cross relaxation rate is dependent on the difference between the double- and zero-quantum transition rates, such that the cross relaxation rate is positive where W_2 dominates and negative where W_0 dominates (Eqns. 23 and 26). From their definitions, clearly W_2 will dominate when a significant fraction of the spectral density can be found at $\omega_{0,I} + \omega_{0,S}$ - in other words fast motion or a short value of τ_c, whereas W_0 will dominate when a more significant fraction of the spectral density is at $\omega_{0,I} - \omega_{0,S}$ - in other words slow motion or a large value of τ_c. Clearly the cross-over point will occur when $W_2 = W_0$. Although the exact rate of cross relaxation is dependent on a number of factors, including field strength, it is typically expected that small organic molecules will exhibit a positive cross-relaxation rate, whereas macromolecules such as proteins will have a negative cross-relaxation rate (Fig. 1.28).

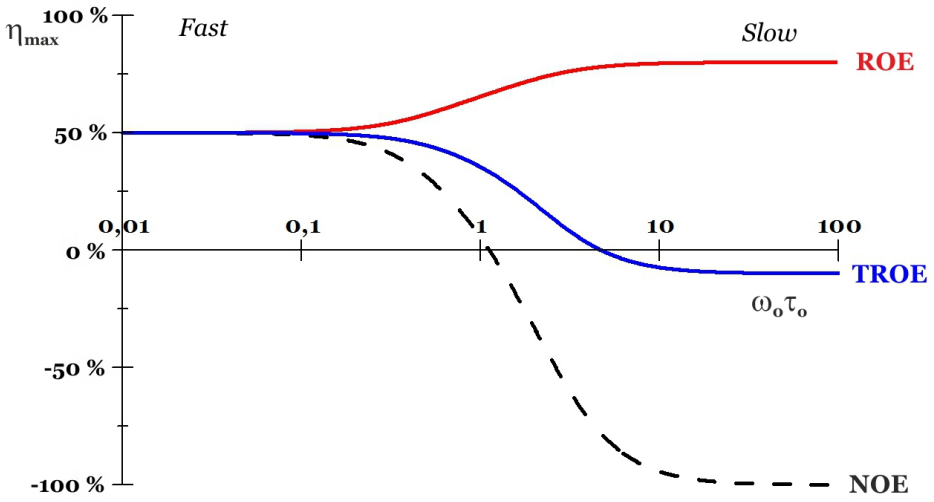

Figure 1.28: The dependency on the NOE enhancement on correlation time (dashed). The enhancement is positive for short correlation times and becomes negative with longer correlation times. Transient rotating frame NOE (blue) and rotating frame NOE (red) are also shown

Although the above equations describe the rate constants for self-relaxation and cross-relaxation, how these rate constants affect the actual longitudinal magnetisation of each spin has not yet been discussed. For this, the Solomon equations are described below:[106]

$$(27)\quad \frac{dI_{Iz}}{dt} = -R_z^I(I_{Iz} - I_{Iz}^0) - \sigma_{IS}(I_{Sz} - I_{Sz}^0)$$

$$(28)\quad \frac{dI_{Sz}}{dt} = -R_z^S(I_{Sz} - I_{Sz}^0) - \sigma_{IS}(I_{Iz} - I_{Iz}^0)$$

As one might expect, the rate of change in the longitudinal magnetisation of each spin is proportional to its own self-relaxation rate constant, given that the spin itself has been perturbed from equilibrium ($I_z \neq I_z^0$). In addition to this, the second term states that the rate of change in the longitudinal magnetisation in the first spin (I) is dependent on the cross-relaxation rate constant between the two spins ($\sigma_{I,S}$) given that the second spin (S) has been perturbed from equilibrium.

The fact that cross-relaxation has an effect on the longitudinal magnetisation as described above is known as the Nuclear Overhauser Effect (NOE), and experiments can be designed such that such an effect can be detected.

1.2.5.3.1 Transient NOE The transient NOE experiment requires two separate pulse sequences. The first starts by selectively inverting just one resonance with a selective 180° pulse, followed by a delay(τ) (Fig. 1.29). During this time, the single resonance relaxes by both self- and cross-relaxation processes - the latter of which perturbs other spins close in space from their equilibrium magnetisations. A non-selective 90° pulse then prepares all spins for detection. The second pulse sequence is simply a reference spectrum generated from a single 90° pulse. Subtracting the reference spectrum from the first experiment produces a difference spectrum in which any observed intensity must be from cross-relaxation. Signals in the difference spectrum may be positive or negative depending on the sign of the cross-relaxation.

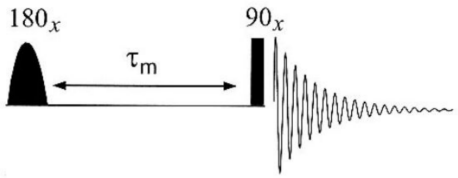

Figure 1.29: The pulse sequence for the transient NOE experiment. A single resonance is inverted with a selective 180° pulse. After a delay period (τ_m) non-selective 90° pulse is applied prior to detection.

If I is the spin that is inverted and S is spin coupled through dipolar interaction, the experiment can be described using an initial rate approximation:

$$(29)\quad \frac{dI_{Sz}}{dt}(init) = -R_z^S(I_{Sz}^0 - I_{Sz}) - \sigma_{IS}(-I_{Iz}^0 - I_{Iz})$$

$$(30)\quad \frac{dI_{Sz}}{dt}(init) = 2\sigma_{IS}I_{Iz}^0$$

This can be integrated to give:

$$(31)\quad I_{Sz}(t) = 2\sigma_{IS}I_{Iz}^0 t + c$$

$$(32)\quad I_{Sz}(t) = 2\sigma_{IS}I_{Iz}^0 t + I_{Sz}^0$$

The constant of integration is known here since at $t = 0$ the z-magnetisation of spin-S ($I_{Sz}(0)$) will be its equilibrium value (I_{Sz}^0). Since the 90° pulse will convert all the z-magnetisation after the delay (τ) into observable signal, the observed signal is therefore proportional

to:

$$(33)\ I_{Sz}(\tau) = 2\sigma_{IS} I^0_{Iz}\tau + I^0_{Sz}$$

whereas in the reference spectrum the observed signal will simply be proportional to I^0_{Sz}. Therefore, overall the so called NOE enhancement (η) can be described as:

$$(34)\ \eta = (2\delta_{IS} I^0_{Iz}\tau + I^0_{Sz}) - I^0_{Sz} = 2\sigma_{IS} I^0_{Iz}\tau$$

Therefore, the cross-relaxation rate constant can be calculated by measuring the the NOE enhancement directly. Since the cross-relaxation is dependent on the distance between the two spins, such measurements are useful in structural calculations in order to produce distance restraints.

1.2.5.3.2 Steady State NOE The scheme for the steady state NOE experiment is similar to that of the transient NOE, except that the inverting 180° pulse is replaced with a continuous low power saturating pulse. This pulse is still selective for a single resonance (I), but instead of inverting the z-magnetisation, the population is equalised such that there is no net magnetisation. Since the self-relaxation pathway is inaccessible due to the constant saturating pulse keeping equalised populations, the only available relaxation pathway is *via* cross relaxation to a nearby spin (S). The z-magnetisation of the S-spin will continue to change until eventually it reaches a steady state in which the cross-relaxation rate from the I-spin equals that of its self-relaxation. In this case, the Solomon equation for spin S is:

$$(35)\ 0 = -R^S_z(I_{Sz,SS} - I^0_{Sz}) - \sigma_{IS}(-I^0_{Iz})$$

$$(36)\ I_{Sz,SS} = \frac{\sigma_{IS}}{R^S_z} I^0_{Iz} + I^0_{Sz}$$

Subtracting the reference spectrum from the saturated spectrum gives an NOE enhancement described as:

$$(37)\ \eta_{SS} = \frac{\sigma_{IS}}{R^S_z}$$

One disadvantage of using the steady state NOE is that the observed enhancement isn't due only to the cross-relaxation but also due to the

self-relaxation. Therefore this approach can only be treated qualitatively, but it does have the advantage that theoretically the maximum enhancement should be larger than that of the transient NOE, since the irradiated spin can only relax *via* cross-relaxation.

1.2.5.3.3 Truncated Driven NOE The steady state NOE described above depends heavily on the self relaxation rate (R_z) of each proton and so cannot be used to directly measure distances. Conversely, the transient NOE is typically too weak to accurately measure distances with practical acquisition times.

A solution to this is to apply the saturating pulse utilised in the steady state experiment for a shorter period of time (typically 100s of ms), such that the initial rate of NOE build up is measured instead if the equilibrium value. This approach is known as the truncated driven NOE (TOE),[107] and is useful since, during the initial build up of NOE, the cross-relaxation rate (σ_{IS}) dominates over the self relaxation rate (R_z).

Since the self relaxation rate is not significant during the initial build up of the NOE, internuclear distances can be extracted with reasonable accuracy by calibrating using a TOE value between two nuclei of known distance (covalently bound). The TOE is advantageous over the transient NOE in that experimental times are significantly lowered.[107] The TOE is also the basis of the STD NMR technique described below.

1.2.6 Saturation Transfer Difference NMR spectroscopy

Saturation Transfer Difference (STD) NMR spectroscopy is mostly employed as a ligand-based NMR experiment for detecting the interaction between a small molecule ligand (e.g. carbohydrate) and a macromolecular receptor (e.g. protein), although it also has other applications, such as detecting the kinetics of internal rotations in small molecules.[108] It is a NOE-based technique that relies on the detection of an intermolecular transfer of magnetisation from a saturated receptor molecule onto a small molecule ligand. To understand this, the

thermodynamics and kinetics of binding must first be described.

1.2.6.1 Thermodynamics and Kinetics of Binding

The simplest description of a bimolecular association process is the one-site model, in which there is a reversible exchange between the free macromolecule (M) and ligand (L) and the complex of the two (ML). The model assumes no other intermediate species, such as rearrangement of either molecule upon binding. The rate of association and dissociation can be described by the kinetic rate constants k_{on} and k_{off} respectively.

$$(38)\quad M + L \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} ML$$

Equilibrium is achieved when rate of association of the two individual species is equal to the rate of dissociation of the complex. The relative proportions of free and bound species present at equilibrium are dependent on the affinity of the two species for forming the complex - that is, for species with high affinity for one another, the complex will be present in greater proportion at equilibrium compared to species with lower affinity for one another. This can be described by the molar dissociation constant:

$$(39)\quad k_{on}[M][L] = k_{off}[ML]$$

$$(40)\quad K_d = \frac{[M][L]}{[ML]} = \frac{k_{off}}{k_{on}}$$

A smaller value for K_d indicates a higher affinity complex. Therefore it is clear to see that smaller values of k_{off} and larger values of k_{on} contribute to a higher affinity complex. This is because k_{off} is inversely proportional to the residence time of the ligand - that is how long the ligand remains bound to the macromolecule - and k_{on} is proportional to the probability of forming the complex. Usually k_{on} is assumed to be limited by diffusion, in which case it can be assumed to be of the order of 10^8 M^{-1} s^{-1}.

Alternatively the interaction can be described by the fraction of bound macromolecule present at equilibrium:

$$(41)\quad f_B^M = \frac{[ML]}{[M]+[ML]} = \frac{[L]}{[L]+K_d}$$

where the final term in the expression is derived from Eqns. 39 and 40, and corresponds to the Langmuir binding isotherm (Fig. 1.30). Here it is clear that the fractional occupancy of the macromolecule can be increased with increasing ligand concentration. The equation follows a hyperbolic function in which the fraction of bound macromolecule initially increases linearly with ligand concentration in the limit of $[L] << K_d$ and then asymptotically approaches fully saturated macromolecule as $[L] >> K_d$.

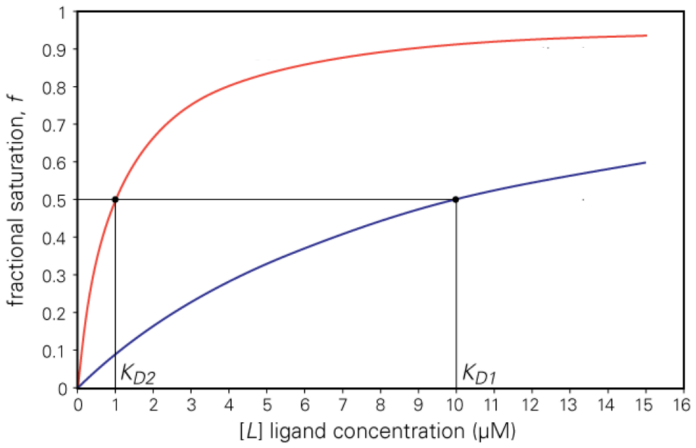

Figure 1.30: Representative Langmuir binding isotherms for two ligands, one with lower affinity (blue, $K_{d,1}$) amd one with higher affinity (red, $K_{d,2}$). The dissociation constant (K_d) values correspond to ligand concentrations at which the macromolecule is 50% saturated with bound ligand at equilibrium.

1.2.6.2 The STD NMR experiment

The STD NMR experiment is essentially equivalent to the truncated driven NOE (TOE) experiment but here the saturating pulse is applied selectively to the macromolecule. Since this selective pulse can only 'hit' a narrow range of resonances, only small subset of macromolecule spins are directly irradiated. However, for large molecules that tumble slowly in solution, cross-relaxation is very efficient and spins that are nearby to those that are directly irradiated are also rapidly saturated. Importantly, it is possible for the indirectly saturated spins - those that become saturated by cross-relaxation of the directly irradiated spins - to produce a so-called relayed NOE, in which they too can relax by cross-relaxation with there neighbours. By this relay of cross-relaxation, all of the spins of the macromolecule are rapidly saturated.

This process is know as spin diffusion.

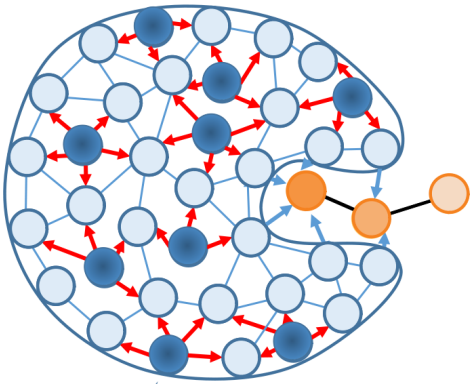

Figure 1.31: Cartoon showing the process of spin diffusion in STD NMR. Certain macromolecule protons are directly saturated by the saturating pulse (dark blue). This saturation is transferred (red arrows) to other protons that are close in space (light blue) and eventually to the bound ligand (orange). Ligand protons in closer proximity to the surface of the macromolecule receive more saturation (darker).

The macromolecule-ligand complex is formed spontaneously between the saturated macromolecule and the unperturbed ligand, allowing for saturation to be transferred intermolecularly to the ligand by intermolecular spin diffusion (Fig. 1.32). This is made possible since, in the bound state, the NMR parameters of the ligand (such as correlation time) become those of a macromolecule. Therefore, cross-relaxation of ligand spins with macromolecule spins within the bound state is very effective. However, it is necessary for the ligand to then dissociate into the free state in order to be detectable as the efficient transverse relaxation (and the low concentration of the complex) precludes the observation of the ligand signals in the bound state. Cross-relaxation in rapidly tumbling small-molecules is relatively ineffective and so the saturated spins can remain saturated for the duration of the experiment. The constant exchange of ligand between the free and the bound state allows for bulk saturation of the ligand, which results in an observable reduction in ligand peak intensity. As with the steady state NOE experiment, subtraction of the saturated spectrum (I_{sat}) from a reference spectrum (I_0) results in a difference spectrum (I_{diff}) in which the intensity of the ligand peaks are proportional to the amount of saturation transferred from the macromolecule to that ligand resonance.

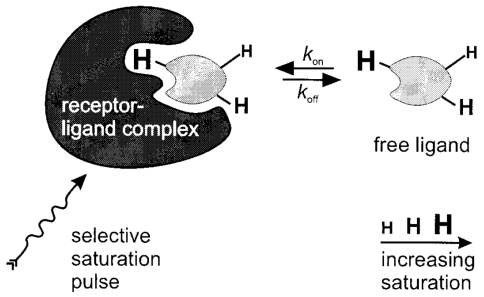

Figure 1.32: Diagram showing the principle of STD NMR spectroscopy. The receptor is selectively saturated, which can transfer magnetisation to any bound ligand through an intermolecular NOE. Ligand protons in closest proximity to the receptor surface receive the most saturation. This saturation accumulates in the bulk free ligand which is detected in the STD NMR difference spectrum. Taken from [109]

$$(42)\quad I_{diff} = I_0 - I_{sat}$$

$$(43)\quad STD(\%) = \frac{I_{diff}}{I_0} \times 100$$

One important consideration for STD NMR spectroscopy is that the exchange between the free and the bound states must be fast - that is the residence time of the ligand within the bound state must be much shorter than the relaxation time of that ligand proton within the bound state. Otherwise the magnetisation transferred to the ligand from the macromolecule would relax before it is able to accumulate in solution. For macromolecules, T_1 is typically relatively long (~1 s) but T_2 may be expected to be as short as 1-10 ms (Fig. 1.22). Typically then one might expect that the high affinity limit would be for complexes with dissociation constants in the micromolar range, giving a residence time of approximately 0.1-10 ms, assuming diffusion controlled association.

Furthermore, the ligand should be in a large excess over the macromolecule. Firstly, in the fast exchange limit, the average NMR properties of the bulk ligand ($\langle Q_L \rangle$) can be assumed to be the weighted sum of its properties in the free (Q_f^L) and the bound state Q_b^L:

$$(44)\quad \langle Q \rangle = f_f^L Q_f^L + f_b^L Q_b^L$$

Therefore, by maintaining a large ligand excess, the bulk ligand behaves as a small molecule and so benefits from sharp peaks and limited spin diffusion. In addition, the large excess ensures that the macromolecule binding sites are fully saturated with ligand, ensuring the

maximum amount of signal can be gained from the experiment. Finally, the large excess minimises the chance of ligand rebinding meaning that most likely only unperturbed ligand will bind to the macromolecule, allowing for more accumulation of saturation on the bulk ligand.[110]

1.2.6.3 Binding epitope mapping by STD NMR

The presence of peaks in the STD NMR difference spectrum is indicative of binding of the small molecule ligand to the macromolecule, assuming that the small ligand is not directly irradiated by the saturating pulse. This is commonly utilised in fragment screening as it provides a relatively facile method for detecting the binding of low affinity fragments to their targets, which may not be accessible to other commonly used techniques[111] (Fig. 1.33).

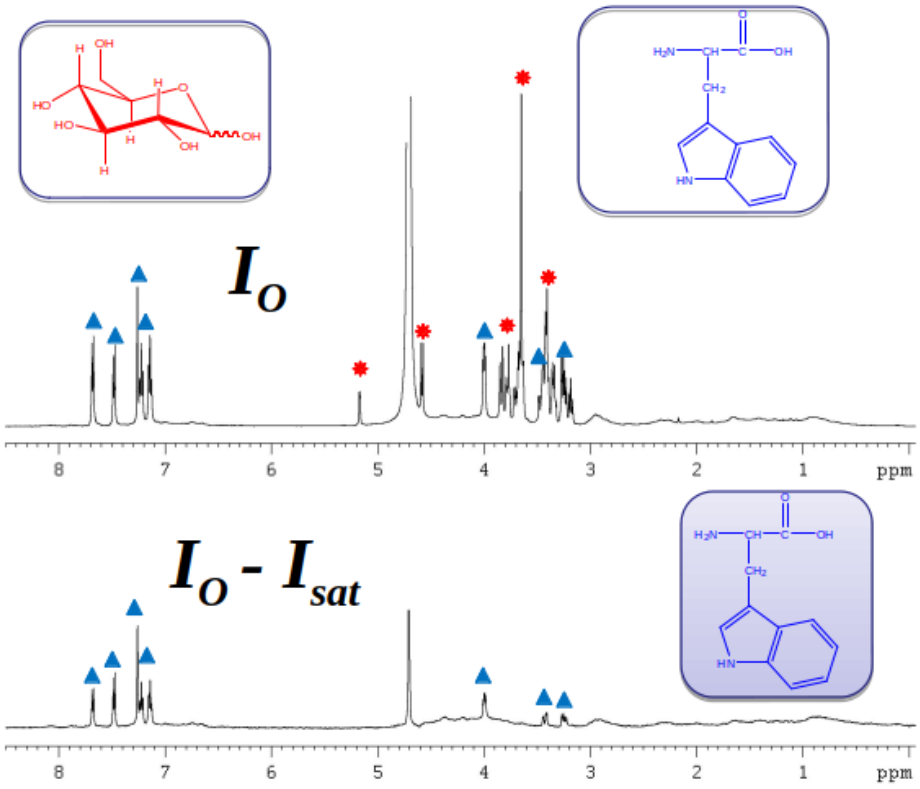

Figure 1.33: Reference (top) and STD NMR difference spectra of glucose (red) and tryptophan (blue) in the presence of bovine serum albumin (BSA). The difference spectrum contains only tryptophan resonances showing that tryptophan binds to BSA, whereas glucose does not.

However, an important advantage of STD NMR spectroscopy is that

since the strength of the dipolar interaction is dependent on the distance between the two spins, those ligand protons that are in closest proximity to the macromolecule surface receive the most saturation and so would be expected to have the strongest peaks in the STD NMR difference spectrum. Unfortunately, the magnitude of the STD intensity cannot be directly correlated with distances because, as with the steady state NOE, the observed cross-relaxation rate is also dependent on the self-relaxation term, as well as the exchange between the free and the bound states. Nevertheless, mapping of the relative STD intensities onto a structure of the small-molecule can give important information about the orientation of the small molecule within the binding site, since the most intense STD values should be in close proximity to the macromolecule surface - as an example see Fig. 1.34.

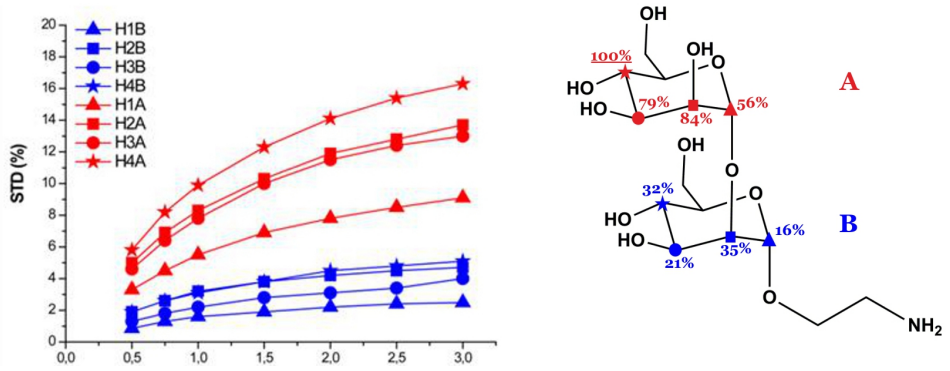

Figure 1.34: Example of binding epitope mapping using Man-α(1,2)-Man-α-O(CH_2)NH~ in the presence of anti-HIV-1 human antibody 2G12. (Left) STD build up curves are constructed by measuring the STD intensity for each proton over a number of saturation times. Fitting to this data gives the initial growth rate, which is normalised and mapped onto a structure of the ligand (Right). Taken from the Doctoral Thesis of Pedro M. Enríquez-Navas (University of Seville, 2011).

In practice, the 'initial growth rates' method is usually used.[112] Using this method the STD intensity of each ligand proton is measured for a range of saturation times (t_{sat}). The resulting 'build-up curve' takes the form of a monoexponential curve that plateaus at a maximum STD value (STD_{max}). This is because at longer saturation times the self-relaxation will be such that no more saturation can accumulate by cross-relaxation. The build-up curve can be modelled using the equation:

$$(45)\quad STD(t_{sat}) = STD_{max}(1 - exp(-k_{sat}t_{sat}))$$

where k_{sat} is the rate constant for saturation transfer. The curve can

be fitted to the experimental data by using STD_{max} and k_{sat} as fitting parameters. Initially the STD intensity increases linearly according to the rate of saturation transfer. Therefore the initial growth rate (STD_0) can be defined as:

$$(46)\quad STD_0 = \lim_{t_{sat}\to 0} \frac{\delta STD(t_{sat})}{\delta t_{sat}} = STD_{max} k_{sat}$$

The initial growth rate is then determined for each ligand proton. These rates are then normalised against the largest growth rate before mapping onto a structure of the ligand (Fig. 1.34).

The main advantage of the initial growth rates method is that it effectively eliminates the effect of self-relaxation on the observed value. This is important because protons with slower longitudinal relaxation times (e.g. typically aromatic protons) accumulate magnetisation more effectively than those with faster relaxation times. Therefore, the observed STD intensity at longer saturation times is significantly skewed by self-relaxation and may lead to misinterpretation of the binding epitope map.

1.3 Molecular Modelling

NMR spectroscopy experiments do not directly produce a set of coordinates that generate a 3D molecular model of the protein-ligand complex. In NMR spectroscopy intensities are measured, and those intensities must be interpreted in the context of a 3D model of the macromolecule and the ligand that matches the experimental observations. In this case we resort to molecular modelling, where we can generate different models of the molecules based on energetic grounds, and combine them with the experimental NMR observations to generate NMR-derived structures. Even if a crystal structure is available for the complex, molecular modelling is typically needed (in the form of molecular dynamics) to give flexibility to the structure to better match the experimental NMR observables. In the last section of this introduction the most commonly used methods for modelling biomolecular structures will be discussed.

1.3.1 Molecular Mechanics

Molecular mechanics (MM) is the application of classical mechanics to describing and predicting the chemical and physical properties of molecular systems, usually utilising the power of modern computational systems to perform the calculations. Often these are atomistic models of the system in question in which each atom is modelled by virtual 'particle' and the relationship between each particle and another is described by a set of terms that model properties such as atomic radii, bonds and charges (which will be discussed in depth later). The number of particles in a system can vary considerably depending on the nature of that system and the particular MM method being utilised, but, for example, an atomistic MM model of a protein can easily exceed 10,000's particles. This can make modelling very computationally demanding and limits the size of system that can reasonably be studied, as well as the timescale (e.g. molecular dynamics). In response to this, further simplified models (atomistic classical mechanics in itself is already a simplified approximation of the true system) have been developed that hope to extend both the size and time dimensions of molecular mechanics models. These include course graining,[113] in which the virtual particles instead represent multiple atoms, or indeed entire molecules or groups of molecules, or simplified descriptions of the particle interactions, such as only considering the Lennard-Jones or torsional interactions.[114] Of course, these also have their limitations, since it becomes increasingly difficult to accurately represent a complicated system with fewer parameters.

1.3.1.1 Molecular Mechanics Forcefields

The total energy of a molecular mechanics (MM) model system can be described as the sum of all the terms that describe the interaction of a particle with the other particles in the system, for all particles of the system. Over time many different MM forcefields have been developed, but they can typically be termed generic or specific forcefields, meaning that they are designed to either work for as broad a range of molecules as possible or a particular class of molecule respectively. Of course, with the generalisation of a forcefield comes a penalty in accuracy since the classical nature of MM cannot inherently describe effects

of a quantum nature. However, specific forcefields can be optimised to reproduce such quantum behaviours. A good example would be GLYCAM[115] a carbohydrate specific parameter set based on the AMBER forcefield[116] The conformation of a carbohydrate ring is heavily influenced by the anomeric effect, which of course can be explained by quantum mechanics. Therefore, generic forcefields cannot accurately reproduce the known ring conformation of many carbohydrates, but GLYCAM is optimised to do so. The drawback to this is that the specific GLYCAM parameters cannot be extended to non-carbohydrates as they do not make physical sense outside of their intended use case.

The exact functional forms of different forcefields varies but typically they can be generalised to the sum of covalent and non-covalent terms:

$$E_{pot} = E_{covalent} + E_{non-covalent} \quad (47)$$

where the covalent potential typically contains terms that describe the bond-length, bond-angle and torsion-angle, whereas non-covalent interactions such as van der Waals and electrostatic forces are described by the Lennard-Jones (LJ) and Coulombic potentials respectively. Some forcefields include additional terms that account for more nuanced or specific effects, such as coordination to metal ions.[117]

1.3.1.1.1 The AMBER Forcefield The AMBER (Assisted Model Building with Energy Refinement) forcefield is a biomolecular forcefield developed primarily for modelling protein systems, although it has since been extended to nucleic acids,[118] lipids[119] and carbohydrates.[115] It has been parametrised primarily from protein X-ray structures and to reproduce backbone dihedral angles of small peptides. It is perhaps one of the most popular forcefields for protein simulations due to its performance at accurately reproducing protein structure, especially at longer timescales (see molecular dynamics) compared to other forcefields. The group responsible for the AMBER forcefield have also produced an excellent suite of software (AmberTools) that makes building, simulating and analysing far more accessible compared to many other modelling programs. It has been commented that the AMBER forcefield does have a bias towards helical secondary structure, although this has been improved in more recent versions based on parametrisation including experimental NMR data.[116]

The covalent potential for the AMBER forcefield is defined as:

$$(48)\ E_{covalent} = \sum_{bonds} K_r(r - r_0)^2 + \sum_{angles} K_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} \frac{V_n}{2}[1 + cos(n\phi - \gamma)]$$

where the first term is in the form of a harmonic potential dependent on the length of each chemical bond (r) compared to an equilibrium length (r_0) and the magnitude of the potential depends on the specific force constant (K_r) for each bond. Similarly the angle (θ) between each particle connected by two bonds is harmonic with a angle force constant, K_θ. Conversely the potential for dihedral angles (ϕ, the angle between particles connected by three bonds) is periodic with a periodicity of n, an offset of γ and a dihedral constant, V_n.

In AMBER, the only non-bonded terms are the LJ and Coulombic potentials:

$$(49)\ E_{non-covalent} = \sum_{LJ,i<j}(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}) + \sum_{Coul.,i<j} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}$$

where A_{ij} and B_{ij} are pairwise parameters that ultimately determine the position and depth of the energy minimum as a function of inter-particle distance (r_{ij}), q is the particle electrostatic change and ϵ_0 is the permittivity of the medium.

1.3.1.1.2 Other Common Biomolecular Forcefields Although AMBER may be one of the most popular forcefields, it may sometimes be better to use an alternative, depending on the particular system and the question that is being asked. For example, OPLS (Optimised Potentials for Liquid Simulation) has been developed as a general, all-atom forcefield that has been parametrised to reproduce structural and thermodynamic properties of liquids, such as hydration free energies.[120] CHARMM (Chemistry at Harvard Molecular Mechanics) is another popular forcefield and has parameters for many biomacromolecules, including proteins,[121] DNA and lipids. It takes a modular approach in that it is parametrised based on small fragments and assumes these can be pieced together to accurately represent whole residues. It is also unique in that it contains a so-called Urey-Bradley term in the forcefield, which is a harmonic potential describing a pseudo-bond between particles with

a 1-3 relationship, which helps account for angle bending. Finally, GROMOS (Groningen Molecular Simulation) is another popular forcefield.[122] Like OPLS, it aims to reproduce the thermodynamic properties of liquid systems. However, it takes a different approach in that it uses a united-atom model. This is a form of course-graining, in which a heavy atom and all of its covalently bound hydrogen atoms are described as a single particle. This has the advantage that many fewer calculations have to be performed due to fewer overall particles in the system, but may have an accuracy cost in assuming that the whole fragment can be described by a single particle.

1.3.1.2 Energetically optimised structures: Energy minimisation

It is often useful to know the lowest energy conformation(s) of a molecule/system in order to determine its native structure and/or prepare it for subsequent simulations or calculations, such as molecular docking, molecular dynamics, or prediction of chemical/physical properties based on the structure. For this reason, energy minimisation algorithms can be applied to find the configuration of the system that gives the lowest potential energy in the MM forcefield.

In an ideal world one would generate every possible configuration of the system in question, then evaluate the energy of each configuration with the MM forcefield. Out of the entire set, the configuration with the lowest energy could then be chosen. This lowest energy configuration would be known as the global energy minimum.

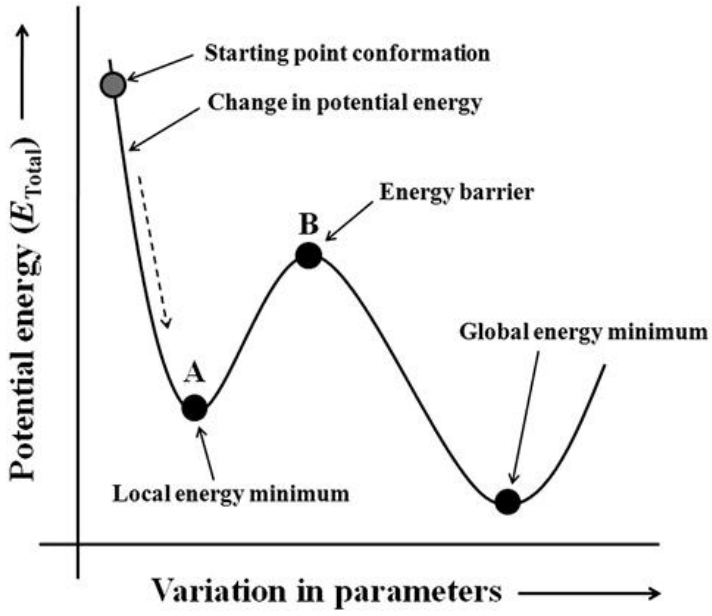

Figure 1.35: Diagram showing the presence of local and global minima on the potential energy surface. Given the shown starting point, energy minimisation would only find the local minimum. The remaining potential energy surface must be explored using techniques such as molecular dynamics to overcome the energy barrier. Source: https://basicmedicalkey.com/computational-chemistry/

The practical problem with this approach is that most model systems contain very many particles and the exact position of each particle in three-dimensional space has an effect on the total potential energy - often described as a hyperdimensional potential energy surface (PES) that is a function of every position coordinate in the system (3n, where n is the number of particles). Therefore it would be computationally impossible to generate configurations that cover the whole PES for most model systems. Instead practically energy minimisation must be performed by starting with an initial configuration and modifying the atomic coordinates slightly until it is not possible to reduce the energy of the system any further. This is normally referred to as the local minimum, since, given the complexity of the PES, it is unlikely that energy minimisation will find the true minimum (Fig. 1.35).

1.3.1.2.1 Steepest Descent Minimisation The steepest descent method is based on the assumption that, from an initial configuration, the most efficient way to reach the local energy minimum is to modify the atomic coordinates in the direction that gives rise to the most negative gradient of the PES. That is, the minimisation will proceed in the direction of the steepest gradient

at each step; the new steepest gradient is then evaluated before travelling in that direction. This is repeated until the local minimum is found (in reality until the gradient changes by less than a given threshold).

$$\boldsymbol{x_{n+1}} = \boldsymbol{x_n} - k_n \nabla f(\boldsymbol{x_n}) \quad (50)$$

where $\boldsymbol{x}$ is the vector that describes the configuration of the system and k is the step size needed to reach the minimum in the given search direction.

While this is perhaps one of the simplest gradient minimisation techniques and each iteration is very fast, overall it is somewhat inefficient (more iterations needed in total) since the new search direction will often be orthogonal to the previous search direction. Therefore, the minimisation will tend to "zig-zag" towards the minimum (Fig. 1.36).

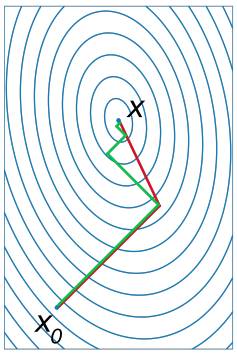

Figure 1.36: Diagram showing the path of the steepest descent algorithm (green) vs. the conjugate gradient algorithm (red) on a two-dimensional potential energy surface (gradient indicated by blue contours). By mixing the direction of the current steepest gradient with the previous steepest gradient, the conjugate gradient algorithm can find the energy minimum (x) in fewer steps.

1.3.1.2.2 Conjugate Gradient Minimisation Perhaps one of the most popular gradient minimisation techniques in MM is the conjugate gradient method. It is based upon the steepest descent method but aims to solve the "zig-zagging" problem that makes steepest descent inefficient.

The first iteration of the conjugate gradient method is identical to the steepest descent method in that the PES is minimised in the direction of the steepest gradient. However, in subsequent iterations the direc-

tion of travel is determined by mixing the previous direction with the direction of the new steepest gradient. By doing so, the minimisation is more likely to travel in a direction towards the minimum, instead of in orthogonal zig-zags (Fig. 1.36).

$$\boldsymbol{x_{n+1}} = \boldsymbol{x_n} - k - n\boldsymbol{h_n} \quad (51)$$

where $\boldsymbol{h_n} = \nabla f(\boldsymbol{x_n}) + \gamma_n \boldsymbol{h_{n-1}}$ and γ is a factor that determines how much of the previous step should be mixed with the current step.

1.3.1.3 Modelling the structures of Protein-Ligand complexes: Molecular Docking

Molecular docking is a computational technique used to generate models of molecular complexes by first searching for a set of possible configurations and then scoring those candidates in some way. There are many published methods for performing the search and scoring, and some will be discussed below. Perhaps the most common use of molecular docking is in the generation of protein-small molecule complexes[123,124] since: (1) it provides a relatively cheap method for finding initial hits and optimising leads in the field of structure based drug discovery; (2) small molecules are computationally the simplest molecules to dock since there are fewer atoms to alter the configuration of during the search and to evaluate in the scoring phase. However, molecular docking can be applied to larger systems, such as modelling protein-protein interactions.[125]

1.3.1.3.1 Search algorithms To obtain a complete set of all possible complex configurations, all possible translations, rotations and conformations of both the receptor and the ligand would need to be searched. Practically this is currently far too demanding computationally and so only a subset of possible complex configurations are generated in reality. This is normally achieved in a number of ways. Firstly, the actual search space can be limited by defining boundaries around a specific region of the receptor in which the ligand can be positioned. Of course, this requires knowledge of which region of the receptor the ligand may bind to, but this is usually accessible from previous structural studies, such as a crystal structure of a lead com-

pound in complex with the receptor.

Secondly, the conformational flexibility of both the ligand and the receptor can be limited. The simplest form of search is known as rigid ligand docking, in which both the ligand and receptor cannot change conformation and only the translation and rotation of the ligand relative to the receptor can be searched. This allows the search to proceed rather rapidly, although may lead to unrealistic or high-energy configurations of the complex since neither the ligand nor the receptor can adapt to the presence of one another. Typically such a search would only be acceptable for high-throughput screening, in which potential binders must be identified from a pool of 10,000s of molecules, and more accurate docking would be performed on those molecules that pass the initial screen.

Perhaps the most popular form of docking is flexible ligand docking, in which the conformation of the ligand is also searched as well as its rotation and translation. The receptor is still considered to be rigid. This of course assumes that the receptor is already in the correct conformation to accept the ligand, but this is usually approximately true since one would be starting from a receptor model that is bound to a related ligand. Furthermore, for many receptor-ligand interactions, there is no significant conformational rearrangement of the receptor and so the lock-and-key approximation is still acceptable.

Finally, it is possible to model the receptor flexibly also. Typically, due to the computational demand, one would only alter the conformation of a small set of receptor atoms. For example, in protein-ligand docking one would alter the conformation of only the protein sidechains within the search space, known as induced fit docking. This method is of course more accurate since it allows the receptor binding site to adapt to form a better shape complementarity to the specific ligand in question, but comes at a significant cost computationally. Furthermore, it still does not address the issue of significant rearrangement of the receptor upon ligand binding.

In order to generate the conformations described above, there are essentially two major options - a systematic or stochastic approach.[123] In the systematic approach, starting from an initial configuration, the various search parameters are varied in a systematic manner. After

each iteration its score can be evaluated and further changes are made until the scoring function reaches an energetic minimum. For the stochastic search, many random configurations are generated and then evaluated for their score. A popular variation of the stochastic method is the genetic algorithm, in which randomised search parameters are encoded in so-called chromosomes.[123] Well-scoring chromosomes are mixed with other well-scoring chromosomes and the process is repeated for a given number of generations.

In general, systematic methods provide an efficient way of finding the nearest minimum energy solution. However, since the result is dependent on the initial configuration, it is likely that only a local minimum and not the global minimum will be found. However, this can be solved by starting the search with multiple starting configurations. Stochastic algorithms are able to cover a much wider part of the search space and so are more likely to find the global minimum. However, as a result the computational cost is high and subsequent iterations may not yield better results.

1.3.1.3.2 Scoring functions Scoring functions aim to estimate the free energy of binding (ΔG_{bind}) for the receptor-ligand complex by considering a number of parameters. They can be largely be split into two families: force field based and empirical. Force field based scoring functions use a molecular mechanics forcefield to evaluate the binding free energy, and so rely on discrete physical properties such as torsion angles and electrostatic interactions. Conversely, empirical functions define the score based on empirical parameters such as hydrogen bonding, hydrophobic interactions and entropic effects. These parameters usually first optimised using a training dataset of known compounds. The free energy of binding for an empirical scoring function is the sum of all the free energies for each parameter (ΔG_x):

$$\Delta G_x = C_x \sum f(x) \quad (52)$$

where C_x is a constant that is fitted during training of the scoring function, and $f(x)$ gives a score between 1 and 0 depending on the distance of that parameter from ideal conditions. An example is that the Glide[126] scoring function for hydrogen bond angles is 1 for angles within 30° of 180° but tends to zero outside this range.

Typically force field based functions may be more generalisable, but empirical functions can be very well tuned for their specific purpose and so can score very well in those cases. However one particular problem of molecular docking is the treatment of solvation potentials. The process of desolvation and resolvation of the receptor and ligand to form the complex plays a large role in determining whether binding is a favourable process or not. Although many functions attempt to model this implicitly, their performance is mediocre at best.[127] More accurate solvation effects could be calculated by considering solvation with explicit solvent atoms, although this is very computationally demanding and not commonly performed.[128]

1.3.1.3.3 The Glide docking protocol The Grid-based Ligand Docking with Energetics (Glide)[126] docking approach is implemented by the Schrodinger Maestro molecular modelling suite and is often cited to outperform many other well-known docking approaches. It uses a semi-empirical approach using both force-field and empirical terms in its scoring function, a grid-based potential for efficiently calculating non-bonded term (similar in concept to that described for the particle mesh Ewald), and filter-based approach in which a large pool of initial conformations is gradually filtered down through several phases of increasingly accurate docking calculations. The protocol is described in more detail below (Fig. 1.37):

1. The grid is first calculated around the region of the receptor determined to be the binding site. This grid describes the shape and non-bonded properties of the receptor using a series of increasingly accurate fields that can interact with the ligand. Since the grid is precomputed, the receptor contribution to the interaction need not be recalculated at each step, increasing efficiency.

2. A conformational search of the ligand is then performed using the OPLS3 forcefield to produce many initial low-energy conformations of the ligand. This step is skipped if rigid ligand docking is used.

3. The translation and rotation of each of these conformations is then systematically searched over the whole volume of the grid and roughly scored using a discrete version of the scoring function in which the contribution of different ligand types to the

score is precalculated with a resolution of 1 $Å^3$ grid boxes. Low scoring poses are then eliminated from the search.

4. The surviving poses are then minimised using the OPLS3 force-field, including a local Monte Carlo search of ligand torsion angles which helps to optimise peripheral groups and relieve internal strain.

5. Finally the full scoring function is applied by calculating the contribution of each ligand individually.

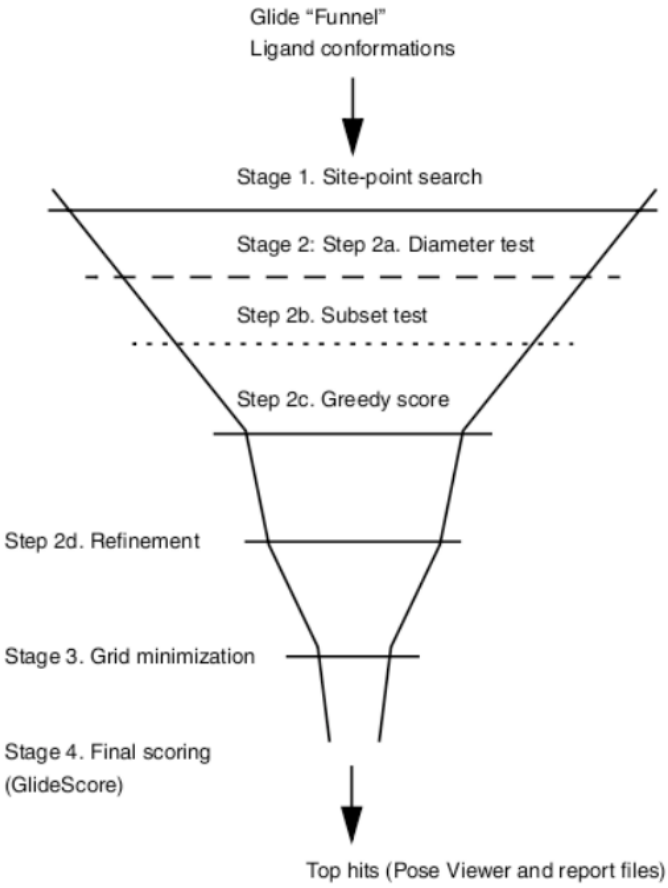

Figure 1.37: Diagram showing the protocol followed by the Glide docking method.

The exact form of the glide scoring function is complex so will not be discussed here. However, it is an empirically based function with terms for hydrophobic interactions, hydrogen bonding, metal ion coordination, as well as force-field based Coulombic and LJ terms. Furthermore, the scoring function takes into account solvation by explicitly docking water molecules into the binding site, which is made practical by the grid-based evaluation of potential. This is a significant improve-

ment on other docking techniques, although the limited sampling of water poses means accuracy is still somewhat limited in that regard.

1.3.1.4 Understanding molecular flexibility and dynamics stability: Molecular Dynamics Simulations

Molecular dynamics is the application of time evolution to a molecular mechanics system using the classical equations of motion.

$$\boldsymbol{f_i} = m\ddot{\boldsymbol{x}}_{\boldsymbol{i}} \tag{53}$$

$$\boldsymbol{f_i} = -\frac{\partial E_{pot}}{\partial \boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{N}}} \tag{54}$$

where $\boldsymbol{f_i}$ is the force acting on a particular particle and $\boldsymbol{x_i}$ is the displacement of that particle. In other words, MD consists of simulating the system at a given finite temperature, in contrast to the energy minimisation calculations. That is, the total energy of the molecule will not only contain the potential energy term, but also the kinetic energy term. Through this the dynamic properties of the system may be observed, such as changes in structural conformation. Bulk physical properties can also be calculated, since given that the simulation is given enough time to thoroughly sample the PES, the time-weighted distribution of states will be equal to the average of the whole ensemble, that is the bulk observable properties. This is known as the Ergodic hypothesis.[129]

Since computers are inherently discrete and computation of the force and the displacement are inter-dependent, it is necessary to calculate them in a stepwise fashion. Commonly the so-called velocity Verlet is used, in which the positions of the particles are updated every time step (τ), and the velocities of those particles on every half timestep.

$$\boldsymbol{x_i}(t+\tau) = \boldsymbol{r_i}(t) + \dot{\boldsymbol{x}}_{\boldsymbol{i}}(t+\frac{\tau}{2})\tau \tag{55}$$

$$\dot{\boldsymbol{x}}_{\boldsymbol{i}}(t+\frac{\tau}{2}) = \dot{\boldsymbol{x}}_{\boldsymbol{i}}(t-\frac{\tau}{2}) + \frac{\boldsymbol{f}_i(t)}{m}\tau \tag{56}$$

Of course calculation of the force can be very costly computation-

ally and ideally one would want to increase the timestep as much as possible to increase performance. However, this has a number of problems: (1) if the time step is close to the time of the observed dynamic property that property will be calculated inaccurately, (2) larger timesteps mean that the particles will move a larger distance between each timestep. At best this will decrease the accuracy of the simulation, since the large movements will break its ergodicity. At worst, it will cause the simulation to crash as the coordinates of the particles become infinitely large. This is a particular problem for hydrogen atoms, since the forces applied to them compared to their mass tends to be very large and so they experience large accelerations. In biomolecular simulations, it has become commonplace to apply constraints to these hydrogen atoms, such as the SHAKE[130] and RATTLE[131] algorithms, which essentially fix the bond lengths of bonds involving hydrogen atoms. In practice a timestep in the order of several femtoseconds (1-2 fs) is reasonable for atomistic simulations of biomolecules employing a hydrogen restraint algorithm.

1.3.1.4.1 Statistical Ensembles Statistical ensembles represent the set of microstates a system can exist in given that a certain set of parameters remain constant. For example, if an isolated system is considered, that is no particles or energy can flow in or out of the system, the system is said to belong to the microcanonical ensemble (NVE) in which the number of particles (N), volume (V) and total energy (E) remain constant. All possible microstates of the system must contain the same total energy, and so any changes to the structure of the system that alter its potential energy must be accounted for by an equal and opposite change in kinetic energy. Since it is not possible for the potential energy of the system to exceed the total energy, which is fixed, it becomes difficult for the system to properly explore the PES, since if any energy barrier exceeds the total energy of the system, the system will become trapped in the same local space.

Alternative ensembles include the canonical ensemble (NVT) or the isobaric-isothermal (NPT) ensemble, which hold volume and temperature or pressure and temperature constant respectively. The NVT ensemble represents a closed system in which heat can be transferred from an external source. However, for production molecular dynamics simulations the NPT ensemble is typically used since it most closely represents the 'real world' scenario of an open flask in which the atmo-

spheric pressure is constant, but the system can expand or contract in volume depending on its kinetic energy. Nevertheless, in both cases, the total energy of the system is not fixed, meaning that large potential energy barriers can be overcome by spontaneous fluctuations in total energy.

1.3.1.4.2 Periodic Boundary Conditions Given the limitations in computational power, it is usual to simulate very small model systems. For example, for studying the dynamics of a protein one may simulate a single protein molecule solvated in a minimal amount of water. This has the consequence that many of the molecules in the system are close to the surface between liquid and vacuum, and may behave differently to what one would expect in reality (i.e. an infinite continuum of liquid). To try to overcome this effect, while still remaining computationally viable, it is usual to employ periodic boundary conditions (PBCs) during the simulation. In such cases, the simulation box is surrounded by identical 'images' of the simulation box - that is the images have exactly the same configuration of particles as the simulation box and when the atomic positions of the particles within the simulation box change, the positions of the particles within the images change in the same manner. However, the image particles are still used to apply forces to the particles in the simulation box, making it appear as if it is within an infinite liquid. Importantly, if a particle within the simulation box travels across the boundary from the simulation box to the image, the equivalent incoming image particle on the opposite face of the box is treated as the new simulation particle (Fig. 1.38).

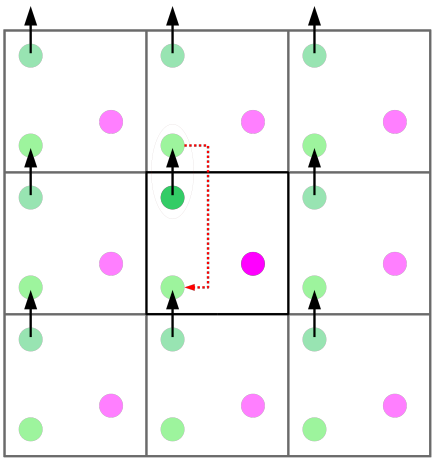

Figure 1.38: Representation of periodic boundary conditions with the simulation box (black border) surrounded by image boxes. Particles that leave the simulation box are replaced by the incoming image particle (dashed red line). Used with permission from https://commons.wikimedia.org/wiki/File:Limiteperiodicite.svg

1.3.1.4.3 Aqueous Solvent Models In reality molecules do not exist isolation, but surrounded by a vast continuum of solvent molecules. The solvent has a profound impact on the properties of the system and so the solvent properties need to be accurately represented. In the context of biomolecular simulation, only water will be considered.

For example, the dielectric screening by a solvent strongly affects the strength of electrostatic interactions, the solvent polarity and any hydrogen bonding with the solute can affect the solute structure, and the viscosity of the solvent affects the diffusion of solute. For example, calculating the desolvation energy of a small molecule binding to a protein is important in calculating the binding affinity of that interaction and so the solvent model must accurately represent that.

1.3.1.4.3.1 Implicit Solvation The simplest solvation model considers the solvent, not to be explicit atomic coordinates, but just a continuum interaction term that models the average bulk properties of the solvent.

A popular model is the Generalised Born (GB) approximation, which models the electrostatic screening by adding an electrostatic interaction between each solute atom and the solvent, where the distance

parameter is replaced with a parametrised value known as the Born radius (f_{GB}) that simulates the whether or not a solute atom is solvent exposed or buried within the solute.

$$(57)\quad E_{GB} = -\frac{1}{8\pi\epsilon_0}(1-\frac{1}{\epsilon})\sum_{i,j}^{N}\frac{q_i q_j}{f_{GB}}$$

where ϵ_0 is the permittivity of free space and ϵ is the permittivity of the solvent.

Since there are no explicit solvent atoms, this method is typically relatively fast. However, clearly this model misses a lot of properties that are important to accurately simulating the solute-solvent interaction. As a result, it has been reported that simulations utilising implicit solvents can produce innaccurate ensembles. For example, in protein simulations, implicit solvents tend to over-stabilise the α-helical conformation.[132]

1.3.1.4.3.2 Explicit Solvation Explicit solvation simulates the solvent by including solvent molecules in the model system thereby allowing the solute to explicitly interact with the solvent. Explicit solvent simulations are typically more accurate than implicit simulations (given correct parametrisation), since effects such as viscosity and hydrogen bonding emerge from the explicit interaction.

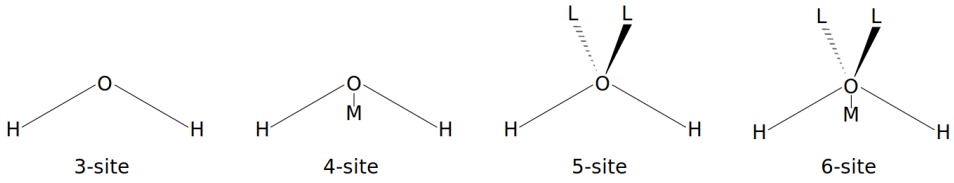

Figure 1.39: Cartoon of common explicit water models, all of which at least include the three atomic coordinates of the oxygen and hydrogen atoms. The 4-site and 6-site models also use a single dummy atom (M) to improve the polarity of the model, whilst the 5-site and 6-site model include additional coordinates for the oxygen lone pair electrons (L). Used with permission from https://en.wikipedia.org/wiki/Water_model#/media/File:Water_models.svg

The obvious choice for an explicit water model involves a 3-site model, with each site corresponding to one of the three nuclei in the water molecule. The SPC and TIP3P models are popular solvation models that utilise the 3-site approach. Both enforce a rigid geometry of

the water molecule with charges on all three atoms. As well as slight differences in the parametrisation of charges and bond lengths, the main difference between the two models is that SPC enforces a tetrahedral bond angle (109°), whereas TIP3P chooses the value actually observed for water (105°). Of the two, TIP3P has been shown to better reproduce many thermodynamic properties,[133] and has been used as the solvent in the parametrisation of a number forcefields, including AMBER[118] and CHARMM.[121]

One major drawback of both TIP3P and SPC are that they significantly underestimate the viscosity of water, and as such diffusion based processes occur more rapidly than in reality. This is largely to do with the dipole moment of the water molecule being inaccurately modelled. To address this issue, the TIP4P model is a 4-site model, in which a fourth dummy atom is placed behind the oxygen atom.[134,135] This dummy atom is massless and does not interact with other atoms except that it carries the electrostatic charge that should be assigned to the oxygen atom. In doing so the polarity of the water model is closer to the true polarity of water and its diffusion properties are improved.

Further, more complex models also exist that take into consideration the lone pair electrons on the oxygen atom (TIP5P, TIP6P).[134] As with TIP4P, these models further increase the accuracy of the water model. However, these come at a computational cost, since including additional coordinates for each solvent molecule can lead to many thousands of additional calculations, since the solvent atoms typical make up the majority of the system. Furthermore, as mentioned previously, many biomolecular forcefields are parametrised based on the TIP3P model. Therefore, although more complex models do have definite benefits, TIP3P is still often favoured for computational efficiency and to ensure the forcefield behaves as expected.

1.3.1.4.4 Dealing with Long Range Interactions The calculation of non-bonded interactions in incredibly costly since every particle in the system interacts with every other particle in the system. Therefore, one would have to calculate N^2 interactions (where N is the number of particles in the system), which could easily be millions of calculations per time step. To reduce the computational burden, we can take advantage of the fact that, although both the LJ and Coulom-

bic interactions continue to infinity, after a certain distance their effect is negligible relative to the inherent error of the simulation, and so they can be ignored. Therefore, a cutoff is usually employed to ignore all long-range interactions after a certain distance. For example, in the AmberTools integrator PMEMD, the cutoff by default is 8 Å for LJ-interactions. However having a hard cutoff in which interactions are immediately ignored after a certain distance, can lead to its own artefacts since the interaction energy either side of the threshold is significantly different. Therefore other cutoff schemes utilise a switching function to taper away the interaction to zero after the cutoff.

While the above method works well for LJ-interactions, electrostatic interactions tend to persist over a longer distance and so need to be treated with a different method. The Particle Mesh Ewald (PME)[136] is most commonly employed for efficiently calculating electrostatic charges in molecular dynamics simulations. Using this method, the electrostatic potential is split into short- and long-range terms. The short range term tends to zero rather rapidly and so can be truncated in the same manner described for LJ-interactions. For the long range interactions, the potential is calculated by representing the electrostatic charge as a grid of electrostatic density values. The force applied to each particle by the long-range electrostatic potential can then be calculated based on its position in the grid.

1.3.1.4.5 Typical Simulation Protocol Assuming one has a suitable model of the system of interest, which may be derived experimentally (X-ray crystallography, NMR) or by modelling (homology modelling, molecular docking), molecular dynamics simulations may be prepared using the following steps:

1. Generation of Model Topology

- The first step in preparing for molecular dynamics simulations involves defining the forcefield parameters for each particle in the system. Such parameters are typically printed in a topology file, that also tends to describe features such as connectivity, and may be prepared manually or more typically by software that recognises the format of the input model. For example, the AmberTools software suite has tools that can automatically produce topology files for proteins, carbohydrates, lipids from

Protein Data Bank (PDB) files.

2. Solvation

- Solvent molecules (usually water for biomolecular simulations) are then added to the system to solvate the model. Typically one would add as few as possible to reduce computation time, but enough so that particles cannot directly interact with their own images. This would depend on the cutoff for long range interactions, and the rule of thumb is to solvate the solute sufficiently that it is at least twice the distance of the cutoff from its image. The shape of the simulation box can also be defined at this point. A cubic box is conceptually the simplest to compute but more recently more complex boxes, such as the truncated octahedron, have been developed which allow fewer total solvent molecules to be used.

3. Neutralisation

- Typically one would want to to ensure that the total electronic charge of the system is zero, since this represents how the system would exist in reality and charged systems could lead to very high potential energies that could destabilise the system. For this reason, counter ions can be placed within the simulation box to neutralise any electronic charge. Most forcefields have parameters for many atomic ions, although for most biomolecular simulations one would usually use either sodium or chlorine ions.

4. Minimisation

- Although the initial model may have been minimised prior to molecular dynamics, the additional of solvent molecules and ions will often put the system into a high energy state due to steric clashes or non-optimal non-bonded interactions. Trying to begin the simulation without first finding the local minimum will often lead to unstable simulations. Usually the minimisation will be spilt into two stages: (1) the solute molecules will be restrained so that only the solvent molecules, which are most likely the furthest from being optimised, can be moved during

minimisation. (2) The minimisation is then repeated with no restraints. This procedure prevents high energy interactions with the solvent from distorting the conformation of the solute atoms significantly.

5. Heating

- Heating is simulated by gradually adding kinetic energy to the system (which initially has no temperature) over a period of time (usually hundreds of picoseconds) until the system reaches the desired simulation temperature. This must be performed gradually since a sudden large increase in kinetic energy will make the simulation highly unstable. Heating is performed using the NVT ensemble, since energy cannot be added to the system using the NVE ensemble, and in the NPT ensemble the increasing kinetic energy would cause the system to continuously expand to maintain the pressure.

6. Equilibration

- Since in production one would usually use the NPT ensemble, a short period of time is required after heating to switch to the NPT ensemble and ensure that the system is at thermodynamic equilibrium, in particular to equilibrate the density of the solvent. The equilibration stage is performed under the same conditions as the production dynamics, except that the trajectory is typically not saved or analysed.

7. Production

- The conditions used during equilibration are continued and now the coordinates are saved at given intervals to create a molecular dynamics trajectory. Production dynamics are typically the longest stage, ranging from 1-1000 ns, depending on the information required and computing power available.

1.3.1.4.6 Enhanced Sampling Methods While molecular dynamics is a useful technique for understanding the dynamic properties of a system, its use is usually limited to transitions that occur on a very short timescale due to the computational restraints on simulation

time. Since overcoming the transition barrier between two states is a stochastic process and the probability of a transition is governed by the energy needed to overcome this barrier, a number of techniques have been derived to increase the probability of such a transition and therefore enabled the simulation to sample more of the PES in the same amount of simulation time. These are known as enhanced sampling methods.

1.3.1.4.6.1 Accelerated Molecular Dynamics Accelerated Molecular Dynamics (aMD) is an enhanced sampling technique that increases the probability of overcoming energy barriers, thereby accelerating the simulation, by applying a boost potential (ΔE_{pot}) that effectively increases the potential energy of the system (E^*_{pot}) and reduces the energy barrier.[137]

$$E^*_{pot} = E_{pot} + \Delta E_{pot} \quad (58)$$

The magnitude of the boost is determined by the difference between the true potential and a predetermined threshold (E_{thres}) and an acceleration factor (α). The energy threshold determines the maximum energy that can have a boost potential applied to it. Exactly how to choose this threshold is discussed elsewhere,[138] but is normally based from finding the average potential energy of the system over the course of a short 'traditional' molecular dynamics trajectory. The acceleration factor has the effect of flattening the PES depending on its value, with small values flattening it more. The flatter the PES, the faster the PES can be sampled, but this comes at a cost in accurately knowing the shape of the PES. For a moderate boost, a value of $0.2\times$ the number of atoms in the system is typically used.

$$\Delta E_{pot} = \frac{(E_{thres} - E_{pot})^2}{\alpha + (E_{thres} - E_{pot})} \quad (59)$$

where $E_{pot} < E_{thres}$

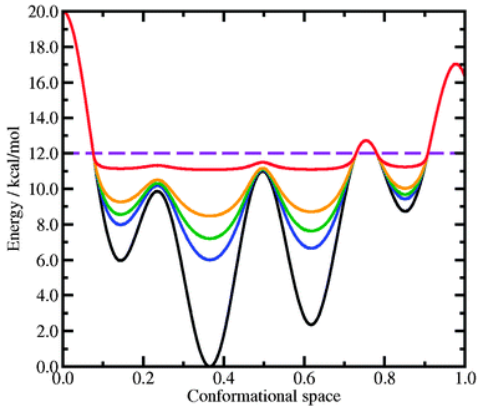

Figure 1.40: The effect of the accelerated molecular dynamics boost potential on the potential energy surface. The true surface is shown (black) compared to increasingly strong boost factors (blue to red). For energies larger than the threshold (dashed purple) no boost is applied. Taken from [139]

While the aMD procedure can be used simply for accelerated conformational sampling, any observable property calculated from the aMD simulation ($A*$) can be reweighted to determine the true value of that observable (A) in real time. For example, the potential energy over some reaction coordinate ($\boldsymbol{r}$) of interest may be reweighted to find the true potential energy.[137]

$$\langle A \rangle = \frac{\langle A * (\boldsymbol{r}) exp(\beta \Delta E_{pot}(\boldsymbol{r})) \rangle}{\langle exp(\beta \Delta E_{pot}(\boldsymbol{r})) \rangle} \quad (60)$$

where $\beta = \frac{1}{k_B T}$. However, it has been demonstrated that such reweighting can be very noisy and as such the calculated observables may not be very accurate, especially for more complex systems such as proteins.[140] More recently a modification to the method has been published, known as Gaussian Accelerated Molecular Dynamics (GaMD), which has been shown to provide a more accurate reweighting by ensuring that the boost potential follows a Gaussian distribution.[141]

The aMD procedure can be applied to either the whole PES, just to the dihedral potential, or using a dual-boost in which boost potentials are applied to both the whole PES and the dihedral potential. The latter has proven particularly effective in exploring the folding of proteins.[138] Unlike some enhanced sampling techniques (e.g. umbrella sampling), aMD does not require any prior knowledge of the system and so can be used when the 'desired transition' is unknown.

1.3.1.4.6.2 Adaptive Steered Molecular Dynamics In steered molecular dynamics (SMD) an external force is applied to the system in order to steer the system along a particular reaction coordinate at a given velocity. This is particularly useful for observing transitions that may occur over time periods much longer than is accessible by conventional molecular dynamics. For example, the technique is commonly used for ligand unbinding experiments,[142] and can be correlated with experimental atomic force microscopy data.[143]

Since the force required to move the system along the reaction coordinate at a given rate is known, the work done on the system can be calculated. Of course, the work calculated for a particular trajectory will be dependent on the initial microstate of the system and the resulting work calculated for multiple trajectories will be different for each trajectory. In such a non-equilibrium system (i.e. the reaction coordinate changes at a finite rate) it has been shown that the ensemble average work ($\langle W \rangle$) over many independent simulations can be related to the equilibrium Helmholtz free energy change (ΔF) between the start and end states, in an equation known as the Jarzynski equality:[144]

$$\langle exp(-\beta W) \rangle = exp(-\beta \Delta F) \quad (61)$$

To calculate the equilibrium free energy change between the two states, many independent trajectories must be run over the entire reaction coordinate in order to ensure that the, so-called Jarzynski average work has converged. Obviously this is very computationally expense and requires vast amounts of computing time. More recently adaptive SMD (ASMD) simulations have been developed which aim to reduce this computational demand by splitting the simulation into stages along the reaction coordinate.[145] After each stage the Jarzynski average work is calculated and the trajectory with whose work is closest to this average is used to seed the next stage of simulations. By doing so, the total amount of simulation time can be reduced, thereby making this technique more accessible computationally.

1.3.1.4.7 Applications of Molecular Dynamics in Structural Biology Molecular dynamics is incredibly useful in the field of structural biology because it gives access to atomic structural and dynamic

detail not accessible to many experiments. It is particularly complementary to X-ray crystallography, which can produce a high-resolution three dimensional model of the structure of interest, but lacks solution state, dynamic information, and so cannot solely be used to understand the function of a molecule.

For example, molecular dynamics can be used to study conformational changes in a protein and the factors that may facilitate that. For example, we have recently demonstrated the pH dependence on the conformation of the binding site of a serine-rich repeat protein (SRRP) from *L. reuteri* by MD,[146] which has be experimentally observed to bind to different ligands at different pHs. We showed that protonation of a key structural sidechain at low pH caused conformational rearrangement of a binding site loop. This was not possible by other means, since the protein could not be crystallised under these conditions. Furthermore, simulations have been used to explain allosteric effects, such as long range conformational changes in response to ligand binding.[147,148]

Molecular dynamics has also found an important role in drug discovery, since other *in silico* techniques such as docking miss important solvation and entropic factors. MD has been successful in discriminating binders from non-binders to a greater degree of accuracy than docking, thereby accelerating the drug discovery process.[149,150]

As computational power grows further, the power of MD simulations will continue to increase, giving access to larger systems over longer timespans. Already microsecond length simulations are accessible,[151] but with many macromolecular processes taking place over the millisecond to second timescales, MD is still limited in that regard. For example, *ab initio* protein folding is a key goal for structural biologists, but as of yet only simulations of small peptide folding have been accessible.[152] Furthermore, large scale simulations of systems such as the HIV capsid have already been successful.[153] In the future, it may be possible to simulate entire cellular processes, making MD an indispensable technique in the study of biological systems.

1.4 General Objectives of the Thesis

Throughout this thesis the techniques of NMR Spectroscopy and Molecular Modelling will be used heavily to understand the structural basis for the molecular recognition of ligands by biologically relevant carbohydrate active proteins. Within the thesis, it is shown how these techniques are highly complementary, to use molecular modelling to provide atomic level detail that explains the experimental observations made by NMR spectroscopy. In particular, we wanted to investigate the power of combining STD NMR spectroscopy with molecular docking, using the CORCEMA-ST software to quantifiably measure the agreement between the experimental and theoretical data. Furthermore, the powerful computing resources available to our group enable the routine use of long molecular dynamics simulations (1 μs) to study large scale protein dynamics.

Overall the objectives are:

- **Differential Interaction of Hyaluronan with the Cell Surface Receptors, CD44 and LYVE-1**
  - Generate a homology model of the LYVE-1 hyaluronan binding domain and experimentally demonstrate its validity.
  - Compare and contrast the binding epitopes of a synthetic hyaluronan tetrasaccharide with CD44 and LYVE-1, using molecular modelling, including molecular dynamics simulations, to explain the similarities and differences in binding.
- **Understanding Ligand Recognition by *Ps*LBP**
  - Validate X-ray crystallography models of the complexes of *Ps*LBP with glucose-1-phosphate and mannose-1-phosphate in the solution state.
  - Understand the the recognition of cognate and non-cognate substrates by *Ps*LBP.
- **Characterisation of the Interaction between the *Salmonella enterica* effector proteins and their Death Domain Substrates**
  - Understand the differences in molecular recognition of target substrates by the effector proteins SseK1 and SseK2
  - Understand the differences in dynamics of the SseK1 and

SseK2 proteins and how this may affect molecular recognition

- Generate a molecular model of the SseK2:FADD complex and experimentally validate the structure

Chapter 2

Differential Interaction of Hyaluronan with the Cell Surface Receptors, CD44 and LYVE-1

2.1 Introduction

2.1.1 CD44

CD44 is a cell-surface, transmembrane receptor involved in cell-cell and cell-matrix interactions.[154–156] Its principal ligand is hyaluronan (HA),[83] an abundant, high-molecular-weight, extracellular matrix glycosaminoglycan (Fig. 2.1) that has a diverse set of functions, including imparting compressibility and lubrication to a tissue,[157] water homeostasis,[158] matrix scaffolding[159] and cell surface interactions. HA is composed of repeating disaccharides of N-acetyl-glucosamine (GlcNAc) and glucuronic acid (GlcA) linked by β(1-4) and β(1-3) glycosidic linkages respectively.[82]

CD44 is expressed ubiquitously - being found in numerous embryonic and adult tissues, as well as a multitude of cell types, including: epithelia,[160] endothelia,[161] smooth muscle cells,[162] fibroblasts and astrocytes. It has been noted that CD44 expression is particularly elevated in regions of active cell growth,[160,163] which agrees with the established role of CD44 plays in cell adhesion and migration.[154–156]

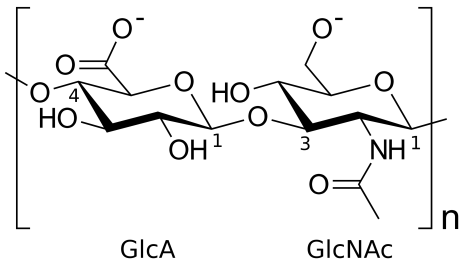

Figure 2.1: Chemical structure of hyaluronan, a high-molecular-weight glycosaminoglycan consisting of repeating disaccharide units of glucuronic acid (GlcA) and N-acetyl-glucosamine (GlcNAc) linked via β(1-3) and β(1-4) glycosidic linkages respectively.

It is then perhaps unsurprising that over-expression of CD44 has been implicated as an important factor in the malignancy of many cancers[156,164,165] and much work has been dedicated to targeting CD44 for the diagnosis and treatment of cancer.[166,167] However, contradictory evidence exists showing that CD44 can both promote and inhibit cancer progression, as reviewed in [168].

This discrepancy may be explained in part by two factors: (1) variable splicing of the CD44 gene, and (2) changes in post-translational modifications to the CD44 protein.

2.1.1.1 Variable splicing of CD44

The CD44 gene consists of 19 exons, at least 12 of which can be variably-spliced (Fig. 2.2),[169] leading to a multitude of potential splice variants. The vast majority of the variable exons are found within the membrane-proximal region of the extracellular domain. The remaining variable exons are found in the cytoplasmic tail. The standard isoform of CD44 (CD44s) is defined as containing none of the variably-spliced exons, whereas those isoforms containing one or more variable exons are known as variant CD44 (CD44v).[165]

In the literature, it has been generally observed that CD44v isoforms, which typically contain a longer extracellular domain, promote aggressive tumour growth and metastasis.[170–174] This is in contrast to CD44s which is more abundantly expressed in non-metastatic tumours and down-regulation of CD44s is associated with increased tumourigenicity.[174–176]

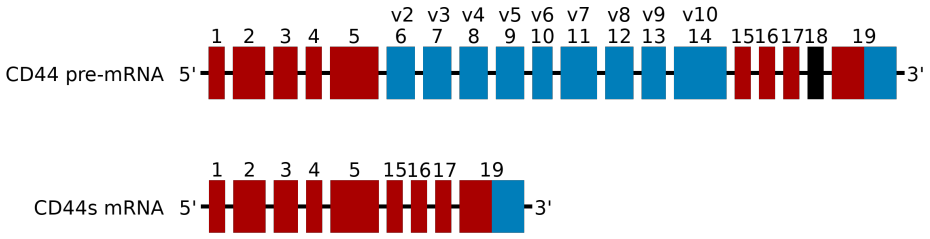

Figure 2.2: Variable splicing of CD44 exons. **Top:** Full CD44 pre-mRNA transcript containing all CD44 exons. Exons 1-5, 15-17 (red) are conserved in all CD44 variants, whereas exons 6-14 (blue) are variably spliced. If exon 18 (black), which contains a premature stop codon, is spliced into the final transcript, the resulting protein will not contain the sequence encoded by exon 19 (red and blue). **Bottom:** mRNA transcript for CD44s, which contains none of the variably spliced exons.

2.1.1.2 Post-translational modification of CD44

Differential post-translational modification of the CD44 protein is also responsible for the variation in its function.[177,178] For example, Ser291 and Ser325 in the cytoplasmic domain of CD44 are known to be phosphorylated by Ca^{2+}/calmodulin-dependent protein kinase II[179] and protein kinase C[180] respectively; the inhibition of either process impairs cell mobility. CD44 is also known to be acylated at Cys286 and Cys295, leading to its association with lipid rafts, which is essential for endocytosis of CD44 and its ligands.[181] Furthermore, CD44 can be cleaved by sheddases,[182,183] a class of protease that cleaves the extracellular domain of transmembrane proteins. Sheddases are often also up-regulated in cancer.[184] Cleavage of the CD44 extracellular domain (into so-called soluble CD44) has been shown to promote cell migration since the cleaved intracellular domain can act as a transcriptional promoter of the CD44 gene.[185,186] However, a separate study has shown that over-expression of soluble CD44 inhibits cell adhesion since the soluble binding domains compete with the membrane-bound binding domains for HA binding.[187]

However, perhaps the most puzzling post-translational modifications made to CD44 are the N- and O-glycosylation of its extracellular domain (Fig. 2.3). It is well known that the covalent modification of proteins with glycans can exert a number of effects,[7] including shielding, regulation and recognition. In the case of CD44, there are numerous confounding reports that both N- and O-glycosylation can

have either an inhibitory or activating effect. For example, multiple studies report that inhibition of either N- or O-glycosylation reduces HA-binding to CD44.[188,189] Conversely, N-glycans terminated by sialic acid residues have been shown to have an inhibitory effect on CD44, whilst treatment with neuraminidase is able to restore receptor function.[190–192] Finally, O-glycosylation of CD44 has also been observed to have an inhibitory effect.[193] Interestingly, most of CD44's predicted O-glycosylation sites are found within the variable exons,[194] suggesting that the degree of O-glycosylation can vary wildly between CD44 isoforms. Conversely, CD44s contains 6 potential N-glycosylation sites, with only an additional 2 being found in some longer isoforms. To further confound the issue, glycosylation is notoriously heterogeneous,[195] with levels of glycosylation and glycan structure varying greatly between cell types and environmental conditions.

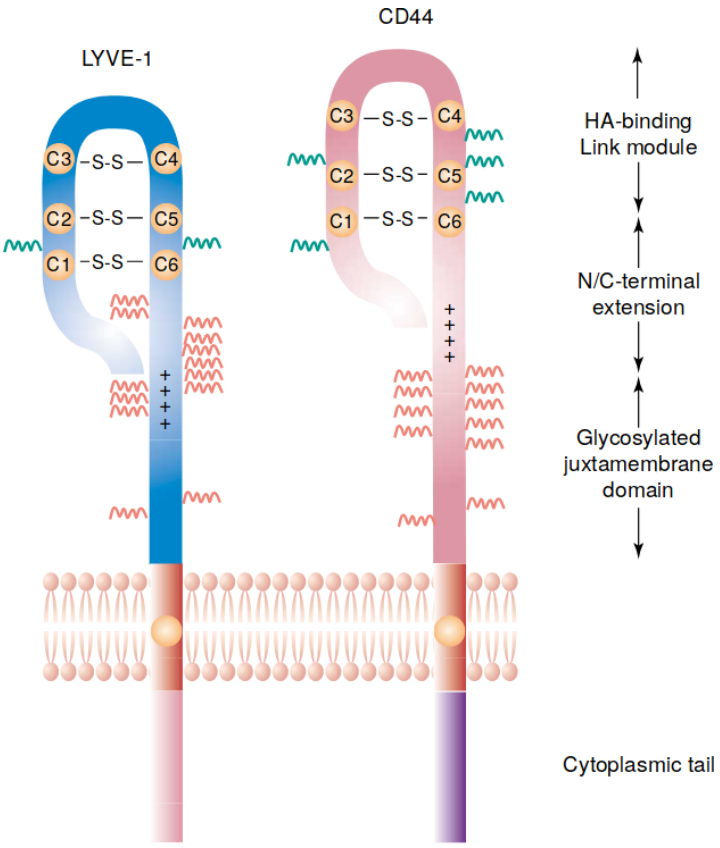

Figure 2.3: Cartoon showing the overall topology of CD44 (right) and LYVE-1 (left) including the location of N- (green zig-zags) and O-glycosylation (red zig-zags) sites and cysteine residues (gold circles). Taken from [196].

In summary, CD44 is an important receptor in the adhesion and migration of cells under healthy and pathological conditions. Its activity is dependent on a number of different factors, including variation in

splicing and variation in post-translational modifications. To further understand its function, CD44 should be studied from a structural point of view.

2.1.1.3 The Structure of CD44

The structure of CD44 can be decomposed into 4 main domains:[169,197] (1) a hyaluronan-binding domain (HABD), (2) a variable membrane-proximal stalk, (3) a transmembrane domain, and (4) a cytoplasmic domain.

2.1.1.3.1 The Hyaluronan-Binding Domain of CD44 The CD44 hyaluronan-binding domain (HABD) contains a consensus amino acid sequence homologous to the so-called Link domain[197] - a structurally conserved domain of approximately 100 residues containing a HA-binding site. The Link domain is found in a number of HA-binding proteins,[198] including the proteoglycans aggrecan[199] and versican,[200] and TSG-6.[201] Apart from CD44, the only other known receptor containing the Link domain is LYVE-1.[198,202]

The three-dimensional structure of the Link domain was first solved using NMR for the human TSG-6 protein.[203] A X-ray crystal structure of the TSG-6 HABD was solved later (Fig. 2.4 i. and iii.).[72] The domain is a compact globular shape with a well-defined, large hydrophobic core. It consists of 2 α-helices and 6 β-strands; the strands $\beta1$, $\beta2$ and $\beta6$ come together to form an antiparallel β-sheet, whilst a second antiparallel β-sheet is formed by $\beta3$, $\beta4$ and $\beta5$. A hook-like loop is formed between the $\beta4$ and $\beta5$ strands and is stabilised by a disulfide bond between a pair of cysteine residues. A second disulfide is formed between the $\alpha1$ helix and the $\beta6$ strand by another pair of cysteine residues. HA binds laterally along a cleft adjacent to the hook-like loop.[204]

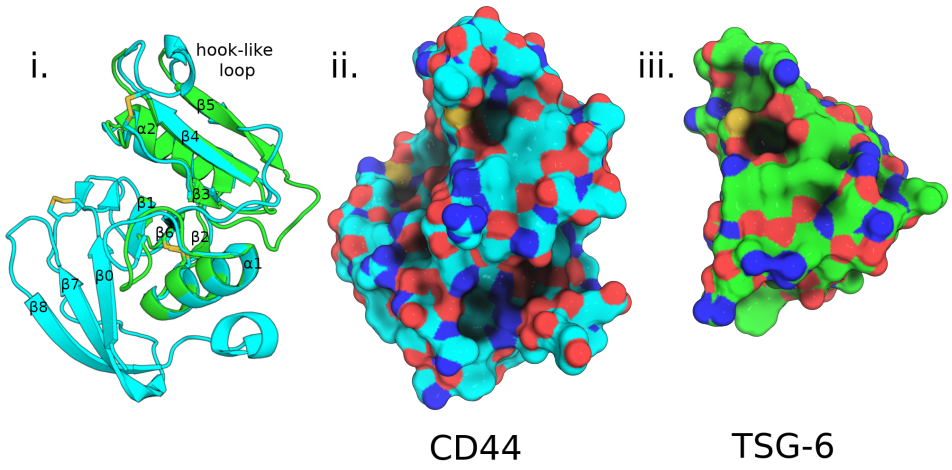

Figure 2.4: Three-dimensional structures of the HABDs of TSG-6 (PDB accession code: 2PF5) and CD44 (PDB accession code: 1UUH). **i:** Cartoon representations of TSG-6 (green) and CD44 (cyan) HABDs. Both contain the consensus Link domain fold, although CD44 HABD requires N- and C-terminal extensions for hyaluronan binding. Disulphide bridges highlighted shown as yellow sticks. **ii & iii:** Surface representations of CD44 (carbons in cyan) and TSG-6 (carbons in green) HABDs respectively.

Across all known Link domains, the primary sequence within secondary structural elements is highly conserved,[205] highlighting the existence of a consensus Link domain fold. The four cysteine residues are also highly conserved, indicating their importance in defining the Link domain fold.

The Link domain of CD44 shares approximately 37% identity with the Link domain of TSG-6.[201] However, the Link domain of CD44 alone is not sufficient to bind HA - elements further downstream are additionally required for HA binding.[197,206–209] The X-ray crystal structure of the human CD44 HABD (Fig. 2.4 i. and ii.)[210] reveals three new β-strands (β0, β7, β8) that extend both the N- and C-terminus of the consensus Link domain and continue the antiparallel β-sheet formed previously by β1, β2 and β6. An additional pair of cysteine residues found in the N- and C-terminal extensions form a third disulfide bond. The C-terminal extension contains a number of basic residues that have been shown to be important for HA binding to CD44.[206–209] However, these residues are far from the HA-binding site of Link domains and the X-ray crystal structure of the murine CD44 HABD bound to a HA octasaccharide[211] suggests that only one of these residues (Arg155) directly contacts HA (Fig. 2.5). A NMR study[212] of the human CD44 HABD shows that, upon binding of HA, the C-terminal extension unfolds, causing the CD44 HABD to transition from an or-

dered (O) to a so-called partially disordered (PD) state (Fig. 2.6). This behaviour is missed by X-ray crystallography since the act of crystallisation selects for low conformational entropy. Molecular dynamics simulations show that the O-to-PD transition frees the basic residues of the C-terminal extension to interact with HA, which enhances binding.[213] Furthermore, CD44 HABD mutants that constitutively adopt the O-state bind HA with lower affinity than the wild-type and the equilibrium between O- and PD-states has been shown to be important in lymphocyte rolling in the high-shear environment of the vascular endothelia.[214] Other studies have tried to explain the interaction of HA with the C-terminal extension through multiple HA binding modes,[210,215] although the evidence for this is not convincing.

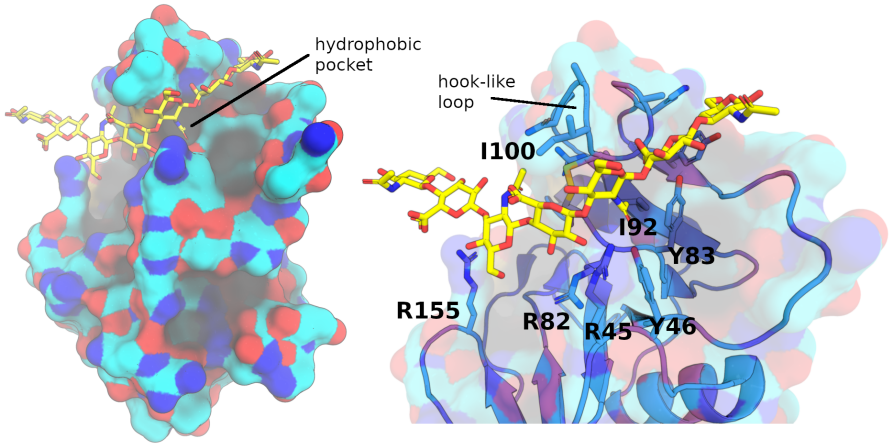

Figure 2.5: The three-dimensional structure of the murine CD44 HABD (cyan) bound to a HA octasaccharide (yellow) (PDB accession code: 2JCR). **Left:** Surface representation of the CD44 HABD, showing the HA-binding groove and deep pocket accommodating the HA GlcNAc methyl group. **Right:** Cartoon representation highlighting residues of interest - R45, Y46, R82, Y83, I92, I100, and R155 in murine CD44 (R41, Y42, R78, Y79, I88, I96, and R150 in human CD44).

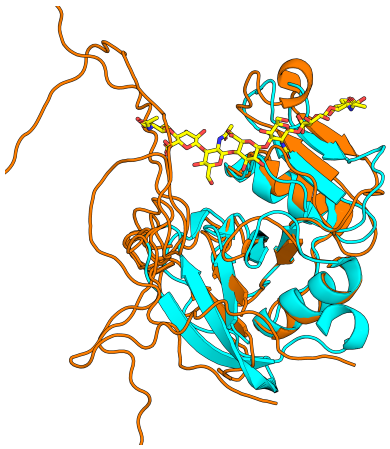

Figure 2.6: The ordered (O) to partially-disordered (PD) transition in the CD44 HABD. The HA (yellow) bound X-ray crystal structure of the CD44 HABD (cyan, PDB accession code: 2JCR) is compared against the PD structure obtained by NMR (orange, PDB accession code: 2l83). For clarity only the lowest energy structure of the NMR-derived ensemble is shown for residues 19-146, and the 5 lowest energy structures are shown for residues 147-178. In several structures the PD C-terminal extension is in close proximity to the HA binding site.

The binding mode of HA observed in the murine CD44 HABD crystal structure[211] (Fig. 2.5), taken together with the understanding that the C-terminal extension undergoes an O-to-PD transition,[212] appears to explain the interaction CD44:HA most successfully. The binding mode of HA positions it along the groove adjacent to the hook-like loop, as observed in TSG-6.[204] The groove is lined by aliphatic, aromatic and basic residues, and the interaction is dominated by a large number of hydrogen bonds. In particular, Arg45 (Arg41 in human CD44) forms several contacts with HA and mutation of this residue completely abolishes HA binding.[207] It has also been suggested that this residue could be involved in affinity switching since it is found in two different conformations in X-ray crystal structures.[211] This hypothesis is also supported by molecular dynamics simulations.[213] The binding-site residues, Tyr42, Arg78 and Tyr79 (human CD44 sequence) have also been shown to be critical for HA-binding.[207] Of these, Tyr79 (Tyr83 in murine CD44) forms one side of a hydrophobic pocket into which the methyl group of a GlcNAc residue of HA is inserted[211] (Fig. 2.5).

2.1.1.3.2 The Membrane-proximal Stalk of CD44 The membrane-proximal stalk of CD44 links the HABD to the transmembrane domain and is expected to adopt an elongated structure.[154] Its

length is highly variable, ranging from only 46 amino acid residues in CD44s, up to 381 amino acid residues in the variant containing all variable exons (CD44v1-10). The stalk contains numerous motifs for post-translational modifications, varying depending on which exons are present, including attachment of O-glycans and proteolytic cleavage as discussed in Section 2.1.1.2. One particularly interesting example is that CD44 variants containing the v3 exon (Fig. 2.2) can be modified with heparan sulfate (HS),[216] which in turn can bind and present HS-binding growth factors - a behaviour which may have implications in the development of cancer.

2.1.1.3.3 The Transmembrane Domain of CD44 The transmembrane domain of CD44 makes a single pass through the cell membrane and consists predominantly of hydrophobic residues, with the exception of a single cysteine residue (Cys286).[154] This cysteine residue, through acylation is important in the formation of lipid rafts, as discussed in Section 2.1.1.2. In addition, this unpaired cysteine is able to induce dimerisation of CD44 molecules through formation of a disulfide bond with the equivalent cysteine of an adjacent CD44 molecule.[217] This dimerisation enhances hyaluronan binding through increased avidity and is important for observing CD44 activity in cells.[218]

2.1.1.3.4 The Cytoplasmic Domain of CD44 The standard form of the cytoplasmic domain of CD44 consists of 72 amino acid residues.[177] This is termed the 'long-tail' form and exists in almost all observed forms of CD44. A 'short-tail' form is possible by inclusion of exon 19, which contains an alternative stop codon, but has only been detected in chondrocytes and has impaired HA-internalisation properties.[219] As discussed in Section 2.1.1.2, the cytoplasmic tail can be modified at several positions, which affects its function. In addition, the cytoplasmic tail can interact with the cell cytoskeleton by binding to a number of adaptor proteins. The CD44 sequence from Arg292-Lys300 constitutes a FERM-binding motif,[177] which facilitates the interaction of CD44 with the actin cytoskeleton mediated by binding of ERM (ezrin/radixin/moesin) proteins.[220] In addition, the sequence from Asn304-Leu318 constitutes an ankyrin-binding site,[221] which also mediates the interaction of CD44 with the actin cytoskeleton. In this way, HA-binding to CD44 can directly affect cell adhesion

and migration by restructuring the actin cytoskeleton.[222,222–225]

2.1.2 LYVE-1

Like CD44, LYVE-1 is a cell-surface, transmembrane receptor for hyaluronan (HA). LYVE-1 is 41% homologous to CD44[226] and is the only other known receptor containing the HA-binding Link domain.[198] It was originally believed to be expressed exclusively in the lymphatic endothelia[226] (hence the name LYmphatic Vessel Endothelial receptor 1) but has since been observed in embryonic vascular endothelia;[227] it is subsequently down-regulated during vascular remodelling. Unlike CD44, which is the predominant HA receptor in many cell types,[83] LYVE-1 appears to be predominantly inactive both *in vitro* and *in vivo*.[228] Therefore it has been of great interest to understand the function of LYVE-1, the factors that affect its activation, and how and why these differ from CD44.

2.1.2.1 Function of LYVE-1

While CD44 appears to be active to some degree under many conditions, for a long time the factors required to activate LYVE-1 were elusive. More recently it has been discovered that LYVE-1 activity is highly dependant on both clustering of LYVE-1 on the cell surface and organisation of HA into polyvalent complexes involving HA crosslinking by HA-binding matrix proteins.[229,230] It has since been shown that cells, including Group A *Streptococci*,[231] macrophages[229] and dendrocytes,[232] that are encapsulated in an organised HA-matrix can be taken up into the lymphatics *via* binding to LYVE-1 on the surface of lymphatic vessels (Fig. 2.7).

This also has important implications in cancer progression, since the lymphatic vessels provide a major route for metastasis[233,234] and LYVE-1 may facilitate the uptake of cancer cells into the lymphatics.[196,230,235–237] Furthermore, the up-regulation of CD44 on the surface of many cancer cells[156,164,165] may help to organise the HA-matrix surrounding the cancer cell. Indeed, it has been shown that CD44, HA and LYVE-1 can form stable ternary

complexes.[226] Furthermore, LYVE-1 appears to play an important role in lymphangiogenesis.[238,239]

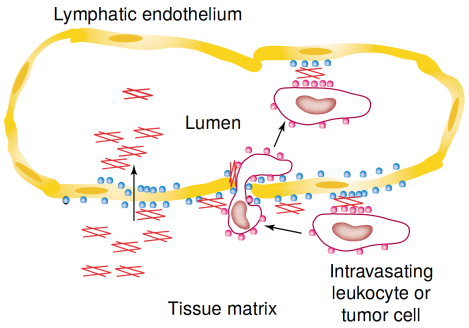

Figure 2.7: Cartoon showing the proposed function of LYVE-1 (blue circles) which is expressed on the surface of lymphatic endothelial cells (yellow). Through binding to hyaluronan (HA, red lines) it can uptake HA into the lymphatic vessels or traffick cells expressing CD44 through formation of a ternary complex. Taken from [196].

2.1.2.2 Structure of LYVE-1

Although LYVE-1 transcripts can be variably spliced, in contrast to CD44 only one is translated into a functional protein containing the Link domain.[226,240] This single functional isoform consists of 322 amino acid residues, shorter even than the 361 residue CD44s, resulting in an extracellular domain that is truncated by 29 residues in comparison.[241] Despite this, the degree of glycosylation on the LYVE-1 extracellular domain is still comparable to CD44s, with 27 potential O-glycosylation sites (32 in CD44s) and 2 N-glycosylation sites (Asn53 and Asn130; 6 sites in CD44s)[194,228] (Fig. 2.3).

Experiments in HEK293T cells have shown that the LYVE-1 N-glycosylation sites are important for effective binding of HA, in particular Asn130.[228] Conversely, the presence of sialylated O-glycans has an inhibitory effect an LYVE-1 binding to HA. It is important to remember however that these experiments represent only one cell type and, as observed for CD44,[188–193] the effect of glycosylation on the function of the receptor can vary drastically depending on the cell type and environmental conditions. However, what is clear is that glycosylation of the LYVE-1 extracellular domain can have a drastic impact on its binding to HA.

2.1.2.2.1 The Hyaluronan-Binding Domain of LYVE-1 The hyaluronan-binding domain (HABD) of LYVE-1 shares 51% homology with the HABD of CD44.[242] Like CD44, the HABD of LYVE-1 consists of a consensus Link module extended at both the N- and C-termini, including the additional disulfide-forming cysteine pair that stabilises the extension.[242] No experimentally derived structure yet exists for the LYVE-1 HABD, but homology modelling suggests that it may share many similar features with the HABD of CD44, including: (1) a HA-binding cleft adjacent to the hook-like loop, (2) a hydrophobic pocket to accept the methyl-group of GlcNAc residues within HA, and (3) interaction of HA with residues far from the HA-binding site.[242]

Perhaps the most distinguishing feature of the LYVE-1 HABD, according to the homology model, is the lack of intra-molecular hydrogen bonds between the LYVE-1 HABD and HA,[242] in direct contrast with CD44.[211] Instead, LYVE-1 appears to mediate its interaction with HA predominantly *via* electrostatics, as indicated by the presence of multiple charged basic residues surrounding the binding site (Arg99, Lys105, Lys108, Lys117, Arg122). This is supported by experimental evidence that the interaction of LYVE-1 with HA is strongly diminished with increasing ionic strength of the buffer.[242] Despite the predominantly electrostatics-driven interaction, LYVE-1 surprisingly conveys a higher specificity for HA even than CD44;[226] LYVE-1 binds solely to HA, whereas CD44 has a minor specificity for chondroitin sulfate as well.[83]

An important factor modulating the affinity of the CD44 HABD for HA is the unfolding of the C-terminal extension (see Section 2.1.1.3.1)[212,214] such that several basic residues within the extension can interact with HA.[213] These residues are not present in the LYVE-1 HABD.[242] Furthermore, binding kinetics experiments show that LYVE-1 HABD monomers can form a binary complex with HA far more rapidly than the CD44 HABD can.[212,214,243,244] Together these suggest that unfolding of the HABD C-terminal extension does not occur in LYVE-1.

2.1.2.2.2 The Membrane-proximal Stalk of LYVE-1 The membrane-proximal stalk of LYVE-1 is severely truncated compared to most CD44 isoforms, yet is still heavily glycosylated (see Section

2.1.2.2.2). Like CD44, LYVE-1 can also be proteolytically cleaved by metalloproteases to form soluble, truncated, membrane-detached, extracellular domains.[245,246] The LYVE-1 cleavage site is found between Phe226 and Glu229 within the membrane-proximal stalk. Cleavage of LYVE-1 extracellular domains has been shown to inhibit lymphangiogenesis.[245,246] This has important consequences in cancer, with high levels of LYVE-1 cleavage associated with lower levels of metastasis and better outcomes for patients.[235,247]

Another interesting feature of the LYVE-1 membrane-proximal stalk is a series of basic residues(RRKK, Arg195-Lys198) preceding a single, unpaired cysteine residue (Cys201).[202,242] The cysteine is redox-labile and is able to form LYVE-1 homodimers under reducing conditions.[243] Furthermore, this dimerisation is essential for for binding to HA *in vivo* and therefore represents a redox-switch that allows LYVE-1 to respond to environmental conditions. The function of the basic residues is not understood, although they must have some importance due to their conservation across murine and human LYVE-1.[202] Although speculation, it may be that they modulate the redox potential of Cys201, since chemically local positive charge will push the redox equilibrium towards oxidation, and positively charged redox agents have been shown to accelerate the formation of disulfide bonds in proteins.[248]

2.1.2.2.3 The Transmembrane Domain of LYVE-1 Similarly to CD44, the transmembrane domain of LYVE-1 is a single-pass anchor containing a single unpaired cysteine residue (Cys257),[202] in conservation with CD44 (Cys286).[154] However, it has been reported that Cys257 in LYVE-1 does not form a disulfide bond to create homodimers,[243] as in CD44 (see Section 2.1.1.3.3).[154] It may be that it is still involved in endocytosis, as observed for CD44, since LYVE-1 does act as an endocytic receptor,[202] although this hypothesis has not been tested.

2.1.2.2.4 The Cytoplasmic Domain of LYVE-1 The cytoplasmic domain of LYVE-1 is involved in signalling pathways that induce lymphangiogenesis and permeability of endothelial cell junctions.[230] The exact nature of this signalling is still unknown since, although the LYVE-1 cytoplasmic domain contains phosphorylation sites, there is no evidence for them becoming phosphorylated.[202,226]

However, LYVE-1 signalling appears to be closely coupled to the MAPK/Erk signalling pathway in both junctional permeability and lymphangiogenesis signalling.[230] Furthermore, LYVE-1 signalling can be blocked through inhibition of growth factor receptors, including EGFR, PDGFR and VEGFR, indicating that LYVE-1 may recruit their kinase activity for downstream signalling.

2.1.3 Summary

Both CD44 and LYVE-1 are important receptors for hyaluronan (HA) that have implications in cell trafficking and some disease states, such as cancer. Despite their high sequence homology, they have many unique characteristics, that may be a result of their differences in tissue expression, thereby allowing them to adapt to their own unique niches.

CD44 has been extensively studied and the factors surrounding its function are well known. This is in part due to its earlier discovery and characterisation, but also due to its ease of expression. For example, the X-ray crystal structure of the CD44 HA-binding domain (HABD) was published over 15 years ago,[210] since high-purity, homogeneous samples of the CD44 HABD can be readily produced from *E. coli*.[197]

In contrast, LYVE-1 was discovered far more recently and so the literature surrounding its function is still in its infancy. Furthermore, attempts to produce samples of the LYVE-1 HABD for structural characterisation have failed. It is believed that the glycans present on native LYVE-1 are essential for proper folding and solubility.

2.1.4 Objectives

The aim of this chapter is to elucidate some of the structural features responsible for the unique function of LYVE-1, and compare them to CD44, using a combination of Saturation Transfer Difference NMR Spectroscopy and Molecular Modelling. Specifically the objectives are:

- To characterise the molecular recognition of HA by CD44 in

solution state.

- To characterise the molecular recognition of HA by LYVE-1 in solution state.

- To carry out a comparative study between both receptors.

2.2 Materials and Methods

2.2.1 Nuclear Magnetic Resonance Spectroscopy

The synthetic hyaluronan tetrasaccharide (HA4S) was synthesised by Dr Jose Luis de Paz (IIQ, CSIC-US, Seville, Spain)[249] and was assigned based on 1D ^{1}H NMR, ^{1}H-^{1}H COSY, ^{1}H-^{1}H TOCSY, ^{1}H-^{13}C HSQC and ^{1}H-^{1}H NOESY NMR experiments. For the TOCSY the mixing time was set to 80 ms and for the NOESY a number of experiments were acquired, with mixing times ranging from 300-1200 ms. The concentration of HA4S was 3.5 mM in 99% D_20 and the assignment experiments were acquired at 298 K. For saturation transfer difference NMR spectroscopy, all samples were prepared in a buffer containing 137 mM NaCl, 10 mM Na_2HPO_4, 2.7 mM KCl and 1 mM KH_2PO_4 at pH 7.4, using D_2O as the solvent. Final protein and ligand concentrations were 130 μM and 2.7 mM respectively. STD NMR experiments were performed using a train of 50 ms Gaussian pulses applied on the f2 channel at either 0.6 ppm (on-resonance) or 40 ppm (off resonance). A spoil sequence was used to destroy unwanted magnetisation and a 40ms spinlock pulse (3.8kHz) was used to suppress protein signals (stddiff.3). A range of saturation times were used to construct the STD NMR build up curves (0.5, 0.75, 1, 1.5, 2, 3, 4, 5, 6, 7, 8 s), with 128 scans per experiment. The recycle delay (d1) was set to 8 s. The experimental temperature for STD NMR experiments was 280 K. All experiments were performed on a Bruker Avance III 800 MHz spectrometer equipped with a 5-mm TXI 800 MHz H-C/N-D-05 Z BTO probe.

2.2.2 Homology Modelling

The coordinates of murine CD44 bound to HA8[210] (PDB accession code: 2JCR) were used as a template for the human CD44 and LYVE-1 HABD sequences. These sequences were aligned to the template using the ClustalW algorithm.[250] The Schrödinger Prime software[251] was used to construct an energy-based all-atom model with the OPLS3 forcefield,[120] keeping the HA8 model from the template. Regions of low sequence homology were refined using extended serial loop sampling in the Prime software.

2.2.3 Molecular Docking

The conformational flexibility of the HA4S ligand was considered by performing a conformational search using Schrödinger's MacroModel software.[252] Conformers were generated by Monte Carlo torsional sampling, eliminating structures with RMSD < 0.5 Å from existing structures or with a potential energy > 5 kcal mol^{-1} from the minimum energy structure according to the OPLS3 forcefield. Restraints of 100 kcal mol^{-1} were applied to all internal ring torsions so that the favoured $^4C_1$ ring conformation was preserved. Generated structures were minimised by the conjugate gradient method, converging on a threshold of 0.05 kcal mol^{-1}. Docking calculations were performed using Schrödinger's Glide software.[126,253] A cubic receptor grid with side length of 30 Å was generated by centring on the HA8 ligand. Glide docking of HA4S conformers was performed using standard precision, enhanced sampling and a distance-dependant dielectric constant of 2. The resulting poses were clustered using the hierarchical agglomerative average linkage algorithm.

2.2.4 Molecular Dynamics Simulations: Equilibration

Complexes were parametrised using AMBER ff14SB[116] and GLYCAM_06j for protein and carbohydrate moieties respectively. Model systems were generated by solvating with explicit TIP3P water

molecules within a truncated octahedron bounding box buffered from the complex by 10 Å (for ASMD, this was increased to 20 Å) and neutralising with Na^+ ions. Conjugate gradient minimisation was run with 20 kcal mol^{-1} $Å^{-2}$ restraints on solute atoms, converging on a threshold of 1 x 10^{-4} kcal mol^{-1} $Å^{-1}$, before repeating with no restraints. The restraints were reapplied to solute atoms before heating at constant volume to 310 K over a period of 500 ps. The system was then equilibrated at constant pressure (1 atm) over a period of 1.3 ns, with the restraints being released slowly over the last 800 ps. In all cases periodic boundary conditions and the particle mesh Ewald method were applied. A Langevin thermostat with a collision frequency of 5 ps^{-1} and a Berendsen barostat with a relaxation time of 2 ps were used. The SHAKE algorithm was used to restrain all bonds involving hydrogen, allowing a timestep of 2 fs to be used. A cutoff of 8 Å was used for non-bonded interactions.

2.2.5 Molecular Dynamics Simulations: Production

For standard production dynamics, parameters were used as described above, with no restraints applied. Simulations were run in triplicate for 100 ns each, saving coordinates every 10 ps.

For adaptive steered molecular dynamics, the distance between the methyl group of GlcNAc3 and the Cδ 1 of I88 (CD44) or I97 (LYVE-1) was increased from 3 to 33 Å over a total of 30 ns using a restraint of 5 kcal mol^{-1}. Each stage was repeated 30 times, each with a random seed to generate a unique trajectory. At the end of each stage, the Jarzynski average[144] was calculated and the trajectory closest to this average was used to initiate the next stage. Coordinates were saved every 2 ps.

2.3 Results

2.3.1 Nuclear Magnetic Resonance Assignment of a Synthetic Hyaluronan Tetrasaccharide

A synthetic tetrasaccharide of hyaluronan (HA4S, Fig. 2.8) was chosen for studying the interaction of hyaluronan (HA) with CD44 and LYVE-1 by Nuclear Magnetic Resonance (NMR) spectroscopy. HA4S contained a N-acetylglucosamine (GlcNAc) non-reducing terminal residue and a para-methoxyphenyl (PhOMe) linker covalently bound to the glucuronic acid (GlcA) reducing terminal residue *via* a β-glycosidic bond. This was chosen for two reasons: (1) the repetitive nature of glycosaminoglycans and inherent low spectral dispersion of carbohydrates[20,254]makes studying a longer construct by NMR spectroscopy extremely challenging, (2) the PhOMe linker retains the anomeric configuration of the reducing terminal residue, thereby deconvoluting the spectrum by reducing the number of species present in solution (absence of anomeric equilibrium).

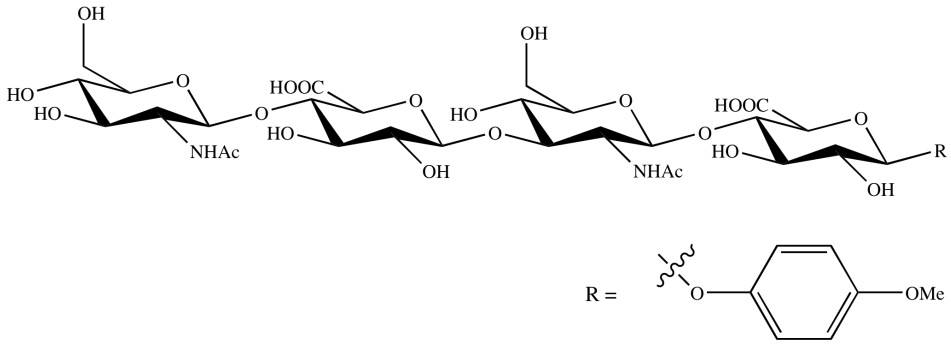

Figure 2.8: Chemical structure of the synthetic hyaluronan tetrasaccharide.

The NMR assignment of HA4S (Table 2.1) was achieved through a combination of Correlation Spectroscopy (COSY), Total Correlation Spectroscopy (TOCSY), Nuclear Overhauser Effect Spectroscopy (NOESY), and ^{1}H-^{13}C Heteronuclear Single Quantum Correlation (HSQC) NMR spectroscopy experiments.

Table 2.1: Assignment of the chemical shifts of ^{1}H and ^{13}C nuclei of HA4S. Experiments were performed in an 800 MHz NMR spectrometer at 298 K. Only non-exchangable proton resonances were observed since the sample was prepared in D_2O.

Residue	Position	δ ^{1}H (ppm)	δ ^{13}C (ppm)
GlcNAc1	1	4.44	100.58
	2	3.61	55.23
	3	3.43	73.81
	4	3.37	75.78
	5	3.37	69.52
	6	3.67	60.50
	6'	3.82	60.50
	Me	1.94/1.96	22.29
GlcA2	1	4.38	103.1
	2	3.26	72.44
	3	3.49	73.42
	4	3.63	76.18
	5	3.65	79.54
GlcNAc3	1	4.49	100.6
	2	3.77	54.09
	3	3.64	82.30
	4	3.41	75.21
	5	3.44	68.31
	6	3.70	60.50
	6'	3.84	60.50
	Me	1.94/1.96	22.29
GlcA4	1	4.91	101.2
	2	3.51	72.44
	3	3.60	73.51
	4	3.74	76.62
	5	3.74	79.80
PhOMe	ortho	7.01	118.0
	meta	6.89	114.9
	Me	3.72	55.65

The TOCSY experiment was used to separate the HA4S resonances into groups corresponding to individual (as yet unclassified) hexopyranose residues (Fig. 2.9). This was possible because the glycosidic bond between residues separates inter-residue protons from one another by

more than 3 bonds - couplings over a greater number of bonds than this are typically too weak to be detected.[255]

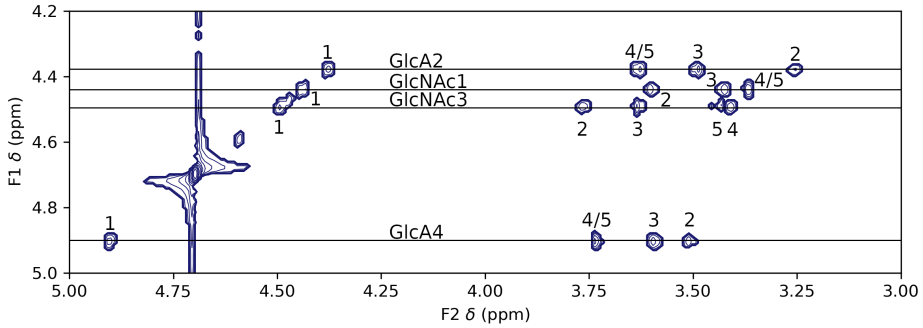

Figure 2.9: Section of the TOCSY spectrum of HA4S, showing the diagonal peaks corresponding to each anomeric proton resonance. Cross peaks correspond to proton resonances within the same spin system, and therefore within the same residue.

Subsequently, the anomeric protons of each residue were easily identifiable since the resonances of these protons typically lay within the δ 4.4-5.5 ppm region,[254] well dispersed from any non-anomeric carbohydrate resonances. The anomeric carbon resonances, observed in the ^{1}H-^{13}C HSQC (Fig. 2.10), are typically also shifted downfield considerably compared to non-anomeric carbon resonances.

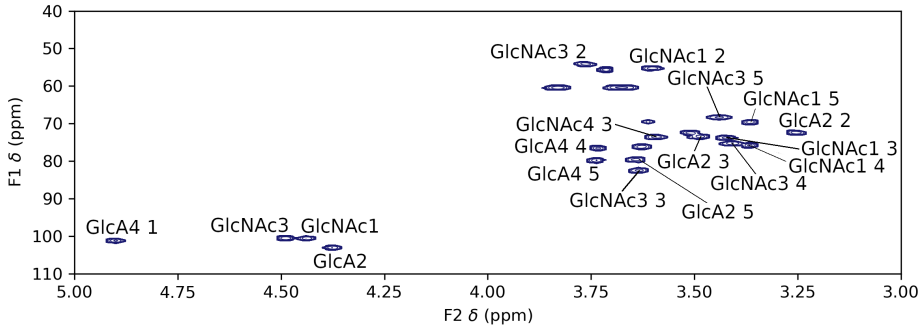

Figure 2.10: Section of the ^{1}H-^{13}C HSQC spectrum of HA4S, with resonances from the PhOMe and GlcNAc methyl groups omitted for clarity. Cross peaks correspond to correlations between proton and carbon resonances coupled through one bond and therefore correspond to each individual position on the pyranose ring.

Due to the dispersion of the anomeric proton resonances it was also straightforward to identify the chemical shift of the H2 protons, since the H1-H2 cross-peaks in the COSY spectrum (Fig. 2.11), which identifies through-bond correlations through ^{3}J-coupling, were well isolated.

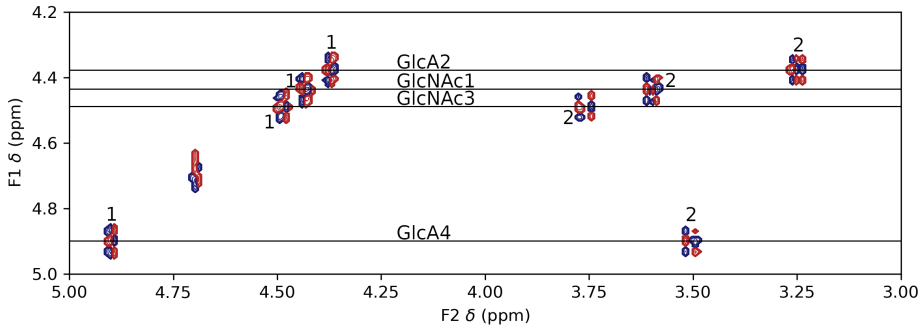

Figure 2.11: Section of the COSY spectrum of HA4S showing the diagonal peaks corresponding to each anomeric proton resonance. Cross peaks correspond to proton resonances coupled to these anomeric protons through three bonds and therefore, in this case, come from protons at position 2 in the pyranose rings.

In the assignment of small molecules, it is common to determine the connectivity of the whole spin system by continuing to trace the cross-peaks from one resonance to another.[255] This too is possible for carbohydrates, but the narrow spectral dispersion of non-anomeric proton resonances make this a challenge. Here, the H3, H4, and H5 resonances of each residue could be assigned by comparing the through-bond COSY and TOCSY experiments with the through-space NOESY experiment. In pyranoses with the same configuration as β-D-glucose, including GlcNAc and GlcA, the CH protons lay in two distinct planes:[27] H1, H3 and H5 below the ring (α face), and H2 and H4 above the ring (β face). Protons within each plane are in close proximity to one another, and therefore give rise to strong cross-peaks in the NOESY spectrum (Fig. 2.12). In this way, the H3, H4, and H5 resonances of each residue were assigned.

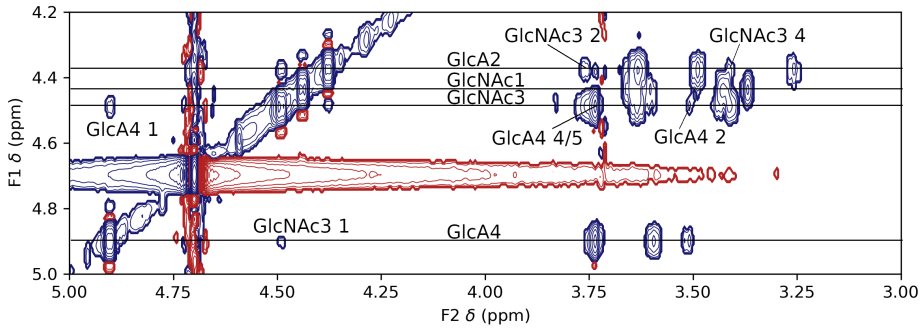

Figure 2.12: Section of the NOESY spectrum of the synthetic hyaluronan tetrasaccharide, showing the diagonal peaks corresponding to each anomeric proton resonance. Cross peaks correspond to proton resonances dipolarly coupled to these anomeric protons through space. These include intra-residue NOEs (and are also observed in the TOCSY experiment) and inter-residue NOES (not observed in the TOCSY experiment). Only well isolated inter-residue cross-peaks are labelled for clarity.

The GlcA residues could then be distinguished from the GlcNAc residues, since GlcA does not contain exocyclic protons on the C6 carbon (H6 and H6'). Furthermore, the C2 carbon atoms of GlcNAc resonate much further upfield than in GlcA due to the presence of the amide of the N-acetyl group of GlcNAc.

The H6 protons of GlcNAc were identified by through-bond correlation with the H5 resonance. The H6 and H6' resonances also gave rise to a unique identifying pattern in the ^{1}H-^{13}C HSQC spectrum (Fig. 2.10) since they show correlation to the same carbon resonance. The methyl of the N-acetyl group of GlcNAc could be easily identified because both the proton and carbon resonances can be found far upfield compared with all other proton and carbon resonances within the carbohydrate residue. However, they could not be distinguished from one another since the methyl group is not within the same spin system as the remaining protons in the same residue. Furthermore, no intra-residue NOESY cross-peaks were clearly identifiable.

The position of each residue within the HA4S sequence could then be assigned using knowledge of the pattern of the glycosidic linkages.[20] For example, the GlcNAc3 residue in HA4S forms a glycosidic bond with the previous GlcA residue *via* the C3 carbon. This bond is not present in GlcNAc1. Therefore the chemical shift of the GlcNAc3 C3 carbon is shifted downfield relative to that of GlcNAc1. For the GlcA residues, their position in the HA4S sequence could be distinguished,

since the anomeric proton resonance of GlcA4 was shifted downfield relative to that of GlcA1 due to the presence of the PhOMe linker.

Finally the chemical shifts of the PhOMe linker could easily be assigned, since the methyl protons resonated far upfield compared to the aromatic protons. The ortho- and meta-protons could be distinguished from one another due to the presence of a NOESY cross-peak between the GlcA4 anomeric proton and the aromatic ortho-proton.

2.3.2 STD NMR Study of the binding of the synthetic hyaluronan tetrasaccharide to CD44 and LYVE-1

To gain structural details of the interactions of HA4S with CD44 and LYVE-1, saturation Transfer Difference (STD) NMR spectroscopy experiments were performed for the samples containing synthetic hyaluronan tetrasaccharide (HA4S) in the presence of either CD44 (Fig. 2.13) or LYVE-1 (Fig. 2.15). Binding of HA4S to both receptors was detected by the STD NMR, which showed unambiguous signals in the difference spectrum. The relatively low ligand-to-protein ratio of 21:1 used for these experiments led to chemical shift perturbations observable in the ligand signals, so that a different pattern of signals were observed for HA4S in the presence of CD44 or LYVE-1, compared to HA4S alone. STD NMR build up curves were constructed (Fig. 2.14A, Fig. 2.16A) by measuring the STD intensity of each resolvable resonance of the STD NMR spectrum for saturation times ranging from 0.5 s to 8 s. Since the STD NMR experiments were based on one-dimensional ^{1}H NMR experiments, it was not possible to resolve all assigned resonances and therefore STD NMR build up curves for several protons are missing. The STD NMR build up curves were used to extract the initial rate of STD intensity build up (STD_0) for each proton resonance by fitting each build up curve to the monoexponential equation described in Eqn. 45 (CD44: Table 2.2, LYVE-1: Table 2.3). Binding epitope maps could then be constructed for HA4S binding to CD44 or LYVE-1 by plotting the normalised STD_0 values onto a structure of HA4S (Fig. 2.14B, Fig. 2.16B).

Analysis of the STD NMR build up curves (Fig. 2.14, Fig. 2.16) reveals that, in general, the STD intensities observed for HA4S binding to either CD44 or LYVE-1 are comparable. In both cases, STD intensity is spread along the length of the tetrasaccharide, indicating a longitudinal mode of binding, compatible with the observation that HA binds across a shallow groove along the CD44 HA binding domain (HABD) (Fig. 2.5), as seen in the X-ray crystal structure of the murine CD44 HABD bound to a HA octasaccharide.[211] These results indicate that HA4S also binds to LYVE-1 in a similar manner. This is the first experimental demonstration of the similarity of binding modes of hyaluronan to LYVE-1 and CD44.

2.3.2.1 The Binding of HA4S to CD44

The binding epitope map of HA4S binding to CD44 (Fig. 2.14) reveals that the majority of close contacts between HA4S and CD44 occur between the central two residues of HA4S (GlcA2-GlcNAc3), in particular the methyl group of GlcNAc3, which receives the largest amount of saturation. Again, this agrees with the X-ray crystal structure of the murine CD44 HABD bound to a HA octasaccharide,[211] which highlights that this central disaccharide does indeed make the greatest number of contacts with the CD44 HABD, and shows the methyl group of the GlcNAc residue buried within a deep hydrophobic pocket (Fig. 2.5). Furthermore, the fact that only a low degree of saturation is observed for the methyl group of GlcNAc1 shows that only one predominant binding mode is present - that is to say that there is no significantly populated binding mode in which the GlcNAc1 methyl is buried within the hydrophobic pocket. This further suggests that the minimum recognition motif of HA for binding to CD44 is at least GlcAβ(1-3)GlcNAc.

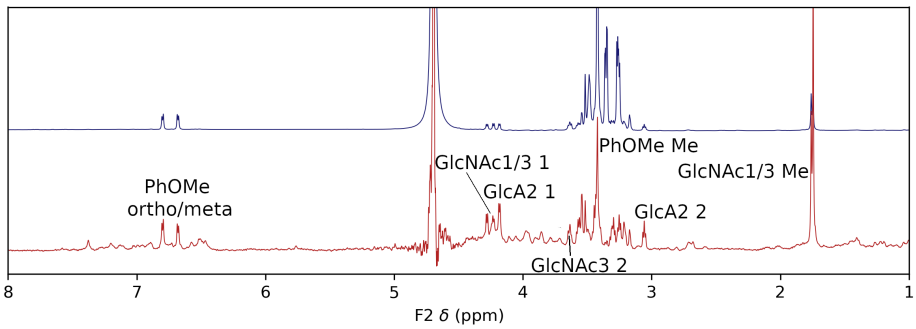

Figure 2.13: Reference (blue) and difference (red) spectra for HA4S in the presence of CD44. Measured using a saturation frequency of 0.6ppm at 800 MHz and a saturation time of 2 s. Presence of additional signals is due to some protein signals not being fully removed by the spinlock.

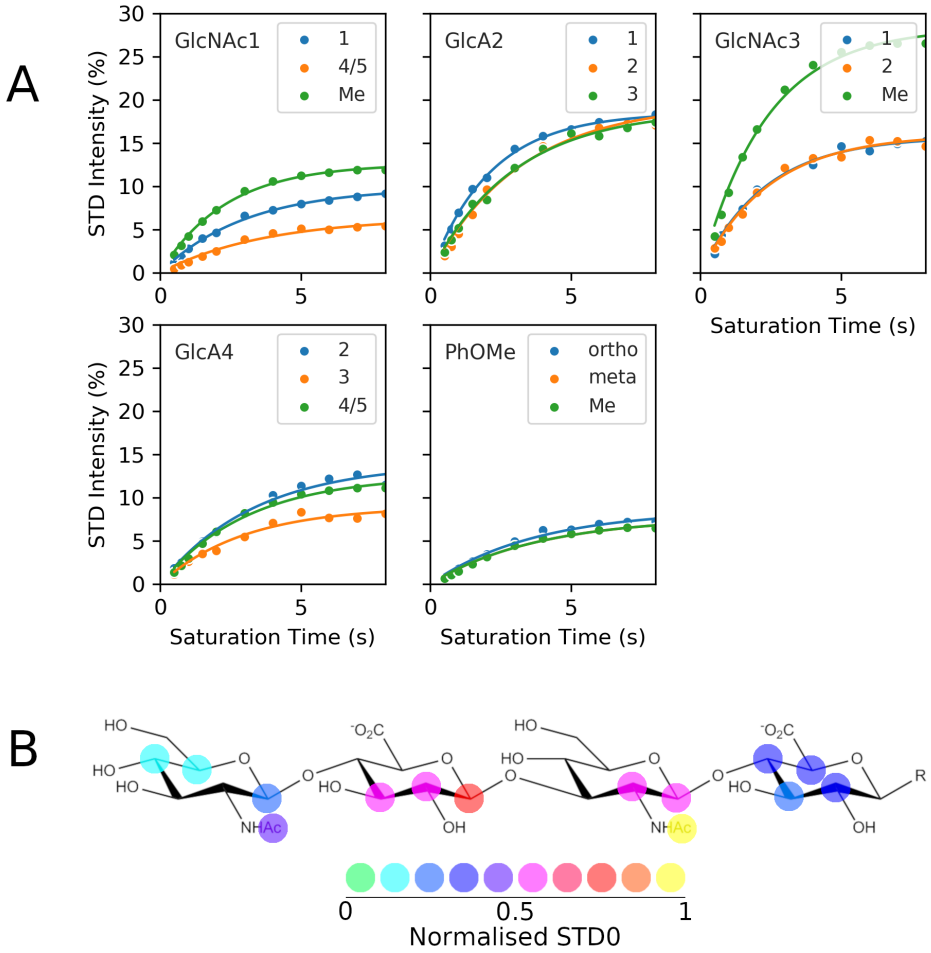

Figure 2.14: **A:** STD NMR build up curves for HA4S in the presence of CD44. Graphs are plotted by residue (GlcNAc1 - PhOMe; top left - bottom right) and positions of each proton resonance are marked on inset legends. Plots show experimentally measured STD intensities (dots) and curves calculated by fitting to Eqn. 45 (lines). **B:** Binding epitope map for the interaction of HA4S with CD44. Colours represent the normalised values of STD_0 for proton resonances at the indicated positions (low - high, cold - hot). The R group represents the PhOMe linker, which is omitted for clarity.

Table 2.2: Initial rate of STD intensity build up (STD_0) for each proton resonance of HA4S in the presence of CD44. Normalised values ($STD_0(norm.)$) are calculated by dividing STD_0 by the largest STD_0 value. [a/b]: STD intensity measured for overlapping resonances.

Residue	Position	k_{sat} (s^{-1})	STD_{max} (%)	STD_0 (% s^{-1})	$STD_0(norm.)$
GlcNAc1	1	0.33	15.8	3.25	0.27
	4	0.27[a]	6.40[a]	1.71[a]	0.14[a]
	5	0.27[a]	6.40[a]	1.71[a]	0.14[a]
	Me	0.43	28.3/12.6	12.1/5.38	1.00/0.44
GlcA2	1	0.47	18.5	8.61	0.71
	2	0.31	19.6	6.08	0.50
	3	0.33	18.9	6.29	0.52
GlcNAc3	1	0.43	15.8	6.83	0.56
	2	0.41	16.0	6.60	0.55
	Me	0.43	28.3/12.6	12.1/5.38	1.00/0.44
GlcA4	2	0.30	14.0	4.17	0.34
	3	0.33	18.9	3.00	0.25
	4	0.32[b]	12.6[b]	4.03[b]	0.33[b]
	5	0.32[b]	12.6[b]	4.03[b]	0.33[b]
PhOMe	ortho	0.27	8.56	2.01	0.17
	meta	0.26	7.74	2.28	0.19
	Me	0.42	2.41	1.03	0.08

2.3.2.2 The Binding of HA4S to LYVE-1

The binding epitope of HA4S binding to LYVE-1 (Fig. 2.16) is intriguing since the only strong contact appears to be the methyl group of GlcNAc3; the remainder of the epitope shows normalised STD values of between 0.1 and 0.4 uniformly distributed across HA4S (Fig. 2.14, Table 2.3). Such an observation is indicative of a transient protein-ligand complex, since in such cases the residence time of the ligand within the binding pocket permits the saturation to spread only to those protons that are in closest proximity to the protein protons. Indeed, this agrees with experimental observations that the LYVE-1-HA system exhibits much faster binding kinetics than CD44-HA.[212,214,243,244] Finally, again the low degree of saturation on the GlcNAc1 methyl group suggests that LVYE-1 also has requirements for a minimum structural

motif, that is at most a trisaccharide of GlcA-GlcNAc-GlcA.

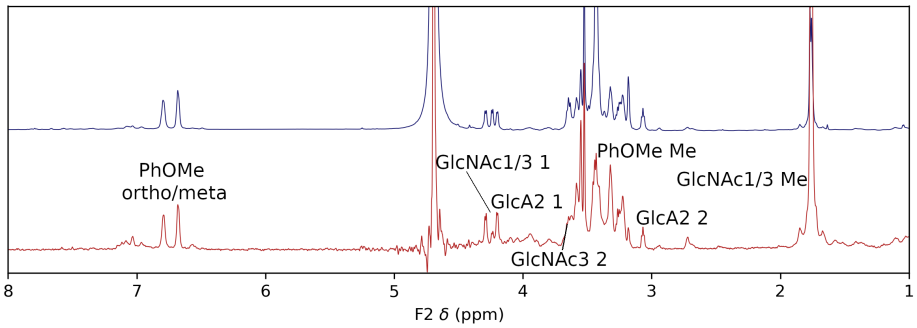

Figure 2.15: Reference (blue) and difference (red) spectra for HA4S in the presence of LYVE-1. Measured using a saturation frequency of 0.6ppm at 800 MHz and a saturation time of 2 s. Presence of additional signals is due to some protein signals not being fully removed by the spinlock.

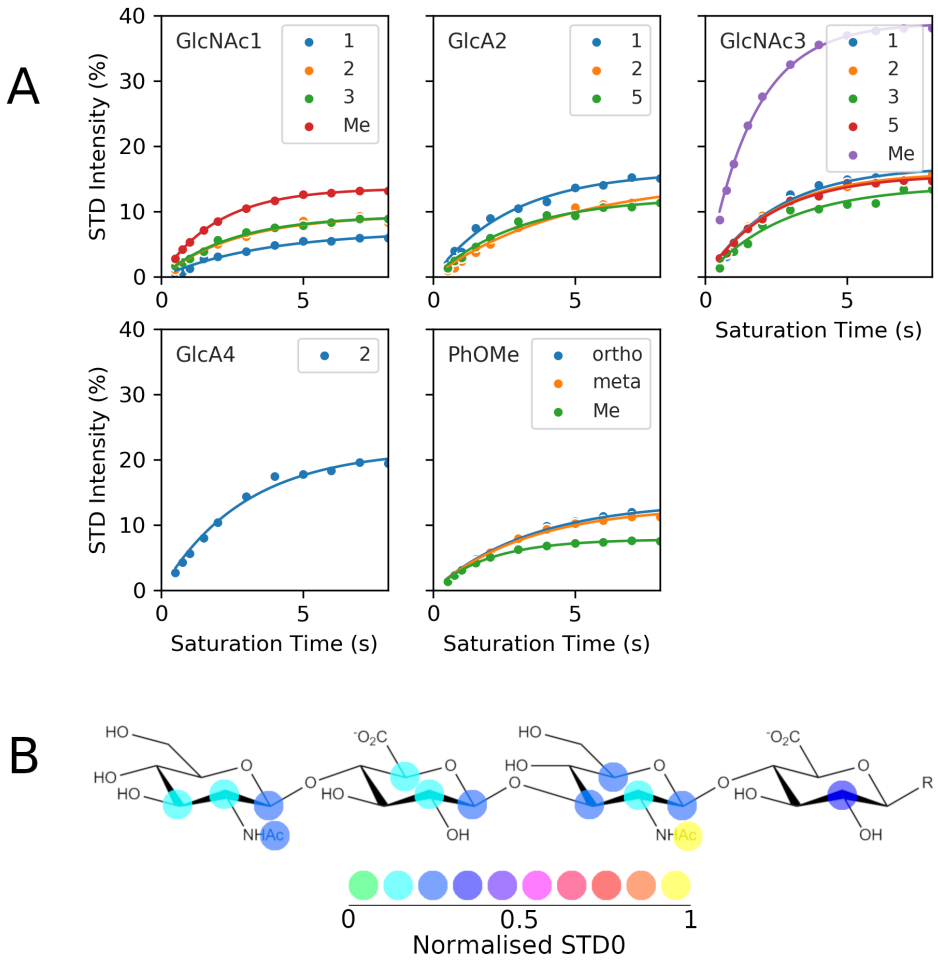

Figure 2.16: **A:** STD NMR build up curves for HA4S in the presence of LYVE-1. Graphs are plotted by residue (GlcNAc1 - PhOMe; top left - bottom right) and positions of each proton resonance are marked on inset legends. Plots show experimentally measured STD intensities (dots) and curves calculated by fitting to Eqn. 45 (lines). **B:** Binding epitope map for the interaction of HA4S with LYVE-1. Colours represent the normalised values of STD_0 for proton resonances at the indicated positions (low - high, cold - hot). The R group represents the PhOMe linker, which is omitted for clarity.

Table 2.3: Initial rate of STD intensity build up (STD_0) for each proton resonance of HA4S in the presence of LYVE-1. Normalised values ($STD_0(norm.)$) are calculated by dividing STD_0 by the largest STD_0 value. [a/b]: STD intensity measured for overlapping resonances.

Residue	Position	k_{sat} (s^{-1})	STD_{max} (%)	STD_0 (% s^{-1})	$STD_0(norm.)$
GlcNAc1	1	0.27	6.95	6.68	0.29
	2	0.37	9.39	3.51	0.15
	3	0.40	9.34	3.70	0.16
	Me	0.49	13.6	22.97/6.69	1.00/0.29
GlcA2	1	0.35	16.2	5.72	0.25
	2	0.22	14.8	3.25	0.14
	5	0.33	12.2	4.02	0.17
GlcNAc3	1	0.39	16.9	6.68	0.29
	2	0.41	16.0	3.51	0.15
	3	0.36	13.9	4.96	0.22
	5	0.42	15.6	6.59	0.29
	Me	0.59	38.9	22.97/6.69	1.00/0.29
GlcA4	2	0.34	21.6	7.28	0.32
PhOMe	ortho	0.29	13.8	3.91	0.17
	meta	0.30	12.9	3.81	0.17
	Me	0.42	7.80	3.90	0.17

2.3.3 Homology modelling of the human CD44 and LYVE-1 Hyaluronan Binding Domains in Complex with hyaluronan

Since human CD44 was used for STD NMR experiments, a homology model of the human CD44 hyaluronan (HA) binding domain (HABD) was produced (Fig. 2.17 A) from the X-ray crystal structure of the murine CD44 HABD bound to a HA octasaccharide.[211] With the murine and human CD44 HABDs sharing 88% identity and 96% similarity, alignment of the two sequences resulted in only a one residue deletion in the human CD44 HABD relative to mouse (Fig. 2.18). The subsequent minimised homology model of the human CD44 HABD bound to a HA octasaccharide differed from the X-ray crystal structure of the murine complex only by an RMSD of 0.173 Å (Cα) (Fig. 2.17), with almost all of this deviation due to the deletion of a residue

in the loop between $\beta 5$ and $\beta 6$ and in the loop between $\beta 6$ and $\beta 7$.

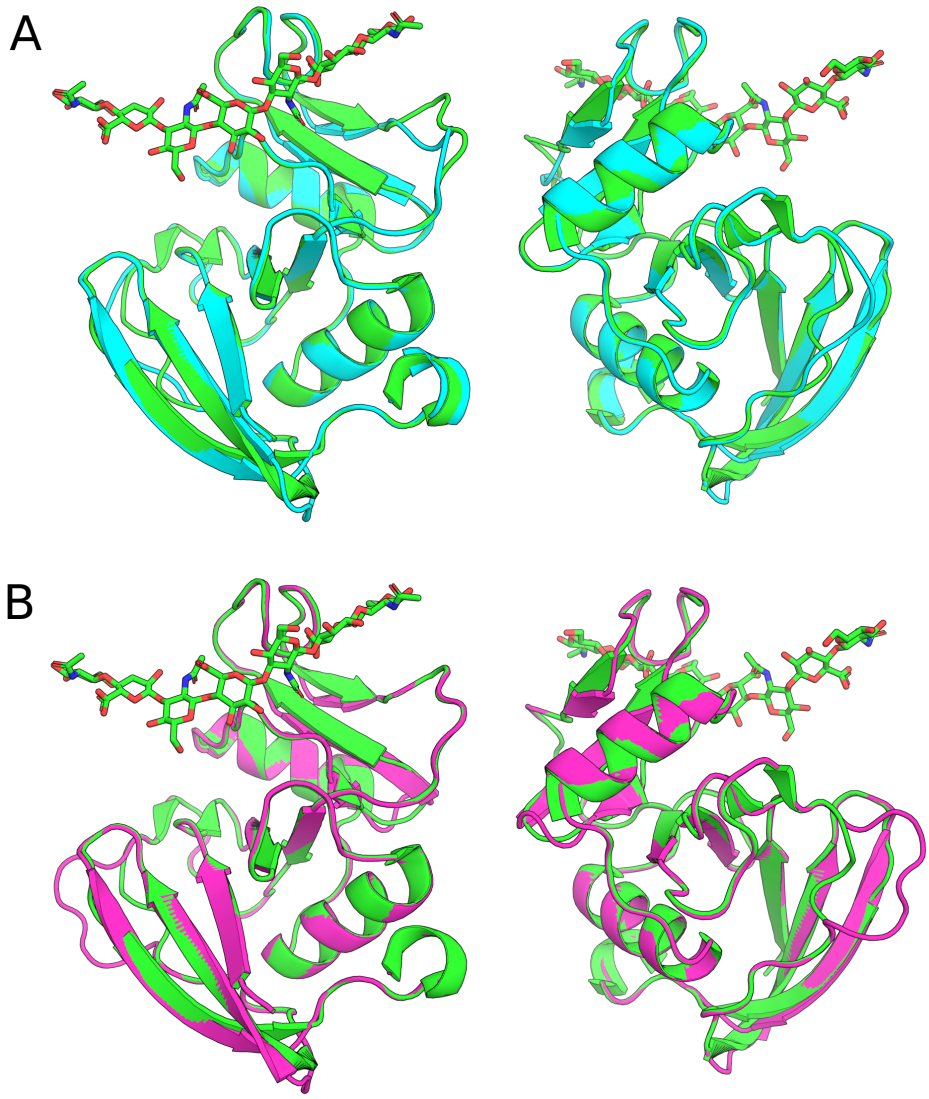

Figure 2.17: Cartoon representation of the X-ray crystal structure of the murine CD44 (green, PDB accession code: 2JCR) superimposed with the homology models of the human CD44 HABD (cyan, **A**) and the human LYVE-1 HABD (pink, **B**).

```
muCD44          QIDLNVTCRYAGVFHVEKNGRYSISRTEAADLCQAFNSTLPTMDQMKLALSKGFETCRYG
huCD44          QIDLNITCRFAGVFHVEKNGRYSISRTEAADLCKAFNSTLPTMAQMEKALSIGFETCRYG
huLYVE-1        ELSIQVSCRIMGITLVSKKANQQLNFTEAKEACRLLGLSLAGKDQVETALKASFETCSYG
                ::.:::**  *:  *.*:.. .:. *** : *: :. :*.   *:: **. .**** **

muCD44          FIEG-NVVIPRIHPNAICAANHTGVYILVTSNTSHYDTYCFNASAPPEEDCT----SVTD
huCD44          FIEG-HVVIPRIHPNSICAANNTGVYILTSN-TSQYDTYCFNASAPPEEDCT----SVTD
huLYVE-1        WVGDGFVVISRISPNPKCGKNGVGVLIWKVPVSRQFAAYCYNSSDTWTNSCIPEIITTKD
                :: .  ***.** **. *. * .** *     : :: :**:*:* .  :.*     :..*

muCD44          LPNSFDGPVTITIVNRDGTRYSKKGEYRTHQEDIDAS
huCD44          LPNAFDGPITITIVNRDGTRYVQKGEYRTNPEDIYPS
huLYVE-1        PIFNTQTATQTTEFIVSDSTYSVASPYSTIPAPTTTP
                     : .   * .  ..: *   . * *      ..
```

Figure 2.18: Multiple sequence alignment of the human CD44 and HABD sequences to murine CD44. Identical (*), strongly similar (:) and weakly similar (.) properties according to the Gonnet PAM matrix[256] are highlighted. Relative to the murine CD44 sequence, the human CD44 HABD contains 1 deletion and the human LYVE-1 HABD contains 5 insertions.

Conversely, the LYVE-1 HABD shares only 37% identity and 54% similarity to the murine CD44 HABD. Nevertheless, the sequence can be aligned with only 1 insertion in the loop between β6 and β7 and 4 insertions in the loop between β3 and β4. Overall, the resulting LYVE-1 homology modelled structure (Fig. 2.17 B) has a RMSD of only 0.164 Å (Cα) from the murine CD44 HABD structure.

According to this model, the topology of the HA-binding groove of the LYVE-1 HABD has some striking differences compared to that of CD44 (Fig. 2.19). Firstly, in agreement with previous descriptions,[242] there are no hydrogen bonding interactions between HA and the sidechains of the LYVE-1 HABD. In addition, there appears to be no interaction between the C-terminal extension and HA. This leaves a wide opening at the non-reducing terminal proximal end of the groove. The reducing terminal proximal end of the groove appears to be somewhat narrower in LYVE-1 than in CD44, due to the positioning of Lys108 and Trp116. This differs slightly from the previously published model,[242] in which the conformation of the loop between β5 and β6 (which contains W116) is modelled differently. However, it should be noted that that model was derived from the X-ray crystal structure of the apo-form human CD44 HABD (PDB accession code: 1UUH), which does adopt a different conformation to the murine HA-bound form. Furthermore, the modelling software used here, Prime, has demonstrated a higher accuracy than other homology modelling programs, particularly in loop regions.[257,258] Unfortunately, there is limited data on the importance of W116 in the HA-binding interaction, since the only mutant produced (δW116Y)

was somewhat conservative and produced little effect on binding.[242] Finally, the hydrophobic residues (Ile100 and Ala103) of the hook-like loop in CD44, which are known to be essential for the interaction with HA, are replaced with charged, flexible lysine residues (Lys105 and Lys108) (Fig. 2.19 right). The only major feature which is conserved, is the hydrophobic binding pocket that accommodates a GlcNAc methyl group - the residues that form this are absolutely conserved (muCD44: Tyr83 and Ile92, huLYVE-1: Tyr87 and Ile97).

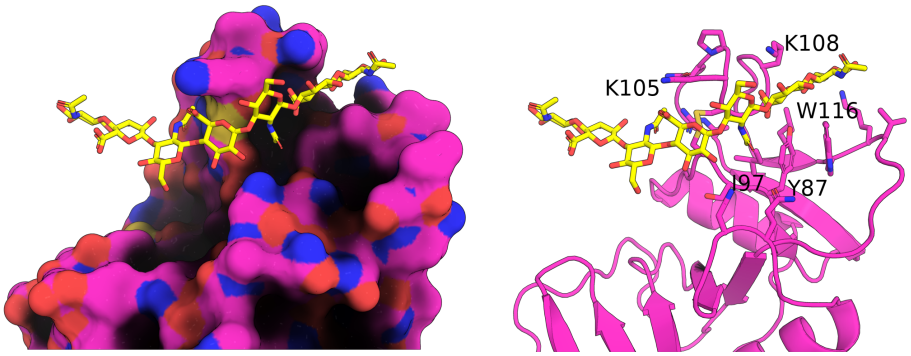

Figure 2.19: **Left:** Surface representation of the homology model of the human LYVE-1 HABD, showing a superposition of an octasaccharide of HA from the X-ray crystal structure of murine CD44 (yellow, sticks). **Right:** Cartoon representation of the homology model of the human LYVE-1 HABD (pink), showing key binding residues (pink, sticks).

2.3.4 Molecular Docking Models of the human CD44 and LYVE-1 Hyaluronan Binding Domains in Complex with a Synthetic Hyaluronan Tetrasaccharide and Validation with CORCEMA-ST

To produce a model of HA4S in complex with the human CD44 and LYVE-1 HABDs, Glide was used to dock HA4S into the binding groove of the models produced above. The resulting docking poses were clustered and the pose closest to the centroid of each cluster was retained for analysis. For CD44, three major clusters were found, each with approximately 28% occupancy (Fig. 2.20, Table 2.4). However, only one of these clusters was close to the HA structure in the X-ray crystal structure of the murine CD44 HABD (Cluster 2, RMSD: 0.963 Å, in place, heavy atoms) and agreed qualitatively with the experimental STD NMR data, as well as being compatible with the known inter-

actions with CD44. Furthermore, this docking pose showed the best docking score (Glide Emodel: -162 arb. units) and had the smallest variance (5.28 Å^2), indicating better convergence. Therefore it was only this docking pose that was considered for further analysis.

Table 2.4: Docking statistics for HA4S docking to homology models of human CD44 and LYVE-1 HABDs. The RMSD is calculated in place for the heavy atoms of HA4S relative to the HA octasaccharide in the X-ray structure of the murine CD44 HABD, and corresponds to the representative HA4S structure closest to the centroid of its cluster. Docking score given in arbitrary units (AU). Variance (σ^2)of the RMSD of cluster members is also shown.

Protein	Cluster	Occupancy (%)	RMSD (Å)	Docking Score (AU)	σ^2 (Å^2)
CD44	1	28.0	14.5	-133	6.99
	2	28.0	0.963	-162	5.28
	3	27.4	6.76	-134	7.12
LYVE-1	1	54.2	0.836	-94.6	2.11

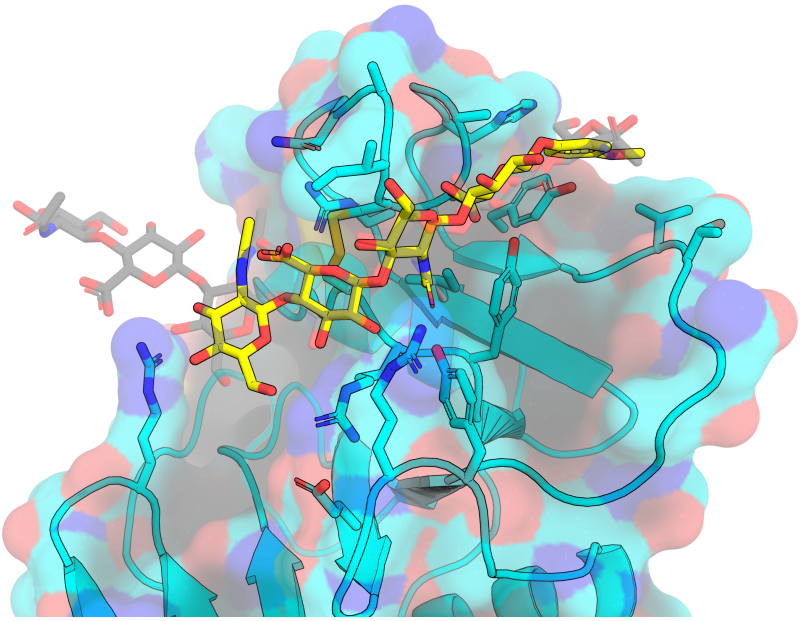

Figure 2.20: Best docking pose of HA4S (yellow) bound to a homology model of the human CD44 HA binding domain (cyan). The HA octasaccharide (grey) from the X-ray crystal structure of the murine CD44 HABD is shown for comparison.

Since the docking pose of HA4S has such significant overlap with the HA octasaccharide in the X-ray crystal structure of the murine CD44 HABD, and the key HA binding residues are conserved between

murine and human CD44, a detailed discussion of the interaction can be found in Section 2.1.1.3.1. However, it should be noted that it is GlcA2 and GlcNAc3 in HA4S that make the apparent key disaccharide pair for interaction with CD44, with the methyl group of GlcNAc being found in the hydrophobic pocket of the CD44 HA binding groove. It is also interesting to note that the PhOMe linker is positioned and oriented such that it may act as a good substitute for the hydrophobic surface of the GlcNAc residue that would occupy that space in a longer HA construct.

In the case of LYVE-1, only one major cluster was observed, with an occupancy of 54% (Fig. 2.21, Table 2.4). This too is very similar to the HA octasaccharide in the X-ray crystal structure of the murine CD44 HABD and has an RMSD of 0.836 Å (in place, heavy atoms). The binding pose of HA4S in the binding groove of the LYVE-1 HABD is comparable to that of HA4S in the binding groove of the CD44 HABD, in particular the GlcNAc3 methyl group being buried in the hydrophobic pocket of the binding groove. This is in excellent agreement with the STD NMR observations (Fig. 2.14, Fig. 2.16). Interestingly, the carboxylate groups of GlcA2 and GlcA4 are located in close proximity to Lys105 and Lys108 respectively. Like in the CD44 model, the PhOMe linker occupies the same space as would be occupied by a GlcNAc residue in a longer HA construct, stacking directly above Trp116. Although the value of the docking score is less negative for LYVE-1 compared to CD44, it should be noted that comparison of docking scores between two different systems can be inaccurate, so their relative values should not be considered too seriously.

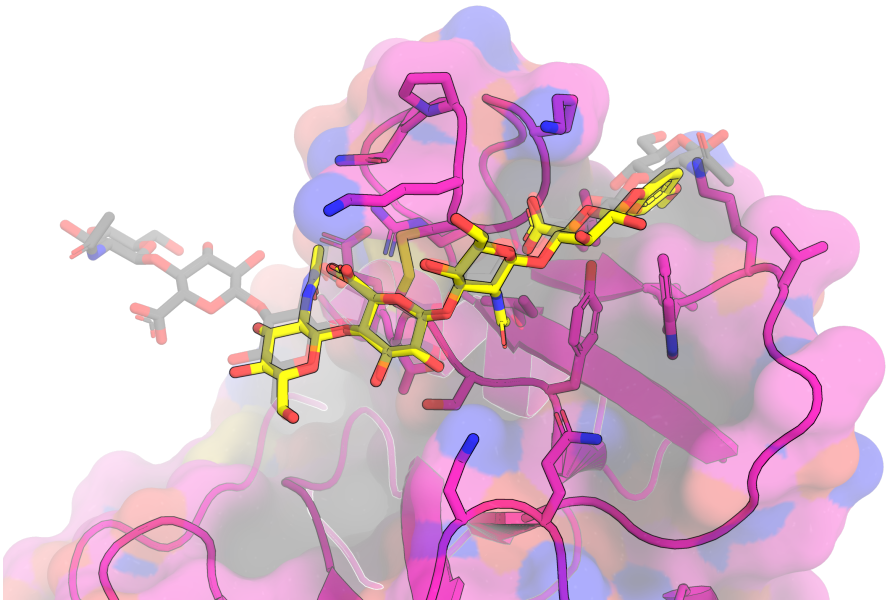

Figure 2.21: Best docking pose of HA4S (yellow) bound to a homology model of the human LYVE-1 HA binding domain (cyan). The HA octasaccharide (grey) from the X-ray crystal structure of the murine CD44 HABD is shown for comparison.

To validate the molecular models of the complexes of HA4S with the HABDs of human CD44 and LYVE-1, the CORCEMA-ST program was used to predict theoretical STD NMR build up curves based on the 3-dimensional models. In the case of the CD44-HA4S complex (Fig. 2.22), the NOE R-factor, which describes the overall agreement between the experimental and theoretical data, was 0.151, indicating an excellent fit between the experimental and theoretical STD NMR build up curves. In general a value of 0.3 is considered a good fit.

For LYVE-1 the overall RNOE was 0.272 (Fig. 2.23), which is considered a good fit, especially considering the the receptor model was a homology model built from a template with only 54% similarity. In fact, if the aromatic moiety is removed from the calculation, the RNOE value is 0.198. This shows that the receptor-carbohydrate interaction is modelled extremely well, and the model complex presented here is very close to the true structure of the complex in the solution state. As for the PhOMe group of HA4S in the model of the LYVE-1-HA4S complex, the predicted STD NMR intensities are slightly stronger than the experimental intensities. There are a couple possible explanations for this: (1) the specific broadening of the PhOMe group in the presence of LYVE-1 suggests that in the HA4S-LYVE-1 complex the T_2 of the PhOMe protons are significantly shorter than the rest of the molecule;

(2) in the model of the HA4S-LYVE-1 complex, the PhOMe group is in close proximity to the loop between $\beta 5$ and $\beta 6$. The chemical properties of the amino acid residues in this loop are quite dissimilar to those in either human or murine CD44 and modelling loops is challenging, so it may be that the conformation of this loop is not in full agreement with the true structure. Nevertheless, this data provides valuable information on the structure of the LYVE-1 HABD, which is yet to be solved experimentally.

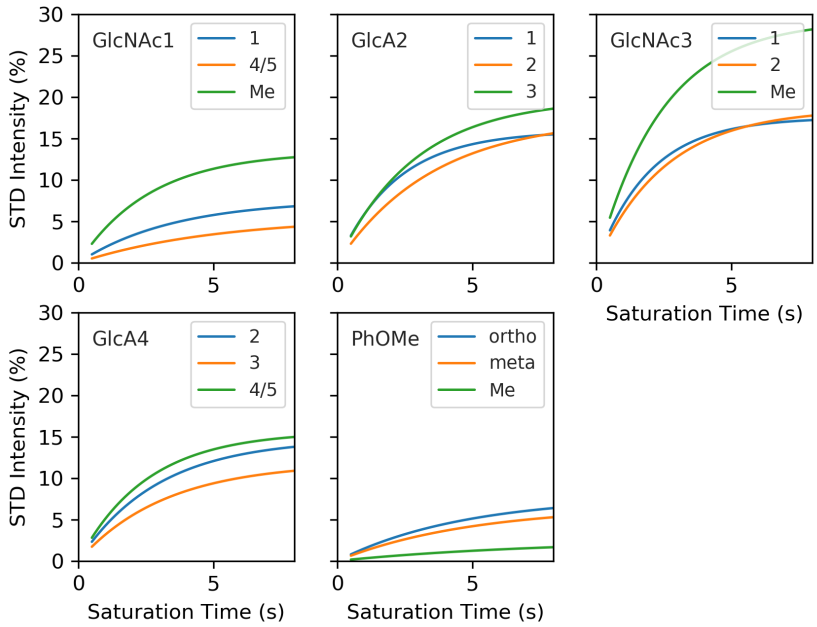

Figure 2.22: Saturation Transfer Difference (STD) Nuclear Magnetic Resonance (NMR) predicted by CORCEMA-ST for the molecular docking model of HA4S in complex with the human CD44 HA binding domain. Graphs are plotted by residue (GlcNAc1 - PhOMe; top left - bottom right) and positions of each proton resonance are marked on inset legends. The RNOE factor compared to the experimental STD NMR data is 0.151.

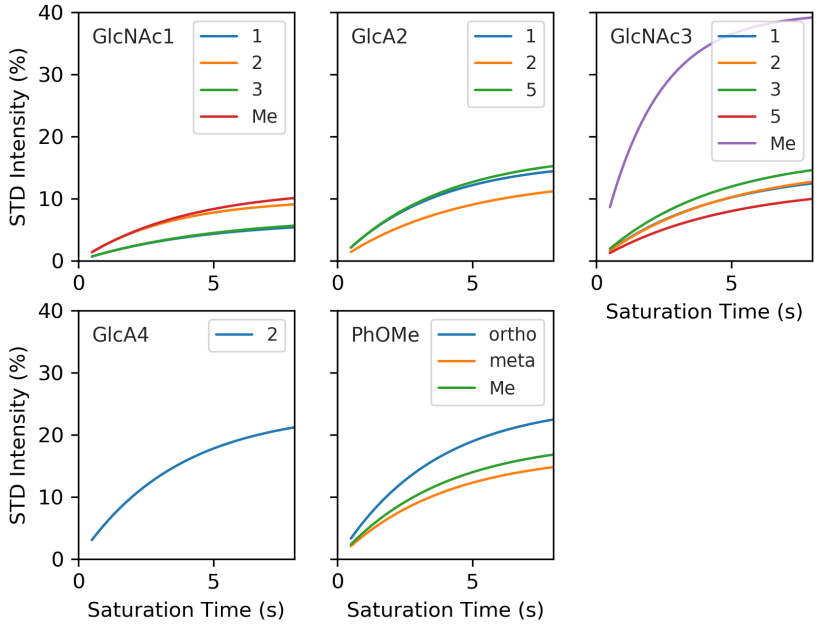

Figure 2.23: Saturation Transfer Difference (STD) Nuclear Magnetic Resonance (NMR) predicted by CORCEMA-ST for the molecular docking model of HA4S in complex with the human LYVE-1 HA binding domain. Graphs are plotted by residue (GlcNAc1 - PhOMe; top left - bottom right) and positions of each proton resonance are marked on inset legends. The RNOE factor compared to the experimental STD NMR data is 0.272 (0.198 discounting the PhOMe residue).

2.3.5 Molecular Dynamics Simulations of the human CD44 and LYVE-1 Hyaluronan Binding Domains in Complex with a Hyaluronan Tetrasaccharide

To assess the stability and understand the dynamics of the modelled complexes, 100 ns molecular dynamics simulations of CD44 and LYVE-1, both in complex with a hyaluronan (HA) tetrasaccharide (HA4) were performed in triplicate. Analysis of the root mean square deviations (RMSD) of CD44 (Fig. 2.24 A), LYVE-1 (Fig. 2.24 B) and HA4 in complex with either receptor (Fig. 2.24 C,D), shows that overall there is little conformational change in either the receptor or the ligand over the course of the simulations. The feature of note for the receptor RMSDs is that there is a relatively large conformational change (~3 Å) in the LYVE-1 HABD for the final 40 ns of run 2. However, since there is no corresponding change in the conformation of the ligand,

it is likely that this change occurs in a region of the protein far from the HA binding site. Indeed, clustering based on the RMSD of the receptor Cα atoms indicates that there are some small changes in the structure of the C-terminal extension. A similar but far smaller (~1.4 Å) change in RMSD is seen for the backbone of CD44 and likewise the change is not accompanied by any conformational change in HA4S, so is likely in a region far from the HA binding site. The RMSD of HA4 was determined using a no-fit calculation, meaning that in each frame of the trajectory HA4S was not superimposed on top of its reference prior to calculation, to take into account rotations and translations relative to its receptor. In CD44, the RMSD values of HA4 remain close to 1.5 Å indicating no significant translation or rotation, whereas they fluctuate to higher RMSD values in complex with LYVE-1.

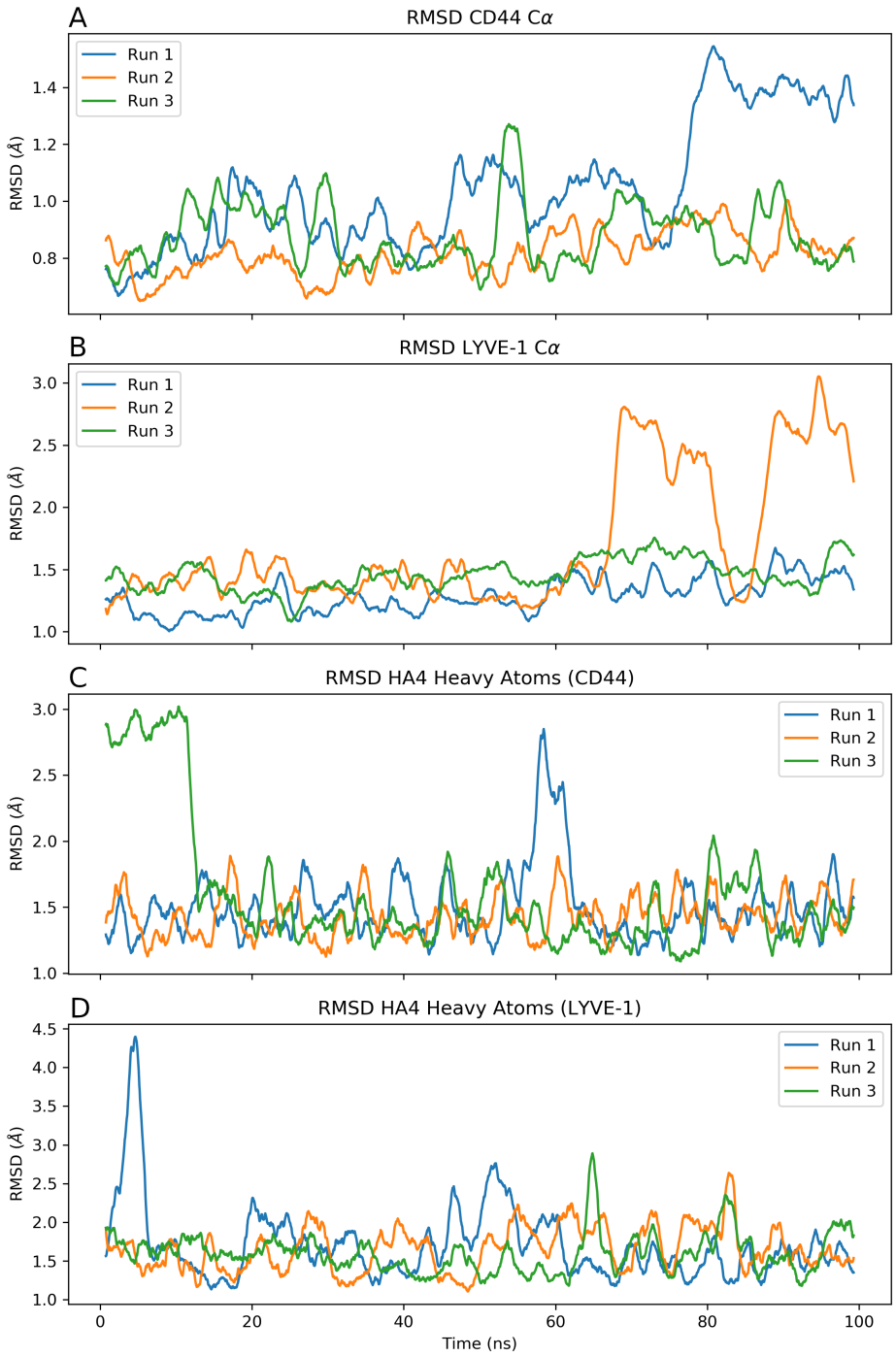

Figure 2.24: Root mean squared deviation (RMSD) plotted as a function of time. **A-B:** RMSD of CD44 (A) and LYVE-1 (B) Cα atoms. **C-D:** No-fit RMSD of HA4 heavy atoms in complex with CD44 (C) or LYVE-1 (D). For all plots, RMSD values are plotted for each of 3 independent replicas (blue, orange, green)

Analysis of the root mean squared fluctuations (RMSF) of HA4 in complex with CD44 or LYVE-1 reveal that generally the binding interaction in LYVE-1 is more flexible relative to CD44 (Fig. 2.25), in line with there being fewer hydrogen bonding interactions unlike that seen in CD44. In particular, the central GlcA2-GlcNAc3 pair, which is

the key binding motif in CD44, is significantly more flexible in LYVE-1 compared to CD44. Indeed, in the case of LYVE-1 there appears to be some translational motion of HA4 within the binding groove (Fig. 2.26), with the only consistent interactions being the methyl group of GlcNAc3 in the hydrophobic pocket of the binding groove, and a hydrogen bond between the backbone amide Gly107 and the carboxylate of GlcA4 - indeed GlcA4 exhibits the lowest RMSF of HA4S in the case of LYVE-1 (Fig. 2.25).

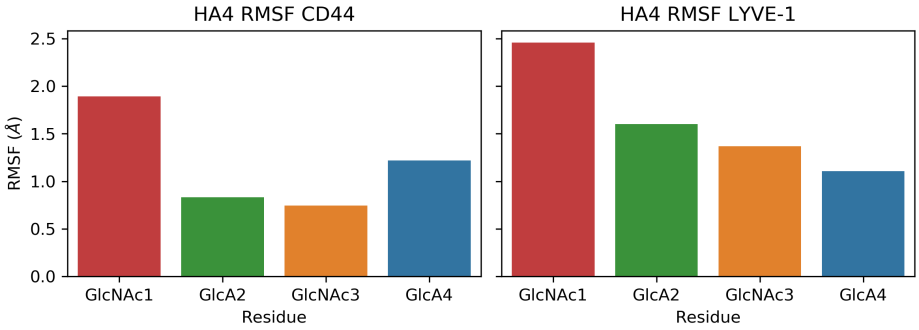

Figure 2.25: Root mean squared fluctuation (RMSF) of HA4 residues in complex with either CD44 (left) or LYVE-1 (right).

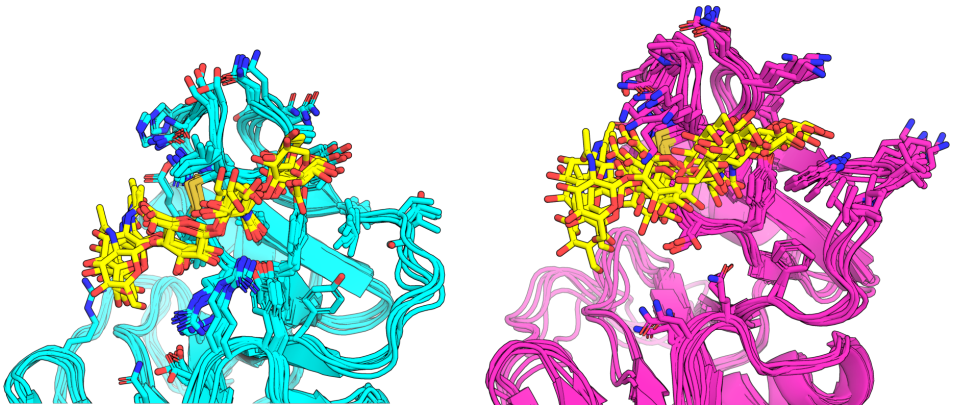

Figure 2.26: Cartoon representation of CD44 (cyan) and LYVE-1 (pink) in complex with the representative structures of HA4 (yellow) for each cluster calculated from k-means clustering. Representative structures chosen as those closest to the centroid of each cluster.

2.3.6 Adaptive Steered Molecular Dynamics of the unbinding of a Hyaluronan Tetrasaccharide from the human CD44 and LYVE-1 Hyaluronan Binding Domains

In order to deepen our understanding about the differences in dynamics and hence in affinity between the complexes of HA and CD44 or LYVE-1, we resorted to Adaptive Steered Molecular Dynamics (ASMD) simulations, which allow the ligand to be virtually pulled from the binding site, and the forces and interactions involved in the process can be monitored. ASMD simulations were performed using models of a hyaluronan tetrasaccharide (HA4) in complex with either the CD44 or LYVE-1 HABDs. In the case of CD44, a large initial force of approximately 400 pN was required to break the majority of hydrogen bonds known to be key in forming the complex of HA and CD44 (Fig. 2.27). After this point, little force is required to pull HA4 further from the binding site, except for a feature at approximately 10 Å, which corresponds to breaking the interaction with Arg41, the key residue known to be essential for HA binding to CD44. Interestingly, the conformation of Arg41 changes from the so-called high-affinity B-form, to the low-affinity A-form as HA4 is removed from the binding site (Fig. 2.28). This may be a mechanism by which HA is directed towards the binding site and would provide the first evidence for the role of Arg41 in affinity switching.

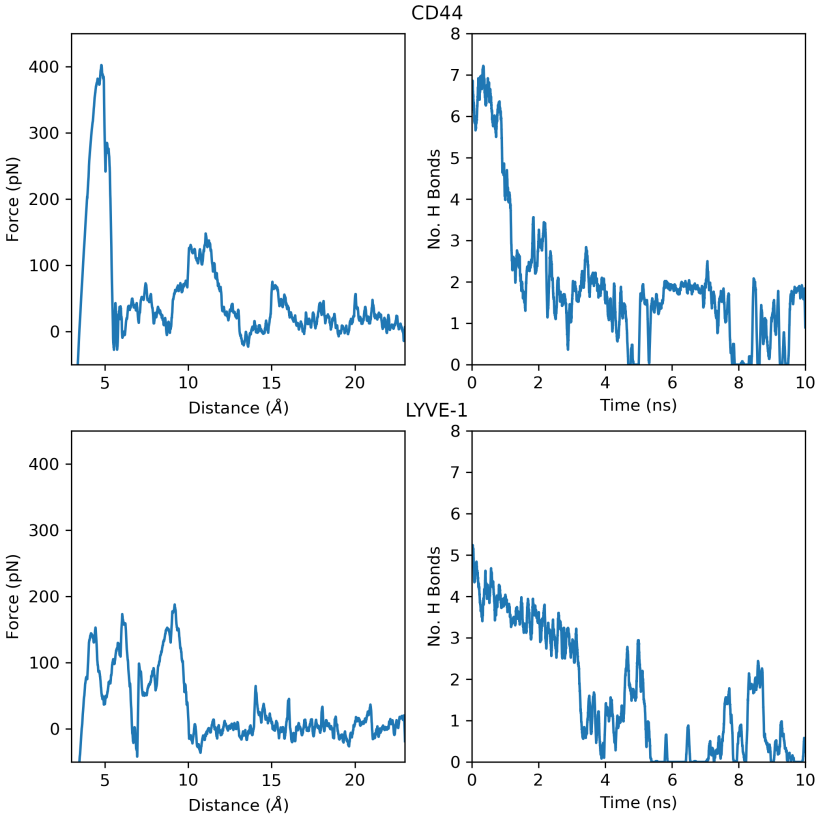

Figure 2.27: **Left:** Force as a function of distance for pulling HA4 from the binding site of CD44(top) and LYVE-1 (bottom). Distance measured as the the distance between the GlcNAc3 methyl group and the center of the hydrophobic binding pocket. **Right:** Number of hydrogen bonds formed between HA4 and CD44 (top) or LYVE-1 (right) as a function of time. HA4 was pulled at a rate of 2 Åns^{-1}.

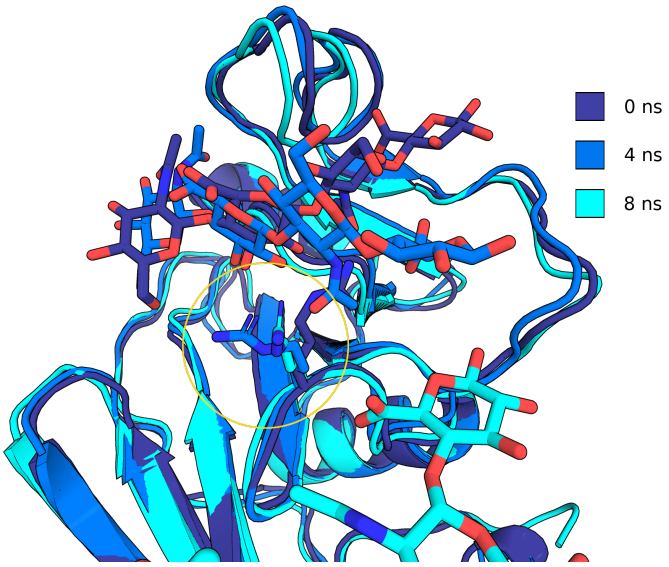

Figure 2.28: Cartoon representation of HA4 unbinding from the human CD44 HABD. Structures shown at 4 ns intervals (dark to light). The key residue Arg41 is shown (yellow circle) highlighting the conformational change from B to A form.

Conversely, the force required to remove HA4 from the binding site of LYVE-1 cannot be described by a single high-force event, but a general force of about 150 pN occurring over a distance of about 10 Å (Fig. 2.27), reflecting the difference in structure of the binding site (Fig. 2.29). This force is required to overcome the electrostatic interactions with Lys105 and Lys108 and the counted hydrogen bonding interactions come predominantly from the interaction of HA4 with these same residues.

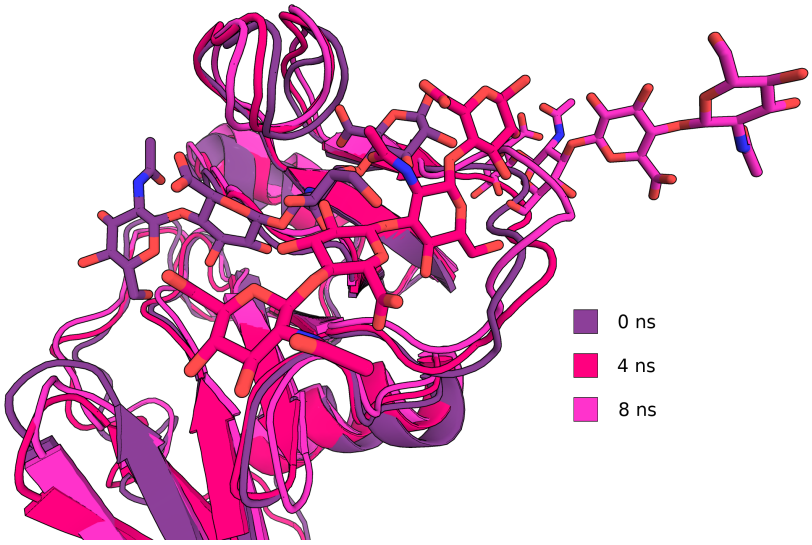

Figure 2.29: Cartoon representation of HA4 unbinding from the human LYVE-1 HABD. Structures shown at 4 ns intervals (dark to light).

2.4 Discussion

The lymphatic hyaluronan (HA) receptor LYVE-1 is of great interest due to its implication in various inflammatory diseases, but particularly in cancer.[233–235] Lymphangiogenesis is becoming better understood as a key player in cancer prognosis, and major route for metastasis is through the lymphatic system - both of which are mediated through the interaction of LYVE-1 on the surface of the lymphatic endothelia[196,238] and HA that is found ubiquitously in the extracellular matrix. Despite this, structural studies have been limited due to challenges in preparing samples of sufficient quality for X-ray crystallography or protein-observed NMR experiments (discussion with D. G. Jackson). In particular, LYVE-1 appears not to be expressed well in hosts that do not possess glycosylation machinery (for example, *E. coli*). Here, for the first time, the interaction between LYVE-1 and HA has been studied by ligand-based NMR in combination with molecular modelling, providing valuable molecular detail of the interaction. Furthermore, we have compared this interaction with that of CD44 and HA, and discuss the fundamental differences between the two protein ligand systems.

A homology model of the LYVE-1 HA binding domain (HABD) has

already been described in a previous study.[242] However, a model of the complex with HA was not produced from that study. Since that model was built from the X-ray crystal structure of the unliganded human CD44 HABD (PDB accession code: 1UUH), a new model was produced here, using the HA-bound X-ray crystal structure of the murine CD44 HABD (PDB accession code: 2JCR). This was chosen since there are some small conformational changes in the structures of the ligand-free and -bound forms of CD44, and it is likely that the HA-bound form of LYVE-1 more closely resembles the structure of HA-bound CD44.

The molecular models of LYVE-1 in complex with HA are validated by excellent agreement with theoretical STD NMR build up curves predicted by CORCMEA-ST (Fig. 2.16 and Fig. 2.23) and show that almost all the key hydrogen bonding interactions, known to be essential for the CD44-HA interaction, are missing in the LYVE-1-HA complex (Fig. 2.19). The only preserved interaction appears to be the burial of a GlcNAc methyl group within the hydrophobic binding pocket of the HA-binding groove. This is supported by the STD NMR data, which shows that only the GlcNAc3 methyl group of the synthetic HA tetrasaccharide (HA4S) receives a significant amount of saturation (Fig. 2.16). This is in contrast to to the STD NMR data for the CD44-HA4S interaction, in which saturation is spread across the central two residues (Glc2A-GlcNAc3) (Fig. 2.14).

Interestingly, the positively-charged residues, Lys105 and Lys108, are positioned in close proximity to the carboxylate groups of GlcA residues in the LYVE-1-HA complex (Fig. 2.19). Given that there are very few other interactions, it appears that the LYVE-1-HA interaction is predominantly electrostatic in nature, which agrees with the observation the the interaction is highly modulated by the ionic strength of the buffer.

Finally, the ASMD simulations performed here, which pull the HA ligand from the binding site, highlight the effect of these profound differences in binding site properties. In particular, the CD44-HA interaction requires a large force to initially to break the network of hydrogen bonding interactions that hold the central GlcA-GlcNAc disaccharide within the binding groove (Fig. 2.27 and Fig. 2.28), whereas, for LYVE-1, a weaker force is required over a longer distance

(Fig. 2.27 and Fig. 2.29). This may be explained by the different niches inhabited by these receptors, since CD44 is known to play an important role in leukocyte attachment to the endothelia in the high-flow vascular environment, whereas the flow-rate of the lymphatics is typically much gentler.

2.5 Conclusions

In this chapter a combination of STD NMR spectroscopy and molecular modelling has been used to generate an experimentally validated homology model of the LYVE-1/HA complex. Using this data, as well as long molecular dynamics simulations, the structure and dynamics of the LYVE-1-HA complex have been compared and contrasted to the CD44/HA complex, which has already been thoroughly studied. Overall the conclusions were:

1. The overall three dimensional structure of the LYVE-1 hyaluronan binding domain (HABD) is comparable to that of CD44 and also possesses a binding groove in which HA binds in a longitudinal fashion.

2. The properties of the binding groove in the LYVE-1 HABD differ dramatically to that in CD44. While CD44 interacts with HA through an ordered hydrogen bonding network presented through the sidechains surrounding the binding groove (as well as the hydrophobic pocket that recognises the N-acetyl group of GlcNAc), the binding groove of LYVE-1 is largely missing these interactions and instead appears to predominantly utilise electrostatic attraction (although the hydrophobic pocket is still present).

3. The LYVE-1/HA interaction appears to be more transient and requires less force to break compared to the CD44/HA interaction. This may be useful in each of the receptors' native environments, since the high flow in the vascular endothelium requires a tighter interaction for adhesion (CD44) compared to the low flow of the lymphatic endothelium (LYVE-1).

Chapter 3

Understanding Ligand Recognition by *Ps*LBP

3.1 Introduction

The synthesis of protein and DNA is a highly reproducible process due to their template driven mechanism - new DNA is synthesised based on complementary base pairing to an existing strand, and proteins are synthesised by translation of mRNA in the ribosome. Both of these processes are tightly controlled and linear in nature - that is, each nucleotide base/amino acid is added sequentially to the growing chain. The result is a homogeneous sample in which, assuming there are no errors or subsequent modifications, one may expect every molecule to be identical.

The same cannot be said of glycan synthesis. Instead, the endoplasmic reticulum (ER) and Golgi are lined with various glycan processing enzymes, that add (glycosyltransferase) or remove (glycosidase) moieties from the glycan chain, each with their own specificity.[259] Since the concentration of these enzymes varies across cell-type and specific environmental conditions of the cell, so too can the structure of the produced glycans vary considerably.[260] This, in part, explains why the field of glycobiology for a long time has fallen behind its counterparts in the study of DNA and proteins.

For the function of glycans and their physiological effects to be fully understood, it will be necessary to produce homogeneous glycan samples. Furthermore, since glycans and glycan-binding proteins are found ubiquitously and play such an important role in functions such as cell mobility and communication,[261] many research groups are exploiting this in order to develop novel diagnostics[262] and therapeutics,[263] utilising carbohydrates to recognise specific cell

types and, in some cases, deliver a therapeutic payload.[264] All of these again call for reproducible synthesis of their carbohydrate moieties.

Unfortunately, the chemical synthesis of carbohydrates is incredibly complex and costly, owing predominantly to the large number of stereocenters and reactive functional groups of the monosaccharide precursors. Typically such reactions take place over very many steps and suffer from poor yields.[265] Therefore recent efforts have focussed on enzymatic synthesis of carbohydrates, which benefits from the innate highly stereo- and regiospecific nature of carbohydrate-active enzymes.[95]

3.1.1 Enzymatic synthesis of carbohydrate derivatives

The vast majority of studies into enzymatic synthesis of carbohydrates has focussed on glycosyltransferases, which catalyse the formation of glycosidic bonds between two carbohydrate moieties, typically using a sugar-nucleotide donor (Fig. 3.1 top).[266] However, such donors tend to be expensive. Furthermore, many glycosyltransferases suffer from low stability, and their narrow substrate specificity limits their usefulness in broad academic and industrial applications.

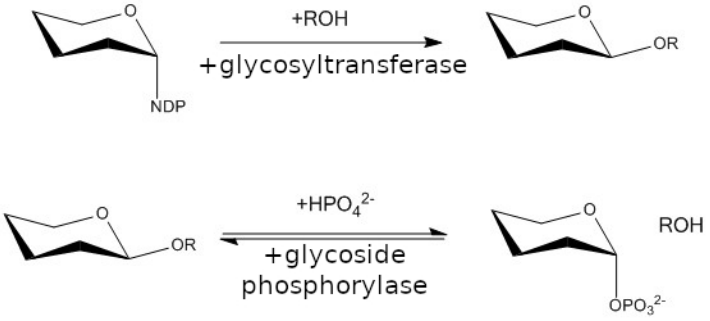

Figure 3.1: Simplified typical mechanism of glycotransferases (top) and glycoside phosphorylases (bottom). NDP = nucleotide diphosphate.

However, more recently, another class of enzymes has emerged as promising candidates for enzymatic synthesis of carbohydrates. The glycoside phosphorylases catalyse the reversible cleavage of glycosidic linkages by transferring inorganic phosphate to the non-reducing sugar

(Fig. 3.1 bottom).[267] The reverse reaction, which results from performing the reaction in the presence of an excess of sugar 1-phosphate as the donor substrate and a corresponding mono- or oligosaccharide as an acceptor substrate, may be used to instead synthesise glycosidic linkages. Such a reaction is synthetically attractive, since the sugar 1-phosphate donors are typically far more accessible and stable than the sugar-nucleotides utilised by glycosyltransferases.

In addition, glycoside phosphorylases typically exhibit far broader acceptor substrate specificity, making them synthetically useful for a large number of applications. For example, cellobiose (Glcβ-1-4Glc) phosphorylase has been shown to be capable of using mannose,[268] xylose and simple alcohols as acceptor substrates,[269] while cellodextrin (Glcβ-1-4Glc oligosaccharides) phosphorylase has been shown to accept both G1P and Gal1P as donor substrates.[270]

Many synthetic monosaccharides can be prepared to contain chemical modifications that may give them more favourable properties compared to their natural counterparts, such as higher binding affinity to specific receptor targets, better bioavailability, or they may be used as markers.[271] Some glycoside phosphorylases have been shown to tolerate such modifications, allowing for a relatively facile pathway to producing oligosaccharides containing such modifications, such as may be utilised in a glycomimetic drug or nanoparticle-based therapeutic/-diagnostic. For example, fluoro-sugars, in which one or more of the hydroxyl groups is replaced with fluorine, seem to be well tolerated by many glucoside phosphorylases.[272] Furthermore, sucrose phosphorylase has been used to transfer a number of non-carbohydrate moieties to glucose, including hydroquinone, cytosine mono phosphate and benzoic acid.[273]

3.1.2 Laminaribiose phosphorylase from Paenibacilus sp. (*Ps*LBP)

Recently, the group of Rob Field (John Innes Centre, Norwich, UK) demonstrated that the laminaribiose (Glcβ-1-3Glc) phosphorylase from *Paenibacillus sp.* (*Ps*LBP) is capable of accepting mannose-1-phosphate (M1P) as a non-cognate donor substrate, leading to the facile synthesis of the unnatural disaccharide Manβ-1-3Glc.[274]

Furthermore, they reported the crystal structure of *Ps*LBP in complex with G1P (PDB ID: 6GH2) and M1P (PDB ID: 6GH3). These are only the second reported structures of a glycoside phosphorylase in complex with its donor substrate, after a sophorose (Glcβ-1-2 Glc) phosphorylase from *Lachnoclostridium phytofermentans* in complex with G1P (PDB ID: 5H42).[275]

Our understanding of donor substrate recognition by glycoside phosphorylases is therefore severely limited and warrants further study. Here we perform STD NMR spectroscopy experiments coupled with molecular modelling to further understand the interaction between *Ps*LBP and its substrates.

3.1.3 Objectives

The aim of this chapter is to provide experimental solution state information for the interaction of the glycoside phosphorylase *Ps*LBP and its substrates, both cognate and non-cognate, to complement and validate the model complexes derived from X-ray crystallography. Specifically the objectives are:

- Use quantitative STD NMR calculations to validate the X-ray crystallography derived models of the *Ps*LBP/G1P and *Ps*LBP/M1P complexes in the solution state.

- Use STD NMR epitope mapping to obtain structural information about the interaction between *Ps*LBP and the acceptor substrate, Glc, as well as the reaction products (reverse reaction) laminaribiose and Manβ-1-3Glc.

3.2 Materials and methods

3.2.1 NMR Spectroscopy

All samples were prepared in D_2O with [D_{11}]Tris (25 mM, pH 7.4) and contained final protein and ligand concentrations of 50 μM and

6 mM respectively (protein omitted in assignment experiments). All experiments were performed at 278 K on a Bruker Avance III 800 MHz spectrometer equipped with a 5 mm TXI 800 MHz H-C/N-D-05 Z BTO probe. All carbohydrate ligands were assigned based on 1D ^{1}H, ^{1}H-^{1}H COSY, ^{1}H-^{1}H TOCSY, ^{1}H-^{13}C HSQC and ^{1}H-^{1}H NOESY NMR experiments. For the TOCSY the mixing time was set to 80 ms and for the NOESY the mixing time was set to 1000 ms. STD NMR spectroscopy experiments were performed by using a train of 50 ms Gaussian pulses applied on the f2 channel at either 0.8 (on-resonance) or 40 ppm (off-resonance). A spoil sequence was used to destroy unwanted magnetisation and a spin lock was used to suppress protein signals (stddiff.3). The recycle delay (d1) was set to 5 s. The total saturation time and number of scans were selected according to Table 3.1. The measured STD intensities were fitted to Eqn 45 to calculate STD_0, which was then used to plot binding epitopes following normalisation against the largest STD_0 value.

Table 3.1: Total saturation time and number of scans selected for STD NMR spectroscopy measurements.

Saturation Time (s)	No. Scans	Saturation Time (s)	No. Scans
0.5	512	2	128
0.75	512	3	128
1	256	5	128
1.5	256		

3.2.2 Preparation of Molecular Models

Crystal structures were imported into Schrödinger Maestro and prepared with the Protein Preparation Wizard. All non-protein or non-ligand atoms were removed. Protons were then added to the model, using PROPKA to predict the protonation state of polar side chains at pH 7.[276] The hydrogen-bonding network was automatically optimised by allowing asparagine, glutamine, and histidine side chains to be flipped. The model was then minimised by using the OPLS3[120] force field and a heavy-atom convergence threshold of 0.3 Å. Because STD NMR spectroscopy experiments were performed in D_2O, the polar protons were removed from the ligand prior to CORCEMA-ST

analysis.[277]

3.2.3 CORCEMA-ST calculations

Protein chemical shifts were predicted by using the SHIFTX2[278] webserver, according to experimental conditions. All protein protons within 15 Å of the ligand were considered in the calculation. The instrument field strength, solvent type, ligand concentration, and protein concentration were set according to experimental values. The free and bound ligand correlation times were optimised to be 0.3 and 300 ns respectively, based on reasonable values for a monosaccharide binding to a 200 kDa protein. The non-specific leakage was also optimised to 0.8 Hz. The internal correlation time was set to 10 ps and the methyl-X order parameter was set to 0.85, according to previously published values.[277] All protein protons with resonances between 0.6 and 1 ppm were considered to be instantaneously saturated to account for line broadening. For G1P, the equilibrium constant and k_{on} were optimised to 25,000 M^{-1} and 10^5 M^{-1} s^{-1} respectively. For M1P, the equilibrium constant was reduced to 16,000 M^{-1}. Both values were in agreement with the affinities typically observed for carbohydrate-binding proteins.

3.3 Results

3.3.1 STD NMR and CORCEMA-ST of G1P and M1P binding to *Ps*LBP

Crystal structures of G1P and the non-cognate donor substrate M1P in complex with *Ps*LBP were obtained by the group of Prof. Rob Field.[274] However, such models only provide a snapshot of the complex in the solid phase, whereas ideally the complex should be studied in the solution phase, with access to dynamic information. Therefore we studied the interaction of these substrates with *Ps*LBP by STD NMR.

Initially experiments were conducted at 298 K. However, at this tem-

perature, G1P was rapidly converted to laminaribiose (LB) (Fig. 3.2). This shows that the enzyme must be capable of hydrolysing G1P to Glc and inorganic phosphate; the pool of Glc can then be used as an acceptor for reaction with the remaining G1P in order to form LB. Therefore, it was necessary to reduce the temperature to 278 K before it was possible to conduct the full STD NMR build-up experiment without any detectable hydrolysis of G1P (Fig. 3.3).

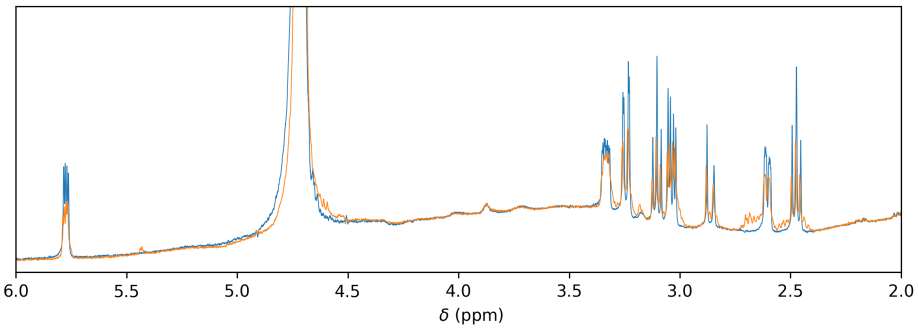

Figure 3.2: Spectra of G1P in the presence of *Ps*LBP before (blue) and after (orange) 18 hours incubation at 298 K. Significant degradation of G1P can be seen by the reduction in intensity of G1P peaks, as well as the appearance of new peaks.

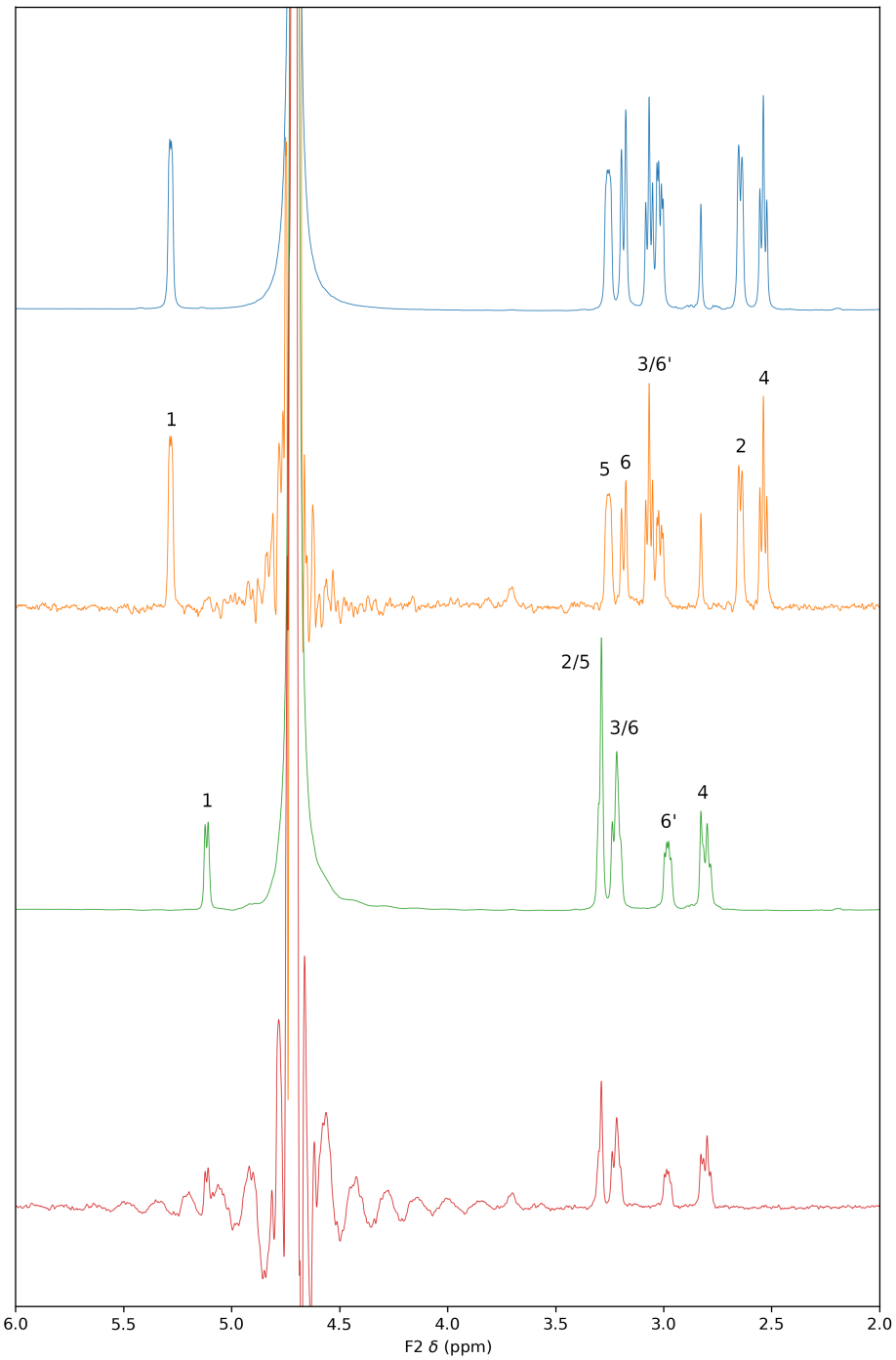

Figure 3.3: Reference and STD NMR difference spectra of G1P (reference: blue, difference: orange) and mannose-1-phosphate (reference: green, difference: red) each in the presenence of *Ps*LBP. Spectra recorded at 800 MHz, 278 K with a saturation time of 2 s. Difference spectra magnified 20x.

The binding epitope map was obtained from the initial slopes of the STD build-up curves of G1P and shows a highly uniform epitope, with relative STD intensities ranging between 84%-100% (Fig. 3.5). Within this epitope the exocyclic H6 protons exhibit the strongest STD intensities, followed by H4, whilst H1 and H2 exhibit the weakest

STD intensities indicating that the most intimate spatial contacts with the enzyme are made by the ligand area encompassing C4 and C6 (Table 3.2). This is in good agreement with the crystal structure of the G1P/*Ps*LBP complex, in which H4 and H6 face the surface of the interior binding cavity. In particular, the H6 protons are in very close proximity to the aromatic sidechains of Trp524 and Phe737 (Fig. 3.4). Conversely, H1 and H2 face the open entrance to the cavity and so receive less saturation.

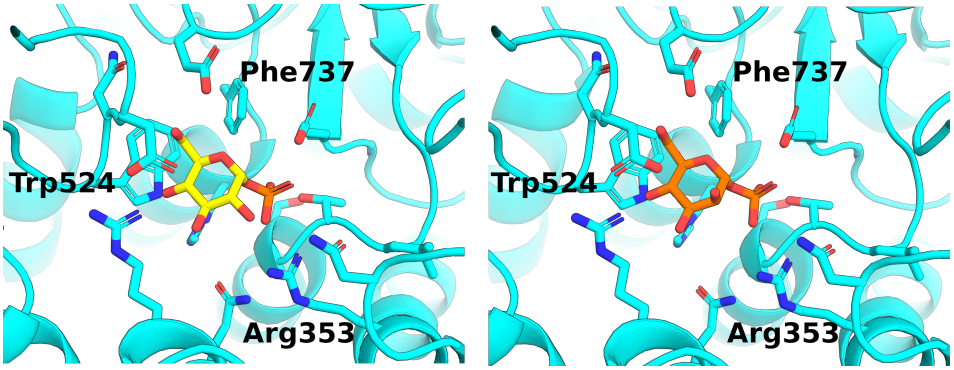

Figure 3.4: Crystal structures of G1P (yellow) and M1P (cyan) in complex with *Ps*LBP.

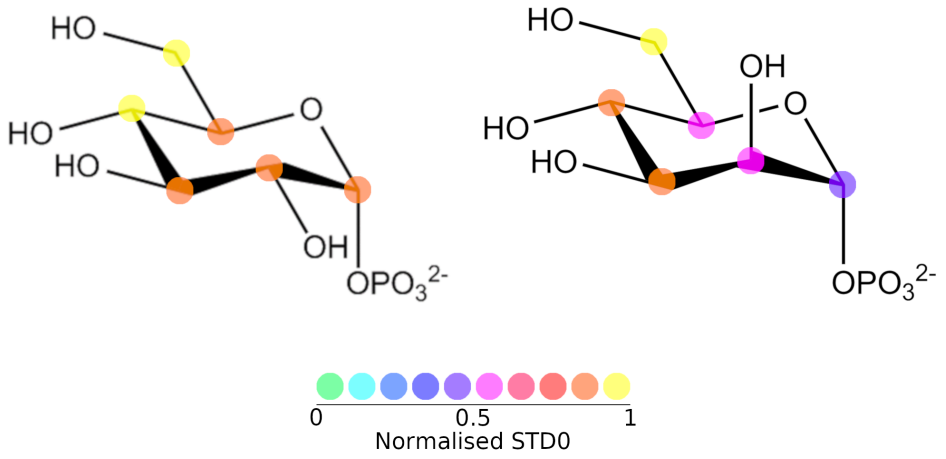

Figure 3.5: Binding epitope maps of G1P (left) and M1P (right), both in the presence of *Ps*LBP. Colours represent the normalised STD values for proton resonances at the indicated positions (low - high, cold - hot).

Table 3.2: Experimental STD intensities measured for G1P binding to *Ps*LBP. *Normalised against the H6' proton.

Proton	δ (ppm)	k_{sat} (% s^{-1})	STD_{max} (%)	STD_0 (%)	Norm.*
H1	5.13	0.49	3.88	1.89	83.8
H2	3.15	0.62	3.03	1.88	83.5
H3	3.46	0.47	4.18	1.95	86.9

Proton	δ (ppm)	k_{sat} (% s^{-1})	STD_{max} (%)	STD_0 (%)	Norm.*
H4	3.07	0.58	3.62	2.08	92.6
H5	3.60	0.78	2.56	2.01	89.1
H6	3.55	1.34	1.63	2.20	97.6
H6'	3.42	1.26	1.78	2.25	100

To confirm the agreement between the solid phase crystal structure and the solution state STD NMR data, CORCEMA-ST[277] was used to predict theoretical STD NMR intensities based on the 3D coordinates of the G1P/*Ps*LBP complex. After optimisation, the NOE R-factor between the experimental and predicted STD NMR datasets was calculated to be 0.09, indicating excellent agreement between the crystal structure and STD NMR data (Fig. 3.6). It should be noted that, while most of the CORCEMA-ST parameters are derived from experimental conditions, in the case of unknown parameters, they can be optimised iteratively to achieve a final working value. For example, the dissociation constant of the G1P/*Ps*LBP complex was unknown, but was optimised for CORCEMA-ST to be 40 μM, which falls within the typical K_d range of protein-carbohydrate interactions.[279,280]

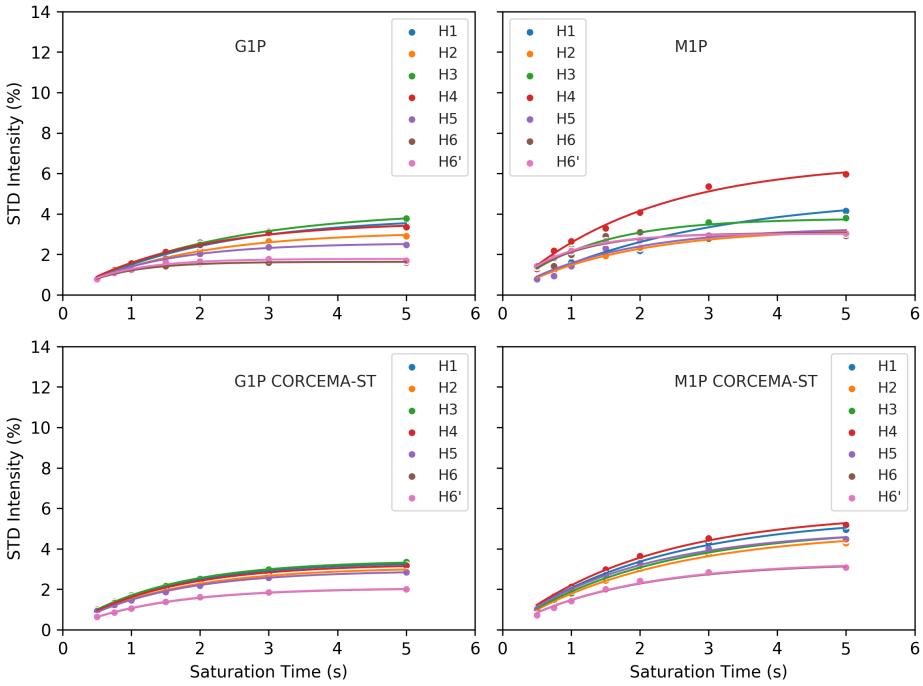

Figure 3.6: Experimentally determined STD build-up curves for glucose-1-phosphate (G1P) and mannose-1-phosphate (M1P) binding to *Ps*LBP (top) and their respective CORCEMA-ST calculated STD intensities (bottom). The NOE R-factors (RNOEs) between the experimental and calculated data for G1P and M1P are 0.09 and 0.23 respectively. For experimental data, circles show observed STD intensities, whereas curves are determined from least squares fitting to Eqn. 45.

In contrast, the binding epitope map of M1P exhibits a far wider range of relative STD intensities, with the weakest intensity, coming from H1, having an intensity of 48% (Fig. 3.5b, Table 3.3). This is indicative of a shorter residence time of M1P in the bound state compared to G1P, since the saturation transferred from *Ps*LBP would be unable to spread as uniformly in a shorter period of time. This is understandable since M1P is not the native substrate, and agrees well with the observation that the catalytic turnover rate is reduced approximately tenfold for M1P compared to G1P.[274]

In a similar manner as G1P, the most intense STD intensities from the M1P binding epitope map come from the H6, H4 and H3 protons (Table 3.3). This suggests that M1P does indeed bind in a similar binding mode to G1P, which is in agreement with the crystal structures (Fig. 3.4). The H2 proton of M1P shows particularly weak STD intensity (55%), especially in comparison to G1P (84%). A difference here is understandable, since the chirality of this stereocentre is inverted such that H2 is equatorial in M1P, compared to axial in G1P, although such

a drastic difference is probably related to the shorter residence time in the bound state.

Table 3.3: Experimental STD intensities measured for M1P binding to *Ps*LBP. *Normalised against the H6' proton.

Proton	δ (ppm)	k_{sat} (% s^{-1})	STD_{max} (%)	STD_0 (%)	Norm.*
H1	5.00	0.38	4.94	1.85	48.2
H2	3.63	0.59	3.29	1.94	50.5
H3	3.57	0.84	3.78	3.17	82.5
H4	3.26	0.49	6.61	3.26	84.9
H5	3.64	0.63	3.34	2.11	55.0
H6	3.58	1.17	3.06	3.59	93.3
H6'	3.40	1.27	3.02	3.84	100

As for G1P, CORCEMA-ST was used to analyse the agreement between the crystal complex of M1P and *Ps*LBP and the associated STD NMR data (Fig. 3.6). All parameters were kept identical to those optimised for G1P, with the exception of the complex dissociation constant. This was optimised to 63 μM, which is in agreement with the proposition that M1P binds to *Ps*LBP with lower affinity than G1P. Despite these optimisations, the best obtainable RNOE factor for the M1P/*Ps*LBP complex was 0.23, which is still considered a very good fit between the crystal structure and STD NMR data. We consider RNOE factors below 0.3 to be in good agreement, whilst many other groups accept values below 0.5.

One particularly interesting difference between the two binding epitope maps is that of H5. In the crystal structures, the position of H5 is identical in both the G1P and M1P complexes and it points towards the aromatic sidechain of Trp524[274] (Fig. 3.4). However, for G1P the relative STD intensity is 89%, whereas it is 55% for M1P. It may be that, in solution, the binding mode of M1P may not be as stable as that of G1P, which would agree with the observation in the crystal structure that Arg353 coordinates O2 of G1P, an interaction that is not possible in M1P since O2 is axial (Fig. 3.4).

With regard to absolute STD intensities, on average those of M1P are higher than those of G1P, with the H6 showing absolute initial growth rates of STD build up of 2.2% s^{-1} and 3.6% s^{-1} for G1P and

M1P respectively. This is indicative of each substrate having a different binding affinity for *Ps*LBP since different STD intensities here would be indicative of different fraction of ligand bound. From other experiments we anticipate the binding affinity of M1P to be lower than G1P, and since this corresponds to an increase in STD intensity, this would suggest that the affinity of the substrates for *Ps*LBP was towards the tighter end (μM) of the detectable affinity range by STD NMR, in agreement with the optimised K_d values obtained from CORCEMA-ST calculations.

This study highlights a very important aspect of the interpretation of STD NMR data. Here the molecular recognition of two different ligands by the same enzyme have been studied by both X-ray crystallography and STD NMR spectroscopy. In this way, the results demonstrate that a simplistic analysis based exclusively on the binding epitopes observed by STD NMR (Fig. 3.5) might lead to the wrong conclusion that the ligands are interacting with the enzyme with different binding modes. It is only after a quantitative analysis, using CORCEMA-ST on the X-ray derived 3D molecular models of the complexes, when the experimental spectroscopic data confirm that indeed in solution both ligands are binding in similar binding mode, as shown in the X-ray structures. Hence, the differences in the binding epitopes are, in this particular case, clearly ascribable to their differences in residence time in the bound state, i.e. in affinity, and not to differences in their interaction modes. Thus, this study emphasises the importance of carrying out quantitative analysis of the STD NMR results whenever it is feasible, using CORCEMA-ST, in order to avoid misinterpretations in comparative binding analysis, as the kinetics of saturation transfer and its relationship to residence time in the bound state are important parameters to take into consideration in structural analysis.

3.3.2 STD NMR of Glc binding to *Ps*LBP

In solution, Glc is spontaneously exchanging between its α and β-anomers such that both species are present in solution (anomeric equilibrium). This precludes such a detailed study as for G1P and M1P, for which the configuration is locked by the presence of the phosphate, since overlapping signals exist from both the α and β-anomers. Nev-

ertheless, the 1D ^{1}H NMR spectrum showed sufficient resolution to measure STD intensities for many of the proton resonances in both the α and β anomers individually.

From the reference NMR spectra, it was clear that both the α and β-anomers of Glc exist in solution in comparable proportions (Fig. 3.7). Despite this, the STD intensity from the α-anomer was very weak (Fig. 3.7, Fig. 3.8). For example, the initial growth rate of STD intensity build up for the anomeric proton of αGlc is about 12% that of βGlc (Table 3.4). This suggests that *Ps*LBP is selective for the β-anomer, which is intriguing since the anomeric configuration of Glc would not impact on the configuration of the enzymatically-formed glycosidic linkage. The binding epitope map of the α-anomer of Glc will not be discussed since the STD intensities were too weak to measure with enough accuracy.

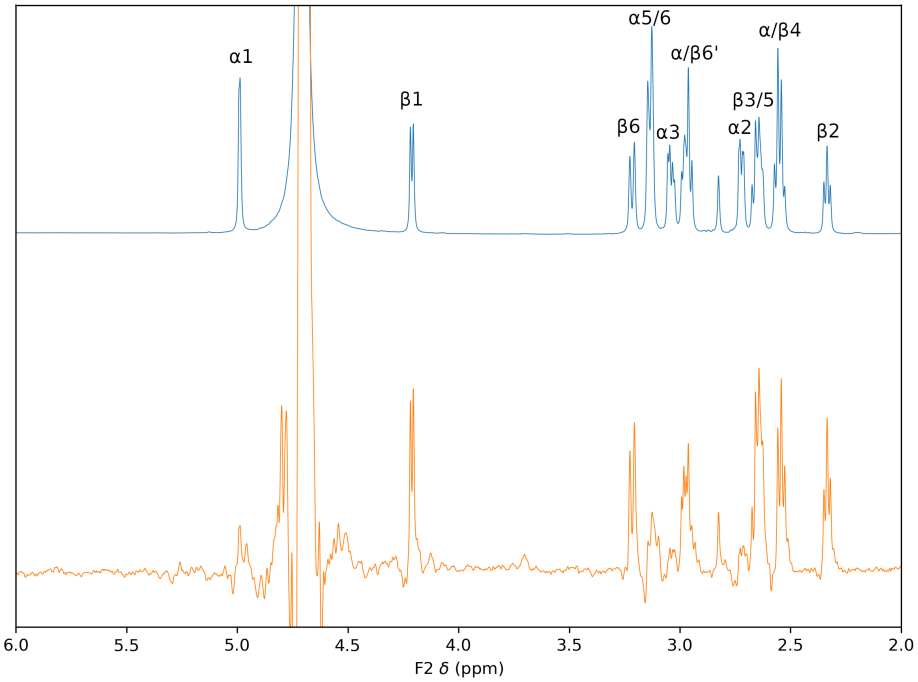

Figure 3.7: Reference and STD NMR difference spectra of glucose (reference: blue, difference: orange) in the presence of *Ps*LBP. Spectra recorded at 800 MHz, 278 K with a saturation time of 2 s. Difference spectrum magnified 50x.

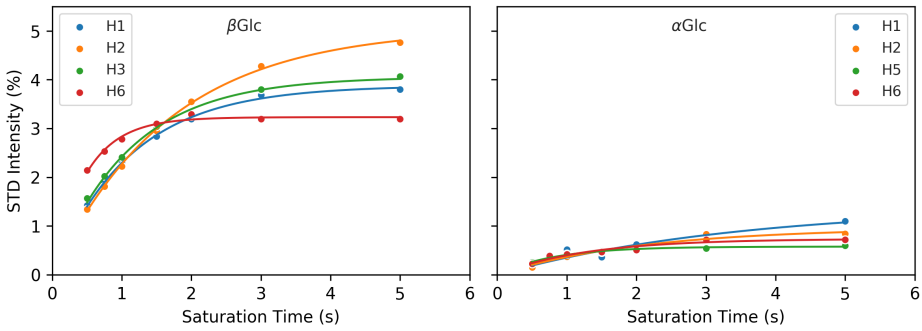

Figure 3.8: Experimentally determined STD build-up curves for β-glucose (left) and α-glucose (right) binding to *Ps*LBP. Circles show observed STD intensities, whereas curves are determined from least squares fitting to Eqn. 45.

Table 3.4: Experimental STD intensities measured for Glc binding to *Ps*LBP. Values of STD_0 for α Glc (left) and β Glc (right) protons shown. $^{\dagger}/^{\ddagger}$ STD intensity measured from same peak. *Normalised against the H6 proton of αGlc and βGlc respectively.

Proton	δ (ppm)	k_{sat} (% s^{-1})	STD_{max} (%)	STD_0 (%)	Norm.*	δ (ppm)	k_{sat} (% s^{-1})	STD_{max} (%)	STD_0 (%)	Norm.*
H1	5.14	0.29	1.41	0.405	57.3	4.56	0.89	3.88	3.47	51.2
H2	3.45	0.47	0.96	0.457	64.6	3.16	0.60	5.06	3.01	44.5
H3	3.63	-	-	-	-	3.39	$0.91^{\ddagger}$	$4.06^{\ddagger}$	$3.68^{\ddagger}$	$54.3^{\ddagger}$
H4	3.32	-	-	-	-	3.32	-	-	-	-
H5	3.75	$1.23^{\dagger}$	$0.58^{\dagger}$	$0.707^{\dagger}$	$100^{\dagger}$	3.37	$0.91^{\ddagger}$	$4.06^{\ddagger}$	$3.68^{\ddagger}$	$54.3^{\ddagger}$
H6	3.75	$1.23^{\dagger}$	$0.58^{\dagger}$	$0.707^{\dagger}$	$100^{\dagger}$	3.81	2.10	3.23	6.77	100
H6'	3.68	0.78	0.74	0.576	81.5	3.64	-	-	-	-

The binding epitope map of βGlc reveals that the strongest STD intensities again come from the H6 protons (Fig. 3.9). However, moderate STD intensities are observed for H1, H3 and H5, whilst the STD intensity of H2 is weak. This suggests a fundamentally different binding mode of βGlc compared to G1P and M1P. This is in agreement with its role as the acceptor substrate, since it would be expected that βGlc would bind to a different, but nearby, binding subsite.

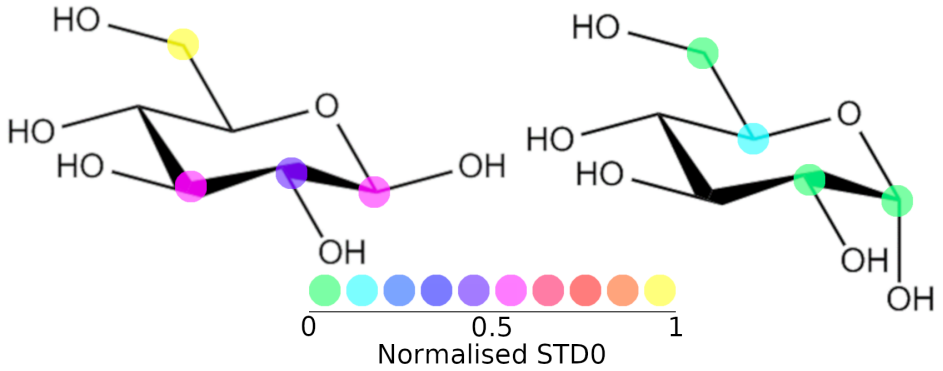

Figure 3.9: Binding epitope maps of β-glucose (left) and α-glucose (right), in the presence of *Ps*LBP. Colours represent the normalised STD values for proton resonances at the indicated positions (low - high, cold - hot).

These data also suggest two other points. Firstly, the donor subsite occupied by G1P and M1P must require a hexose-1-phosphate, otherwise it would be expected that αGlc would bind well. Finally, the structure of the acceptor subsite must be such that α, with its axial C1-hydroxyl group, is unable to bind with good affinity, perhaps due to steric interactions.

3.3.3 STD NMR of laminaribiose and Manβ-1-3-Glc binding to *Ps*LBP

The products of the *Ps*LBP-catalysed reaction between either G1P or M1P and Glc are LB and Manβ-1-3-Glc respectively. Like Glc, the reducing termini of LB and Manβ-1-3-Glc may spontaneously exchange between the α and β-anomers, leading to two distinct species in solution. The NMR spectra are further convoluted by the fact that we are now dealing with disaccharides. In both cases, differences in anomeric configuration lead to unique resolvable chemical shifts for each proton of the reducing sugar, as well as the anomeric proton of the non-reducing sugar.

Similar to Glc, the α and β configurations of LB and Manβ-1-3-Glc are present in comparable concentrations in solution (Fig. 3.10), yet the STD intensities from αLB and Manβ-1-3-Glcα are very weak in comparison to their respective β-configurations (Fig. 3.11). This shows again that *Ps*LBP is selective for the β-configuration of the reducing sugar and this reducing sugar seems to occupy the same subsite as the monosaccharide Glc, which would agree with the known reaction. Furthermore, since the non-reducing sugar, which retains the same configuration in both species, exhibits negligible STD intensities for the α-configurations of both LB and Manβ-1-3-Glc, it appears that the reducing sugar is most important for interaction of the disacharides with *Ps*LBP. This also agrees with the hypothesis that the subsite specific for G1P and M1P, which is also expected to be occupied by the non-reducing sugar of LB and Manβ-1-3-Glc, facilitates the interaction predominantly through interaction with the phosphate group of the phosphorylated donor substrates.

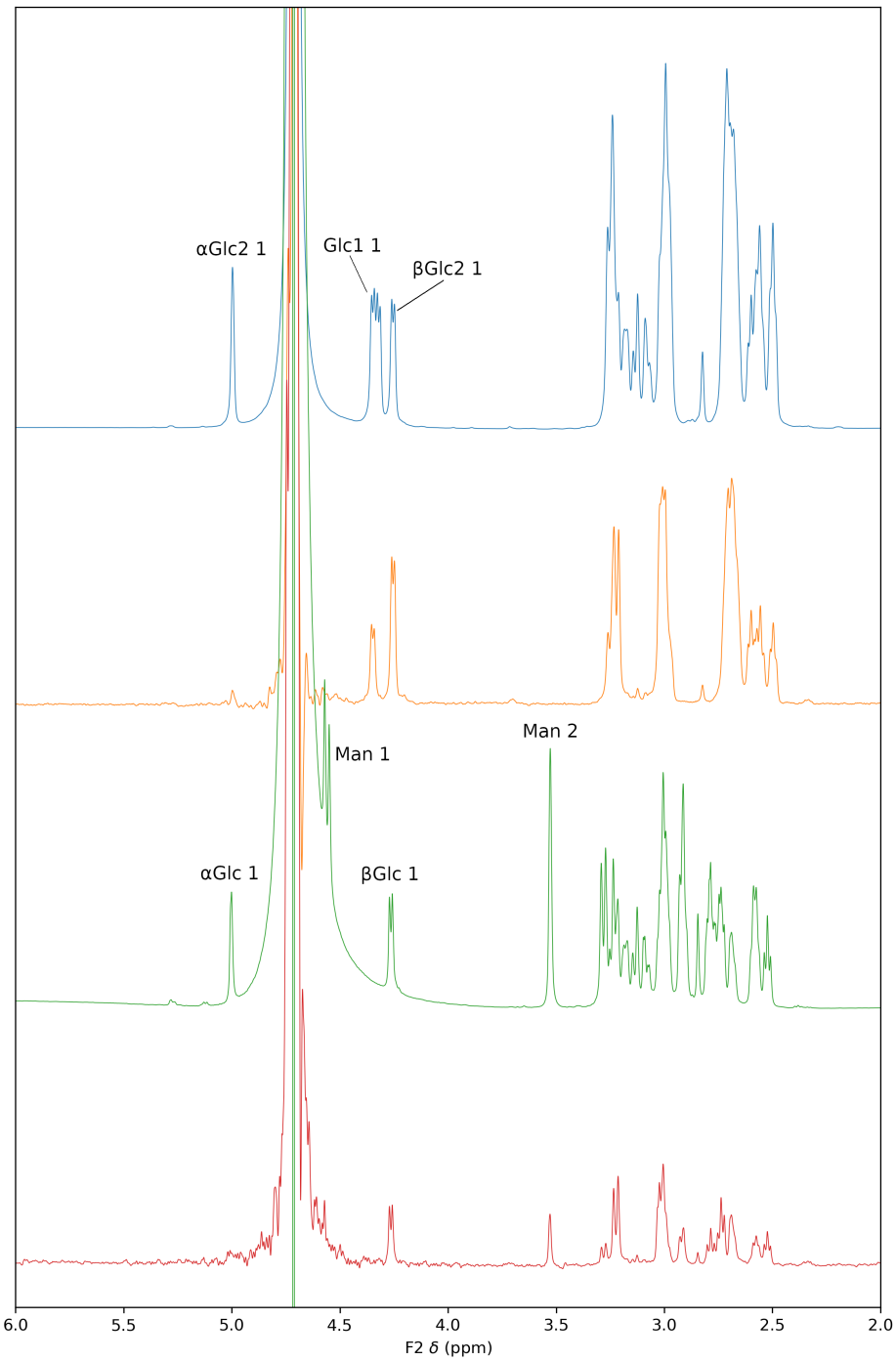

Figure 3.10: Reference and STD NMR difference spectra of laminaribiose (reference: blue, difference: orange) and Manβ-1-3-Glc (reference: green, difference: red) each in the presence of *Ps*LBP. Spectra recorded at 800 MHz, 278 K with a saturation time of 2 s. Difference spectra magnified 10x.

Table 3.5: Experimental STD intensities of LB Glc1 measured for LB binding to *Ps*LBP. *STD_0 normalised against H6 of Glc2 in βLB. †Values correspond to Glc1 H1 in αLB and βLB respectively.

Proton	δ (ppm)	k_{sat} (% s^{-1})	STD_{max} (%)	STD_0 (%)	Norm.*
H1	4.39/4.41†	0.76/0.59†	6.98/1.26†	0.740/5.33†	3.67/26.5†
H2	3.04	0.51	6.07	3.10	15.4
H3	3.20	-	-	-	-
H4	3.16	-	-	-	-
H5	3.09	0.63	6.69	4.23	21.0
H6	3.60	1.24	3.70	4.59	22.8
H6'	3.39	-	-	-	-

Table 3.6: Experimental STD intensities of LB Glc2 measured for LB binding to *Ps*LBP. The columns show STD intensities for Glc2α (left) and Glc2β (right). *STD_0 normalised against H6 of Glc2 in βLB.

Proton	δ (ppm)	k_{sat} (% s^{-1})	STD_{max} (%)	STD_0 (%)	Norm.*	δ (ppm)	k_{sat} (% s^{-1})	STD_{max} (%)	STD_0 (%)	Norm.*
H1	4.91	0.47	1.52	0.714	3.54	4.35	0.77	14.3	11.1	54.9
H2	3.39	-	-	-	-	3.11	0.53	10.5	5.57	27.6
H3	3.58	-	-	-	-	3.39	-	-	-	-
H4	3.54	0.84	1.48	1.24	6.14	3.20	-	-	-	-
H5	3.19	-	-	-	-	3.18	0.86	9.10	7.79	38.7
H6	3.50	1.14	1.22	1.39	6.92	3.57	1.46	13.8	20.1	100
H6'	3.46	1.18	1.18	1.40	6.93	3.41	1.26	12.6	15.8	78.7

Table 3.7: Experimental STD intensities measured of Manβ-1-3-Glc Man for Manβ-1-3-Glc binding to *Ps*LBP. *STD_0 normalised against H6' of Glc in Manβ-1-3Glcβ. †Values correspond to Man1 H1 in Manβ-1-3Glcα and Manβ-1-3Glcβ respectively.

Proton	δ (ppm)	k_{sat} (% s^{-1})	STD_{max} (%)	STD_0 (%)	Norm.*
H1	4.55/4.57†	0.54†	3.64†	0.61/1.95†	4.85/15.6†
H2	3.79	0.60	2.85	1.70	13.6
H3	3.33	0.62	2.85	1.76	14.1
H4	3.23	0.73	3.32	2.44	20.0
H5	3.08	0.75	3.02	2.27	18.1
H6	3.56	1.38	5.47	7.52	60.2
H6'	3.53	1.45	8.64	12.5	100

Table 3.8: Experimental STD intensities measured of Manβ-1-3-Glc Glc for Manβ-1-3-Glc binding to *Ps*LBP. The columns show STD intensities for Glcα (left) and Glcβ (right). *STD_0 normalised against H6' of Glc2 in Manβ-1-3Glcβ.

Proton	δ (ppm)	k_{sat} (% s^{-1})	STD_{max} (%)	STD_0 (%)	Norm.*	δ (ppm)	k_{sat} (% s^{-1})	STD_{max} (%)	STD_0 (%)	Norm.*
H1	4.90	0.75	1.33	1.00	8.01	4.34	0.55	7.75	4.28	34.3
H2	3.32	0.62	2.29	1.41	11.3	3.04	0.47	5.70	2.66	21.3
H3	3.57	-	-	-	-	3.39	1.05	4.91	5.14	41.1
H4	3.19	-	-	-	-	3.19	0.87	6.61	5.75	46.0
H5	3.53	1.03	0.88	0.903	7.23	3.16	0.76	8.20	6.24	49.9
H6	3.48	1.43	0.86	1.23	9.86	3.40	1.12	7.80	9.34	74.7
H6'	-	-	-	-	-	-	-	-	-	-

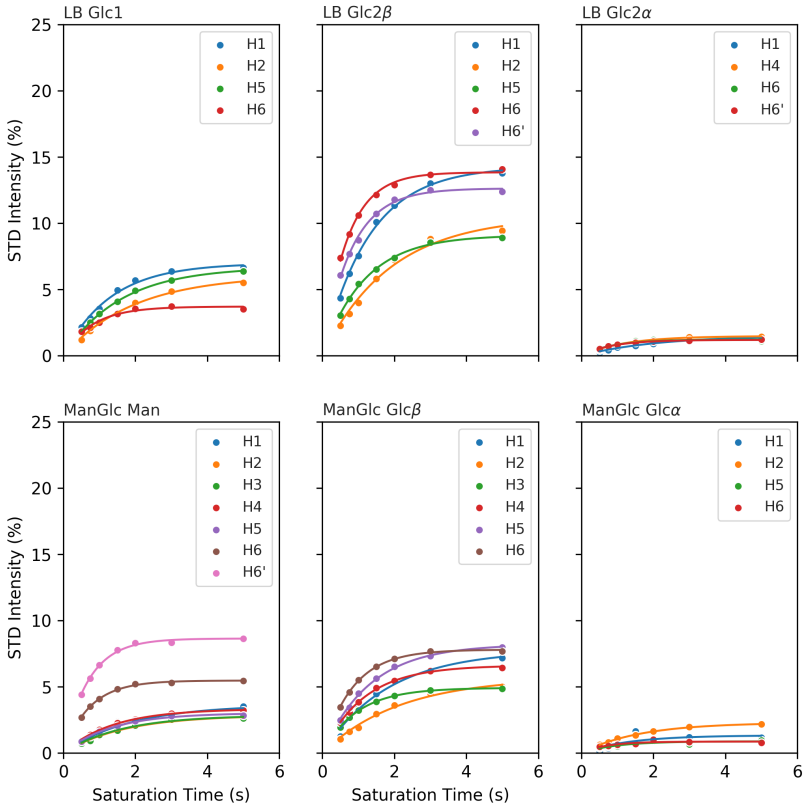

Figure 3.11: Experimentally determined STD build-up curves for laminaribiose (top) and Manβ-1-3-Glc (bottom) binding to *Ps*LBP. Circles show observed STD intensities, whereas curves are determined from least squares fitting to Eqn. 45.

The binding epitope map of βLB shows that the strongest STD intensities are present on the reducing sugar (Glc2), in particular the H6 protons again agreeing with the key relevance of the glucose reducing ring for binding (Fig. 3.12). Moderate STD intensity is seen for the H1 of the reducing sugar, whilst H2, H4 and H5 of the same residue exhibit weak STD intensities. This follows a similar pattern to βGlc, for which the H1 and H6 also make the most significant contacts with *Ps*LBP. This adds further evidence to suggest that the reducing sugar of βLB binds to the same subsite as the monosaccharide βGlc and that it is this residue that is most important for specificity and binding.

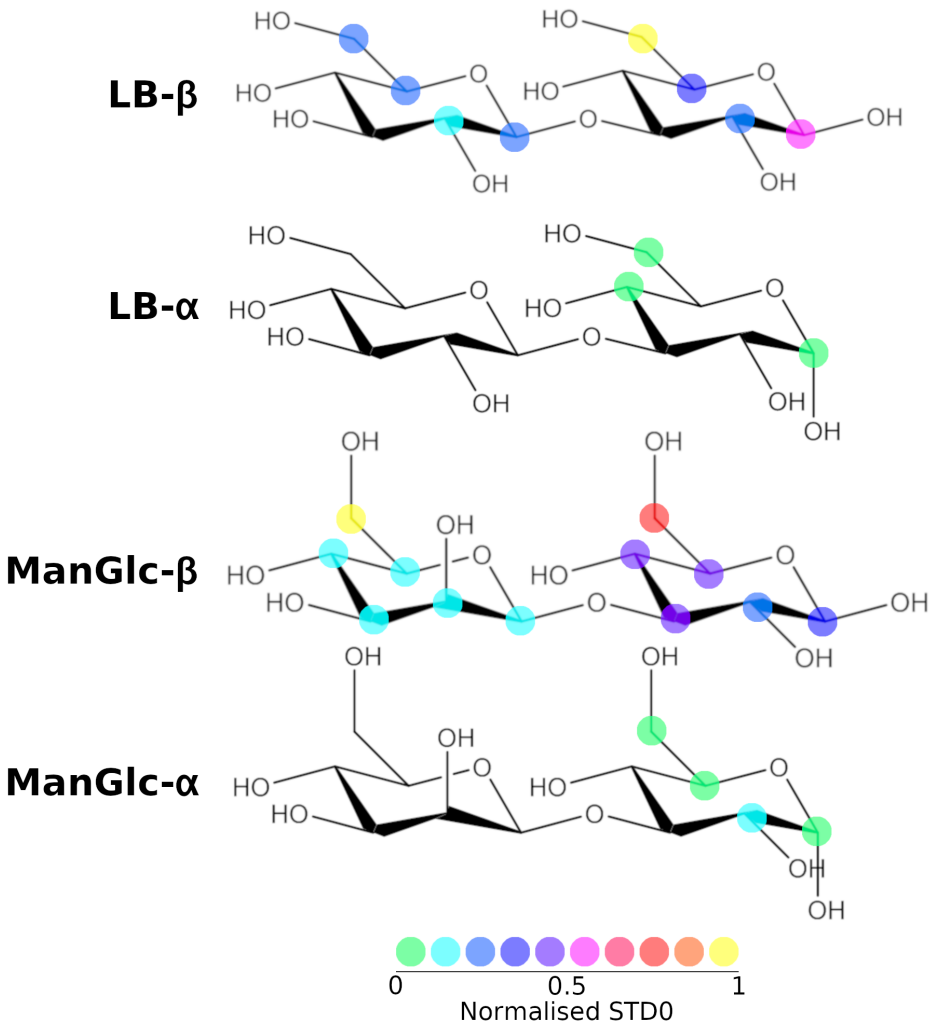

Figure 3.12: Binding epitope maps of laminaribiose (top) and Manβ-1-3-Glc (bottom), both in the presence of *Ps*LBP. Colours represent the normalised STD values for proton resonances at the indicated positions (low - high, cold - hot).

Interestingly, the binding epitope map of the non-reducing sugar (Glc1) is quite disimilar to the epitope map of G1P, despite the expectation that they should occupy the same binding site. In βLB, this sugar exhibits the strongest STD intensities for H1 and H3, unlike the H4 and H6 of G1P. It is important to note that the non-reducing sugar of βLB is in the β-configuration, whilst G1P is in the α-configuration. Therefore, in βLB, the H1 of the non-reducing sugar is axial and would therefore point towards the protein surface instead of towards the empty cavity. However, clearly there is some rearrangement within the binding subsite of the non-reducing sugar of βLB compared to G1P such that the H4 and H6 protons are tilted further from the *Ps*LBP surface and such that H3 makes more intimate contact. This is understandable since the phosphate,

thought to be essential for binding to this subsite, is no longer present.

The binding epitope map for Manβ-1-3-Glcβ shows some similarities to that of βLB (Fig. 3.12), with the H6 of the non-reducing sugar (Man) making the most significant contacts with *Ps*LBP, whilst the STD intensities exhibited by the non-reducing sugar are generally weak, suggesting that Manβ-1-3-Glcβ binds to *Ps*LBP in a similar manner. However, it is important to highlight that the H6 protons of the non-reducing sugar of Manβ-1-3-Glcβ show very strong STD intensities, unlike that observed for βLB. The result indicates that the inversion of the configuration at C2 of the non-reducing Man of Manβ-1-3-Glcβ relative to that of the non-reducing Glc of βLB causes Man to adopt a different binding orientation with its subsite to facilitate more favourable contacts with the *Ps*LBP sidechains.

It is also interesting that the magnitude of the absolute STD intensities observed for βLB and Manβ-1-3-Glcβ are comparable (Tables 3.5, 3.5, 3.7, and 3.8), suggesting similar binding affinities. This is in agreement with the hypothesis that the reducing sugar (Glc2 in LB, Glc in Manβ-1-3-Glcβ) is most important for binding, since the reducing sugar is βGlc in both cases. It also suggests that the non-reducing sugar does not contribute significantly to the binding affinity, which further agrees with the donor subsite requiring a hexose-1-phosphate for recognition.

3.4 Discussion

The chemical synthesis of oligo- and poly-saccharides is extremely challenging due to the large number of stereocenters and reactive functional groups present in these molecules. Enzymatic synthesis provides an attractive alternative because it provides a facile, highly stereospecific, one-step method for forming glycosidic bonds. Glycosidase phosphorylases (GPs) are attractive candidates because, compared to glycosyl transferases (GTs) which have previously been extensively studied for the same purpose, they are typically more stable, utilise cheaper substrates, and are more promiscuous with regard to substrate specificity, increasing their utility over a wider range of syntheses.

The Laminaribiose phosphorylase from *Paenibacilus sp.* (*Ps*LBP) is one such enzyme and has been shown to tolerate the non-cognate donor substrate, mannose-1-phosphate (M1P), as well as its native donor substrate, glucose-1-phosphate (G1P). Here we have used a combination of STD NMR spectroscopy and molecular modelling to confirm the solution structure of the complexes between *Ps*LBP with the donor substrates G1P and M1P. The results highlight excellent agreement between experimental STD NMR data and STD values predicted from the crystal structures of the complexes.

In addition, our results suggest a shorter residence time in the bound state of the non-cognate donor M1P within the donor subsite compared to G1P. This is understandable since the crystal structure shows that the interaction between the C2 hydroxyl of the donor and the sidechain of Arg353 in *Ps*LBP is broken by inversion of the stereochemistry at C2 in M1P. Nevertheless, this inversion is tolerated since the C2 of the donor faces the opening of the donor subsite cavity. This suggests that further substitutions at this position could be tolerated, although the related cellobiose phosphorylase from *Cellvibrio gilvus* (CgCBP) fails to tolerate N-acetylglucosamine 1-phosphate (GlcNAc1P).[281] However, chitobiose phosphorylase from *Vibrio proteolyticus* (VpChBP) does use GlcNAc1P as a donor[282] and the architecture of the donor subsite differs only in the positioning of the arginine that interacts with the C2 substituent (Arg353 in *Ps*LBP, Arg343 in VpChBP), suggesting that modification of *Ps*LBP could be possible in order to accommodate larger C2 substituents.

It is unlikely that modifications to the sugar 1-phosphate donor at positions other than the C2 would be tolerated by *Ps*LBP, since both the crystal structures and the STD NMR data show strong contacts at the C4 and C6 positions, both of which face the rear of the donor subsite cavity so any modification would hinder binding due to steric interactions. This is supported by the fact that Gal1P, in which the stereochemistry about C4 is inverted, is incapable of acting as a donor substrate.[274]

It was not possible to obtain crystal structures of *Ps*LBP in complex with the acceptor, Glc, nor the reaction products, LB or Manβ-1-3-Glc. However, using STD NMR it was possible to unravel some details of these complexes. This underscores the usefulness of STD NMR to

gain structural details of such weak protein-carbohydrate interactions that don't produce complexes stable enough as to be crystallised. In particular, the β-anomer of Glc and the reducing Glc residue in the reaction products bound to *Ps*LBP with much higher affinity than the α-anomer. This is interesting because it highlights two details: (1) the acceptor subsite tolerates only the β-configuration of the acceptor, and (2)the donor subsite has no or low affinity for non-phosphorylated sugars.

It is intriguing as to why *Ps*LBP would select for only the β-anomer of the acceptor since the configuration here should have little effect on the enzymatic mechanism. Therefore, it is likely that this selection is circumstantial and a consequence of the tight β-hairpin gate proposed to limit the degree of polymerisation to disaccharides.[274] The low affinity for non-phosphorylated sugars in the donor subsite highlights that the phosphate is an essential recognition element for this subsite and may prevent the catalytic activity from being hampered by an excess of the acceptor substrate.

As of yet, the specificity of the acceptor subsite has not been tested. Given that the acceptor subsite is intentionally restrictive to prevent polymerisation past a disaccharide, it may be difficult to add larger substituents to the acceptor substrate. Furthermore, the STD NMR data highlight strong contacts with the enzyme particularly at the C6 position suggesting that modifications here may be unsuccessful. The STD intensity for the C2 position is particularly weak for both the acceptor Glc and reducing Glc residue in the reaction products, suggesting that modification here may be possible.

This work provides fundamental structural details that will aid our understanding of the substrate specificity of glycoside phosphorylases and helps pave the way for enzymatic synthesis of a broad range of carbohydrates.

3.5 Conclusions

In this chapter, STD NMR spectroscopy, including quantitative analysis using the CORCEMA-ST software, was used to provide solution

state structural information about the interaction between *Ps*LBP. Overall the conclusions were:

1. The structures of the complexes of *Ps*LBP with G1P and M1P derived from X-ray crystallography are in excellent agreement with the quantitative STD NMR analysis described here, showing that the structures described in the respective models are valid in describing the interaction in the solution state.

2. The acceptor substrate binding site of *Ps*LBP preferentially binds to glucose residues in the β anomeric configuration as shown by very weak STD intensites for the α anomer.

3. The STD NMR binding epitopes combined with the x-ray crystal structures of *Ps*LBP in complex with G1P and M1P suggest the *Ps*LBP may tolerate further substitutions at the C2 position of the donor substrate, making the enzyme synthetically versatile.

Chapter 4

Characterisation of the Interaction between the *Salmonella enterica* effector proteins and their Death Domain Substrates

4.1 Introduction

Salmonella enterica is a species of pathogenic intracellular Gram-negative bacteria responsible for over 1 billion cases of infection every year.[283] Over 2500 serovars have been identified,[284] although they can broadly be categorised as either Typhoidal or Non-Typhoidal serovars.[285] Typhoidal species are restricted to human hosts and are responsible for causing enteric fever, a systemic infection resulting in severe abdominal pain and diarrhoea. Enteric fever is responsible for over 200,000 deaths per year, predominantly in underdeveloped regions, particularly in southern Asia and sub-Saharan Africa.[285] Non-Typhoidal serovars are typically host generalists, that is to say that they are capable of infecting numerous species of host organisms. This is a particular problem for industrialised countries, in which *Salmonella* is transmitted through industrially produced food. The most common form of Non-Typhoidal *Salmonella* infection is gastroenteritis, an infection of the gastrointestinal tract.[284] Gastroenteritis causes symptoms such as diarrhoea and vomiting, and is usually far more acute with a lower mortality rate than enteric fever.

4.1.1 Invasion of Host Cells by *Salmonella enterica*

Like other intracellular pathogens, the success of *Salmonella enterica* is largely due to its ability to survive and reproduce within the cells of the host, thereby evading the humoral immune response[286,287] (Fig. 4.1). Once inside the host, intracellular pathogens can form a protective host-derived vacuole, for *Salmonella* called the *Salmonella*-Containing Vacuole (SCV) to further shield themselves from intracellular mechanisms, such as degradation in the lysosome. A unique feature of *Salmonella enterica* is their ability to infect and survive within the harsh environment of leukocytes such as macrophages,[288,289] which contributes to the immunosuppression of the host.

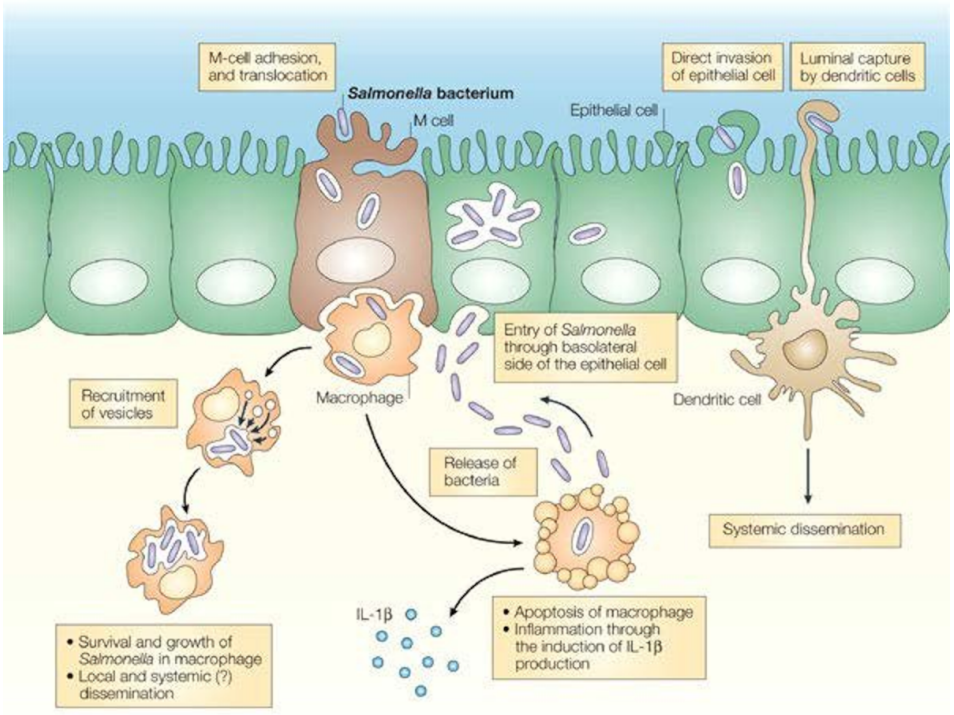

Figure 4.1: The lifecycle of *Salmonella enterica* inside a host organism. Reused with permission of Cota *et al*[290] under the Creative Commons license (https://creativecommons.org/licenses/by/4.0/)

Invasion of host cells by *Salmonella enterica* is mediated predominantly through the Type III Secretion System (T3SS) (Fig. 4.2), although recent studies have suggested that *Salmonella* also possess T3SS-independent invasion mechanisms.[291,292] The T3SS is a large complex of bacterial proteins forming a needle-like structure that traverses both the inner and outer bacterial membranes of *Salmonella*

and protrudes into the extracellular space (Fig. 4.2).[293] It is capable of penetrating the host membrane and translocating material directly from the bacterial cytoplasm to the host cytoplasm. The proteins translocated from the bacterial pathogen to the host by the T3SS are known as effector proteins and are responsible for modulating host cell function, both during invasion and intracellular survival.[293] In fact, *Salmonella enterica* encodes two T3SSs, denoted here as T3SS-SPI1 and T3SS-SPI2 (SPI = *Salmonella Pathogenicity Island*).[293]

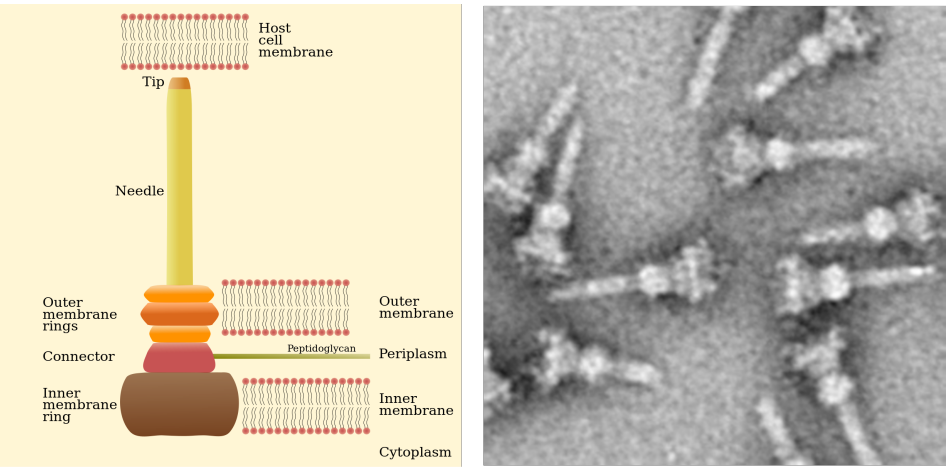

Figure 4.2: **Left:** Cartoon of the structure of the Type III Secretion System (T3SS). **Right:** Transmission electron micrograph of the T3SS complex isolated from *Salmonella enterica*. Reused with permission from Wikimedia Commons under the Creative Commons license (https://creativecommons.org/licenses/by/2.5/deed.en)

The T3SS-SPI1 is responsible for translocating effectors involved in the invasion of the intestinal epithelia.[294] These effectors are predominantly involved in remodelling the host actin cytoskeleton[295–298] and destabilising the host membrane[299] in order to endocytose the *Salmonella* bacterium. Furthermore, after entry into the host cell the T3SS-SPI1 effector SptP is responsible for deactivating host Rac-1 and Cdc42, which allows the host cell morphology to return to its native state.[300] It has been suggested that this acts to maintain the viability of the infected host cell, since prolonged expression of Rac-1 and Cdc42 can be harmful to the cell.

The T3SS-SPI2, on the other hand, is responsible for mediating host cell function after invasion by the *Salmonella* bacterium and its effectors are essential for survival and proliferation of the bacterium within the host cell.[301] In particular, the effectors SseB, SseC, SseD and spiC are responsible for the formation of an actin shell surround-

ing the SCV,[302] which helps to maintain SCV membrane integrity.[303] Furthermore, the effectors SseF and SseG redirect Golgi vesicles to the SCV, which have been shown to be important for *Salmonella* proliferation within the host.[304]

More recently a new family of *Salmonella enterica* T3SS-SPI2 effectors have been identified, its members being labelled SseK1, SseK2, and SseK3.[305,306] The SseK family were originally postulated as translocated effectors based on their homology to the NleB effector from *Citrobacter rodentium*,[307] which was known to be a secreted effector. Homologous effectors have also been found in enteropathogenic *Escherichia coli* (EPEC) and enterohemorrhagic *Escherichia coli* (EHEC).[308]

The exact function and purpose of the SseK effectors is still being elucidated, although it is clear that they are not essential for the virulence of *Salmonella enterica*. Knockouts of both SseK1 and SseK2 had no effect on the formation of the SCV and pathogenesis in infected mice apparently proceeded as for wild type *Salmonella*.[305] Furthermore, experiments in macrophages showed that apoptosis of infected macrophages was SseK independent,[309] although SseK1 and SseK3 were able to inhibit necroptotic cell death, mediated by the Tumour Necrosis Factor(TNF)-α induced NF-κB pathway. This is intriguing since their homologues in *E. coli* and *C. rondentium*, the NleB effectors, have a strong effect on the colonisation abilities of the pathogen.[310,311]

4.1.2 Function of the SseK effectors from *Salmonella enterica*

As well as the intrigue of understanding the role that the SseK proteins may play in the lifecycle of *Salmonella enterica*, these enzymes are also of interest due to their novel enzymatic mechanism. Along with the NleB effectors,[310,311] the SseK effectors have been shown to function as glycosyltransferases (GTs), transferring N-acetylglucosamine residues (GlcNAc) onto arginine sidechains of their targets. Before this discovery, only one other example of arginine glcosylation had been reported[312] and the mechanism by which this occurs is not well understood. What is clear is that the GT activity of the SseK effectors

is dependent on a DXD-motif (Asp-X-Asp)[306,309], common to many GTs, that chelates a divalent manganese ion (Mn^{2+}).

In GTs, the enzymatic catalysis can proceed *via* either a so-called retaining or inverting mechanism.[85] For retaining enzymes, the anomeric configuration of the donor substrate is retained in the formation of the glycosidic linkage with the acceptor substrate, whereas the configuration becomes inverted in the reaction product for inverting enzymes (Fig. 4.3). The SseK effectors share some similarity with some known retaining GTs[313] suggesting they too may follow a retaining mechanism. However, it is usually expected that both aspartate residues of the DXD-motif will chelate the Mn^{2+} in retaining GTs, whereas only one of these aspartate residues (the downstream of the two) typically chelates the Mn^{2+} in inverting GTs.[314] In the X-ray crystal structures of the SseK effectors (discussed below), only the latter aspartate residue chelates the Mn^{2+}, pointing towards an inverting mechanism. One group claims experimental evidence for the SseKs possessing a retaining mechanism.[313] However, their reasoning is factually incorrect, stating that the hydrolysis of UDP-GlcNAc yields only α-GlcNAc, which ignores the chemistry of monosaccharides, in which mutarotation causes an equilibrium of α- and β-anomers in solution. The lack of observed signal in the ^{1}H NMR spectrum for the β-anomer of GlcNAc is explained simply by the fact the this proton typically resonates at approximately 4.7 ppm in β-GlcNAc,[315] and so is obscured by the water signal. Therefore there still exists no experimental evidence for the enzymatic mechanism of these enzymes.

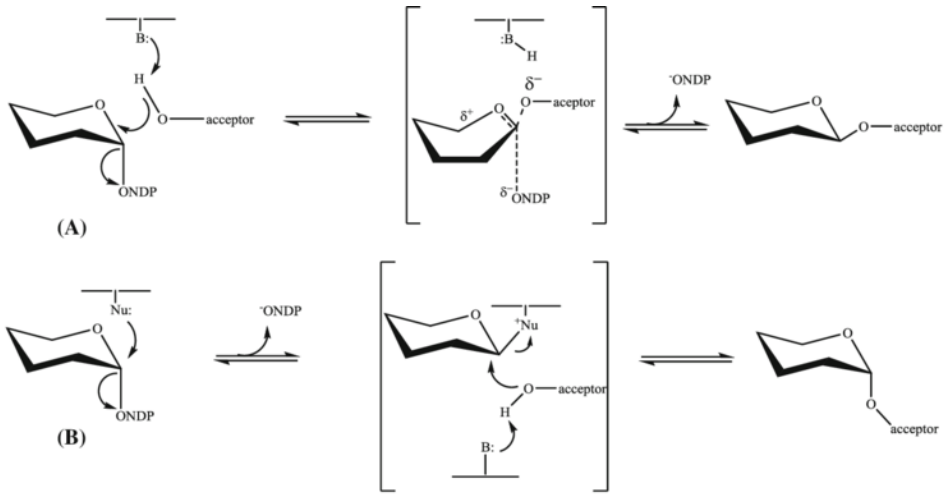

Figure 4.3: General mechanisms for inverting and retaining glycosyltransferases. Top: The inverting mechanism showing the backside attack of the acceptor substrate leading to inversion of anomeric configuration. Bottom: The retaining mechanism showing the formation of a donor-protein intermediate before attack by the acceptor substrate, leading to overall retention of anomeric configuration. Used with permission of Dr.Brock.Schuman under the Creative Commons license.

4.1.3 Structures of the SseK effectors

Recently, X-ray crystal structures have been published for each of SseK1, SseK2, and SseK3.[313,316] These structures consist of UDP-bound SseK1 (PDB: 5H60); apo- (PDB: 5H61), UDP-bound (PDB: 5H62), and UDP-GlcNAc-bound SseK2 (PDB: 5H63); and apo- (PDB: 6EYR) and hydrolysed UDP-GlcNAc-bound SseK3 (PDB: 6EYT).

Each of the SseK enzymes have close structural homology, and each consist of a core catalytic domain and a short (approximately 40 residues) domain consisting of two α-helices joined by a short loop (HLH-domain) that protrudes from the core (Fig. 4.4). The catalytic domain consists of a fold typical of a GT-A fold type, whilst the HLH-domain is unique to the SseK effectors. Interestingly, the core catalytic domains of SseK1, SseK2 and SseK3 have a high sequence identity, with most of the sequence variation between the effectors being found in the HLH-domain.[313,316] This has led to the proposition that the HLH-domain may confer substrate specificity, despite its location far from the catalytic site (Fig. 4.4).

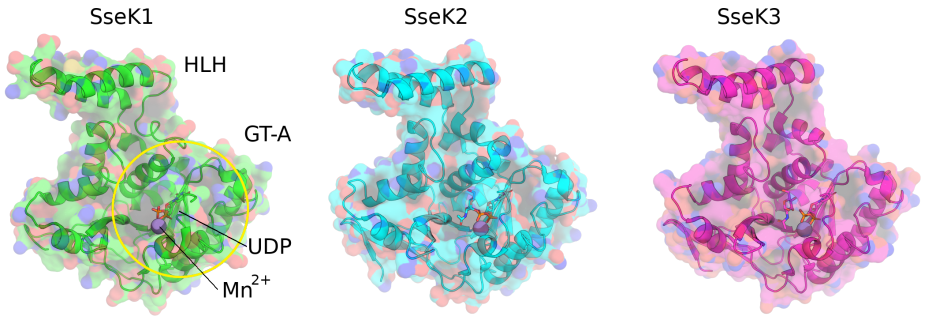

Figure 4.4: Topology of the SseK effectors, SseK1 (green), SseK2 (cyan) and SseK3 (pink). The glycosyltransferase type A (GT-A) domain fold and helix-loop-helix (HLH) protrusion are labelled. The manganese ion (Mn^{2+}) (purple sphere) and ligands are shown (sticks, SseK1: UDP, SseK2: UDP-GlcNAc, SseK3: hydrolysed UDP-GlcNAc)

In each case, the donor substrate binding site consists of a deep, open cavity within the core catalytic domain. The diphosphate moiety of UDP chelates the Mn^{2+} ion and the uridine base is stacked between two aromatic residues (SseK1: Trp51 and Phe187, SseK2: Trp65 and Phe203, SseK3: Trp52 and Phe190) (Fig. 4.5). In both the SseK2 and SseK3 structures containing a UDP-GlcNAc and hydrolysed UDP-GlcNAc respectively, the GlcNAc moiety interacts with a number of absolutely conserved residues; in SseK2 these are Asp204, Arg207, Asp239, and Arg348 (Asp191, Arg194, Asp226, and Arg335 in SseK3) (Fig. 4.6). In both cases, the methyl group of GlcNAc is buried in a pocket adjacent to the Mn^{2+} ion. Interestingly, the GlcNAc residue in SseK3 (hydrolysed UDP-GlcNAc) is rotated somewhat relative to the GlcNAc moiety in SseK2 (UDP-GlcNAc). Furthermore, the ring of the GlcNAc moiety in SseK2 is distorted somewhat compared to that in SseK3, and in either structure a nearby asparagine residue (SseK2: Asn272, SseK3: Asn259) adopts a different rotameric conformation (the only significant difference in the catalytic site structure between the two models), hinting at its involvement in the catalytic mechanism.

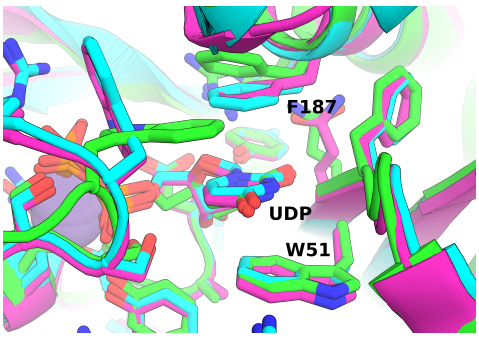

Figure 4.5: Stacking of the UDP uridine base between the Trp (SseK1: Trp51, SseK2: Trp65, SseK3: Trp52) and Phe (SseK1: Phe187, SseK2: Phe203, SseK3: Phe190) residues of SseK1 (green), SseK2 (cyan) and SseK3 (pink) effectors. Figure labels correspond to SSeK1 numbering.

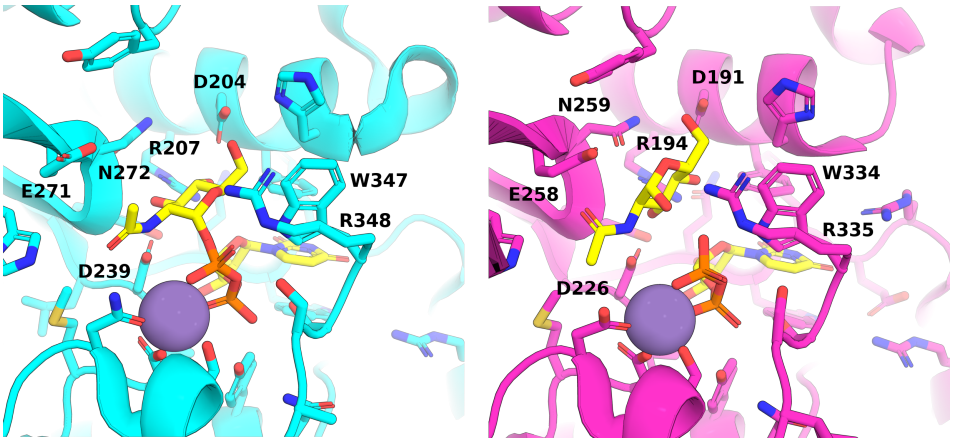

Figure 4.6: Structure of the SseK2 (left, cyan) and SseK3 (right, pink) donor substrate binding site. Shown in the UDP-GlcNAc (SseK2)/hydrolysed UDP-GlcNAc (SSeK3) closed conformation.

Finally, comparison of the apo- and holo-structures of SseK2 and SseK3 reveals that, in the apo state the C-terminal loop is highly flexible (Fig. 4.7), whereas in the holo-enzyme it closes over the catalytic site with the tail arginine residue (SseK2: Arg348, SseK3: Arg335) interacting directly with the β-phosphate of UDP and the anomeric carbon of GlcNAc. For this reason, it has been proposed that the C-terminal loop may act as a so-called lid domain to obscure the catalytic site from the bulk solvent in order to prevent hydrolysis of the donor.[313,316] It should also be noted however, that in 2 of the 4 SseK2 molecules in the asymmetric unit of the X-ray crystal structure (UDP-GlcNAc bound) the "lid" is still in an open conformation, suggesting that it is still somewhat flexible in the holo-enzyme.

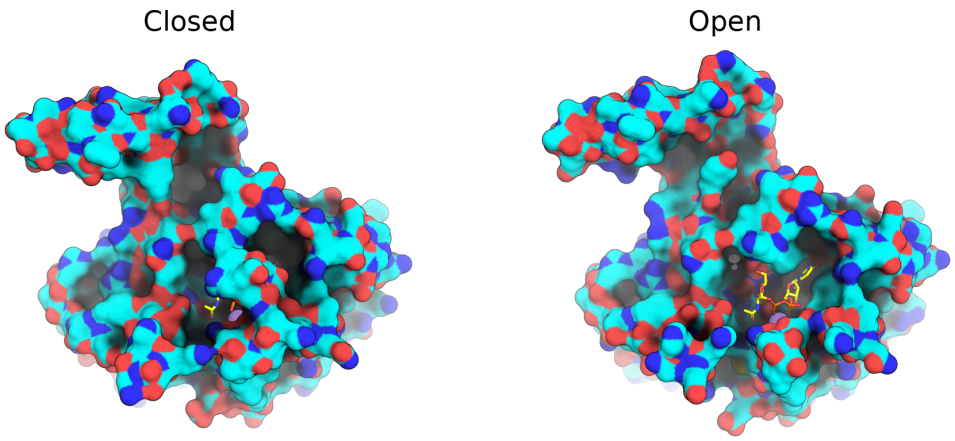

Figure 4.7: The closed (left) and open (right) conformations of the C-terminal "lid" domain of SseK2 (cyan) bound to UDP-GlcNAc (yellow) and manganese (Mn^{2+}, purple sphere). Structures taken from chain A (closed) and chain B (open) of the X-ray crystal structure of SseK2 bound to UDP-GlcNAc (PDB: 5H63).

4.1.4 Substrate specificity of the SseK effectors

The glycosylation targets of the NleB effectors from EPEC, EHEC and *C. rodentium* had previously been elucidated as TRADD, FADD, RIPK1, and TNFR1, which are involved directly in the TNF-α pathway,[310,317] and GAPDH, which affects nF-κB signalling *via* TRAF2 activation.[311] Based on this it was found that SseK1 targets TRADD and GAPDH, SseK2 targets FADD, and SseK3 targets TRADD[309,318] (Table 4.1). SseK1 also bound to FADD but did not glycosylate it, even though the death domains of each TRADD and FADD are structurally homologous. Intriguingly, although SseK2 glycosylated FADD no interaction could be detected, suggesting a highly transient or unstable interaction.

Table 4.1: Binding and glycosylation of the host target proteins FADD, TRADD and GAPDH by the SseK effectors. b. = binding, n.b. = no binding, g. = glycosylation, n.g. = no glycosylation, n.d. = no data

Effector	FADD	TRADD	GAPDH
SseK1	b./n.g.	n.d./g. b	./g.
SseK2	n.b./g.	n.d.	n.b./n.g.
SseK3	n.b./n.g.	n.d./g. n	.b./n.g.

Clearly then, the SseK effectors must possess some as yet unknown mechanism by which they can bind to and glycosylate their targets, both of which appear to be a distinct process. Given the homology of the GT-A-like domains of the SseK effectors it is intriguing to understand the role the HLH-domain may play in ligand recognition and how this affects catalysis. The purpose of this chapter is to investigate the molecular mechanisms that are involved in acceptor substrate recognition and glycosylation and to model their interactions, using a combination of Nuclear Magnetic Resonance spectroscopy and molecular modelling.

4.1.5 Objectives

The aim of this chapter is to provide structural insight into the interactions of SseK1/2 with their acceptor substrates, FADD, TRADD and GAPDH, using a combination of NMR spectroscopy and molecular modelling. Specifically the objectives are:

- Provide structural information about the binding mode of the acceptor substrates using STD NMR spectroscopy to determine the binding epitope maps of the short substrate peptides, $FADD_{110\text{-}118}$, $TRADD_{229\text{-}237}$ and $GAPDH_{195\text{-}203}$.

- To understand how the differences in SseK1 and SseK2 may lead to differential recognition/glycosylation of substrates.

- Generate an experimentally validated model of the SseK2/FADD complex using a combination of NMR spectroscopy and molecular modelling.

- Determine whether SseK1/2 follow an inverting or retaining enzymatic mechanism.

4.2 Material and Methods

4.2.1 Peptide assignment and STD NMR

All experiments were performed at 288 K on a Bruker Avance III 800 MHz spectrometer equipped with a 5-mm TXI 800 MHz H-C/N-D-05 Z BTO probe. $FADD_{110\text{-}118}$, $TRADD_{229\text{-}237}$ and $GAPDH_{195\text{-}203}$ (Genscript) samples were prepared at 1 mM in 90% H_2O/ 10% D_2O and assigned using standard COSY (cosydfesgpph), TOCSY (mlevphpr), and ^{1}H-^{13}C HSQC (hsqctgpsp) experiments. Apoenzyme samples (Dr Hurtado Guerrero, Zaragoza, Spain; Prof. Hardwidge, Kansas, USA) were prepared with 1 mM ligand peptide and 25 μM enzyme in either 25 mM Tris-d_{11} (SseK1) or 10 mM PBS (SseK2); both at pH 7.4 in D_2O. Holoenzyme samples were prepared in the same way, with the addition of 25 μM $MnSO_4$ and 25 μM UDP. The residual water signal was used as a reference for chemical shifts.[319] STD NMR experiments were performed using a train of 50 ms Gaussian pulses (0.4 mW, B_1 field strength 78 Hz) applied on the f2 channel at either 0 ppm (on-resonance) or 40 ppm (off-resonance). A spoil sequence (2 trim pulses of 2.5 and 5 ms followed by a 40% z-gradient applied for 3 ms at the beginning of the experiment) was used to destroy unwanted xy-magnetization from previous scan and a spinlock (1.55 W, 40 ms) was used to suppress protein signals (stddiff.3). The saturation time (d20) was set to 2 s and the recycle delay (d1) was set to 5 s.

4.2.2 Configuration of GlcNAc in glycosylated $GAPDH_{187\text{-}203}$

GlcNAcylated peptide $GAPDH_{187\text{-}203}$ was prepared in the laboratory of Dr Hurtado Guerrero (Zaragoza, Spain) by adding 7.7 mM $GAPDH_{187\text{-}203}$ to 50 mM UDP-GlcNAc, 40μM SseK1, and 2 mM $MnCl_2$ in 25 mM Tris pH 7.5. The reaction was left to proceed at 37 °C for 24 h before purifying the resulting glycopeptide using an Amicon Ultra 10 KDa MWCO centrifuge filter. NMR experiments for glycosylated GAPDH187-203 were performed at 298 K, and consisted of a decoupled ^{1}H-^{13}C HSQC (hsqcetgpsi), and TOCSY with water suppression (mlevgpph19) at 800 MHz, and a non-decoupled

Perfect-CLIP-HSQC[320] at 500 MHz (with a digital resolution of 1.6 Hz, to determine the $^{1}J_{C,H}$ coupling of the anomeric carbon of the transferred GlcNAc residue). The HSQC recycle delay was 1.5 s.

4.2.3 Molecular docking calculations for FADD-SseK2

Crystal structures of SseK1 (PDB: 5H60), SseK2 (PDB: 5H63), and FADD (PDB: 3EZQ) were imported into Schrödinger Maestro and prepared with the Protein Preparation Wizard.[321] All buffer molecules and non-bridging waters were removed. Hydrogen atoms were then added to the model, using PROPKA to predict the protonation state of polar sidechains at pH 7.[276] The hydrogen-bonding network was automatically optimized by sampling asparagine, glutamine, and histidine rotamers. The model was then minimized using OPLS3[120] force field and a heavy atom convergence threshold of 0.3 Å.

A model of the $FADD_{110\text{-}118}$ peptide was created by truncation of the FADD crystal structure (PDB: 3EZQ). Conformers were generated in MacroModel using torsional sampling with the OPLS3 force field, constraining all backbone atoms. Redundant conformers were eliminated using an RMSD cutoff of 0.5 Å. Any conformer with an energy 5 kcal mol^{-1} greater than the lowest energy structure was also eliminated. Resulting conformers were then minimized using the conjugate gradient method, converging on a threshold of 0.05 kcal mol^{-1}. Docking of $FADD_{110\text{-}118}$ to SseK2 was then performed using Glide.[126,253] A cubic grid, suitable for peptide docking, was generated. It was centred on UDP-GlcNAc, with an outer box length of 45 Å and an inner box length of 40 Å. To account for flexibility, van der Waals potentials of all receptor and ligand atoms were scaled by 0.5. All ligand conformers were docked to the receptor using rigid sampling with the SP algorithm. The resulting complexes were then clustered by heavy atom RMSD to eliminate redundant poses, keeping the structure closest to the cluster centroid from each cluster. All SseK2 sidechains within 5 Å of the ligand were then optimized before minimizing using Prime.[251] A second round of docking was performed, as described above, on the receptor structures resulting from the minimisation step. The resulting complexes were clustered by heavy atom RMSD, and the lowest energy representative structure was chosen for analysis.

A model of SseK2 in complex with full length FADD was generated by aligning the backbone atoms of residues 110-118 in the full-length structure to the backbone of the docked $FADD_{110\text{-}118}$ structure. Prime optimization and minimization within 5 Å of the contact surface was used to eliminate an atomic overlap.

4.2.4 Molecular Dynamics Simulations

UDP charges for use with UDP-GlcNAc were derived using the RESP fitting method implemented on the RED server.[322] The UDP fragment was generated by replacing the GlcNAc with a methyl group. In accordance with GLYCAM,[115] the HF/6-31G* level of theory was used with a weight factor of 0.01 and all aliphatic protons were constrained to a charge of 0. The total charge of the UDP fragment was set to −2. The charge of the methyl group was set to 0.194 before removing to give a final fragment with net charge −2.194, in keeping with the modularity of GLYCAM.

Molecular dynamics simulations of SseK1, SseK2, and the SseK2:FADD complex were performed using the Amber PMEMD software.[323] Protein atoms were parametrised using the Amber ff11SB forcefield and the Mn^{2+} ion was modelled using 12-6-4 LJ-type parameters (Amber ions234lm_1264_tip3p). UDP-GlcNAc was parametrised with GLYCAM 06j and GAFF. Each system was solvated in a truncated octahedral box of TIP3P water, with at least 10 Å between the solute and the edge of the box, before neutralizing with Na^+ ions. The system was minimized using the conjugate gradient algorithm, converging on a threshold of 10^{-4} kcal mol^{-1} Å^{-1}, first with 20 kcal mol^{-1} Å^{-2} restraints on solute atoms, before repeating with no restraints. The system was slowly heated to 310 K over 500 ps (NVT), before equilibrating the pressure to 1 atm (NPT) over a further 500 ps. In both cases with 20 kcal mol^{-1} Å^{-2} restraints were used on solute atoms. These restraints were then slowly released over 800 ps before performing Gaussian accelerated molecular dynamics (GaMD) simulations (800 ns SseK1/SseK2, 500 ns SseK2:FADD complex), as implemented in AMBER, using a boost potential on both the dihedral and total potential energies. Here, the simulation was split into 4 distinct stages. First, conventional dynamics were run for 2 ns to automatically calculate an initial boost potential. The

calculated boost potential was then applied and fixed for 400 ps before allowing it to adapt for a further 5.6 ns. The resulting boost potential was then fixed before performing production dynamics for 800 ns (SseK1/SseK2) or 500 ns (SseK2:FADD complex), saving coordinates every 100 ps. In all cases, the SHAKE algorithm was used to restrain all bonds involving hydrogen, allowing for a time step of 2 fs. A Langevin thermostat was used with a collision frequency of 5 ps^{-1} and the barostat (1 atm) used an isotropic Berendsen algorithm with a relaxation time of 1 ps. In all cases, periodic boundary conditions were used, using the particle mesh Ewald to calculate electrostatics.

4.2.5 Production of ^{15}N-labelled FADD and NMR titration of FADD/SseK2

The full length FADD protein was cloned into the pET15b plasmid and transformed into *E. coli* BL21-DE3 cells. Transformed cells were selected by growing on a LB-agarose medium containing 100 μg mL^{-1} ampicillin. A starter culture was produced from a single colony by growing in 5 mL LB containing 100 μg mL^{-1} ampicillin overnight at 37 °C with 200 rpm shaking before inoculating into 1 L ^{15}N minimal medium described in [324]. The culture was grown at 37 °C with 200 rpm shaking until the OD_{600} reached 0.6-0.8. The culture was then induced with 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) before incubating for a further 4 hours. The bacterial pellet was then collected by centrifuging at 10,000 x g for 10 min at 4 °C.

The pellet was resuspended in 20 mL pH 7.9 lysis buffer containing 20 mM Tris, 250 mM NaCl, 2.5 mM imidazole which was then lysed by sonication before centrifuging at 16,000 x g for 20 min at 4 °C. The resulting suppernatant was then purified using a Ni-NTA column by washing with 40 mL lysis buffer then 40 mL wash buffer (as lysis buffer with 3 mM imidazole). The purified protein was then eluted with an elution buffer (as wash buffer with 1 M imidazole).

Purified ^{15}N-labelled FADD was concentrated to 3 mM and exchanged to a buffer of 25 mM Tris, 150 mM NaCl pH 7.4 in a solvent of 90% H_2O/10% D_2O. ^{1}H-^{15}N HSQC NMR experiments were performed in a 500 MHz spectrometer at 298 K, both in the absence and presence of SseK2 (3 mM, Dr Hurtado Guerrero, Zaragoza, Spain).

4.3 Results

4.3.1 Assignment of the Acceptor Substrate Peptides by Nuclear Magnetic Resonance Spectroscopy

To study the binding interaction between SseK1/2 and their acceptor substrates by Saturation Transfer Difference (STD) Nuclear Magnetic Resonance (NMR) spectroscopy, short peptides of the target regions of these substrates were constructed, these were: residues 110-118 of FADD ($FADD_{110\text{-}118}$), residues 229-237 of TRADD ($TRADD_{229\text{-}237}$), and residues 195-203 of GAPDH ($GAPDH_{195\text{-}203}$). These peptides were assigned using a combination of Correlation Spectroscopy (COSY), Total Correlation Spectroscopy (TOCSY), Nuclear Overhauser Effect Spectroscopy (NOESY), and ^{1}H-^{13}C Heteronuclear Single Quantum Correlation (HSQC) NMR spectroscopy experiments (1D ^{1}H NMR spectra are shown in Fig. 4.8, expansions of TOCSY experiments in Fig. 4.9, and chemical shift are in Fig. 4.8, Table 4.2, Table 4.3, Table 4.4). As with the NMR chemical shift assignment of carbohydrates, peptides can be separated into distinct spin systems due to the inter-residue separation of protons by more than 3 bonds by the backbone amide group. Assignments were made using water as the solvent to take advantage of the additional amide proton resonances.

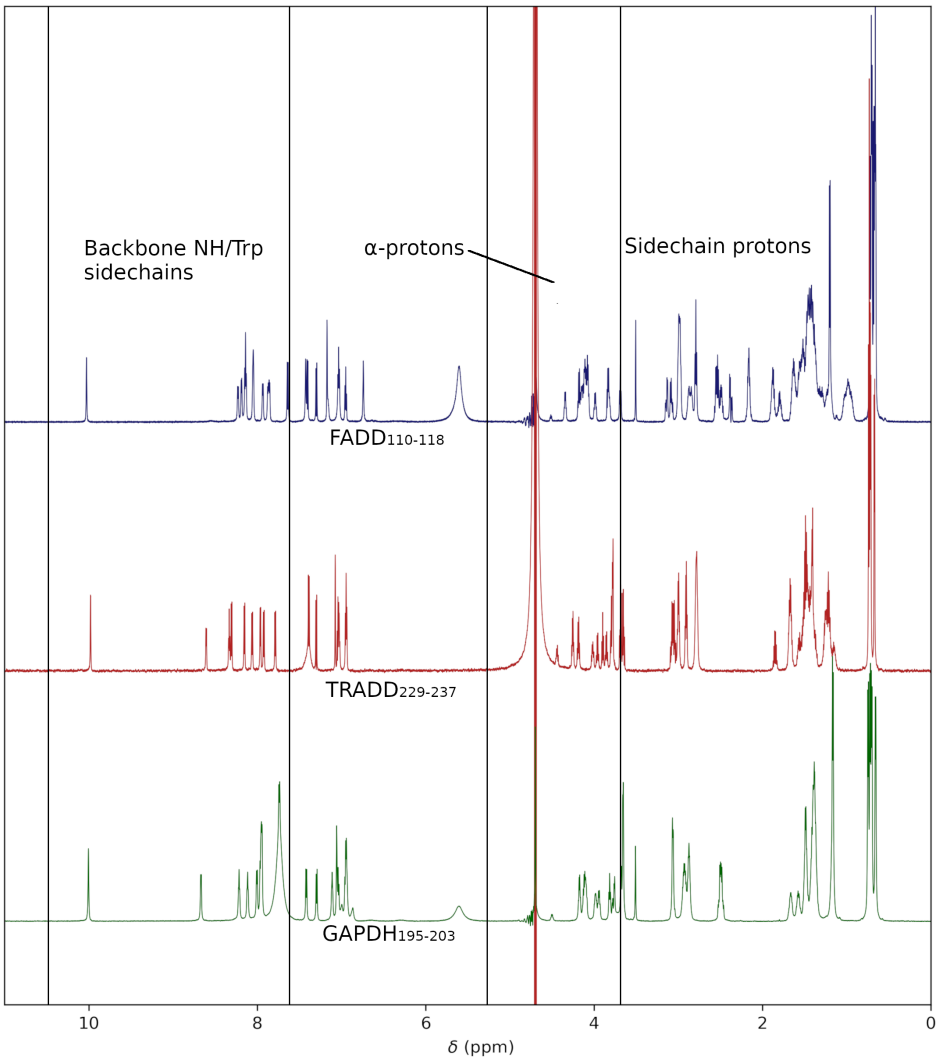

Figure 4.8: 1D ^{1}H NMR spectra of the SseK1/SseK2 acceptor substrate peptides, $FADD_{110\text{-}118}$ (blue), $TRADD_{229\text{-}237}$ (red), $GAPDH_{195\text{-}203}$ (green). Experiments performed at 288 K with 1 mM peptide in a solvent of 90% H_2O/10% D_2O.

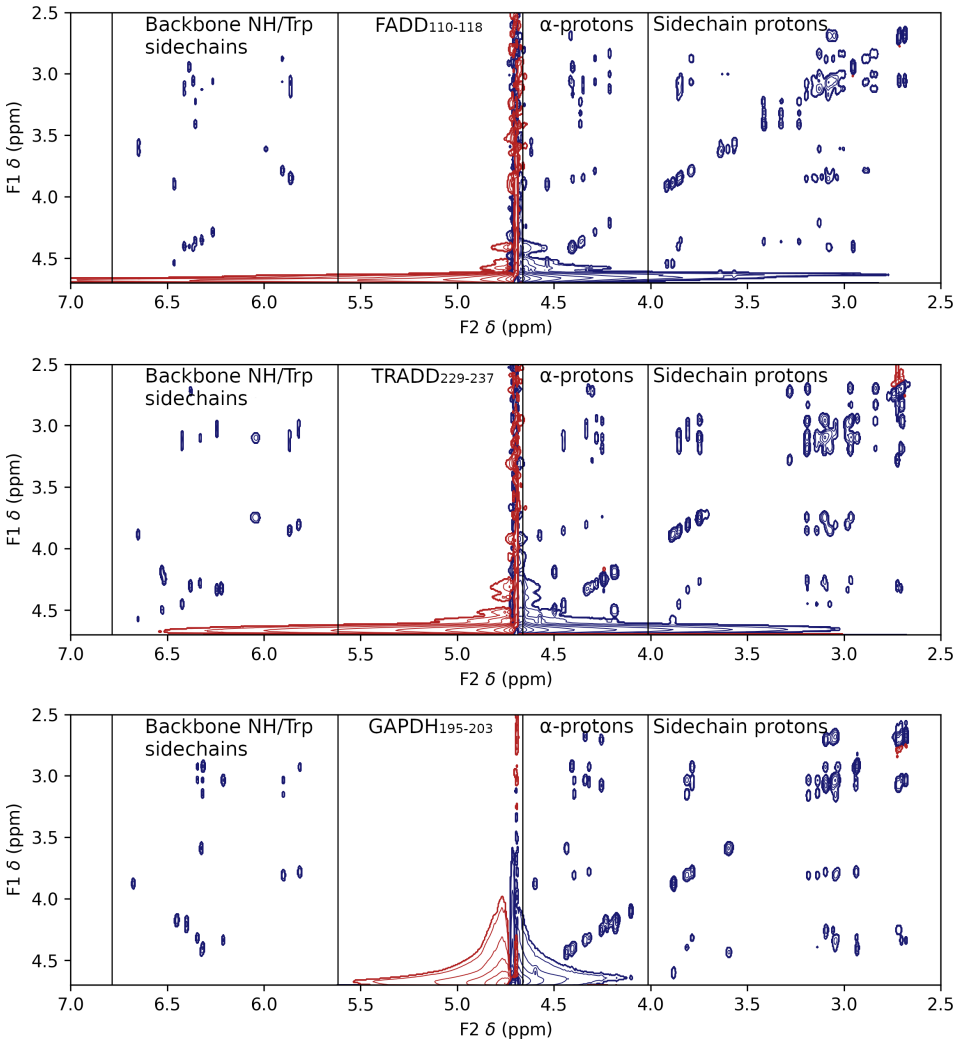

Figure 4.9: Expansions of TOCSY spectra of the $FADD_{110\text{-}118}$ (top), $TRADD_{229\text{-}237}$ (middle) and $GAPDH_{195\text{-}203}$ (bottom) peptides. Marked regions correspond to the proton type of the diagonal peaks. Cross peaks show protons within the same spin system as the diagonal peak, which mostly corresponds to whole amino acid residues. Exceptions occur when the coupling is interrupted by heteroatoms (eg. Trp).

Table 4.2: Assignment of the Nuclear Magnetic Resonance (NMR) chemical shifts of ^{1}H and ^{13}C nuclei of $FADD_{110\text{-}118}$. Experiments were performed in an 800 MHz NMR spectrometer at 288 K.

Residue	Position	δ ^{1}H (ppm)	δ ^{13}C (ppm)
Lys110	α	3.71	52.6
	β	1.54	30.0
	γ	0.98	21.0
	δ	1.30	26.1
	ϵ	2.54	38.7
Asp111	α	4.52	50.6
	β	2.57	37.6
Trp112	α	4.37	54.7

Residue	Position	δ ^{1}H (ppm)	δ ^{13}C (ppm)
	β	3.12	26.5
	$\delta 1$	7.02	122.7
	$\epsilon 3$	7.40	117.9
	$\zeta 2$	7.29	111.6
	$\zeta 3$	7.14	124.6
	$\eta 2$	6.94	119.0
Arg113	α	3.87	53.5
	β	1.40	27.5
	γ	1.06	23.7
	δ	2.87	40.2
Arg114	α	3.98	53.3
	β	1.43	24.0
	γ	1.36	24.2
	δ	2.99	40.2
Leu115	α	4.12	52.0
	β	1.43	24.0
	$\delta 1$	0.71	21.9
	$\delta 2$	0.65	20.3
Ala116	α	4.09	49.8
	β	1.19	16.2
Arg117	α	4.00	53.7
	β	1.94	27.2
	γ	1.74	27.1
	δ	2.12	31.2
Gln118	α	4.10	53.3
	β	1.59	27.6
	γ	2.99	40.2

Table 4.3: Assignment of the Nuclear Magnetic Resonance (NMR) chemical shifts of ^{1}H and ^{13}C nuclei of TRADD$_{229\text{-}237}$. Experiments were performed in an 800 MHz NMR spectrometer at 288 K. Some ^{13}C resonances are missing due to ambiguity of the recorded dataset.

Residue	Position	δ ^{1}H (ppm)	δ ^{13}C (ppm)
Lys229	α	3.78	52.76
	β	1.67	-
	γ	1.21	-
	δ	1.48	-
	ϵ	2.77	39.1
Trp230	α	4.43	55.1

Residue	Position	δ ^{1}H (ppm)	δ ^{13}C (ppm)
	β	3.05	26.8
	δ	7.06	124.3
	$\epsilon 3$	7.38	117.9
	$\zeta 2$	7.29	111.7
	$\zeta 3$	7.04	121.6
	$\eta 2$	6.94	119.3
Arg231	α	3.95	52.5
	β	1.40	-
	γ	1.26	-
	δ	2.89	40.5
Lys232	α	3.85	53.5
	β	1.49	-
	γ	1.16	-
	δ	1.49	-
	ϵ	2.77	39.1
Val233	α	3.89	59.5
	β	1.88	30.1
	γ	0.72	17.9
Gly234	α	3.77	42.0
Arg235	α	4.18	53.2
	β	1.58	-
	γ	1.44	-
	δ	2.99	40.4
Ser236	α	4.24	55.5
	β	3.66	60.9
Leu237	α	4.01	53.7
	β	1.40	40.1
	γ	1.41	24.2
	$\delta 1$	0.70	22.2
	$\delta 2$	0.66	20.5

Table 4.4: Assignment of the Nuclear Magnetic Resonance (NMR) chemical shifts of ^{1}H and ^{13}C nuclei of $GAPDH_{195\text{-}203}$. Experiments were performed in an 800 MHz NMR spectrometer at 288 K.

Residue	Position	δ ^{1}H (ppm)	δ ^{13}C (ppm)
Leu195	α	3.80	51.5
	γ	1.47	39.8
	$\delta 1$	0.74	21.5

Residue	Position	δ ^{1}H (ppm)	δ ^{13}C (ppm)
	$\delta 2$	0.70	20.8
Trp196	α	4.49	55.0
	β	3.06	26.7
	δ	7.04	124.2
	ϵ	7.41	117.9
	$\zeta 2$	7.28	111.6
	$\zeta 3$	7.02	121.7
	$\eta 3$	7.41	117.9
Arg197	α	3.92	53.0
	β	1.48	28.0
	γ	1.16	23.9
	δ	2.86	40.4
Asp198	α	4.16	51.2
	β	2.48	37.8
Gly199/ Gly201	α	3.75/3.64	42.6/42.4
Arg200	β	1.66	27.6
	γ	1.38	24.0
	δ	2.92	40.4
Ala202	α	4.10	49.3
	β	1.15	16.5
Leu203	α	3.97	53.6
	β	1.37	40.2
	γ	0.71	20.6
	$\delta 1$	0.69	22.2
	$\delta 2$	0.65	20.6

The chemical shift assignments were then used to interpret the binding experiments by STD NMR, allowing the different saturation transfer intensities to be ascribed to the specific protons of the acceptor peptides. This is fundamental to generate the map of contacts of the ligands with the protein receptor, as explained in the following section.

4.3.2 Saturation Transfer Difference NMR Spectroscopy the Acceptor Substrate Peptides

In saturation transfer difference (STD) NMR spectroscopy it is generally regarded as best practice to construct STD NMR build up curves by repeating the experiments over a number of different saturation times. However, in this case, the peptides were only stable over a short period of time and so it was only possible to acquire one time point for each experiment. Therefore the results should only be treated qualitatively. A saturation time of 2 s was chosen as the best compromise between signal sensitivity while still remaining short enough to minimise any artefacts. For each acceptor substrate peptide and for each enzyme (SseK1/SseK2), the STD NMR intensity was measured for each observable proton for three separate experiments: (1) in the presence of SseK1/SseK2, (2) in the presence of SseK1/SseK2 and Mn^{2+}, and (3) in the presence of SseK1/SseK2, Mn^{2+} and UDP.

Surprisingly, each of the $FADD_{110\text{-}118}$, $TRADD_{229\text{-}237}$ and $GAPDH_{195\text{-}203}$ peptides bound to both SseK1 and SseK2 under all conditions (apo-enzyme, plus Mn^{2+}, and plus Mn^{2+} and UDP) (Fig. 4.10, Fig. 4.13, Fig. 4.16). This means that the differences in enzyme specificities for glycosylation of the peptide substrate might not be explained by differences in molecular recognition of the ligands, as all the peptides are recognised by both enzymes. Hence, the specificity seems to be related to the catalytic step where the specific composition of the peptide determines whether glycosylation of the arginine residue will proceed or not. The discrepancy between the STD NMR data and the data presented in Table 4.1 can be explained by the fact that STD NMR is optimised for weak affinity binders that fall below the detection threshold of the high throughput techniques used to generate these data.

For each system, binding epitopes were constructed by normalising the STD intensity of each proton at 2 s against the $\zeta 2$-proton of the tryptophan residue of each peptide, since this proton received the most saturation in most of the experiments and normalising against the same proton allows relative comparisons to be made between each system.

4.3.2.1 Molecular recognition of $FADD_{110\text{-}118}$ by SseK1 and SseK2

^{1}H STD NMR spectra characterising the binding of $FADD_{110\text{-}118}$ to SseK1 and SseK2, under different conditions, are shown in Fig. 4.10. The experimentally determined STD NMR intensities are collected in Tables 4.5 and 4.6, for the binding of $FADD_{110\text{-}118}$ to SseK1 and SseK2, respectively.

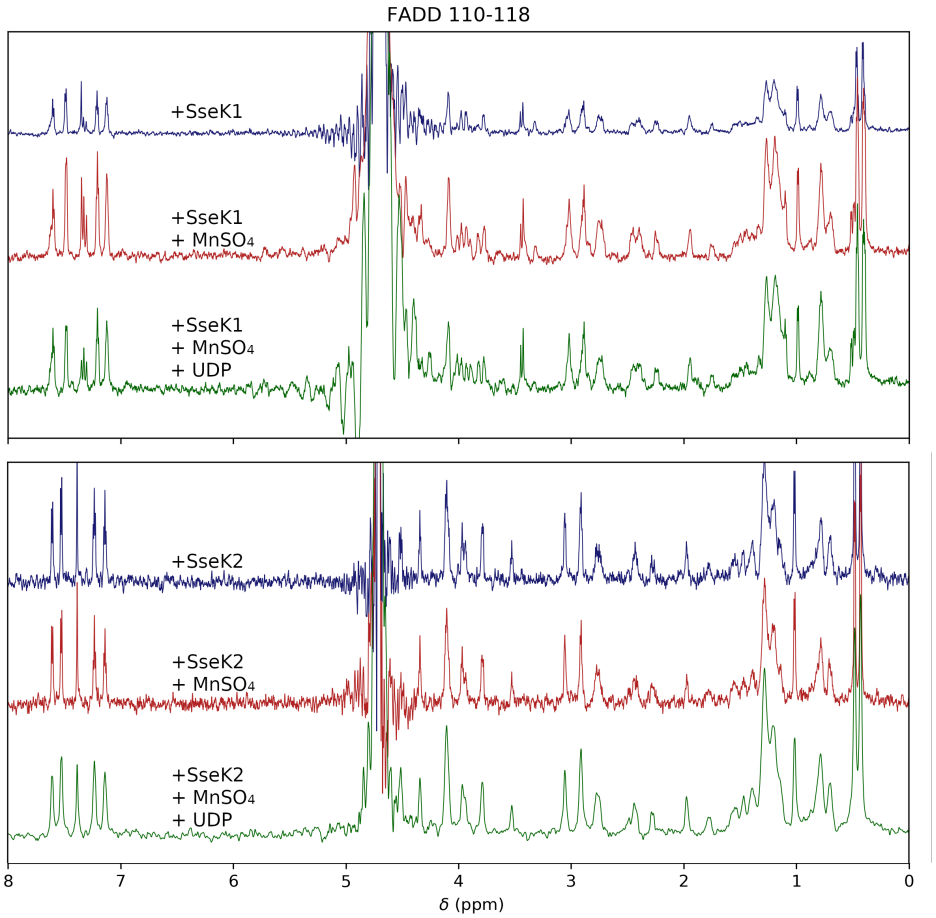

Figure 4.10: STD NMR difference spectra for $FADD_{110\text{-}118}$ binding to SseK1 (top) or SseK2 (bottom). Spectra recorded for the apoenzyme (blue), in the presence of 25 μM Mn^{2+} (red) and in the presence of 25 μM Mn^{2+} and 25 μM UDP (green). Measured using a saturation frequency of 0 ppm at 800 MHz and a saturation time of 2 s.

In the case of $FADD_{110\text{-}118}$ binding to SseK1, STD intensity is concentrated around Trp112 and Leu115 for both the apo-SseK1 and Mn^{2+}-bound SseK1 systems (Fig. 4.11). Upon addition of UDP, the STD intensities of these two residues increase relative to the Trp112 ζ2-proton. Furthermore, the STD intensities of the Arg113 and Arg114

sidechains increase dramatically. Together this suggests that binding of the donor analogue, UDP, in some way induces a conformational rearrangement that positions the arginine sidechains of $FADD_{110\text{-}118}$ in closer proximity to the SseK1 surface.

Comparatively, in the binding of $FADD_{110\text{-}118}$ to SseK2, in all cases (apo, plus Mn^{2+}, plus Mn^{2+} and UDP) the main contact residues of $FADD_{110\text{-}118}$ appear to be Trp112, Arg113 and Leu115 (Fig. 4.12). However, there is no significant change upon addition of Mn^{2+} and/or UDP.

Table 4.5: STD intensities for a saturation time of 2 s for each proton resonance of $FADD_{110\text{-}118}$ in the presence of SseK1.

Residue	Position	STD (%, SseK1)	STD (%, SseK1 + Mn^{2+})	STD (%, SseK1 + Mn^{2+} + UDP)
Lys110	α	4.76	4.11	4.40
	β	4.11	4.62	4.81
	γ	7.46	7.61	7.91
	δ	1.78	2.38	2.25
	ϵ	1.73	2.65	2.55
Asp111	β	2.58	3.48	3.52
Trp112	β	3.42	4.11	3.99
	δ	6.64	7.91	6.02
	$\epsilon 3$	5.57	6.08	6.32
	$\zeta 2$	7.39	7.83	5.79
	$\zeta 3$	5.79	7.04	6.08
	$\eta 2$	6.44	6.57	5.51
Arg113	γ	4.58	5.10	6.20
	δ	3.35	3.77	3.45
Arg114	γ	4.03	5.30	5.51
	δ	2.43	3.10	3.35
Leu115	$\delta 1$	5.68	6.26	6.77
	$\delta 2$	6.38	7.91	8.98
Ala116	β	2.65	3.10	3.16
Arg117	α	5.1	5.51	4.24
	β	1.98	1.74	1.78
	δ	1.69	1.58	1.87
	γ	1.73	2.20	1.88
Gln118	β	2.55	2.73	2.48
	γ	1.98	2.53	2.76

Table 4.6: STD intensities for a saturation time of 2 s for each proton resonance of $FADD_{110\text{-}118}$ in the presence of SseK2.

Residue	Position	STD (%, SseK2)	STD (%, SseK2 + Mn^{2+})	STD (%, SseK2 + Mn^{2+} + UDP)
Lys110	α	1.17	1.30	1.23
	β	1.94	1.63	1.83
	γ	4.24	4.15	4.03
	δ	1.43	1.15	1.18
	ϵ	0.52	0.85	0.71
Asp111	α	1.45	0.76	1.21
	β	1.54	1.35	1.41
Trp112	α	1.98	2.12	1.87
	β	1.96	2.08	1.81
	$\epsilon 3$	2.70	2.87	2.41
	δ	2.73	2.50	2.22
	$\zeta 2$	3.38	3.22	2.98
	$\zeta 3$	3.69	3.22	2.92
	$\eta 2$	3.19	2.81	2.50
Arg113	α	2.65	2.45	2.43
	δ	1.71	2.02	1.64
	γ	2.76	3.07	2.70
Arg114	α	1.28	1.02	1.01
	γ	2.53	2.38	2.27
	δ	1.21	1.04	1.03
Leu115	$\delta 1$	2.65	2.70	2.68
	$\delta 2$	2.98	3.04	3.01
Ala116	β	1.35	1.41	1.34
Arg117	α	1.90	1.98	1.61
	β	1.23	1.02	0.96
	γ	1.48	1.13	1.13
	δ	0.77	0.64	0.81
Gln118	β	1.90	2.02	1.98
	γ	0.93	0.85	0.92

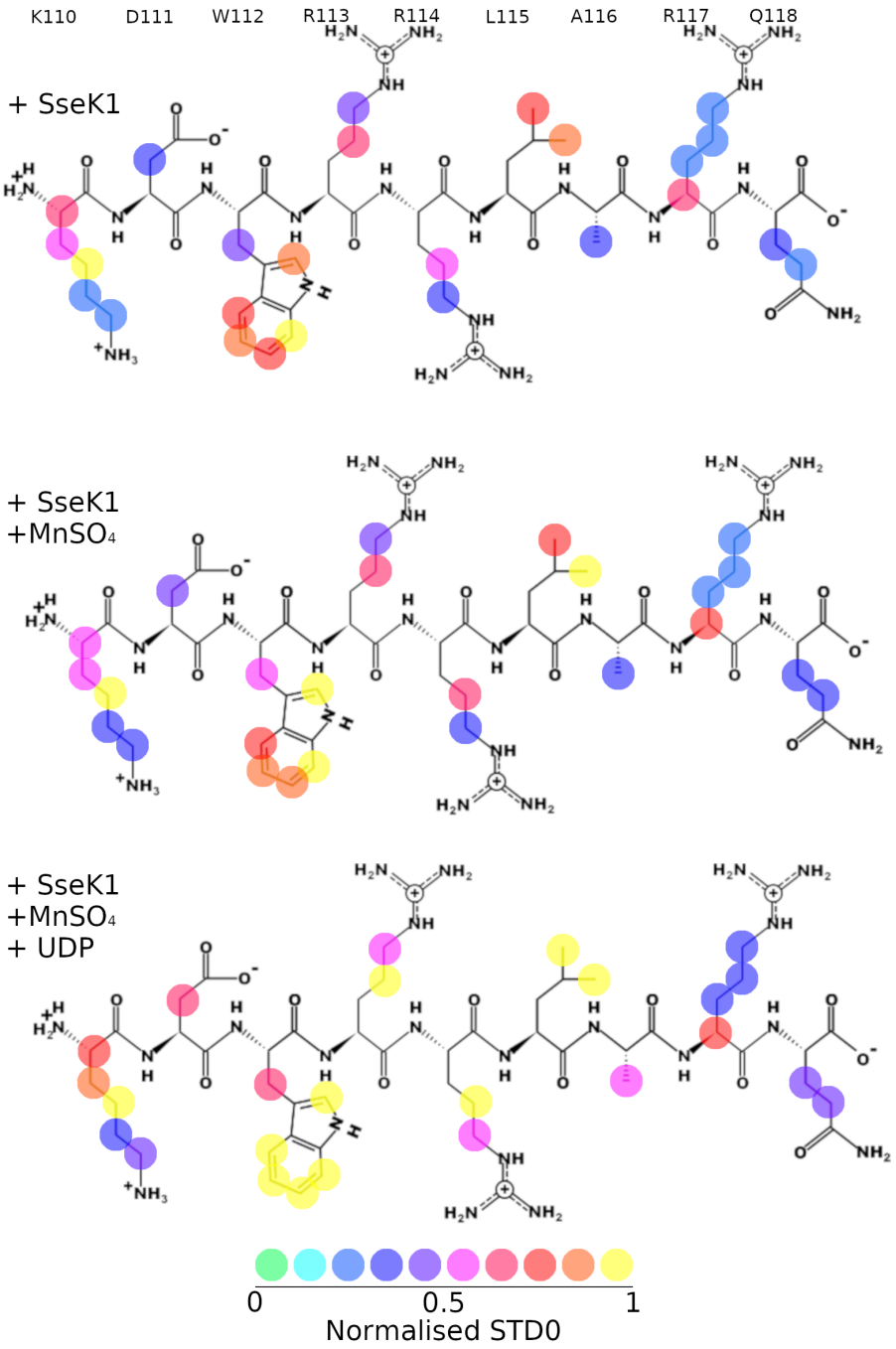

Figure 4.11: Binding epitope maps for the interaction of $FADD_{110\text{-}118}$ in the presence of SseK1, either in the apo-form (top), in the presence of Mn^{2+} (middle) or in the presence of Mn^{2+} and UDP (bottom). Colours represent the normalised STD values for proton resonances at the indicated positions (low - high, cold - hot).

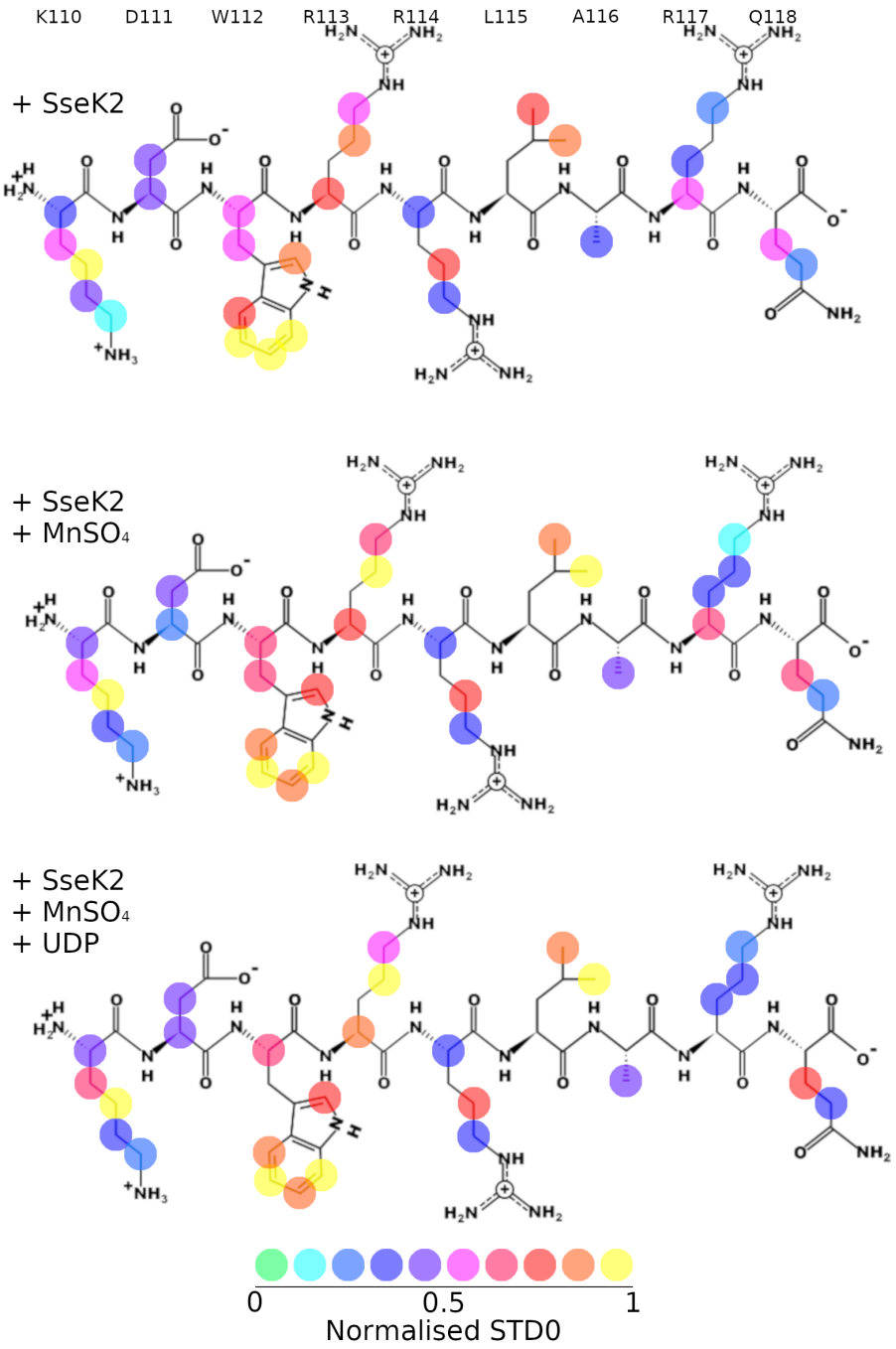

Figure 4.12: Binding epitope maps for the interaction of $FADD_{110\text{-}118}$ in the presence of SseK2, either in the apo-form (top), in the presence of Mn^{2+} (middle) or in the presence of Mn^{2+} and UDP (bottom). Colours represent the normalised STD values for proton resonances at the indicated positions (low - high, cold - hot).

4.3.2.2 Molecular recognition of $TRADD_{229\text{-}237}$ by SseK1 and SseK2

^{1}H STD NMR spectra characterising the binding of $TRADD_{229\text{-}237}$ to

SseK1 and SseK2, under different conditions, are shown in Fig. 4.13. The experimentally determined STD NMR intensities are collected in Tables 4.7 and 4.8, for the binding of $TRADD_{229\text{-}237}$ to SseK1 and SseK2, respectively.

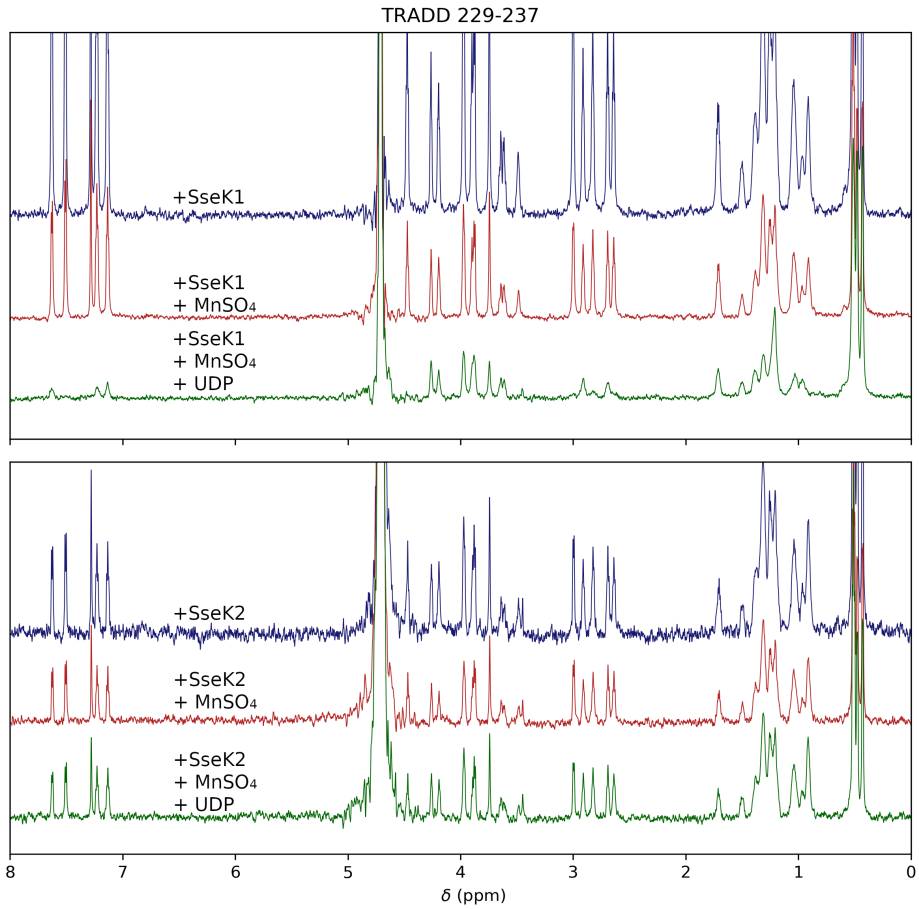

Figure 4.13: STD NMR difference spectra for $TRADD_{229\text{-}237}$ binding to SseK1 (top) or SseK2 (bottom). Spectra recorded for the apoenzyme (blue), in the presence of 25 μM Mn^{2+} (red) and in the presence of 25 μM Mn^{2+} and 25 μM UDP (green). Measured using a saturation frequency of 0 ppm at 800 MHz and a saturation time of 2 s.

For $TRADD_{229\text{-}237}$ binding to SseK1, strong STD intensity is only observed on the Trp230 residue in both the apo-SseK1 and Mn^{2+}-bound forms of SseK1 (Fig. 4.14). Upon addition of UDP, the STD intensity across $TRADD_{229\text{-}237}$ increases, but particularly across Trp230, Arg231 and Val233. Again, this suggests that binding of SseK1 to UDP induces some conformational change that positions Arg231 and Val233 in closer proximity to the surface of SseK1. Furthermore, the spread of saturation across $TRADD_{229\text{-}237}$ suggests higher affinity of $TRADD_{229\text{-}237}$ to the SseK1-Mn^{2+}-UDP complex compared to the apo- and Mn^{2+}-bound forms.

In the case of $TRADD_{229\text{-}237}$ and SseK2, like for $FADD_{1110\text{-}118}$, there appears to be no significant change in STD intensity upon addition of Mn^{2+} and/or UDP (Fig. 4.15). Instead, in all cases the STD intensity is strongest on Trp230 in all cases, with moderate STD intensity found concentrated around Lys229-Val233.

Table 4.7: STD intensities for a saturation time of 2 s for each proton resonance of $TRADD_{229\text{-}237}$ in the presence of SseK1.

Residue	Position	STD (%, SseK1)	STD (%, SseK1 + Mn^{2+})	STD (%, SseK1 + Mn^{2+} + UDP)
Lys229	β	5.57	5.9	7.17
	γ	7.24	7.53	12.03
Trp230	α	15.66	12.39	-
	β	12.63	11.69	16.77
	δ	19.79	18.49	26.26
	$\epsilon 3$	19.79	18.49	20.37
	$\zeta 2$	25.02	23.14	17.77
	$\zeta 3$	23.59	21.82	18.84
	$\eta 2$	22.47	20.58	19.59
Arg231	α	10.70	10.39	12.51
	γ	8.39	8.14	16.13
	δ	7.39	7.1	9.16
Lys232	γ	6.57	7.32	4.62
Val233	β	11.24	11.57	16.77
	γ	9.25	9.43	18.13
Arg235	α	8.07	8.22	9.15
	β	6.57	6.77	8.98
	γ	6.02	5.84	-
	δ	5.00	4.81	5.96
Ser236	α	8.55	8.31	7.68
	β	5.25	4.58	5.41
Leu237	α	9.07	8.22	8.47
	$\delta 1$	6.51	6.70	11.35
	$\delta 2$	6.83	7.10	12.03

Table 4.8: STD intensities for a saturation time of 2 s for each proton resonance of $TRADD_{229\text{-}237}$ in the presence of SseK2.

Residue	Position	STD (%, SseK2)	STD (%, SseK2 + Mn^{2+})	STD (%, SseK2 + Mn^{2+} + UDP)
Lys229	β	2.18	2.68	2.81

Residue	Position	STD (%, SseK2)	STD (%, SseK2 + Mn^{2+})	STD (%, SseK2 + Mn^{2+} + UDP)
	γ	2.90	3.35	3.59
Trp230	α	3.29	3.52	3.52
	β	2.14	2.60	2.76
	δ	4.86	5.68	5.62
	$\epsilon 3$	4.07	5.00	4.67
	$\zeta 2$	5.68	6.64	6.64
	$\zeta 3$	4.58	5.46	5.68
	$\eta 2$	4.68	5.15	5.35
Arg231	α	2.73	3.32	3.07
	δ	1.66	1.76	2.00
	γ	2.84	3.42	3.52
Lys232	γ	2.55	3.55	4.67
Val233	β	3.29	3.84	3.77
	γ	3.96	4.32	4.81
Arg235	α	2.73	2.29	3.10
	β	2.38	2.90	3.19
	δ	1.28	1.38	1.57
	γ	2.08	2.29	3.13
Ser236	α	2.48	2.84	3.04
	β	1.2	1.34	1.51
Leu237	α	2.53	2.78	3.25
	$\delta 1$	2.70	3.01	3.52
	$\delta 2$	2.95	3.32	3.84

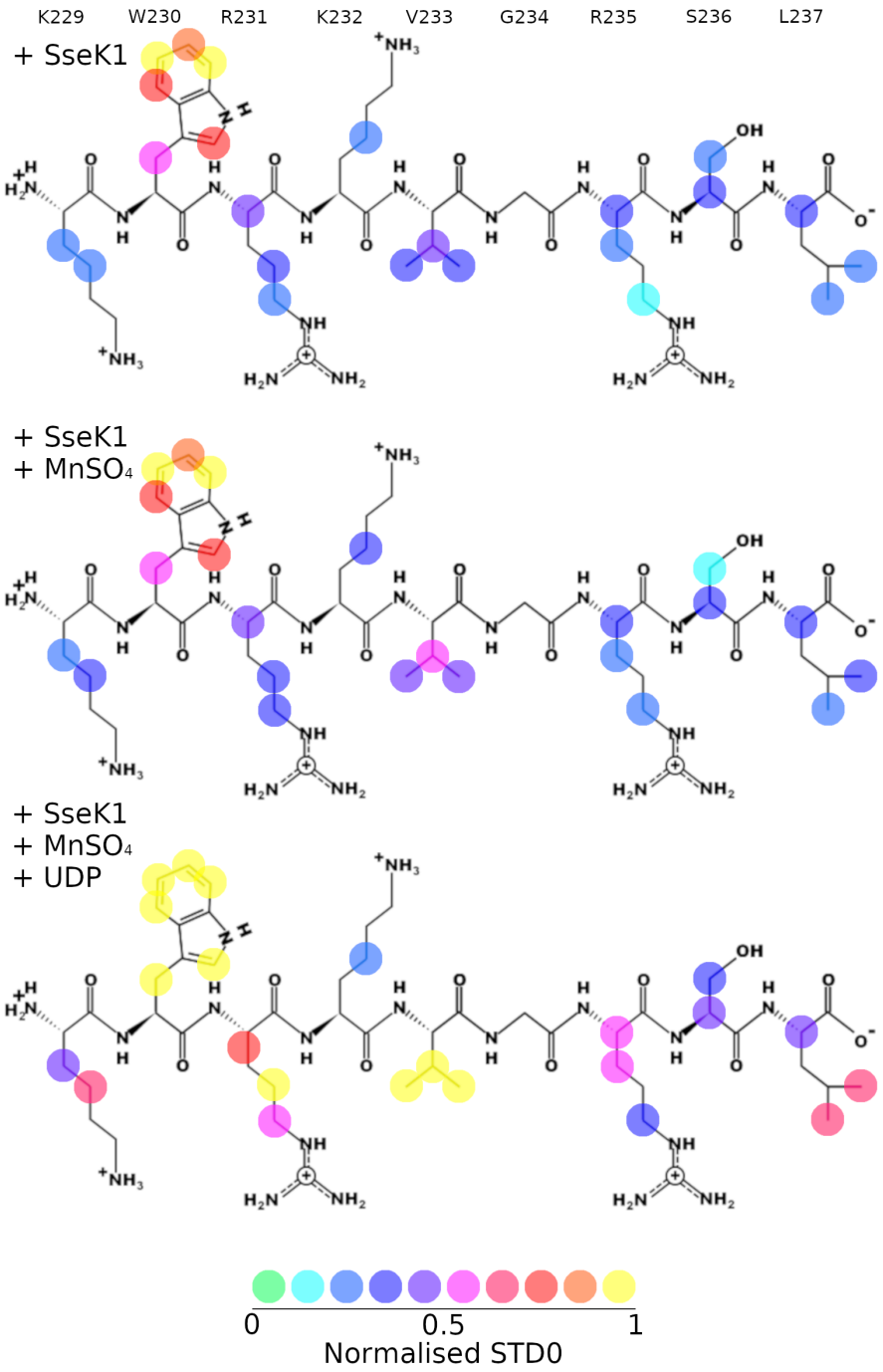

Figure 4.14: Binding epitope maps for the interaction of $TRADD_{229\text{-}237}$ in the presence of SseK1, either in the apo-form (top), in the presence of Mn^{2+} (middle) or in the presence of Mn^{2+} and UDP (bottom). Colours represent the normalised STD values for proton resonances at the indicated positions (low - high, cold - hot).

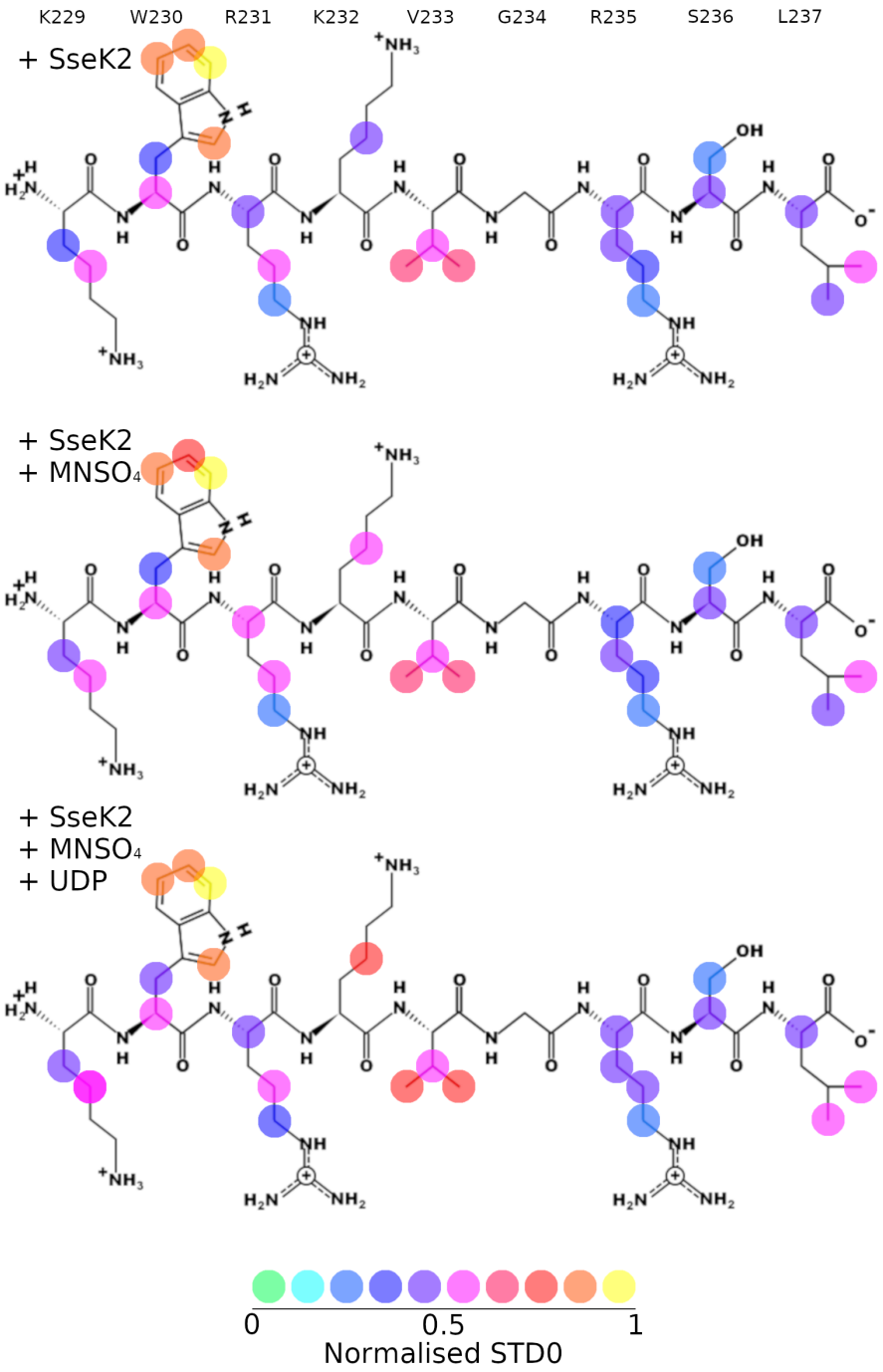

Figure 4.15: Binding epitope maps for the interaction of $TRADD_{229\text{-}237}$ in the presence of SseK2, either in the apo-form (top), in the presence of Mn^{2+} (middle) or in the presence of Mn^{2+} and UDP (bottom). Colours represent the normalised STD values for proton resonances at the indicated positions (low - high, cold - hot).

4.3.2.3 Molecular recognition of $GAPDH_{195\text{-}203}$ by SseK1 and SseK2

^{1}H STD NMR spectra characterising the binding of $GAPDH_{195\text{-}203}$ to

SseK1 and SseK2, under different conditions, are shown in Fig. 4.16. The experimentally determined STD NMR intensities are collected in Tables 4.9 and 4.10, for the binding of $GAPDH_{195\text{-}203}$ to SseK1 and SseK2, respectively.

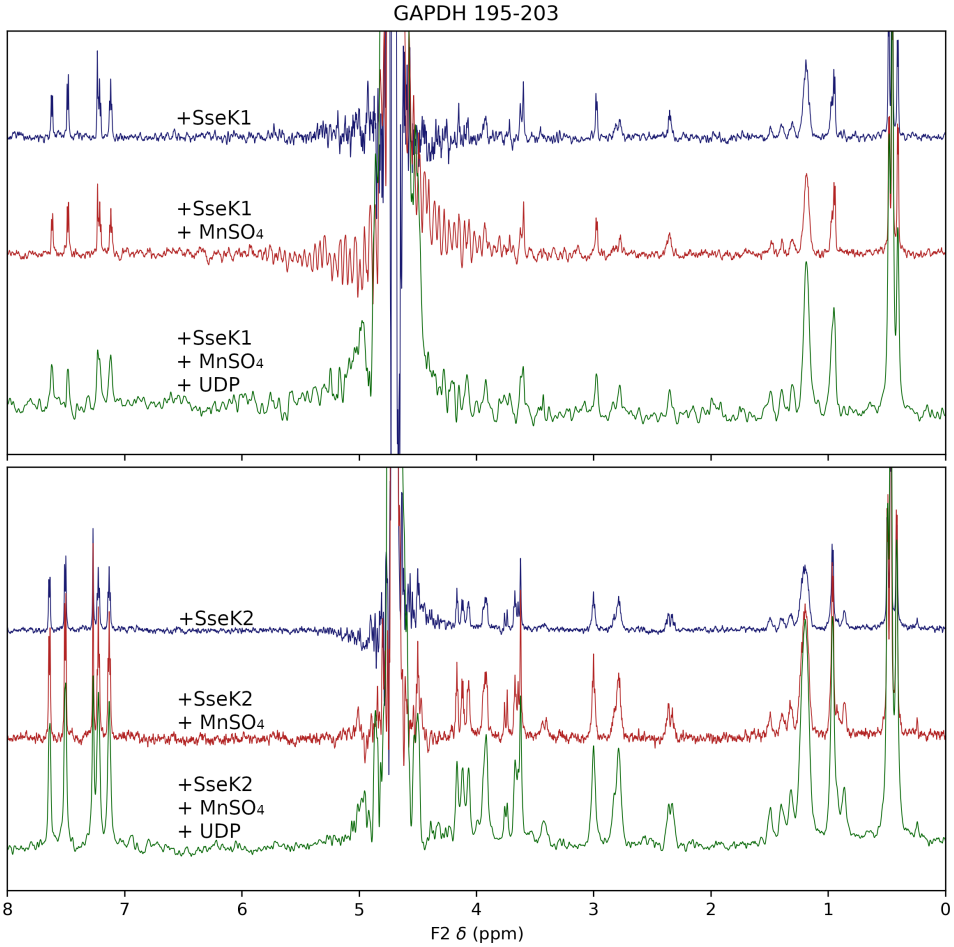

Figure 4.16: STD NMR difference spectra for $GAPDH_{195\text{-}203}$ binding to SseK1 (top) or SseK2 (bottom). Spectra recorded for the apoenzyme (blue), in the presence of 25 μM Mn^{2+} (red) and in the presence of 25 μM Mn^{2+} and 25 μM UDP (green). Measured using a saturation frequency of 0 ppm at 800 MHz and a saturation time of 2 s.

Finally, for $GAPDH_{195\text{-}203}$ binding to SseK1 in the apo-state, the only strong STD intensities are found on Trp196 (Fig. 4.17). Upon addition of Mn^{2+}, there is a small increase in the STD intensity of Arg200. Upon addition of UDP, moderate STD intensities are observed along the whole of $GAPDH_{195\text{-}203}$, although the β proton of Arg200 displays a particularly intense STD value.

In the case of $GAPDH_{195\text{-}203}$ binding to SseK2, strong STD intensities are only observed on Trp196 and there are no significant changes upon addition of Mn^{2+} and/or UDP (Fig. 4.18).

Table 4.9: STD intensities for a saturation time of 2 s for each proton resonance of $GAPDH_{195\text{-}203}$ in the presence of SseK1.

Residue	Position	STD (%, SseK1)	STD (%, SseK1 + Mn^{2+})	STD (%, SseK1 + Mn^{2+} + UDP)
Leu195	α	2.12	1.78	2.08
	γ	2.84	2.92	3.55
Trp196	β	3.22	2.68	2.50
	δ	4.90	3.77	3.29
	$\epsilon 3$	4.40	4.03	3.73
	$\zeta 2$	6.70	5.30	4.11
	$\zeta 3$	6.40	5.05	4.24
	$\eta 2$	6.63	4.85	4.40
Arg197	α	2.38	2.12	2.53
	β	3.95	3.62	4.72
	δ	1.45	1.48	1.98
Asp198	β	2.00	2.06	1.81
Arg200	β	2.78	3.35	3.96
	δ	1.83	1.19	1.73
Ala202	β	2.02	2.33	2.53

Table 4.10: STD intensities for a saturation time of 2 s for each proton resonance of $GAPDH_{195\text{-}203}$ in the presence of SseK2.

Residue	Position	STD (%, SseK2)	STD (%, SseK2 + Mn^{2+})	STD (%, SseK2 + Mn^{2+} + UDP)
Leu195	α	7.24	6.44	6.63
Trp196	β	6.51	7.83	7.03
	δ	10.70	10.70	10.50
	$\epsilon 3$	10.81	11.57	11.35
	$\zeta 2$	13.79	14.07	14.34
	$\zeta 3$	13.10	13.66	13.40
	$\eta 2$	12.76	13.40	13.40
Arg197	α	6.57	7.24	7.68
	δ	4.86	4.90	4.58
Asp198	α	7.17	7.53	6.02
	β	5.46	6.57	5.51
Arg200	α	6.64	6.97	6.44
	β	5.84	5.46	6.08
	δ	4.24	4.36	4.58
Ala202	α	5.73	6.08	5.62

Residue	Position	STD (%, SseK2)	STD (%, SseK2 + Mn^{2+})	STD (%, SseK2 + Mn^{2+} + UDP)
	β	4.62	4.90	4.81
Leu203	α	5.15	5.19	4.90

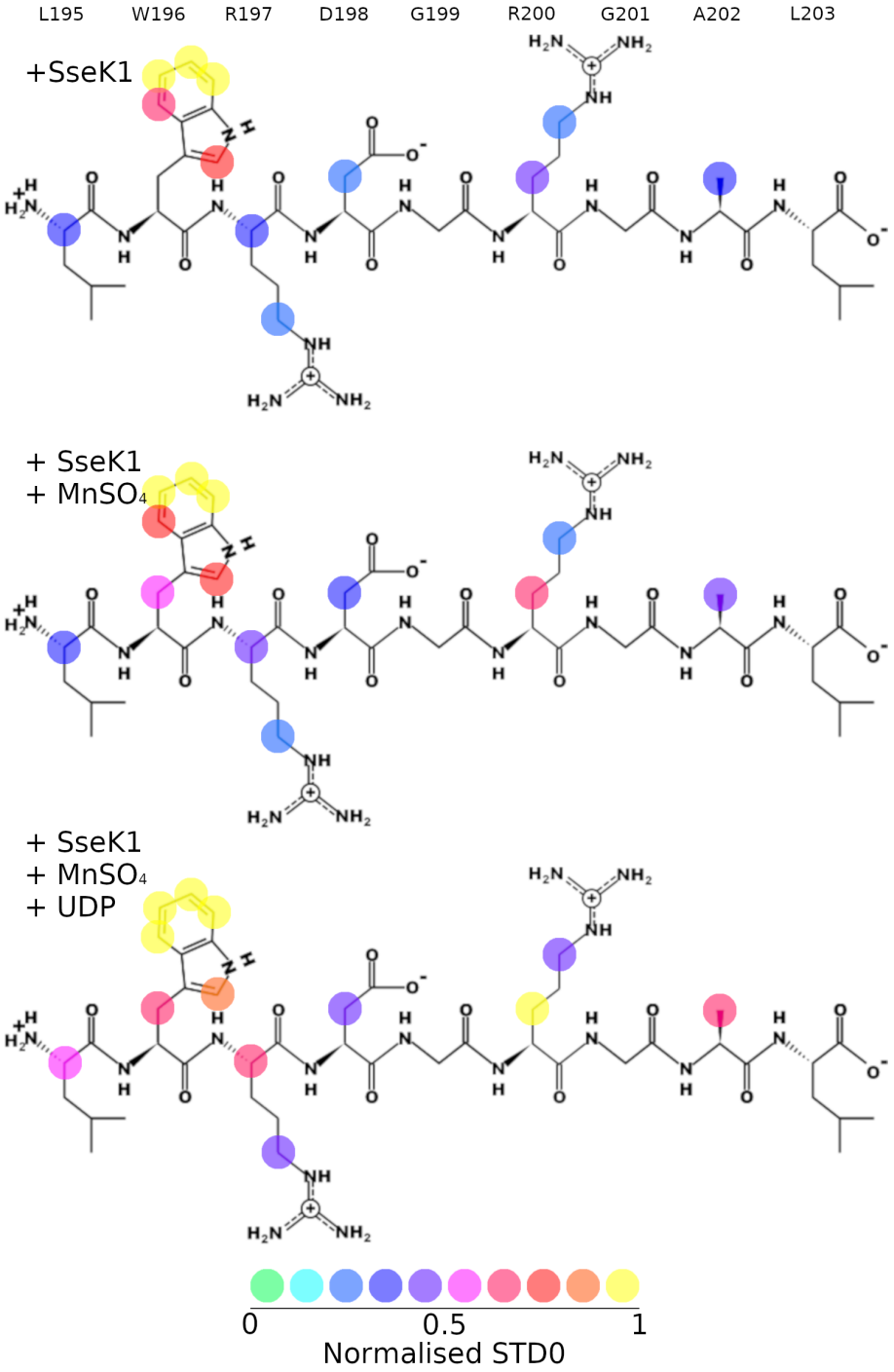

Figure 4.17: Binding epitope maps for the interaction of $GAPDH_{195\text{-}203}$ in the presence of SseK1, either in the apo-form (top), in the presence of Mn^{2+} (middle) or in the presence of Mn^{2+} and UDP (bottom). Colours represent the normalised STD values for proton resonances at the indicated positions (low - high, cold - hot).

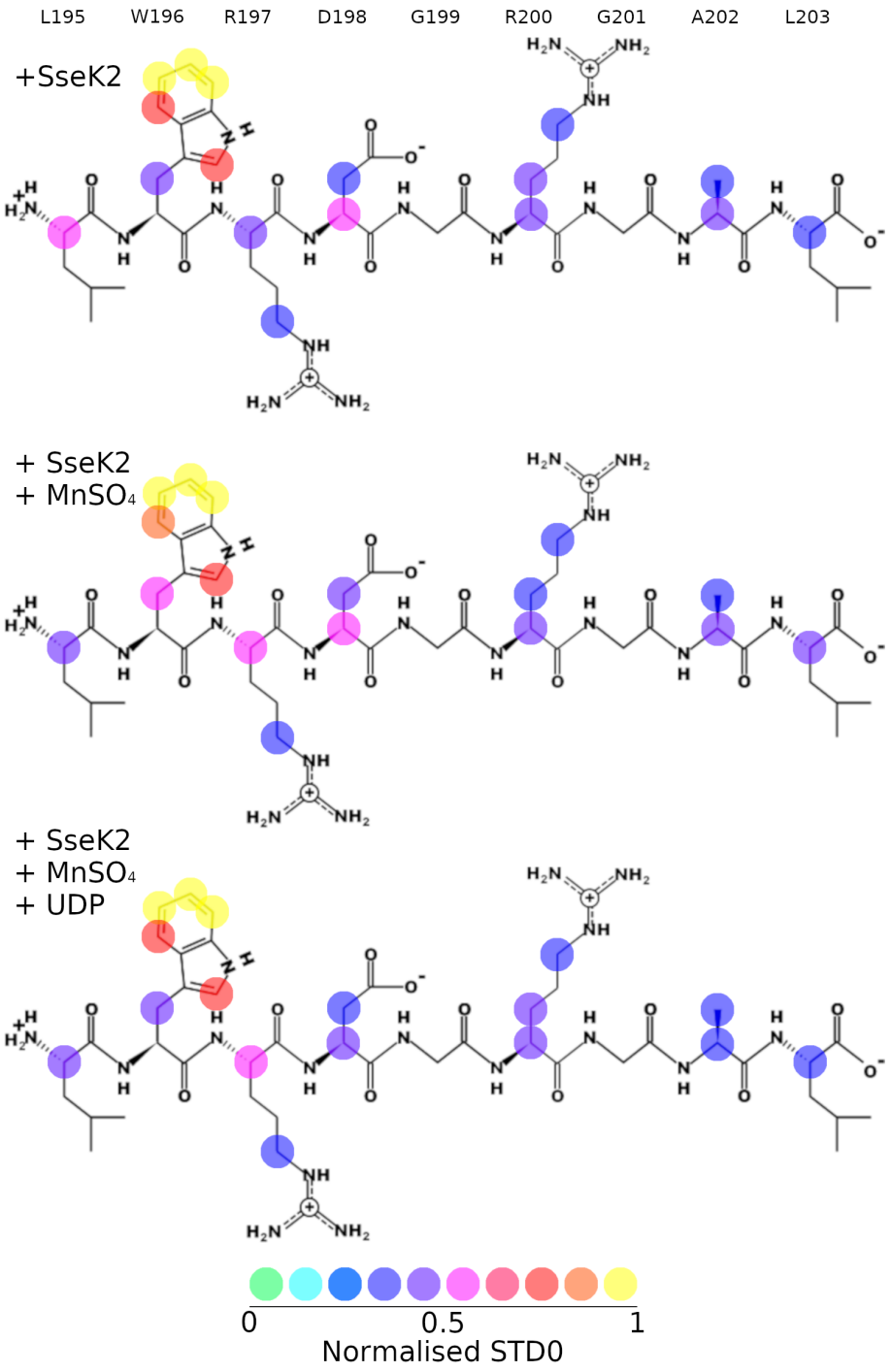

Figure 4.18: Binding epitope maps for the interaction of $GAPDH_{195\text{-}203}$ in the presence of SseK2, either in the apo-form (top), in the presence of Mn^{2+} (middle) or in the presence of Mn^{2+} and UDP (bottom). Colours represent the normalised STD values for proton resonances at the indicated positions (low - high, cold - hot).

Overall these results show that both SseK1 and SseK2 are capable of binding to each of the $FADD_{110\text{-}118}$, $TRADD_{229\text{-}237}$ and $GAPDH_{195\text{-}203}$ peptides, despite the differences in glycosylation of their full-length counterparts. This suggests that structure outside of these regions is responsible for determining whether or not the target becomes glyco-

sylated. In all cases, the tryptophan residue of the peptide displayed particularly strong STD intensities, indicating that this residue may be important for recognition. Finally, it appears that, in SseK1, the binding of Mn^{2+} and UDP induces some sort of conformational rearrangement that allows for tighter binding of the target substrate, but no such rearrangement occurs in the case of SseK2.

4.3.3 Accelerated Molecular Dynamics of SseK1 and SseK2

Accelerated molecular dynamics simulations were performed on the X-ray crystal models of SseK1 and SseK2 both bound to UDP-GlcNAc and manganese.[316] The root mean squared deviation (RMSD) of the SseK1 and SseK2 Cα atoms follow a rather jagged trajectory indicating some transient change in backbone conformation that occur many times over the period of the simulation (Fig. 4.19). However, the fact that on average the RMSDs do not change over the course of the simulations indicates that the structures are stable and the observed changes represent reversible conformational mobility of SseK1 and SseK2 respectively. The changes in RMSD are typically larger for SseK1, indicating a more significant conformational change.

In the case of SseK1, the bound UDP-GlcNAc also appears to have some degree of conformational mobility, although RMSD values between 1-2 Å most likely indicate small shifts in atomic displacement, not significant changes in binding mode or unbinding from SseK1. Conversely, the conformation of the bound UDP-GlcNAc in SseK2 is incredibly stable over the course of the simulation, with an RMSD of around 0.6 Å from the average structure that remains constant over the majority of the trajectory. There is a brief event just after 600 ns in which the RMSD increases to approximately 1.2 Å, but this rapidly reverts back to its previous conformation.

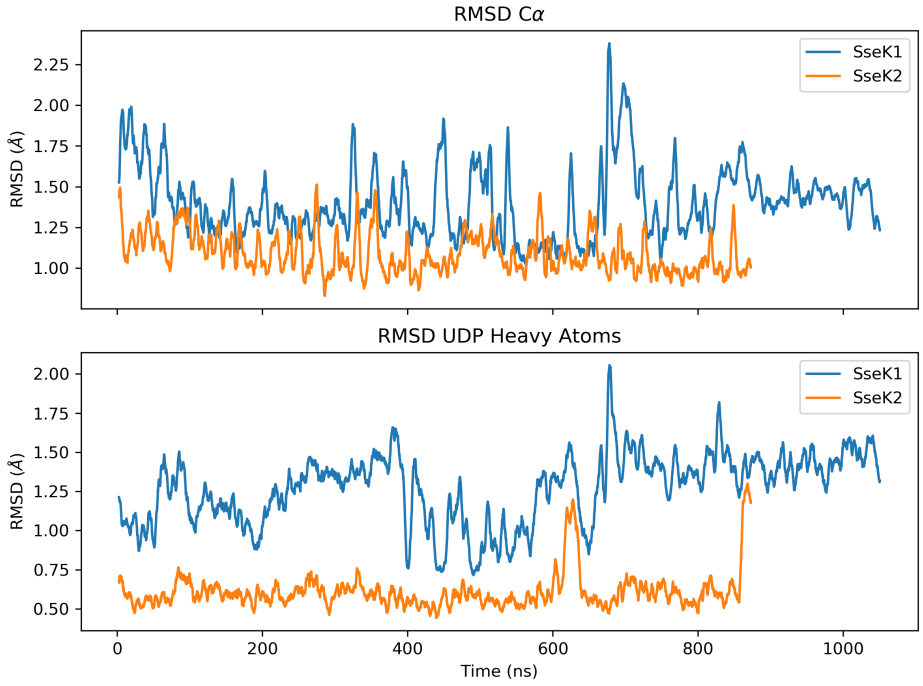

Figure 4.19: Root mean squared deviation (RMSD) as a function of time for SseK1 and SseK2 Cα atoms (top) and for UDP-GlcNAc heavy atoms bound to either SseK1 or SseK2 (bottom). For UDP-GlcNAc heavy atoms, a no-fit RMSD calculation was performed after first fitting to protein Cα atoms, in order to capture both conformational and translational motions.

To investigate the apparent conformational changes in the SseK1 and SseK2 backbones further, the root mean squared fluctuations (RMSF) of the Cα atoms were calculated (Fig. 4.20). It is immediately clear that the N- and C-terminal regions of SseK1 are significantly more mobile than in SseK2 and this may account for larger changes in RMSD observed for SseK1 compared with SseK2. Other than this, the RMSFs of SseK1 and SseK2 are comparable, although it appears that a larger number of residues within the region between residues 150-200 have heightened flexibility in SseK1 compared to SseK2. Mapping these fluctuations onto models of SseK1 and SseK2 (Fig.4.21) reveals that, in both cases there are a large number of residues with high RMSFs in the HLH region, although these fluctuations are slightly more intense in SseK1. For SseK1, the large fluctuations in the N- and C-terminal residues is also visible. For SseK2, there is a loop between two β-strands (hereafter referred to as beta-loop-beta; BLB) close to the donor substrate binding site that exhibits a large degree of conformational flexibility.

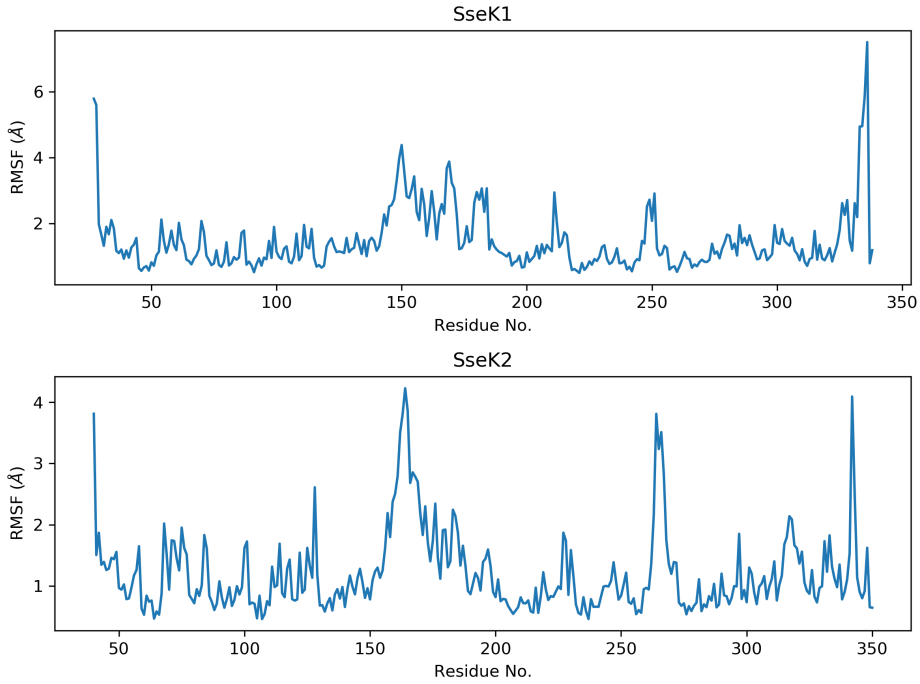

Figure 4.20: Root mean squared fluctuations (RMSF) of Ssek1 (top) and SseK2 Cα atoms calculated on a per residue basis.

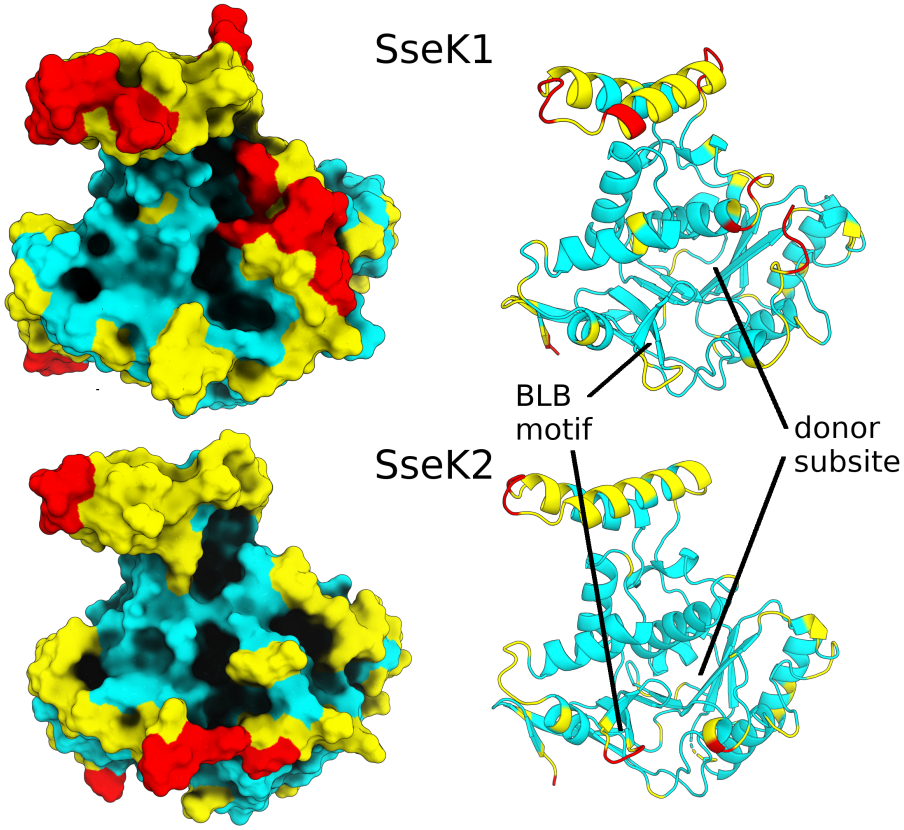

Figure 4.21: Per-residue root mean squared fluctuations of SseK1 (top) and SseK2 (bottom) mapped onto their respective X-ray crystal structures. Both surface (left) and cartoon representations are shown. Map shows RMSF values of larger than 3 Å(red), between 3 and 1.5 Å(yellow) and less than 1.5 Å(cyan).

Principal component analysis (PCA) can further elucidate the motions of a system by transforming the Cartesian space into 3N (where N is the number of atoms) eigenvectors (principal components; PCs), where each subsequent eigenvector describes less of the variance in the system. Importantly, projection of these vectors back into Cartesian space, allows the concerted motions of the system to be analysed. In the case of these simulations, the motions of the Cα atoms were considered. In both cases, the first three eigenvectors accounted for more than 40% of the variance in the backbone motion and the first principal component (PC1) of both systems accounted for 25% of the variance respectively (Fig. 4.22). Subsequent eigenvectors accounted for less than 1% of the motion each and so were not considered further.

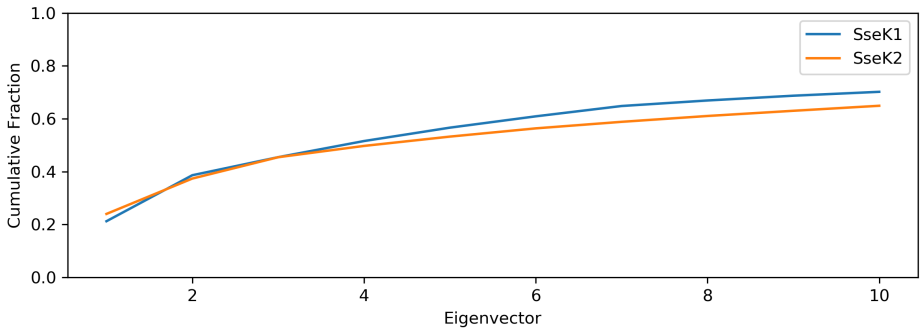

Figure 4.22: Cumulative fraction of total motion accounted for by the first 10 eigenvectors found by PCA for SseK1 and SseK2 Cα atom motion. For each system, all eigenvectors were calculated (3N where N is the number of Cα atoms) and the contribution of each eigenvector was calculated by dividing its eigenvalue by the sum of all eigenvalues.

For both SseK1 and SseK2, the predominant motion (PC1) is rotation of the HLH domain towards the catalytic site (Fig. 4.23). For SseK1, the next two dominant motions are movement of the N-terminal (PC2) and C-terminal (PC3) regions. In fact, PC3 indicates a distinct opening of the lid such that the donor substrate binding site is exposed. Conversely, the RMSF plots and PCA projections of SseK2 show that there is no significant movement of the C-terminal lid domain. It is interesting that in these two enzymes, the lid apparently opens over different timescales (unobservable for SseK2) as this may have an impact on their enzymatic activity. The opening of the SseK1 C-terminal lid may also explain the larger changes in RMSD of the UDP-GlcNAc ligand in SseK1 compared to SseK2 due to the lack of stabilising interactions with the C-terminal lid. For SseK2, PC2 and PC3 involve tilting of the HLH domain and BLB motif towards the catalytic site.

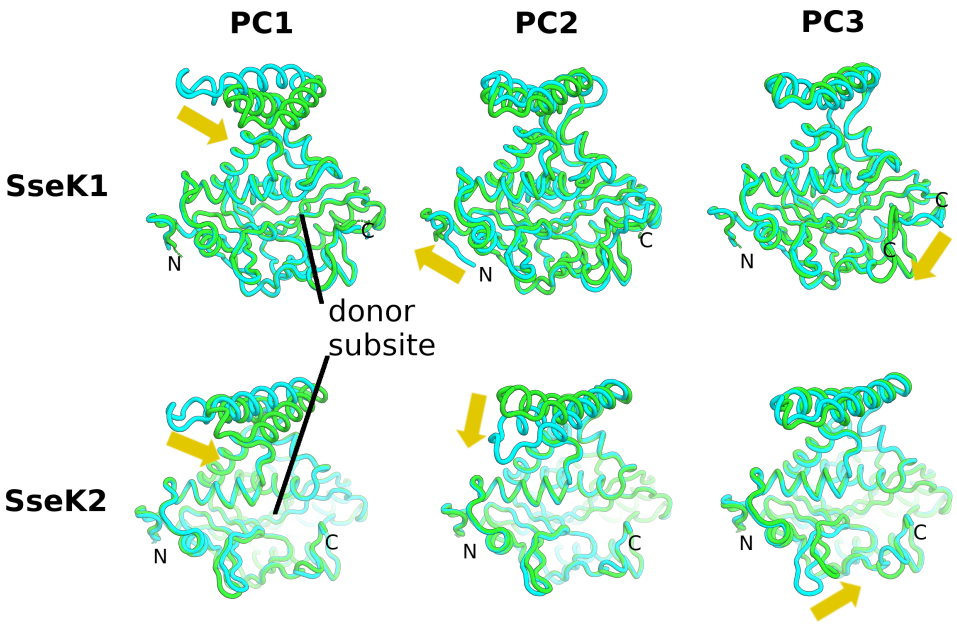

Figure 4.23: The first three principal components (PCs) of PCA analysis for SseK1 (top) and SseK2 (bottom) projected back onto Xray crystal structures of either protein. Motions of each PC are highlighted (yellow arrow)

4.3.4 Molecular Docking of FADD to SseK2

In the crystal structure of SseK2, it is clear that the C-terminal lid domain of SseK2 exists in either a closed or an open conformation. Since the donor substrate binding site is completely obscured from the solvent in the closed conformation, a model of the open conformation was produced my modelling the missing C-terminal residues in the open conformation. The model of the $FADD_{110\text{-}118}$ peptide, produced by truncation of an X-ray crystal structure model of FADD (PDB: 3EZQ), was then docked to the open conformation of SseK2 (which alos contained the coordinates of Mn^{2+} and UDP-GlcNAc in the donor binding site).

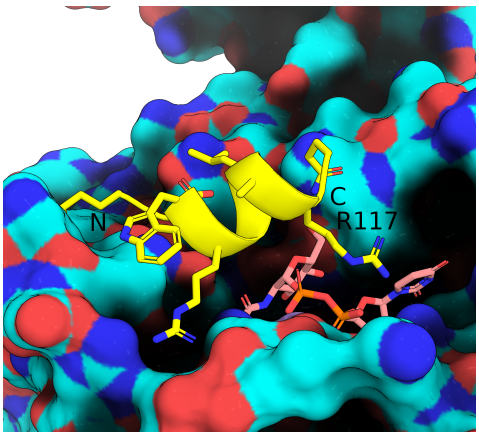

Figure 4.24: Cartoon representation of the docking model of the $FADD_{110\text{-}118}$ peptide (yellow) in complex with SseK2 (cyan surface).

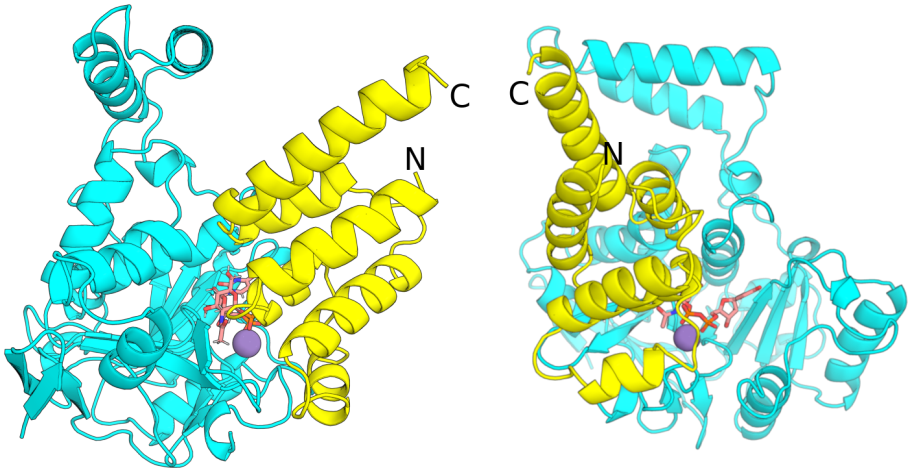

Figure 4.25: Cartoon representation of the model complex of FADD (yellow) and SseK2 (cyan) derived from molecular docking of the $FADD_{110\text{-}118}$ peptide.

The representative structure of the most populated cluster from molecular docking of $FADD_{110\text{-}118}$ to SseK2 shows the peptide bound close to the UDP-GlcNAc donor, in a longitudinal mode with the N-terminal tilted towards SseK2 (Fig. 4.24). This positions Trp112 in close proximity to the protein surface, as well as the other N-terminal residues of $FADD_{110\text{-}118}$, while the C-terminal residues are far from the SseK2 surface. This is in qualitative agreement with the STD-NMR data, which shows that saturation is concentrated around the N-terminal residues of $FAD_{110\text{-}118}$, in particular Trp112. Furthermore, Arg117 of $FADD_{110\text{-}118}$ is positioned in close proximity to the UDP-GlcNAc donor substrate in the docking model, which is known to be the target residue for glycosylation.[317]

To produce a model of the full length FADD protein in complex with SseK2 the backbone atoms of residues 110-118 of the full length FADD model (PDB 3EZQ) were superimposed onto the docked peptide. After refinement, this produced a complex with no atomic clashes. Interestingly, the C-terminal helix of FADD was positioned in close proximity to the HLH of SseK2.

4.3.5 Molecular Dynamics Simulations of the SseK2:FADD complex

Long accelerated molecular dynamics simulations were run on the model of the SseK2/FADD complex derived from molecular docking (Fig. 4.25). The root mean squared deviation (RMSD) of the SseK2 Cα atoms did not change significantly over the course of the simulation (Fig. 4.26 top) indicating that the SseK2 model was not destabilised by the presence of FADD and that there is not a significant structural rearrangement of the SseK2 backbone due to the presence of FADD (other than the opening of the C-terminal lid, which was performed during the docking stage). The sharp spikes in RMSD are comparable to those seen in the simulation of SseK2 in the absence of FADD, and are therefore likely due to rotation of the HLH previously observed.

The RMSD of FADD Cα atoms was calculated by first calculating the RMSD of SseK2 Cα atoms and then performing a no-fit calculation on the FADD Cα atoms, in order to observe any translations or rotations of FADD relative to SseK2. The decrease in RMSD from the average structure over the first 400 ns indicates that the binding pose of FADD derived from docking was not optimal and the it took approximately 400 ns to equilibrate the complex (Fig. 4.26 middle). Therefore further analysis will only consider frames from the trajectory after 400 ns.

In a similar manner the RMSD of UDP-GlcNAc heavy atoms appears to decrease over the first 400 ns before stabilising (Fig. 4.26 bottom), indicating that the presence of FADD does cause some small change to either the position or conformation of UDP-GlcNAc.

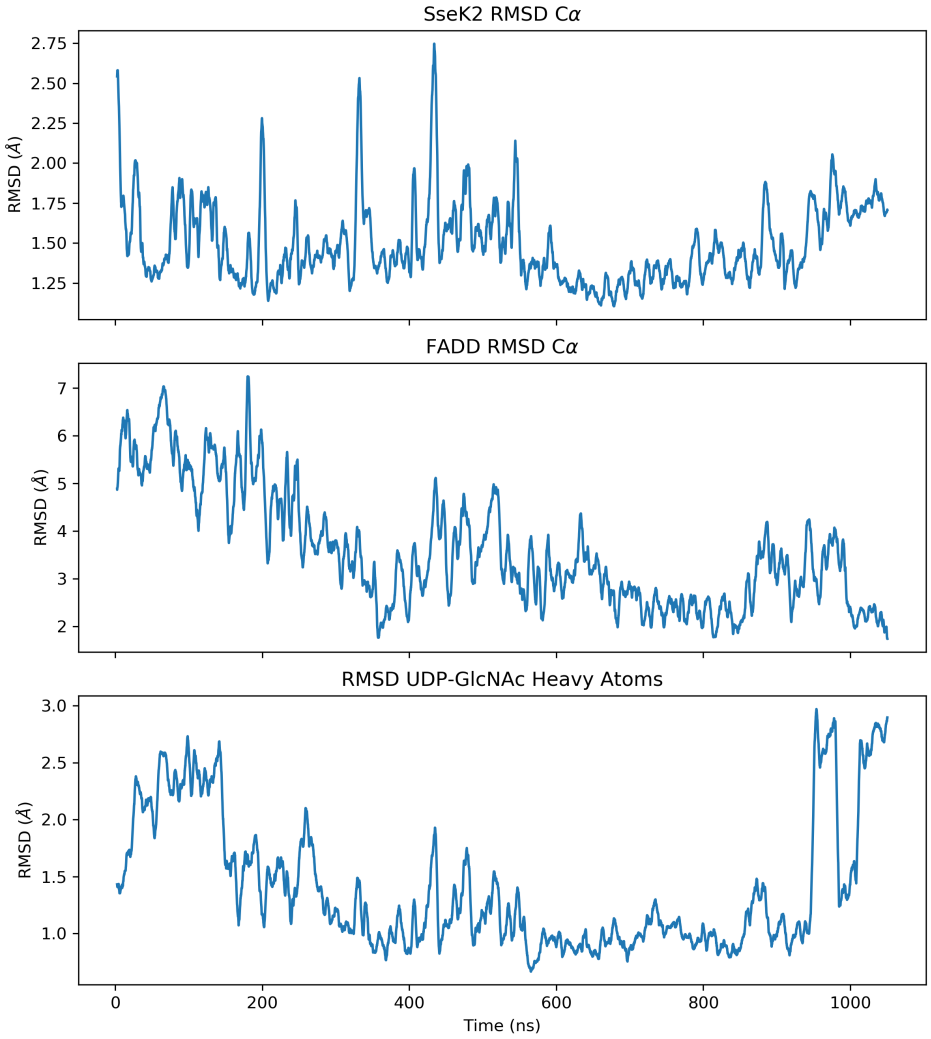

Figure 4.26: Root mean squared deviation (RMSD) as a function of time for SseK2 Cα atoms (top), FADD Cα atoms (middle) and for UDP-GlcNAc heavy atoms (bottom). For FADD Cα and UDP-GlcNAc heavy atoms, a no-fit RMSD calculation was performed after first fitting to SseK2 Cα atoms, in order to capture both conformational and translational motions.

Clustering based on the Cα atoms of SseK2 and FADD was performed and it was found that 5 clusters was optimal (Fig. 4.27). Together the first two clusters accounted for 59% of the trajectory after 400 ns. In these clusters the HLH was in close proximity to the C-terminal helix of FADD (Fig. 4.28), the only significant difference being in that in the most populated cluster there is a small kink in the FADD C-terminal helix. In the less populated clusters, the HLH is rotated away from FADD.

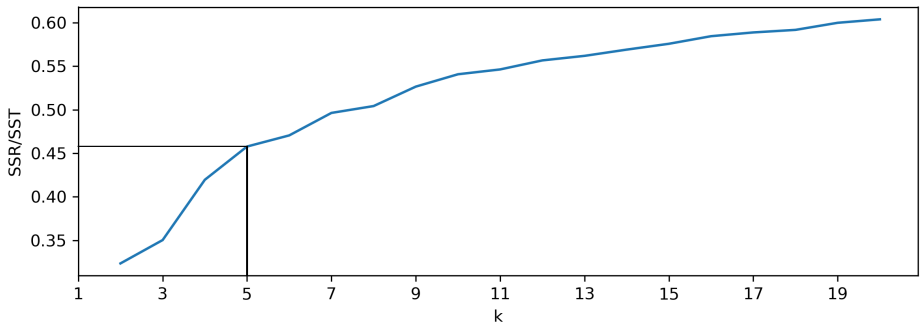

Figure 4.27: Elbow plot showing the SSR/SST (SSR: sum of squared residuals, SST: total sum of squares) ratio for different number of clusters in k-means clustering of SseK2 and FADD Cα heavy atoms. The optimal number of clusters is that which gives the best tradeoff between fraction of variance explained (SSR/SST) and number of clusters, which is at the 'elbow' of the plot - in this case 5.

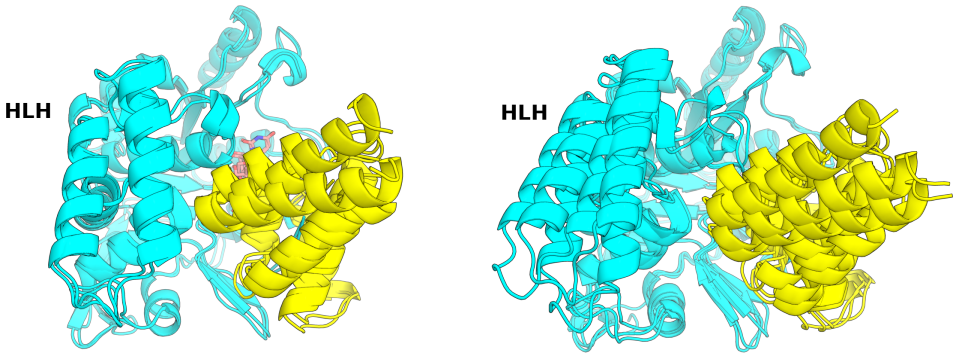

Figure 4.28: Cartoon representation of the centroid structures of the clusters from k-means clustering (left: clusters 1-2, right: clusters 3.5) of the simulation of the SseK2 (cyan) FADD (yellow) complex.

Analysis of the most populated cluster reveals a large amount of complementarity between adjacent residues of SseK2 and FADD (Fig. 4.29). For example, in the HLH domain of SseK2 there is a hydrophobic patch (Val169) followed by several charged residues (Lys176, Asp180) that form complementarity interacts with FADD (Val172, Asp175, Arg166). Furthermore, within the target region of FADD (residue 110-118) the residues Arg113 and Arg114 interact with Glu271 and Asp299 of SseK2 respectively. There are also several positively charged residues (Arg263, Lys264) within the BLB region of SseK2 that interact with complementary residues on FADD (Asp131 and Glu130). Curiously Asp124 of FADD appears to chelate the manganese ion of SseK2. Of the three arginine residues present in the acceptor region of FADD Arg117 appears to be in close proximity to the anomeric carbon and the β-phosphate of UDP-GlcNAc showing that it may be positioned correctly for glycosylation (Fig. 4.30). We

also monitored the distances of the three arginine residues to His260 and Glu271, which are implicated in the enzymatic mechanism, although no stable interactions are observed here.

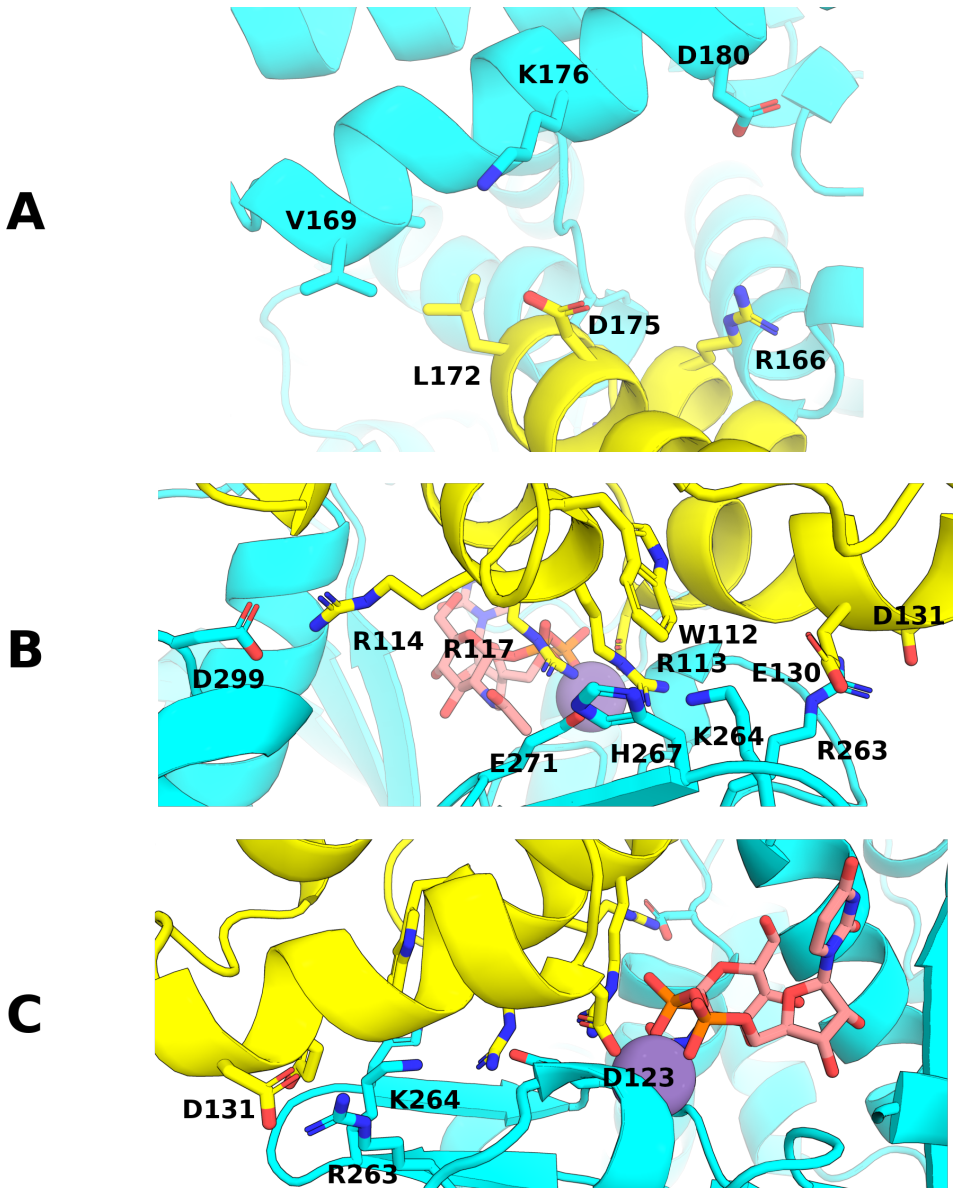

Figure 4.29: Cartoon representation of the centroid structure of the most populated cluster from k-means clustering of the simulation of the SseK2 (cyan) FADD (yellow) complex. Complementary interactions between SseK2 and FADD highlighted (sticks).

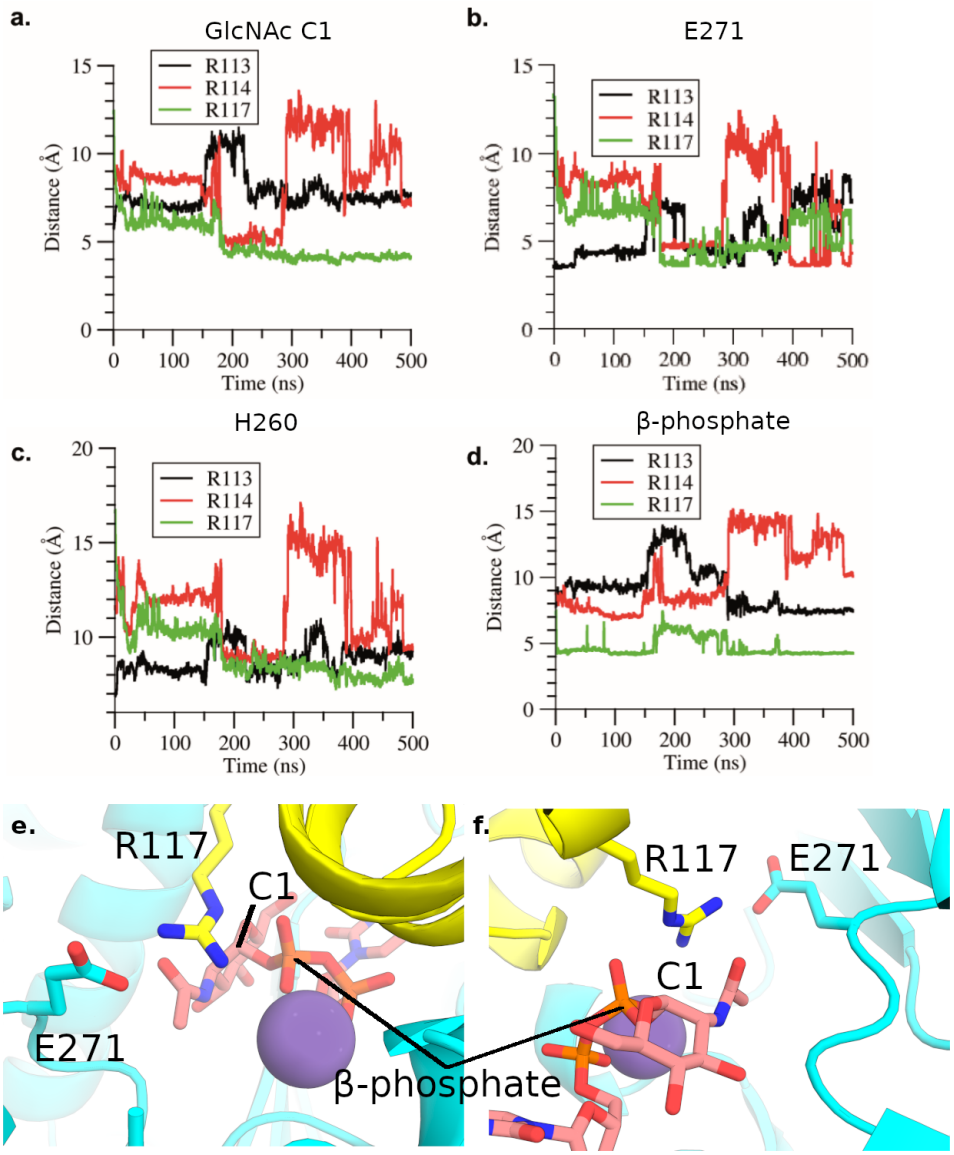

Figure 4.30: Molecular dynamics of the docking model of the ternary SseK2:UDP-GlcNAc:FADD complex show significant conformational rearrangements of Arg113, Arg114, and Arg117 of FADD, orienting Arg117 for a front face attack to GlcNAc, and support the relevance of Glu271 and His260 in acceptor FADD substrate binding. Calculated distances (Å) for contacts between the center of mass of the guanidinium groups of Arg113 (black lines), Arg114 (red lines), and Arg117 (green lines) of FADD and a, the anomeric carbon of GlcNAc of UDP-GlcNAc, b, the center of mass of the carboxylate group of Glu271 of SseK2, c, the center of mass of the imidazole group of His260 of SseK2, d, the center of mass of the beta-phosphate of UDP-GlcNAc. On average, Arg117 of FADD is the residue from FADD closest to the anomeric carbon of GlcNAc and is the only residue establishing close contacts with His260, Glu271, and the beta-phosphate of UDP-GlcNAc. e, Representative structure of the ternary complex of SseK2 (cyan), FADD (yellow) and UDP-GlcNAc (pink) from GaMD simulations, showing that Arg117 is properly oriented for a front face attack with close contact with the anomeric carbon of the UDP-GlcNAc donor substrate. Manganese is shown as a purple sphere and hydrogen atoms are omitted for clarity.

The residue pairs within the HLH were monitored over the course of the trajectory to determine the stability of their interactions (Fig. 4.31). Early in the simulation, the distances between these residues was large, before coming together after about 600 ns. After this point the interaction appears to be stable. This further suggests that the HLH may be involved in recognition since it is able to find and form stable interactions with the acceptor substrate. Furthermore, the interaction between the FADD arginine residues within the target region (Arg113, Arg114 and Arg 117) and their partners in SseK2 (Glu271, Asp299 and UDP-GlcNAc C1) was monitored. Again the interaction was stabilised after approximately 600 ns. Finally it appears that the interaction between the Mn^{2+} ion of SseK2 and Asp124 was stable much earlier in the simulation and remains so for the duration of the simulation.

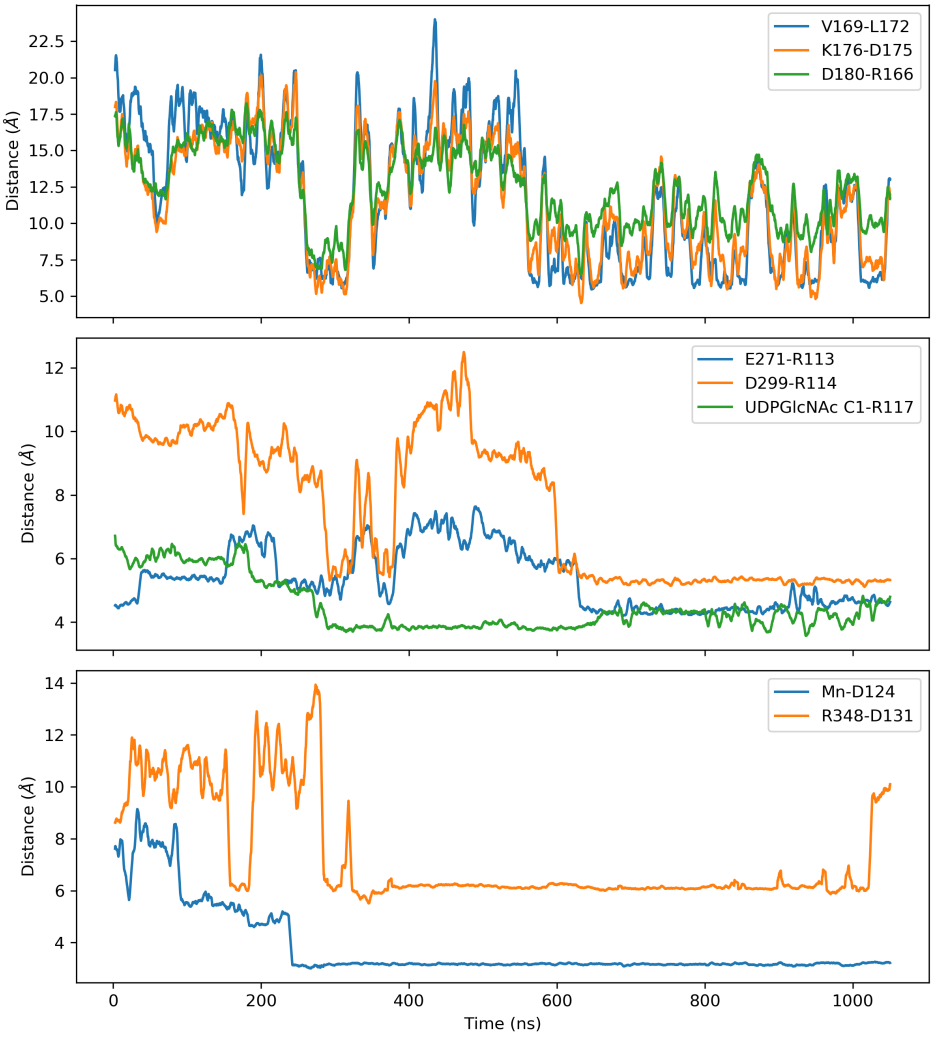

Figure 4.31: Distance measurements for residues involved in complementary interactions between SseK2 and FADD as a function of time. Residues labelled as:SseK2 residue - FADD residue.

4.3.6 NMR spectroscopy of ^{15}N-labelled FADD in Complex with SSeK2

To add to the experimental evidence of the proposed complex of SseK2 and FADD, ^{15}N-labelled FADD was first expressed from a recombinant plasmid in *E. coli*, following a previously published protocol.[316] FADD was purified using Ni-NTA affinity column and appeared on the SDS-PAGE gel at approximately 14 kDa, close to the expected value of 12 kDa (Fig. 4.32).

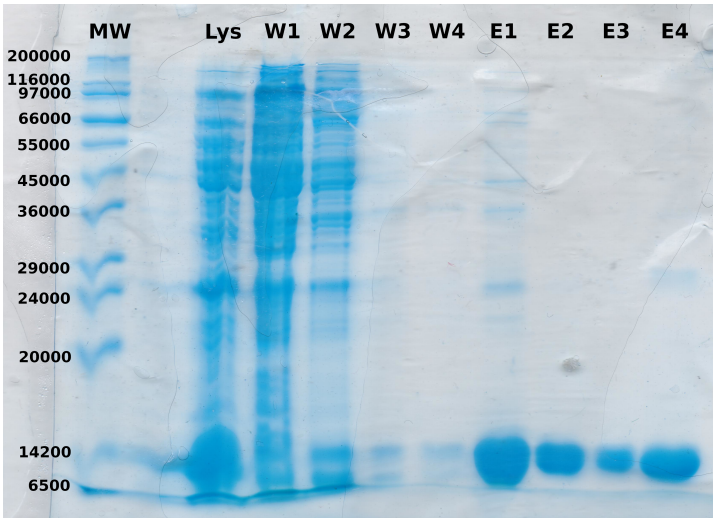

Figure 4.32: SDS-PAGE gel of ^{15}N-FADD expression and purification showing the soluble lysate fraction, four wash fractions (W1-4) and four elution fractions (E1-4). FADD is visible at approximately 14 kDa in the lysate and elution fractions.

^{1}H-^{15}N HSQC experiments were performed for ^{15}N-FADD in the absence and presence of SseK2. Due to solubility issues of SseK2, it was not possible to exceed a protein:ligand ratio of 1:1, meaning that the FADD binding sites would not have been fully saturated. Still there are significant differences between the two spectra, which can be used to qualitatively describe the residues of FADD that are affected by SseK2 binding. Overall the NMR signal of ^{15}N-FADD was reduced in the presence of SseK2 (Fig. 4.33). This is expected, since the increased correlation time of the complex would shorten the transverse relaxation time (T_2) of the ^{15}N-FADD resonances and therefore lead to reduced signal. However, there are some residues that are in the ^{15}N-FADD spectrum that are strongly affected by the presence of SseK2 and disappear from the spectrum entirely. This is indicative of restricted mobility of these residues within the complex and there-

fore suggests that these residues are at the interface between FADD and SseK2 in the complex. The FADD residues affected most significantly by binding to SseK2 are Gly93, Val103, Asn107, Gly109, Arg113, Arg114, Leu115, Asp123, Thr124, Ile129, Tyr133, Leu137, Glu139, Arg140, Val141, Glu152, Glu154, Thr157, Cys168, Asn171, Val173 and Gly191.

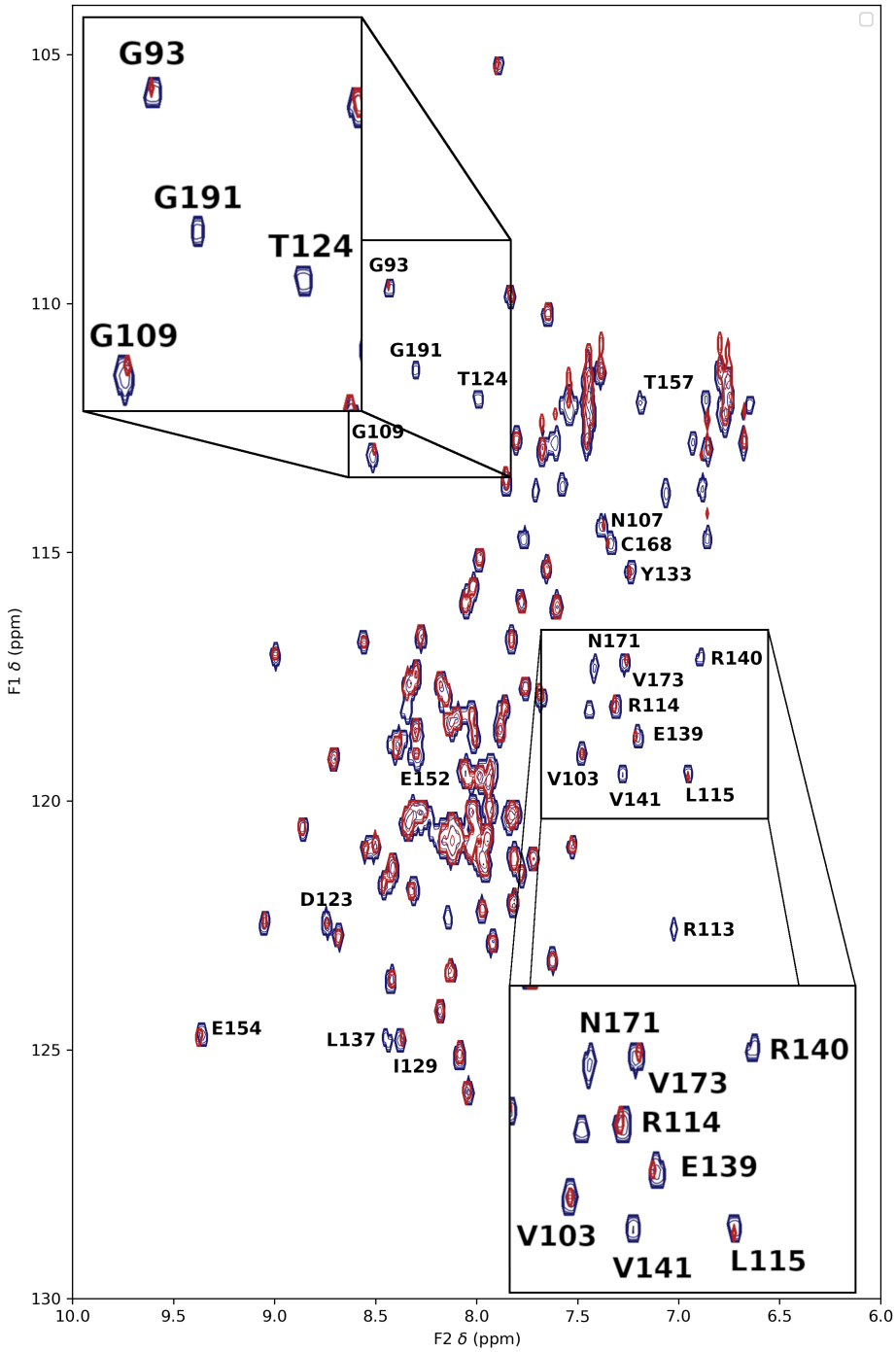

Figure 4.33: The ^{1}H-^{15}N HSQC NMR spectra of ^{15}N-labelled FADD in the absence (blue) or presence of SseK2 (red). Resonances that are significantly perturbed by the presence of SseK2 are labelled with their corresponding residues. ^{15}N-FADD present at 0.3 mM with an equimolar concentration of SseK2. Spectra acquired at 500 MHz.

These residues were then mapped onto the model of the SseK2:FADD complex (Fig. 4.34). The most significantly perturbed residues of FADD were those in the model of the FADD/SseK2 complex close to either the catalytic site or the HLH. Overall the mapping of the FADD residues affected by SseK2 binding provides good experimental evidence for the model of the SseK2:FADD complex.

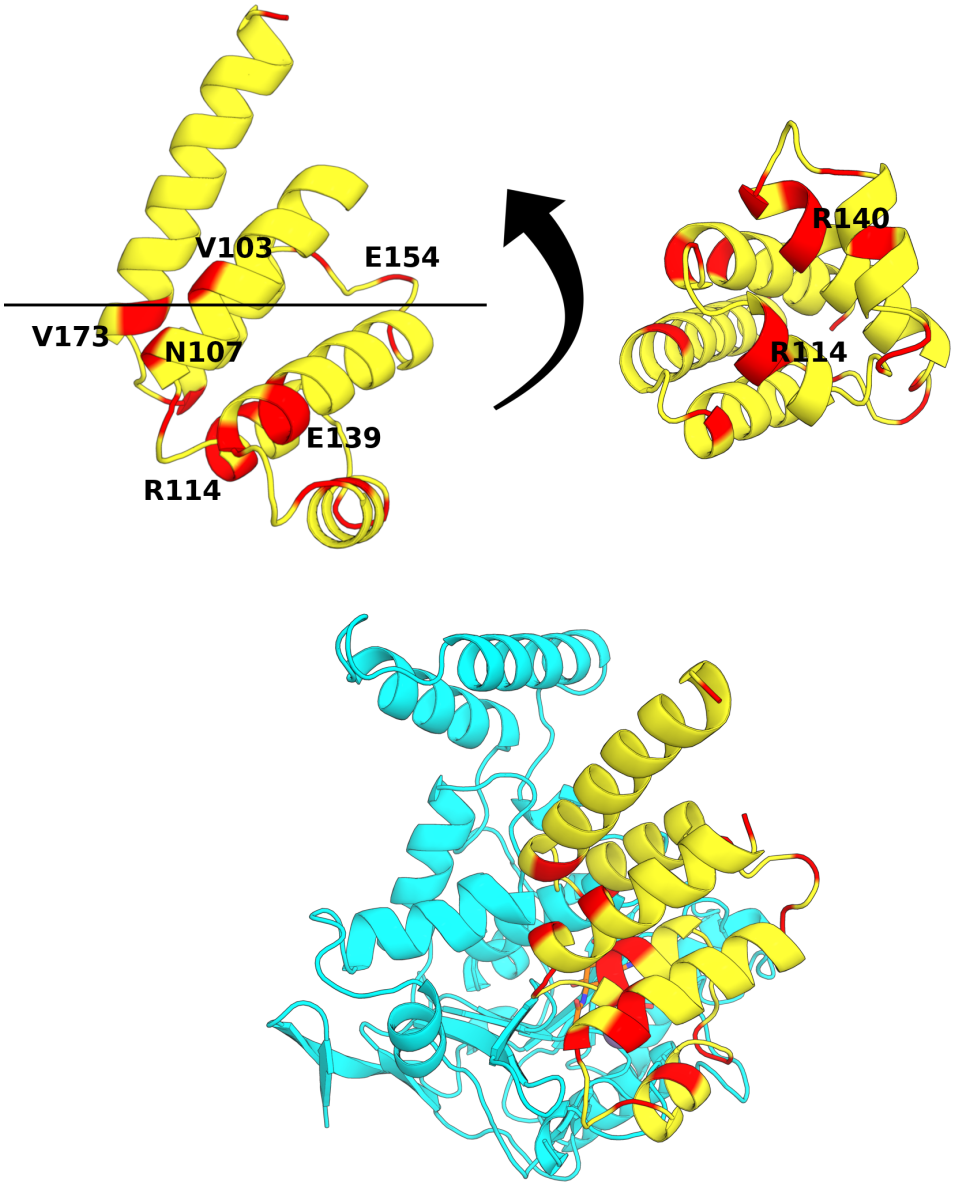

Figure 4.34: Cartoon representation of FADD (yellow) showing residues affected in the ^{1}H-^{15}N HSQC spectrum (red) by FADD binding to SseK2 (cyan). Shown on a model of FADD only (top, PDB: 3EZQ) and on a model of the FADD/SseK2 complex.

4.3.7 Mechanism of Glycosylation of $GAPDH_{187\text{-}203}$ by SseK1

To determine whether SseK1 is an inverting or retaining enzyme, $GAPDH_{187\text{-}203}$ was reacted with SseK1 and the unpurified reaction mixture was analysed by NMR. The shorter $GAPDH_{195\text{-}203}$ construct was tested prior to this, but affinity was too low to produce enough glycosylated $GAPDH_{195\text{-}203}$ for the NMR experiment.

The decoupled ^{1}H-^{13}C HSQC NMR spectra of the reaction mixture (Fig. 4.35a) showed the presence of four resonances in the anomeric region of the spectrum. Three of these could be assigned to UDP-GlcNAc, αGlcNAc and βGlcNAc, whilst the fourth previously uncharacterised peak was assigned to GlcNAc covalently bound to an arginine sidechain. This resonance is significantly shifted upfield in the ^{13}C dimension relative to the other anomeric resonances, showing a more electron-rich environment, compatible with being bound to the conjugated π system of an arginine sidechain. Furthermore, a comparable upfield shift has been observed for the anomeric carbon of rhamnose glycosidically linked to arginine.[325]

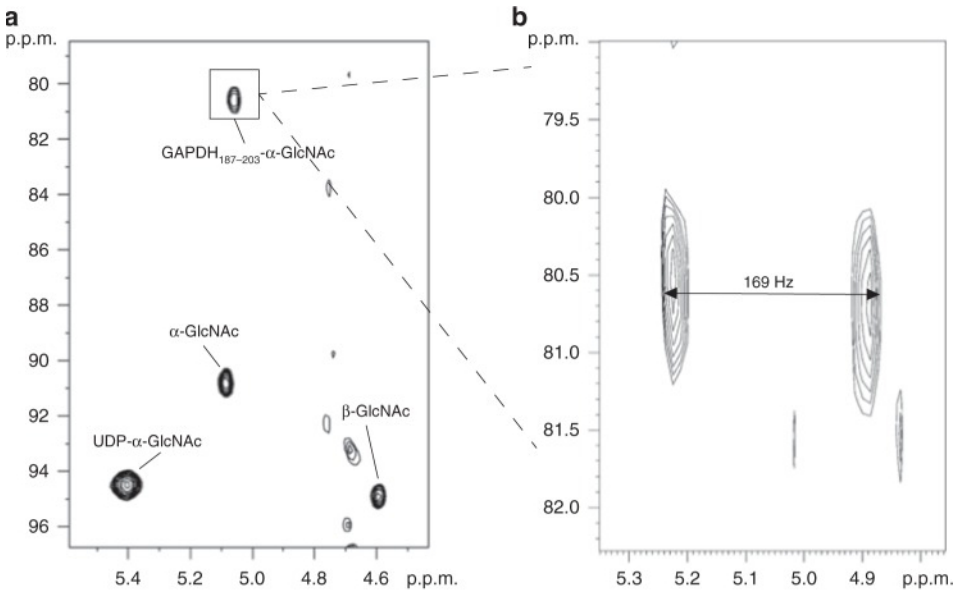

Figure 4.35: SseK1 is a retaining-glycosyltransferase. NMR spectra showing the reaction product of $GAPDH_{187\text{-}203}$ with SseK1. **a**) Decoupled ^{1}H-^{13}C HSQC spectrum (800 MHz) showing the anomeric region, highlighting the presence of αGlcNAc-$GAPDH_{187\text{-}203}$, with a large ^{13}C upfield shift relative to the free species. **b**) Expansion of ^{1}H-^{13}C CLIP HSQC spectrum (500 MHz) with no decoupling to measure the anomeric $^{1}J_{CH}$ coupling inαGlcNAc-$GAPDH_{187\text{-}203}$. A value of 169 Hz indicates an α-configuration.

The ^{1}H resonance frequency of 5.06 ppm for the $GAPDH_{187\text{-}203}$-GlcNAc anomeric proton is typical of an α glycosidic linkage and is close to the resonance frequency for the anomeric protein of free αGlcNAc. To confirm this an non-decoupled ^{1}H-^{13}C HSQC experiment was performed on the sample. The measured $^{1}J_{CH}$ value for the anomeric position of $GAPDH_{187\text{-}203}$-αGlcNAc was 169 Hz, which is typical of an α glycosidic linkage also.[326] Together these data show that the glycosidic linkage of $GAPDH_{187\text{-}203}$-GlcNAc is in the α configuration and therfore SseK1 functions as a retaining enzyme for the glycosyl transfer reaction to the $GAPDH_{187\text{-}203}$ acceptor.

4.4 Discussion

The SseK effectors provide an interesting topic of discussion both for their putative role in *Salmonella* infection and their novel enzymatic mechanism. While none of these effectors appear to be absolutely essential to the virulence of *Salmonella*, SseK1 and SseK3 have been shown to inhibit necroptotic cell death,[309] which is a mechanism of programmed cell death that causes the intracellular contents of the host cell to leak into the extracellular space, causing an inflammatory immune response. Mediation of necroptotic cell death by bacterial pathogens has been documented in a number of species.[327,328] It is likely that, while not essential, SseK1 and SseK3 provide an advantage to *Salmonella* by reducing inflammation at the site of infection. It is still unclear what the role of SseK2 is, although glycosylation of FADD does inhibit downstream nF-κB signalling.

The target sequences of FADD, TRADD and GAPDH all share similar sequences - in particular a conserved WR-motif. It is therefore perhaps unsurprising that all three target peptides ($FADD_{110\text{-}118}$, $TRADD_{229\text{-}237}$, $GAPDH_{195\text{-}203}$) bind to both SseK1 and SseK2. Therefore, while this motif appears to be important for binding (the tryptophan of each peptide always shows the strongest STD intensity), clearly it is not sufficient to confer specificity between SseK1 and SseK2.

To this end, a model of the SseK2/FADD complex was produced and validated experimentally by NMR. This complex was chosen for a

number of reasons. 1) From a practical standpoint FADD is the only SseK effector target small enough to be reasonably accessible to conventional NMR techniques. 2) FADD was the only target that could be reasonably modelled. TRADD consists of two domains, a death domain and a TRAF-binding domain.[329] Models exist only of the death domain of TRADD and preliminary docking results suggested that actually the TRAF-binding domain of FADD would be involved in the interaction with SseK1. Furthermore, the target region of GAPDH is located within an extended loop region, making accurate modelling challenging. 3) The interaction between SseK2 and FADD is very weak compared to the interaction of TRADD and GAPDH with SseK1[309,318] and therefore is probably the least likely complex to be solved experimentally. Therefore this method of molecular modelling and NMR spectroscopy of the SseK2/FADD complex provides molecular detail most likely otherwise inaccessible for this complex.

The model complex shows that the target region of FADD (residues 110-118) interacts directly with SseK2 residues surrounding the catalytic site. It particular Arg113 of FADD interacts directly with Glu271 of SseK2, a residue that has been identified as being essential for catalysis to occur in both SseK1 (Glu255) and SseK2 (Glu271). It is known that it is not Arg113 but Arg117 that becomes glycosylated in FADD.[317] Therefore it may be that the interaction between Glu271 (SseK2) and Arg113 (FADD) orients the target in such a way that Arg117 is accessible for glycosylation. Indeed, in this model Arg117 is position in close proximity to the anomeric carbon of UDP-GlcNAc. Furthermore this may explain why the WR-motif is conserved across all three targets, if recognition of the arginine residue immediately downstream of the tryptophan is necessary for proper orientation of the target.

Finally here is the first experimental evidence that the SseK effectors work as retaining enzymes, here demonstrated for the GlcNAc transfer to the $GAPDH_{187\text{-}203}$ substrate. Several studies had made claims either way (retaining or inverting), based on either similarity or very questionable interpretation of data. However, here the configuration of the reaction product of $GAPDH_{187\text{-}203}$ with SseK1 was detected directly by measuring the $^{1}J_{CH}$ coupling constant at the anomeric, which was found to be α as in the UDP-GlcNAc donor. Still, there is much

to be discovered about this unique mechanism, such as the catalytic residues involved and any intermediates that may exist during the reaction. Furthermore, it is currently unknown why glycosylation of these particular target residues is able to induce downstream signalling. One particularly interesting question that is yet to be answered is what the pKa of the glycosylated arginine is and whether it is charged at physiological pH, as unmodified arginine is.

This study has implications particularly in understanding the mechanism of substrate recognition and catalysis for this unique group of enzymes. Furthermore, while these effectors are not essential for infection, targeting them therapeutically may still be of interest since impairing the ability of *Salmonella* to suppress inflammation may help the host better combat the infection and in combination with other therapies may provide an effective treatment.

4.5 Conclusions

In this chapter a combination of NMR spectroscopy and molecular modelling was used to provide structural insight into the interaction of SseK1/2 and their acceptor substrates FADD, TRADD and GAPDH, and to provide some insight into their glycosylation mechanism. Overall the conclusions were:

1. Both SseK1 and SseK2 are capable of binding to each of the short acceptor peptides, $FADD_{110\text{-}118}$, $TRADD_{229\text{-}237}$ and $GAPDH_{195\text{-}203}$, showing that recognition of the full length acceptors must come from interactions outside of the active site.

2. A model of the SseK2/FADD complex was produced by molecular modelling and validated using STD NMR epitope mapping and chemical shift perturbation analysis using ^{15}N-labelled FADD.

3. The differential recognition of acceptor substrates is most likely due to differences in the HLH domains of SseK1 and SseK2, as shown by protein sequence analysis and molecular modelling.

4. The enzymatic mechanism of SseK1 (and likely SseK2 also) follows a retaining mechanism, as shown by analysis of the $^{1}J_{CH}$ coupling constant at the anomeric position of the reaction product between $GAPDH_{187\text{-}203}$ and UDP-GlcNAc catalysed by SseK1.

Appendix 1

List of Publications

The following list shows publications resulting from work performed during this period of research:

- **Walpole, S.**, de Paz, J.L., Nieto, P.M., Banerji, S., Jackson, J., Angulo, J. "NMR analysis of the differential recognition of hyaluronan by the cell surface receptors CD44 and LYVE-1". 2019. In preparation for Frontiers in Chemistry

- Beekman, A. M., Cominetti, M. M. D., **Walpole, S. J.**, Prabhu, S., O'Connell, M. A., Angulo, J. & Searcey, M., Identification of selective protein-protein interaction inhibitors using efficient in silico peptide-directed ligand design, 2019, Chemical Science. 10, 16, p. 4502-4508

- Bidula, S. M., Cromer, B. A., **Walpole, S.**, Angulo, J. & Stokes, L., Mapping a novel positive allosteric modulator binding site in the central vestibule region of human P2X7, 2019, Scientific Reports. 9, 1, 3231

- Nepravishta, R., **Walpole, S.**, Tailford, L., Juge, N. & Angulo, J., Deriving ligand orientation on weak protein-ligand complexes by DEEP-STD NMR in the absence of protein chemical shift assignment, 2019, ChemBioChem. 20, 3, p. 340-344

- Kuhaudomlarp, S., **Walpole, S.**, Stevenson, C. EM., Nepogodiev, S. A., Lawson, D. M., Angulo, J. & Field, R. A., Unravelling the Specificity of Laminaribiose Phosphorylase from Paenibacillus sp. YM-1 towards Donor Substrates Glucose/Mannose 1-Phosphate by Using X-ray Crystallography and Saturation Transfer Difference NMR Spectroscopy, 2019, ChemBioChem. 20, 2, p. 181-192

- Dhuna, K., Felgate, M., Bidula, S., **Walpole, S.**, Bibic, L.,

Cromer, B., Angulo, J., Sanderson, J., Stebbing, M. & Stokes, L., Ginsenosides act as positive modulators of P2X4 receptors, 2019, Molecular Pharmacology. 95, 1

- Watt, J., Hughes, G., **Walpole, S.**, Monaco, S., Stephenson, G., Bulman Page, P., Hemmings, A., Angulo, J. & Chantry, A., Discovery of Small Molecule WWP2 Ubiquitin Ligase Inhibitors, 2018, Chemistry - A European Journal. 24, 67, p. 17677-17680

- Park, J. B., Kim, Y. H., Yoo, Y., Kim, J., Jun, S-H., Cho, J. W., El-Qaidi, S., **Walpole, S.**, Monaco, S., Garcia-Garcia, A. A., Wu, M., Hays, M. P., Hurtado-Guerrero, R., Angulo, J., Hardwidge, P. R., Shin, J-S. & Cho, H-S., Structural basis for arginine glycosylation of host substrates by bacterial effector proteins, 2018, Nature Communications. 9, 4283

- **Walpole, S.**, Monaco, S., Nepravishta, R. & Angulo, J., STD NMR as a Technique for Ligand Screening and Structural Studies, 2018, Methods in Enzymology. Elsevier, Vol. 608

- Sequeira, S., Kavanaugh, D., Mackenzie, D., Suligoj, T., **Walpole, S.**, Leclaire, C., Gunning, A. P., Latousakis, D., Willats, W., Angulo, J., Dong, C. & Juge, N., Structural basis for the role of Serine-Rich Repeat Proteins from Lactobacillus reuteri in gut microbe-host interactions, 2018, Proceedings of the National Academy of Sciences USA. 115, 12, p. E2706-E2715

List of Abbreviations

^{1}H NMR	1D Proton NMR
ASMD	Adaptive Steered MD
AMBER	Assisted Model Building with Energy Refinement (biomolecular MM forcefield)
CORCEMA	Complete Relaxation and Conformational Exchange Matrix
COSY	Correlation Spectroscopy
FADD	Fas-associated Death Domain
G1P	Glucose 1-phosphate
GAG	Glycosaminoglycan
GaMD	Gaussian Accelerated MD
GAPDH	Glyceraldehyde 3-phosphate Dehydrogenase
Glc	Glucose
GlcA	Glucuronic Acid
GlcNAc	N-acetylglucosamine
Glide	Grid-based Ligand Docking with Energetics
GLYCAM	Carbohydrate-specific MM forcefield
GP	Glycoside Phosphorylase
GT	Glycosyltransferase

GT-A	Glycosyltransferase Type-A fold
GT-B	Glycosyltransferase Type-B fold
HA	Hyaluronan
HABD	HA Binding Domain
HLH	Helix-loop-helix
HSQC	Heteronuclear Single-quantum Correlation
k_{sat}	Saturation Transfer Rate Constant
LB	Laminaribiose
LBP	LB Phosphorylase
LYVE-1	Lymphatic Vessel Endothelial Hyaluronan Receptor 1
M1P	Mannose 1-phosphate
Man	Mannose
MD	Molecular Dynamics
MM	Molecular Mechanics
NMR	Nuclear Magnetic Resonance
NOE	Nuclear Overhauser Effect
NOESY	NOE Spectroscopy
NPT	Isobaric-Isothermal Ensemble
NVE	Microcanonical Ensemble
NVT	Canonical Ensemble

OPLS	Optimised Potentials for Liquid Simulation (general MM forcefield)
PDB	Protein Data Bank
PES	Potential Energy Surface
RMSD	Root Mean Squared Deviation
RMSF	Root Mean Squared Fluctuations
STD_0	Initial Rate of STD Intensity Growth
STD NMR	Saturation Transfer Difference NMR
STD_{max}	Maximum STD Intensity
TOCSY	Total Correlation Spectroscopy
TRADD	Tumor Necrosis Factor Receptor Type-1-associated Death Domain
UDP	Uridine Diphosphate

Bibliography

1. Crick, F. Central dogma of molecular biology. Nature 1970, 227(5258), 561–563.

2. Stanley, P.; Kornfield, S. Historical background and overview. In Essentials of glycobiology; n.d.

3. Varki, A.; Taniguchi, N.; Aebi, M. N-glycans. In Essentials of glycobiology; n.d.

4. Brockhausen, I.; Stanley, P. O-galnac glycans. In Essentials of glycobiology; n.d.

5. Gabius, H.J. The sugar code: Why glycans are so important. BioSystems 2018, 164, 102–111.

6. Rademacher, T.W.; Parekh, R.B.; Dwek, R.A. Glycobiology. Annu. Rev. Biochem. 1988, 57, 785–838.

7. Varki, A.; Gagneux, P. Biological functions of glycans. In Essentials of glycobiology; n.d.

8. Schauer, R. Sialic acids and their role as biological masks. Trends in Biochemical Sciences Elsevier BV: 1985, 10(9), 357–360 10.1016/0968-0004(85)90112-4.

9. Abu-Shakra, M.; Buskila, D.; Shoenfeld, Y. Molecular mimicry between host and pathogen: examples from parasites and implication. Immunol. Lett. 1999, 67(2), 147–152.

10. Gabius, H.J.; Andre, S.; Jimenez-Barbero, J.; Romero, A.; Solis, D. From lectin structure to functional glycomics: principles of the sugar code. Trends Biochem. Sci. 2011, 36(6), 298–313.

11. Solis, D.; Bovin, N.V.; Davis, A.P.; Jimenez-Barbero, J.; Romero, A.; Roy, R.; et al. A guide into glycosciences: How chemistry, biochemistry and biology cooperate to crack the sugar code. Biochim.

Biophys. Acta 2015, 1850(1), 186–235.

12. Freeze, H.; Schachter, H.; Kinoshita, T. Genetic disorders of glycosylation. In Essentials of glycobiology; n.d.

13. Carchon, H.; Van Schaftingen, E.; Matthijs, G.; Jaeken, J. Carbohydrate-deficient glycoprotein syndrome type IA (phosphomannomutase-deficiency). Biochim. Biophys. Acta 1999, 1455(2-3), 155–165.

14. Dube, D.H.; Bertozzi, C.R. Glycans in cancer and inflammation–potential for therapeutics and diagnostics. Nat Rev Drug Discov 2005, 4(6), 477–488.

15. Bull, C.; Stoel, M.A.; Brok, M.H. den; Adema, G.J. Sialic acids sweeten a tumor's life. Cancer Res. 2014, 74(12), 3199–3204.

16. Stanley, P. Glycosylation mutants of animal cells. Annu. Rev. Genet. 1984, 18, 525–552.

17. Ludovic, M.; Eric, N.-O.; Maxime, G.; Abdoul-Salam, K.; Follet-Gueye, M.-L.; Azeddine, D.; et al. O-glycosylation in plant and mammal cells: The use of chemical inhibitors to understand the biosynthesis and function of o-glycosylated proteins. Plant Science Today Horizon E-Publishing Group: 2015, 2(2), 43–51 10.14719/pst.2015.2.2.67.

18. Haab, B.B. Antibody-lectin sandwich arrays for biomarker and glycobiology studies. Expert Rev Proteomics 2010, 7(1), 9–11.

19. Harvey, D.J. Analysis of carbohydrates and glycoconjugates by matrix-assisted laser desorption/ionization mass spectrometry: An update for 2011-2012. Mass Spectrom Rev 2017, 36(3), 255–422.

20. Duus, J.; Gotfredsen, C.H.; Bock, K. Carbohydrate structural determination by NMR spectroscopy: modern methods and limitations. Chem. Rev. 2000, 100(12), 4589–4614.

21. Hart, G.W.; Copeland, R.J. Glycomics hits the big time. Cell 2010, 143(5), 672–676.

22. Bush, C.A.; Martin-Pastor, M.; Imberty, A. Structure and con-

formation of complex carbohydrates of glycoproteins, glycolipids, and bacterial polysaccharides. Annu Rev Biophys Biomol Struct 1999, 28, 269–293.

23. Carmen Fernandez-Alonso, M. del; Diaz, D.; Berbis, M.A.; Marcelo, F.; Canada, J.; Jimenez-Barbero, J. Protein-carbohydrate interactions studied by NMR: from molecular recognition to drug design. Curr. Protein Pept. Sci. 2012, 13(8), 816–830.

24. Roldos, V.; Canada, F.J.; Jimenez-Barbero, J. Carbohydrate-protein interactions: a 3D view by NMR. Chembiochem 2011, 12(7), 990–1005.

25. Kogelberg, H.; Solis, D.; Jimenez-Barbero, J. New structural insights into carbohydrate-protein interactions from NMR spectroscopy. Curr. Opin. Struct. Biol. 2003, 13(5), 646–653.

26. Poveda, A.; Jiménez-Barbero, J. NMR studies of carbohydrate-protein interactions in solution. Chemical Society Reviews Royal Society of Chemistry (RSC): 1998, 27(2), 133 10.1039/a827133z.

27. Seeberger, P. Monosaccharide diversity. In Essentials of glycobiology; n.d.

28. Takahashi, K.; Ono, S. Calorimetric studies on the mutarotation of D-galactose and D-mannose. J. Biochem. 1973, 73(4), 763–770.

29. Kirby, A.J.; Williams, N.H. Anomeric and gauche effects. In ACS symposium series; American Chemical Society: 1993, 55–69 10.1021/bk-1993-0539.ch004.

30. Ferro, D.R.; Provasoli, A.; Ragazzi, M.; Torri, G.; Casu, B.; Gatti, G.; et al. Evidence for conformational equilibrium of the sulfated l-iduronate residue in heparin and in synthetic heparin mono- and oligosaccharides: NMR and force-field studies. Journal of the American Chemical Society American Chemical Society (ACS): 1986, 108(21), 6773–6778 10.1021/ja00281a052.

31. Brameld, K.A.; Goddard, W.A. Substrate distortion to a boat conformation at subsite -1 is critical in the mechanism of family 18

chitinases. Journal of the American Chemical Society American Chemical Society (ACS): 1998, 120(15), 3571–3580 10.1021/ja972282h.

32. Zou, J.y.; Kleywegt, G.J.; Stahlberg, J.; Driguez, H.; Nerinckx, W.; Claeyssens, M.; et al. Crystallographic evidence for substrate ring distortion and protein conformational changes during catalysis in cellobiohydrolase Ce16A from trichoderma reesei. Structure 1999, 7(9), 1035–1045.

33. Cremer, D.; Pople, J.A. General definition of ring puckering coordinates. Journal of the American Chemical Society American Chemical Society (ACS): 1975, 97(6), 1354–1358 10.1021/ja00839a011.

34. Pechenaya, V.I. Conformational flexibility of the furanose ring in DNA and its dipole moment. J. Biomol. Struct. Dyn. 1989, 7(2), 381–388.

35. Pattabiraman, N.; Rao, M.J. Flexibility of the furanose ring in nucleic acids: A compilation of crystal data. International Journal of Biological Macromolecules Elsevier BV: 1982, 4(2), 91–98 10.1016/0141-8130(82)90031-9.

36. Xu, B.; Unione, L.; Sardinha, J.; Wu, S.; Etheve-Quelquejeu, M.; Pilar Rauter, A.; et al. gem-Difluorocarbadisaccharides: restoring the exo-anomeric effect. Angew. Chem. Int. Ed. Engl. 2014, 53(36), 9597–9602.

37. Marchessault, R.H.; Perez, S. Conformations of the hydroxymethyl group in crystalline aldohexopyranoses. Biopolymers Wiley: 1979, 18(9), 2369–2374 10.1002/bip.1979.360180925.

38. Bock, K.; Duus, J.Ø. A conformational study of hydroxymethyl groups in carbohydrates investigated by1H NMR spectroscopy. Journal of Carbohydrate Chemistry Informa UK Limited: 1994, 13(4), 513–543 10.1080/07328309408011662.

39. Kirschner, K.N.; Woods, R.J. Solvent interactions determine carbohydrate conformation. Proc. Natl. Acad. Sci. U.S.A. 2001, 98(19), 10541–10545.

40. Yu, H.; Chen, X. Carbohydrate post-glycosylational modifications. Org. Biomol. Chem. 2007, 5(6), 865–872.

41. Bucala, R.; Model, P.; Cerami, A. Modification of DNA by reducing sugars: a possible mechanism for nucleic acid aging and age-related dysfunction in gene expression. Proc. Natl. Acad. Sci. U.S.A. 1984, 81(1), 105–109.

42. Lee, A.T. The nonenzymatic glycosylation of DNA by reducing sugars in vivo may contribute to DNA damage associated with aging. AGE Springer Nature: 1987, 10(4), 150–155 10.1007/bf02432163.

43. Lee, A.T.; Cerami, A. Elevated glucose 6-phosphate levels are associated with plasmid mutations in vivo. Proc. Natl. Acad. Sci. U.S.A. 1987, 84(23), 8311–8314.

44. Malhotra, R. Membrane glycolipids: Functional heterogeneity: A review. Biochemistry & Analytical Biochemistry OMICS Publishing Group: 2012, 1(2) 10.4172/2161-1009.1000108.

45. Holmgren, J.; Lonnroth, I.; Mansson, J.; Svennerholm, L. Interaction of cholera toxin and membrane GM1 ganglioside of small intestine. Proc. Natl. Acad. Sci. U.S.A. 1975, 72(7), 2520–2524.

46. Alonso, M.A.; Millan, J. The role of lipid rafts in signalling and membrane trafficking in T lymphocytes. J. Cell. Sci. 2001, 114(Pt 22), 3957–3965.

47. Pike, L.J. Lipid rafts: bringing order to chaos. J. Lipid Res. 2003, 44(4), 655–667.

48. Raetz, C.R.; Whitfield, C. Lipopolysaccharide endotoxins. Annu. Rev. Biochem. 2002, 71, 635–700.

49. Alexander, C.; Rietschel, E.T. Bacterial lipopolysaccharides and innate immunity. J. Endotoxin Res. 2001, 7(3), 167–202.

50. Rosenfeld, Y.; Shai, Y. Lipopolysaccharide (Endotoxin)-host defense antibacterial peptides interactions: role in bacterial resistance and prevention of sepsis. Biochim. Biophys. Acta 2006, 1758(9),

1513–1522.

51. Pinho, S.S.; Reis, C.A. Glycosylation in cancer: mechanisms and clinical implications. Nat. Rev. Cancer 2015, 15(9), 540–555.

52. Lannoo, N.; Van Damme, E.J. Review/N-glycans: The making of a varied toolbox. Plant Sci. 2015, 239, 67–83.

53. Xu, C.; Ng, D.T. Glycosylation-directed quality control of protein folding. Nat. Rev. Mol. Cell Biol. 2015, 16(12), 742–752.

54. Scheiffele, P.; Peranen, J.; Simons, K. N-glycans as apical sorting signals in epithelial cells. Nature 1995, 378(6552), 96–98.

55. Carraway, K.L.; Hull, S.R. O-glycosylation pathway for mucin-type glycoproteins. Bioessays 1989, 10(4), 117–121.

56. Praissman, J.L.; Wells, L. Mammalian O-mannosylation pathway: glycan structures, enzymes, and protein substrates. Biochemistry 2014, 53(19), 3066–3078.

57. Moloney, D.J.; Haltiwanger, R.S. The O-linked fucose glycosylation pathway: identification and characterization of a uridine diphosphoglucose: fucose-beta1,3-glucosyltransferase activity from Chinese hamster ovary cells. Glycobiology 1999, 9(7), 679–687.

58. Perez-Vilar, J.; Hill, R.L. The structure and assembly of secreted mucins. J. Biol. Chem. 1999, 274(45), 31751–31754.

59. Bond, M.R.; Hanover, J.A. A little sugar goes a long way: the cell biology of O-GlcNAc. J. Cell Biol. 2015, 208(7), 869–880.

60. Schaefer, L.; Schaefer, R.M. Proteoglycans: from structural compounds to signaling molecules. Cell Tissue Res. 2010, 339(1), 237–246.

61. Couchman, J.R.; Pataki, C.A. An introduction to proteoglycans and their localization. J. Histochem. Cytochem. 2012, 60(12), 885–897.

62. Taylor, K.R.; Gallo, R.L. Glycosaminoglycans and their proteogly-

cans: host-associated molecular patterns for initiation and modulation of inflammation. FASEB J. 2006, 20(1), 9–22.

63. Mulloy, B.; Rider, C.C. Cytokines and proteoglycans: an introductory overview. Biochem. Soc. Trans. 2006, 34(Pt 3), 409–413.

64. Cummings, R.; Schnaar, R.; Esko, J.; Drickamer, K.; Taylor, M. Principles of glycan recognition. In Essentials of glycobiology; n.d.

65. Weis, W.I.; Drickamer, K. Structural basis of lectin-carbohydrate recognition. Annu. Rev. Biochem. 1996, 65, 441–473.

66. Lemieux, R.U. How water provides the impetus for molecular recognition in aqueous solution. Accounts of Chemical Research American Chemical Society (ACS): 1996, 29(8), 373–380 10.1021/ar9600087.

67. Ghazarian, H.; Idoni, B.; Oppenheimer, S.B. A glycobiology review: carbohydrates, lectins and implications in cancer therapeutics. Acta Histochem. 2011, 113(3), 236–247.

68. Dambuza, I.M.; Brown, G.D. C-type lectins in immunity: recent developments. Curr. Opin. Immunol. 2015, 32, 21–27.

69. Maureen, T.; Drickamer, K.; Schnaar, R.; Etzler, M.; Varki, A. Discovery and classification of glycan-binding proteins. In Essentials of glycobiology; n.d.

70. Feinberg, H.; Mitchell, D.A.; Drickamer, K.; Weis, W.I. Structural basis for selective recognition of oligosaccharides by DC-SIGN and DC-SIGNR. Science 2001, 294(5549), 2163–2166.

71. Blundell, C.D.; Almond, A.; Mahoney, D.J.; DeAngelis, P.L.; Campbell, I.D.; Day, A.J. Towards a structure for a TSG-6.hyaluronan complex by modeling and NMR spectroscopy: insights into other members of the link module superfamily. J. Biol. Chem. 2005, 280(18), 18189–18201.

72. Higman, V.A.; Blundell, C.D.; Mahoney, D.J.; Redfield, C.; Noble, M.E.; Day, A.J. Plasticity of the TSG-6 HA-binding loop and

mobility in the TSG-6-HA complex revealed by NMR and X-ray crystallography. J. Mol. Biol. 2007, 371(3), 669–684.

73. E., C.; Elling, L. Galectins: Structures, binding properties and function in cell adhesion. In Biomaterials - physics and chemistry; InTech: 2011, 10.5772/24647.

74. Rabinovich, G.A.; Toscano, M.A. Turning 'sweet' on immunity: galectin-glycan interactions in immune tolerance and inflammation. Nat. Rev. Immunol. 2009, 9(5), 338–352.

75. Esko, J.; Prestegard, J.; Linhardt, R. Proteins that bind sulfated glycosaminoglycans. In Essentials of glycobiology; n.d.

76. Esko, J.D.; Lindahl, U. Molecular diversity of heparan sulfate. J. Clin. Invest. 2001, 108(2), 169–173.

77. Canales, A.; Angulo, J.; Ojeda, R.; Bruix, M.; Fayos, R.; Lozano, R.; et al. Conformational flexibility of a synthetic glycosylaminoglycan bound to a fibroblast growth factor. FGF-1 recognizes both the (1)C(4) and (2)S(O) conformations of a bioactive heparin-like hexasaccharide. J. Am. Chem. Soc. 2005, 127(16), 5778–5779.

78. Angulo, J.; Nieto, P.M.; Martin-Lomas, M. A molecular dynamics description of the conformational flexibility of the L-iduronate ring in glycosaminoglycans. Chem. Commun. (Camb.) 2003, (13), 1512–1513.

79. Munoz-Garcia, J.C.; Chabrol, E.; Vives, R.R.; Thomas, A.; Paz, J.L. de; Rojo, J.; et al. Langerin-heparin interaction: two binding sites for small and large ligands as revealed by a combination of NMR spectroscopy and cross-linking mapping experiments. J. Am. Chem. Soc. 2015, 137(12), 4100–4110.

80. Olson, S.T.; Richard, B.; Izaguirre, G.; Schedin-Weiss, S.; Gettins, P.G. Molecular mechanisms of antithrombin-heparin regulation of blood clotting proteinases. A paradigm for understanding proteinase regulation by serpin family protein proteinase inhibitors. Biochimie 2010, 92(11), 1587–1596.

81. Hascall, V.; Esko, J. Hyaluronan. In Essentials of glycobiology; n.d.

82. Laurent, T.C.; Laurent, U.B.; Fraser, J.R. The structure and function of hyaluronan: An overview. Immunol. Cell Biol. 1996, 74(2), 1–7.

83. Aruffo, A.; Stamenkovic, I.; Melnick, M.; Underhill, C.B.; Seed, B. CD44 is the principal cell surface receptor for hyaluronate. Cell 1990, 61(7), 1303–1313.

84. Turley, E.A.; Noble, P.W.; Bourguignon, L.Y. Signaling properties of hyaluronan receptors. J. Biol. Chem. 2002, 277(7), 4589–4592.

85. Rini, J.M.; Esko, J.D. Glycosyltransferases and glycan-processing enzymes. In Essentials of glycobiology; n.d.

86. Angulo, J.; Langpap, B.; Blume, A.; Biet, T.; Meyer, B.; Krishna, N.R.; et al. Blood group B galactosyltransferase: insights into substrate binding from NMR experiments. J. Am. Chem. Soc. 2006, 128(41), 13529–13538.

87. Pedersen, L.C.; Dong, J.; Taniguchi, F.; Kitagawa, H.; Krahn, J.M.; Pedersen, L.G.; et al. Crystal structure of an alpha 1,4-N-acetylhexosaminyltransferase (EXTL2), a member of the exostosin gene family involved in heparan sulfate biosynthesis. J. Biol. Chem. 2003, 278(16), 14420–14428.

88. Chen, C.Y.; Jan, Y.H.; Juan, Y.H.; Yang, C.J.; Huang, M.S.; Yu, C.J.; et al. Fucosyltransferase 8 as a functional regulator of nonsmall cell lung cancer. Proc. Natl. Acad. Sci. U.S.A. 2013, 110(2), 630–635.

89. GlycosylTransferase family classification; n.d.

90. Lairson, L.L.; Henrissat, B.; Davies, G.J.; Withers, S.G. Glycosyltransferases: structures, functions, and mechanisms. Annu. Rev. Biochem. 2008, 77, 521–555.

91. Henrissat, B.; Sulzenbacher, G.; Bourne, Y. Glycosyltransferases, glycoside hydrolases: surprise, surprise! Curr. Opin. Struct. Biol.

2008, 18(5), 527–533.

92. Rye, C.S.; Withers, S.G. Glycosidase mechanisms. Curr Opin Chem Biol 2000, 4(5), 573–580.

93. Palcic, M.M. Biocatalytic synthesis of oligosaccharides. Current Opinion in Biotechnology Elsevier BV: 1999, 10(6), 616–624 10.1016/s0958-1669(99)00044-0.

94. Bucke, C. Review oligosaccharide synthesis using glycosidases. Journal of Chemical Technology & Biotechnology Wiley: 1996, 67(3), 217–220 10.1002/(sici)1097-4660(199611)67:3<217::aid-jctb558>3.0.co;2-8.

95. Muthana, S.; Cao, H.; Chen, X. Recent progress in chemical and chemoenzymatic synthesis of carbohydrates. Current Opinion in Chemical Biology Elsevier BV: 2009, 13(5-6), 573–581 10.1016/j.cbpa.2009.09.013.

96. Rather, M.; Mishra, S. β-glycosidases: An alternative enzyme based method for synthesis of alkyl-glycosides. Sustainable Chemical Processes Springer Nature: 2013, 1(1), 7 10.1186/2043-7129-1-7.

97. Shim, J.-H.; Chen, H.-M.; Rich, J.R.; Goddard-Borger, E.D.; Withers, S.G. Directed evolution of a -glycosidase from agrobacterium sp. To enhance its glycosynthase activity toward c3-modified donor sugars. Protein Engineering Design and Selection Oxford University Press (OUP): 2012, 25(9), 465–472 10.1093/protein/gzs045.

98. Mackenzie, L.F.; Wang, Q.; Warren, R.A.J.; Withers, S.G. Glycosynthases: mutant glycosidases for oligosaccharide synthesis. Journal of the American Chemical Society American Chemical Society (ACS): 1998, 120(22), 5583–5584 10.1021/ja980833d.

99. Teze, D.; Daligault, F.; Ferrières, V.; Sanejouand, Y.-H.; Tellier, C. Semi-rational approach for converting a GH36 α-glycosidase into an α-transglycosidase. Glycobiology Oxford University Press (OUP): 2014, 25(4), 420–427 10.1093/glycob/cwu124.

100. Levitt, M.H. Spin dynamics: Basics of nuclear magnetic reso-

nance; Wiley: 2015.

101. Atkins, P.W.; Paula, J.D. Atkins physical chemistry; Oxford University Press: 2014.

102. Hore, P.J. Nuclear magnetic resonance; Oxford University Press: 2015.

103. Keeler, J. Understanding nmr spectroscopy; Wiley: 2012.

104. Aljohani, H.A. Role of carboxylate ligands in the synthesis of aunps: Size control, molecular interaction and catalytic activity. KAUST Research Repository: 2016, 10.25781/kaust-w23l0.

105. Keshari, K.R.; Wilson, D.M. Chemistry and biochemistry of 13C hyperpolarized magnetic resonance using dynamic nuclear polarization. Chem Soc Rev 2014, 43(5), 1627–1659.

106. Solomon, I. Relaxation processes in a system of two spins. Physical Review American Physical Society (APS): 1955, 99(2), 559–565 10.1103/physrev.99.559.

107. Wagner, G.; Wuthrich, K. Truncated driven nuclear overhauser effect (TOE). A new technique for studies of selective 1H 1H overhauser effects in the presence of spin diffusion. Journal of Magnetic Resonance (1969) Elsevier BV: 1979, 33(3), 675–680 10.1016/0022-2364(79)90180-x.

108. Quiros, M.T.; Angulo, J.; Munoz, M.P. Kinetics of intramolecular chemical exchange by initial growth rates of spin saturation transfer difference experiments (SSTD NMR). Chem. Commun. (Camb.) 2015, 51(50), 10222–10225.

109. Mayer, M.; Meyer, B. Group epitope mapping by saturation transfer difference NMR to identify segments of a ligand in direct contact with a protein receptor. Journal of the American Chemical Society American Chemical Society (ACS): 2001, 123(25), 6108–6117 10.1021/ja0100120.

110. Angulo, J.; Enriquez-Navas, P.M.; Nieto, P.M. Ligand-receptor

binding affinities from saturation transfer difference (STD) NMR spectroscopy: the binding isotherm of STD initial growth rates. Chemistry 2010, 16(26), 7803–7812.

111. Kobayashi, M.; Retra, K.; Figaroa, F.; Hollander, J.G.; Ab, E.; Heetebrij, R.J.; et al. Target immobilization as a strategy for NMR-based fragment screening. Journal of Biomolecular Screening SAGE Publications: 2010, 15(8), 978–989 10.1177/1087057110375614.

112. Mayer, M.; James, T.L. NMR-based characterization of phenothiazines as a RNA binding scaffold. J. Am. Chem. Soc. 2004, 126(13), 4453–4460.

113. Ingólfsson, H.I.; Lopez, C.A.; Uusitalo, J.J.; Jong, D.H. de; Gopal, S.M.; Periole, X.; et al. The power of coarse graining in biomolecular simulations. Wiley Interdisciplinary Reviews: Computational Molecular Science Wiley: 2013, 4(3), 225–248 10.1002/wcms.1169.

114. Stein, E.G.; Rice, L.M.; Brunger, A.T. Torsion-angle molecular dynamics as a new efficient tool for NMR structure calculation. J. Magn. Reson. 1997, 124(1), 154–164.

115. Kirschner, K.N.; Yongye, A.B.; Tschampel, S.M.; Gonzalez-Outeirino, J.; Daniels, C.R.; Foley, B.L.; et al. GLYCAM06: a generalizable biomolecular force field. Carbohydrates. J Comput Chem 2008, 29(4), 622–655.

116. Maier, J.A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K.E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J Chem Theory Comput 2015, 11(8), 3696–3713.

117. Hay, B.P.; Clement, O.; Sandrone, G.; Dixon, D.A. A molecular mechanics (MM3(96)) force field for metal-amide complexes. Inorganic Chemistry American Chemical Society (ACS): 1998, 37(22), 5887–5894 10.1021/ic980641j.

118. Perez, A.; Marchan, I.; Svozil, D.; Sponer, J.; Cheatham, T.E.; Laughton, C.A.; et al. Refinement of the AMBER force field for nu-

cleic acids: improving the description of alpha/gamma conformers. Biophys. J. 2007, 92(11), 3817–3829.

119. Dickson, C.J.; Madej, B.D.; Skjevik, A.A.; Betz, R.M.; Teigen, K.; Gould, I.R.; et al. Lipid14: The Amber Lipid Force Field. J Chem Theory Comput 2014, 10(2), 865–879.

120. Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J.Y.; et al. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. J Chem Theory Comput 2016, 12(1), 281–296.

121. Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; Groot, B.L. de; et al. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. Nat. Methods 2017, 14(1), 71–73.

122. Oostenbrink, C.; Villa, A.; Mark, A.E.; Gunsteren, W.F. van A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. J Comput Chem 2004, 25(13), 1656–1676.

123. Ferreira, L.; Santos, R. dos; Oliva, G.; Andricopulo, A. Molecular docking and structure-based drug design strategies. Molecules MDPI AG: 2015, 20(7), 13384–13421 10.3390/molecules200713384.

124. Ruyck, J. de; Brysbaert, G.; Blossey, R.; Lensink, M.F. Molecular docking as a popular tool in drug design, an in silico travel. Adv Appl Bioinform Chem 2016, 9, 1–11.

125. Vakser, I.A. Protein-protein docking: from interaction to interactome. Biophys. J. 2014, 107(8), 1785–1793.

126. Friesner, R.A.; Banks, J.L.; Murphy, R.B.; Halgren, T.A.; Klicic, J.J.; Mainz, D.T.; et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J. Med. Chem. 2004, 47(7), 1739–1749.

127. Muniz, H.S.; Nascimento, A.S. Towards a critical evaluation of an empirical and volume-based solvation function for ligand docking. PLOS ONE D. Roccatano, Ed. Public Library of Science (PLoS):

2017, 12(3), e0174336 10.1371/journal.pone.0174336.

128. Tripathi, A.; Bankaitis, V.A. Molecular Docking: From Lock and Key to Combination Lock. J Mol Med Clin Appl 2017, 2(1).

129. Moore, C.C. Ergodic theorem, ergodic theory, and statistical mechanics. Proc. Natl. Acad. Sci. U.S.A. 2015, 112(7), 1907–1911.

130. Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H.J.C. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. Journal of Computational Physics Elsevier BV: 1977, 23(3), 327–341 10.1016/0021-9991(77)90098-5.

131. Andersen, H.C. Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations. Journal of Computational Physics Elsevier BV: 1983, 52(1), 24–34 10.1016/0021-9991(83)90014-1.

132. Zhou, R. Free energy landscape of protein folding in water: explicit vs. implicit solvent. Proteins 2003, 53(2), 148–161.

133. Mark, P.; Nilsson, L. Structure and dynamics of the TIP3P, SPC, and SPC/e water models at 298 k. The Journal of Physical Chemistry A American Chemical Society (ACS): 2001, 105(43), 9954–9960 10.1021/jp003020w.

134. Dick, T.J.; Madura, J.D. Chapter 5 a review of the TIP4P, TIP4P-ew, TIP5P, and TIP5P-e water models. In Annual reports in computational chemistry; Elsevier: 2005, 59–74 10.1016/s1574-1400(05)01005-4.

135. Huggins, D.J. Correlations in liquid water for the TIP3P-Ewald, TIP4P-2005, TIP5P-Ewald, and SWM4-NDP models. J Chem Phys 2012, 136(6), 064518.

136. Darden, T.; York, D.; Pedersen, L. Particle mesh ewald: An n·log(N) method for ewald sums in large systems. The Journal of Chemical Physics AIP Publishing: 1993, 98(12), 10089–10092 10.1063/1.464397.

137. Hamelberg, D.; Mongan, J.; McCammon, J.A. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. J Chem Phys 2004, 120(24), 11919–11929.

138. Shaw, D.E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R.O.; Eastwood, M.P.; et al. Atomic-level characterization of the structural dynamics of proteins. Science 2010, 330(6002), 341–346.

139. Markwick, P.R.L.; McCammon, J.A. Studying functional dynamics in bio-molecules using accelerated molecular dynamics. Physical Chemistry Chemical Physics Royal Society of Chemistry (RSC): 2011, 13(45), 20053 10.1039/c1cp22100k.

140. Shen, T.; Hamelberg, D. A statistical analysis of the precision of reweighting-based simulations. J Chem Phys 2008, 129(3), 034103.

141. Miao, Y.; Feher, V.A.; McCammon, J.A. Gaussian Accelerated Molecular Dynamics: Unconstrained Enhanced Sampling and Free Energy Calculation. J Chem Theory Comput 2015, 11(8), 3584–3595.

142. Patel, J.S.; Berteotti, A.; Ronsisvalle, S.; Rocchia, W.; Cavalli, A. Steered molecular dynamics simulations for studying protein-ligand interaction in cyclin-dependent kinase 5. J Chem Inf Model 2014, 54(2), 470–480.

143. Mikulska, K.; Strzelecki, J.; Nowak, W. Nanomechanics of Î²-rich proteins related to neuronal disorders studied by AFM, all-atom and coarse-grained MD methods. J Mol Model 2014, 20(3), 2144.

144. Jarzynski, C. Nonequilibrium equality for free energy differences. Physical Review Letters American Physical Society (APS): 1997, 78(14), 2690–2693 10.1103/physrevlett.78.2690.

145. Ozer, G.; Valeev, E.F.; Quirk, S.; Hernandez, R. Adaptive steered molecular dynamics of the long-distance unfolding of neuropeptide y. Journal of Chemical Theory and Computation American Chemical Society (ACS): 2010, 6(10), 3026–3038 10.1021/ct100320g.

146. Sequeira, S.; Kavanaugh, D.; MacKenzie, D.A.; ?uligoj, T.; Walpole, S.; Leclaire, C.; et al. Structural basis for the role of serine-

rich repeat proteins from Lactobacillus reuteri in gut microbe-host interactions. Proc. Natl. Acad. Sci. U.S.A. 2018, 115(12), E2706–E2715.

147. Boczek, E.E.; Reefschlager, L.G.; Dehling, M.; Struller, T.J.; Hausler, E.; Seidl, A.; et al. Conformational processing of oncogenic v-Src kinase by the molecular chaperone Hsp90. Proc. Natl. Acad. Sci. U.S.A. 2015, 112(25), E3189–3198.

148. Deganutti, G.; Cuzzolin, A.; Ciancetta, A.; Moro, S. Understanding allosteric interactions in G protein-coupled receptors using Supervised Molecular Dynamics: A prototype study analysing the human A3 adenosine receptor positive allosteric modulator LUF6000. Bioorg. Med. Chem. 2015, 23(14), 4065–4071.

149. Śledź, P.; Caflisch, A. Protein structure-based drug design: From docking to molecular dynamics. Current Opinion in Structural Biology Elsevier BV: 2018, 48, 93–102 10.1016/j.sbi.2017.10.010.

150. Priya, R.; Sumitha, R.; Doss, C.G.; Rajasekaran, C.; Babu, S.; Seenivasan, R.; et al. Molecular Docking and Molecular Dynamics to Identify a Novel Human Immunodeficiency Virus Inhibitor from Alkaloids of Toddalia asiatica. Pharmacogn Mag 2015, 11(Suppl 3), S414–422.

151. Salomon-Ferrer, R.; Gotz, A.W.; Poole, D.; Le Grand, S.; Walker, R.C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. J Chem Theory Comput 2013, 9(9), 3878–3888.

152. Miao, Y.; Feixas, F.; Eun, C.; McCammon, J.A. Accelerated molecular dynamics simulations of protein folding. J Comput Chem 2015, 36(20), 1536–1549.

153. Perilla, J.R.; Schulten, K. Physical properties of the HIV-1 capsid from all-atom molecular dynamics simulations. Nature Communications Springer Nature: 2017, 8, 15959 10.1038/ncomms15959.

154. Ponta, H.; Sherman, L.; Herrlich, P.A. CD44: from adhesion molecules to signalling regulators. Nat. Rev. Mol. Cell Biol. 2003,

4(1), 33–45.

155. Goodison, S.; Urquidi, V.; Tarin, D. CD44 cell adhesion molecules. MP, Mol. Pathol. 1999, 52(4), 189–196.

156. Naor, D.; Sionov, R.V.; Ish-Shalom, D. CD44: structure, function, and association with the malignant process. Adv. Cancer Res. 1997, 71, 241–319.

157. Hlavacek, M. The role of synovial fluid filtration by cartilage in lubrication of synovial joints. J Biomech 1993, 26(10), 1145–1160.

158. DAY, T.D. Connective tissue permeability and the mode of action of hyaluronidase. Nature 1950, 166(4227), 785–786.

159. Dicker, K.T.; Gurski, L.A.; Pradhan-Bhatt, S.; Witt, R.L.; Farach-Carson, M.C.; Jia, X. Hyaluronan: a simple polysaccharide with diverse biological functions. Acta Biomater 2014, 10(4), 1558–1570.

160. Mackay, C.R.; Terpe, H.J.; Stauder, R.; Marston, W.L.; Stark, H.; Gunthert, U. Expression and modulation of CD44 variant isoforms in humans. J. Cell Biol. 1994, 124(1-2), 71–82.

161. Fox, S.B.; Fawcett, J.; Jackson, D.G.; Collins, I.; Gatter, K.C.; Harris, A.L.; et al. Normal human tissues, in addition to some tumors, express multiple different CD44 isoforms. Cancer Res. 1994, 54(16), 4539–4546.

162. Picker, L.J.; Nakache, M.; Butcher, E.C. Monoclonal antibodies to human lymphocyte homing receptors define a novel class of adhesion molecules on diverse cell types. J. Cell Biol. 1989, 109(2), 927–937.

163. Alho, A.M.; Underhill, C.B. The hyaluronate receptor is preferentially expressed on proliferating epithelial cells. J. Cell Biol. 1989, 108(4), 1557–1565.

164. Karousou, E.; Misra, S.; Ghatak, S.; Dobra, K.; Gotte, M.; Vigetti, D.; et al. Roles and targeting of the HAS/hyaluronan/CD44 molecular system in cancer. Matrix Biol. 2017, 59, 3–22.

165. Prochazka, L.; Tesarik, R.; Turanek, J. Regulation of alternative splicing of CD44 in cancer. Cell. Signal. 2014, 26(10), 2234–2239.

166. Mattheolabakis, G.; Milane, L.; Singh, A.; Amiji, M.M. Hyaluronic acid targeting of CD44 for cancer therapy: from receptor biology to nanomedicine. J Drug Target 2015, 23(7-8), 605–618.

167. Yan, Y.; Zuo, X.; Wei, D. Concise Review: Emerging Role of CD44 in Cancer Stem Cells: A Promising Biomarker and Therapeutic Target. Stem Cells Transl Med 2015, 4(9), 1033–1043.

168. Louderbough, J.M.; Schroeder, J.A. Understanding the dual nature of CD44 in breast cancer progression. Mol. Cancer Res. 2011, 9(12), 1573–1586.

169. Screaton, G.R.; Bell, M.V.; Jackson, D.G.; Cornelis, F.B.; Gerth, U.; Bell, J.I. Genomic structure of DNA encoding the lymphocyte homing receptor CD44 reveals at least 12 alternatively spliced exons. Proc. Natl. Acad. Sci. U.S.A. 1992, 89(24), 12160–12164.

170. Ni, J.; Cozzi, P.J.; Hao, J.L.; Beretov, J.; Chang, L.; Duan, W.; et al. CD44 variant 6 is associated with prostate cancer metastasis and chemo-/radioresistance. Prostate 2014, 74(6), 602–617.

171. Banky, B.; Raso-Barnett, L.; Barbai, T.; Timar, J.; Becsagh, P.; Raso, E. Characteristics of CD44 alternative splice pattern in the course of human colorectal adenocarcinoma progression. Mol. Cancer 2012, 11, 83.

172. Guo, W.; Frenette, P.S. Alternative CD44 splicing in intestinal stem cells and tumorigenesis. Oncogene 2014, 33(5), 537–538.

173. Lau, W.M.; Teng, E.; Chong, H.S.; Lopez, K.A.; Tay, A.Y.; Salto-Tellez, M.; et al. CD44v8-10 is a cancer-specific marker for gastric cancer stem cells. Cancer Res. 2014, 74(9), 2630–2641.

174. Yae, T.; Tsuchihashi, K.; Ishimoto, T.; Motohara, T.; Yoshikawa, M.; Yoshida, G.J.; et al. Alternative splicing of CD44 mRNA by ESRP1 enhances lung colonization of metastatic cancer cell. Nat Commun 2012, 3, 883.

175. Erb, U.; Megaptche, A.P.; Gu, X.; Buchler, M.W.; Zoller, M. CD44 standard and CD44v10 isoform expression on leukemia cells distinctly influences niche embedding of hematopoietic stem cells. J Hematol Oncol 2014, 7, 29.

176. Choi, S.H.; Takahashi, K.; Eto, H.; Yoon, S.S.; Tanabe, K.K. CD44s expression in human colon carcinomas influences growth of liver metastases. Int. J. Cancer 2000, 85(4), 523–526.

177. Thorne, R.F.; Legg, J.W.; Isacke, C.M. The role of the CD44 transmembrane and cytoplasmic domains in co-ordinating adhesive and signalling events. J. Cell. Sci. 2004, 117(Pt 3), 373–380.

178. Skelton, T.P.; Zeng, C.; Nocks, A.; Stamenkovic, I. Glycosylation provides both stimulatory and inhibitory effects on cell surface and soluble CD44 binding to hyaluronan. J. Cell Biol. 1998, 140(2), 431–446.

179. Legg, J.W.; Lewis, C.A.; Parsons, M.; Ng, T.; Isacke, C.M. A novel PKC-regulated mechanism controls CD44 ezrin association and directional cell motility. Nat. Cell Biol. 2002, 4(6), 399–407.

180. Lewis, C.A.; Townsend, P.A.; Isacke, C.M. Ca(2+)/calmodulin-dependent protein kinase mediates the phosphorylation of CD44 required for cell migration on hyaluronan. Biochem. J. 2001, 357(Pt 3), 843–850.

181. Thankamony, S.P.; Knudson, W. Acylation of CD44 and its association with lipid rafts are required for receptor and hyaluronan endocytosis. J. Biol. Chem. 2006, 281(45), 34601–34609.

182. Bazil, V.; Strominger, J.L. Metalloprotease and serine protease are involved in cleavage of CD43, CD44, and CD16 from stimulated human granulocytes. Induction of cleavage of L-selectin via CD16. J. Immunol. 1994, 152(3), 1314–1322.

183. Okamoto, I.; Kawano, Y.; Tsuiki, H.; Sasaki, J.; Nakao, M.; Matsumoto, M.; et al. CD44 cleavage induced by a membrane-associated metalloprotease plays a critical role in tumor cell migration. Oncogene 1999, 18(7), 1435–1446.

184. Miller, M.A.; Sullivan, R.J.; Lauffenburger, D.A. Molecular Pathways: Receptor Ectodomain Shedding in Treatment, Resistance, and Monitoring of Cancer. Clin. Cancer Res. 2017, 23(3), 623–629.

185. Kajita, M.; Itoh, Y.; Chiba, T.; Mori, H.; Okada, A.; Kinoh, H.; et al. Membrane-type 1 matrix metalloproteinase cleaves CD44 and promotes cell migration. J. Cell Biol. 2001, 153(5), 893–904.

186. Okamoto, I.; Kawano, Y.; Murakami, D.; Sasayama, T.; Araki, N.; Miki, T.; et al. Proteolytic release of CD44 intracellular domain and its role in the CD44 signaling pathway. J. Cell Biol. 2001, 155(5), 755–762.

187. Pall, T.; Pink, A.; Kasak, L.; Turkina, M.; Anderson, W.; Valkna, A.; et al. Soluble CD44 interacts with intermediate filament protein vimentin on endothelial cell surface. PLoS ONE 2011, 6(12), e29305.

188. Bartolazzi, A.; Nocks, A.; Aruffo, A.; Spring, F.; Stamenkovic, I. Glycosylation of CD44 is implicated in CD44-mediated cell adhesion to hyaluronan. J. Cell Biol. 1996, 132(6), 1199–1208.

189. Rodgers, A.K.; Nair, A.; Binkley, P.A.; Tekmal, R.; Schenken, R.S. Inhibition of CD44 N- and O-linked glycosylation decreases endometrial cell lines attachment to peritoneal mesothelial cells. Fertil. Steril. 2011, 95(2), 823–825.

190. English, N.M.; Lesley, J.F.; Hyman, R. Site-specific de-N-glycosylation of CD44 can activate hyaluronan binding, and CD44 activation states show distinct threshold densities for hyaluronan binding. Cancer Res. 1998, 58(16), 3736–3742.

191. Katoh, S.; Miyagi, T.; Taniguchi, H.; Matsubara, Y.; Kadota, J.; Tominaga, A.; et al. Cutting edge: an inducible sialidase regulates the hyaluronic acid binding ability of CD44-bearing human monocytes. J. Immunol. 1999, 162(9), 5058–5061.

192. Katoh, S.; Zheng, Z.; Oritani, K.; Shimozato, T.; Kincade, P.W. Glycosylation of CD44 negatively regulates its recognition of hyaluronan. J. Exp. Med. 1995, 182(2), 419–429.

193. Bennett, K.L.; Modrell, B.; Greenfield, B.; Bartolazzi, A.; Stamenkovic, I.; Peach, R.; et al. Regulation of CD44 binding to hyaluronan by glycosylation of variably spliced exons. J. Cell Biol. 1995, 131(6 Pt 1), 1623–1633.

194. Azevedo, R.; Gaiteiro, C.; Peixoto, A.; Relvas-Santos, M.; Lima, L.; Santos, L.L.; et al. CD44 glycoprotein in cancer: a molecular conundrum hampering clinical applications. Clin Proteomics 2018, 15, 22.

195. Rudd, P.M.; Dwek, R.A. Glycosylation: heterogeneity and the 3D structure of proteins. Crit. Rev. Biochem. Mol. Biol. 1997, 32(1), 1–100.

196. Jackson, D.G.; Prevo, R.; Clasper, S.; Banerji, S. LYVE-1, the lymphatic system and tumor lymphangiogenesis. Trends Immunol. 2001, 22(6), 317–321.

197. Banerji, S.; Day, A.J.; Kahmann, J.D.; Jackson, D.G. Characterization of a functional hyaluronan-binding domain from the human CD44 molecule expressed in Escherichia coli. Protein Expr. Purif. 1998, 14(3), 371–381.

198. Day, A.J.; Prestwich, G.D. Hyaluronan-binding proteins: tying up the giant. J. Biol. Chem. 2002, 277(7), 4585–4588.

199. Neame, P.J.; Christner, J.E.; Baker, J.R. The primary structure of link protein from rat chondrosarcoma proteoglycan aggregate. J. Biol. Chem. 1986, 261(8), 3519–3535.

200. Zimmermann, D.R.; Ruoslahti, E. Multiple domains of the large fibroblast proteoglycan, versican. EMBO J. 1989, 8(10), 2975–2981.

201. Lee, T.H.; Wisniewski, H.G.; Vilcek, J. A novel secretory tumor necrosis factor-inducible protein (TSG-6) is a member of the family of hyaluronate binding proteins, closely related to the adhesion receptor CD44. J. Cell Biol. 1992, 116(2), 545–557.

202. Prevo, R.; Banerji, S.; Ferguson, D.J.; Clasper, S.; Jackson, D.G. Mouse LYVE-1 is an endocytic receptor for hyaluronan in lymphatic

endothelium. J. Biol. Chem. 2001, 276(22), 19420–19430.

203. Kohda, D.; Morton, C.J.; Parkar, A.A.; Hatanaka, H.; Inagaki, F.M.; Campbell, I.D.; et al. Solution structure of the link module: a hyaluronan-binding domain involved in extracellular matrix stability and cell migration. Cell 1996, 86(5), 767–775.

204. Higman, V.A.; Briggs, D.C.; Mahoney, D.J.; Blundell, C.D.; Sattelle, B.M.; Dyer, D.P.; et al. A refined model for the TSG-6 link module in complex with hyaluronan: use of defined oligosaccharides to probe structure and function. J. Biol. Chem. 2014, 289(9), 5619–5634.

205. Day, A.J. The structure and regulation of hyaluronan-binding proteins. Biochem. Soc. Trans. 1999, 27(2), 115–121.

206. Bajorath, J.; Greenfield, B.; Munro, S.B.; Day, A.J.; Aruffo, A. Identification of CD44 residues important for hyaluronan binding and delineation of the binding site. J. Biol. Chem. 1998, 273(1), 338–343.

207. Peach, R.J.; Hollenbaugh, D.; Stamenkovic, I.; Aruffo, A. Identification of hyaluronic acid binding sites in the extracellular domain of CD44. J. Cell Biol. 1993, 122(1), 257–264.

208. Liao, H.X.; Lee, D.M.; Levesque, M.C.; Haynes, B.F. N-terminal and central regions of the human CD44 extracellular domain participate in cell surface hyaluronan binding. J. Immunol. 1995, 155(8), 3938–3945.

209. Takeda, M.; Terasawa, H.; Sakakura, M.; Yamaguchi, Y.; Kajiwara, M.; Kawashima, H.; et al. Hyaluronan recognition mode of CD44 revealed by cross-saturation and chemical shift perturbation experiments. J. Biol. Chem. 2003, 278(44), 43550–43555.

210. Teriete, P.; Banerji, S.; Noble, M.; Blundell, C.D.; Wright, A.J.; Pickford, A.R.; et al. Structure of the regulatory hyaluronan binding domain in the inflammatory leukocyte homing receptor CD44. Mol. Cell 2004, 13(4), 483–496.

211. Banerji, S.; Wright, A.J.; Noble, M.; Mahoney, D.J.; Campbell,

I.D.; Day, A.J.; et al. Structures of the Cd44-hyaluronan complex provide insight into a fundamental carbohydrate-protein interaction. Nat. Struct. Mol. Biol. 2007, 14(3), 234–239.

212. Takeda, M.; Ogino, S.; Umemoto, R.; Sakakura, M.; Kajiwara, M.; Sugahara, K.N.; et al. Ligand-induced structural changes of the CD44 hyaluronan-binding domain revealed by NMR. J. Biol. Chem. 2006, 281(52), 40089–40095.

213. Favreau, A.J.; Faller, C.E.; Guvench, O. CD44 receptor unfolding enhances binding by freeing basic amino acids to contact carbohydrate ligand. Biophys. J. 2013, 105(5), 1217–1226.

214. Suzuki, T.; Suzuki, M.; Ogino, S.; Umemoto, R.; Nishida, N.; Shimada, I. Mechanical force effect on the two-state equilibrium of the hyaluronan-binding domain of CD44 in cell rolling. Proc. Natl. Acad. Sci. U.S.A. 2015, 112(22), 6991–6996.

215. Vuorio, J.; Vattulainen, I.; Martinez-Seara, H. Atomistic fingerprint of hyaluronan-CD44 binding. PLoS Comput. Biol. 2017, 13(7), e1005663.

216. Bennett, K.L.; Jackson, D.G.; Simon, J.C.; Tanczos, E.; Peach, R.; Modrell, B.; et al. CD44 isoforms containing exon V3 are responsible for the presentation of heparin-binding growth factor. J. Cell Biol. 1995, 128(4), 687–698.

217. Liu, D.; Sy, M.S. Phorbol myristate acetate stimulates the dimerization of CD44 involving a cysteine in the transmembrane domain. J. Immunol. 1997, 159(6), 2702–2711.

218. Liu, D.; Sy, M.S. A cysteine residue located in the transmembrane domain of CD44 is important in binding of CD44 to hyaluronic acid. J. Exp. Med. 1996, 183(5), 1987–1994.

219. Jiang, H.; Knudson, C.B.; Knudson, W. Antisense inhibition of CD44 tailless splice variant in human articular chondrocytes promotes hyaluronan internalization. Arthritis Rheum. 2001, 44(11), 2599–2610.

220. Mori, T.; Kitano, K.; Terawaki, S.; Maesaki, R.; Fukami, Y.; Hakoshima, T. Structural basis for CD44 recognition by ERM proteins. J. Biol. Chem. 2008, 283(43), 29602–29612.

221. Lokeshwar, V.B.; Fregien, N.; Bourguignon, L.Y. Ankyrin-binding domain of CD44(GP85) is required for the expression of hyaluronic acid-mediated adhesion function. J. Cell Biol. 1994, 126(4), 1099–1109.

222. Wang, Y.; Yago, T.; Zhang, N.; Abdisalaam, S.; Alexandrakis, G.; Rodgers, W.; et al. Cytoskeletal regulation of CD44 membrane organization and interactions with E-selectin. J. Biol. Chem. 2014, 289(51), 35159–35171.

223. Brown, K.L.; Birkenhead, D.; Lai, J.C.; Li, L.; Li, R.; Johnson, P. Regulation of hyaluronan binding by F-actin and colocalization of CD44 and phosphorylated ezrin/radixin/moesin (ERM) proteins in myeloid cells. Exp. Cell Res. 2005, 303(2), 400–414.

224. Tsukita, S.; Oishi, K.; Sato, N.; Sagara, J.; Kawai, A.; Tsukita, S. ERM family members as molecular linkers between the cell surface glycoprotein CD44 and actin-based cytoskeletons. J. Cell Biol. 1994, 126(2), 391–401.

225. Zhu, D.; Bourguignon, L.Y. Interaction between CD44 and the repeat domain of ankyrin promotes hyaluronic acid-mediated ovarian tumor cell migration. J. Cell. Physiol. 2000, 183(2), 182–195.

226. Banerji, S.; Ni, J.; Wang, S.X.; Clasper, S.; Su, J.; Tammi, R.; et al. LYVE-1, a new homologue of the CD44 glycoprotein, is a lymph-specific receptor for hyaluronan. J. Cell Biol. 1999, 144(4), 789–801.

227. Gordon, E.J.; Gale, N.W.; Harvey, N.L. Expression of the hyaluronan receptor LYVE-1 is not restricted to the lymphatic vasculature; LYVE-1 is also expressed on embryonic blood vessels. Dev. Dyn. 2008, 237(7), 1901–1909.

228. Nightingale, T.D.; Frayne, M.E.; Clasper, S.; Banerji, S.; Jackson, D.G. A mechanism of sialylation functionally silences the hyaluronan receptor LYVE-1 in lymphatic endothelium. J. Biol. Chem. 2009,

284(6), 3935–3945.

229. Lawrance, W.; Banerji, S.; Day, A.J.; Bhattacharjee, S.; Jackson, D.G. Binding of Hyaluronan to the Native Lymphatic Vessel Endothelial Receptor LYVE-1 Is Critically Dependent on Receptor Clustering and Hyaluronan Organization. J. Biol. Chem. 2016, 291(15), 8014–8030.

230. Jackson, D.G. Hyaluronan in the lymphatics: The key role of the hyaluronan receptor LYVE-1 in leucocyte trafficking. Matrix Biol. 2018.

231. Lynskey, N.N.; Banerji, S.; Johnson, L.A.; Holder, K.A.; Reglinski, M.; Wing, P.A.; et al. Rapid Lymphatic Dissemination of Encapsulated Group A Streptococci via Lymphatic Vessel Endothelial Receptor-1 Interaction. PLoS Pathog. 2015, 11(9), e1005137.

232. Johnson, L.A.; Banerji, S.; Lawrance, W.; Gileadi, U.; Prota, G.; Holder, K.A.; et al. Dendritic cells enter lymph vessels by hyaluronan-mediated docking to the endothelial receptor LYVE-1. Nat. Immunol. 2017, 18(7), 762–770.

233. Farnsworth, R.H.; Achen, M.G.; Stacker, S.A. The evolving role of lymphatics in cancer metastasis. Curr. Opin. Immunol. 2018, 53, 64–73.

234. Karaman, S.; Detmar, M. Mechanisms of lymphatic metastasis. J. Clin. Invest. 2014, 124(3), 922–928.

235. Nunomiya, K.; Shibata, Y.; Abe, S.; Inoue, S.; Igarashi, A.; Yamauchi, K.; et al. Relationship between Serum Level of Lymphatic Vessel Endothelial Hyaluronan Receptor-1 and Prognosis in Patients with Lung Cancer. J Cancer 2014, 5(3), 242–247.

236. Schmaus, A.; Klusmeier, S.; Rothley, M.; Dimmler, A.; Sipos, B.; Faller, G.; et al. Accumulation of small hyaluronan oligosaccharides in tumour interstitial fluid correlates with lymphatic invasion and lymph node metastasis. Br. J. Cancer 2014, 111(3), 559–567.

237. Du, Y.; Liu, H.; He, Y.; Liu, Y.; Yang, C.; Zhou, M.; et al. The

interaction between LYVE-1 with hyaluronan on the cell surface may play a role in the diversity of adhesion to cancer cells. PLoS ONE 2013, 8(5), e63463.

238. Wu, M.; Du, Y.; Liu, Y.; He, Y.; Yang, C.; Wang, W.; et al. Low molecular weight hyaluronan induces lymphangiogenesis through LYVE-1-mediated signaling pathways. PLoS ONE 2014, 9(3), e92857.

239. Johnson, L.A.; Prevo, R.; Clasper, S.; Jackson, D.G. Inflammation-induced uptake and degradation of the lymphatic endothelial hyaluronan receptor LYVE-1. J. Biol. Chem. 2007, 282(46), 33671–33680.

240. Zerbino, D.R.; Achuthan, P.; Akanni, W.; Amode, M.R.; Barrell, D.; Bhai, J.; et al. Ensembl 2018. Nucleic Acids Res. 2018, 46(D1), D754–D761.

241. Consortium, T.U. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017, 45(D1), D158–D169.

242. Banerji, S.; Hide, B.R.; James, J.R.; Noble, M.E.; Jackson, D.G. Distinctive properties of the hyaluronan-binding domain in the lymphatic endothelial receptor Lyve-1 and their implications for receptor function. J. Biol. Chem. 2010, 285(14), 10724–10735.

243. Banerji, S.; Lawrance, W.; Metcalfe, C.; Briggs, D.C.; Yamauchi, A.; Dushek, O.; et al. Homodimerization of the Lymph Vessel Endothelial Receptor LYVE-1 through a Redox-labile Disulfide Is Critical for Hyaluronan Binding in Lymphatic Endothelium. J. Biol. Chem. 2016, 291(48), 25004–25018.

244. Raman, P.S.; Alves, C.S.; Wirtz, D.; Konstantopoulos, K. Distinct kinetic and molecular requirements govern CD44 binding to hyaluronan versus fibrin(ogen). Biophys. J. 2012, 103(3), 415–423.

245. Nishida-Fukuda, H.; Araki, R.; Shudou, M.; Okazaki, H.; Tomono, Y.; Nakayama, H.; et al. Ectodomain Shedding of Lymphatic Vessel Endothelial Hyaluronan Receptor 1 (LYVE-1) Is Induced by Vascular Endothelial Growth Factor A (VEGF-A). J. Biol. Chem. 2016, 291(20), 10490–10500.

246. Wong, H.L.; Jin, G.; Cao, R.; Zhang, S.; Cao, Y.; Zhou, Z. MT1-MMP sheds LYVE-1 on lymphatic endothelial cells and suppresses VEGF-C production to inhibit lymphangiogenesis. Nat Commun 2016, 7, 10824.

247. Dollt, C.; Becker, K.; Michel, J.; Melchers, S.; Weis, C.A.; Schledzewski, K.; et al. The shedded ectodomain of Lyve-1 expressed on M2-like tumor-associated macrophages inhibits melanoma cell proliferation. Oncotarget 2017, 8(61), 103682–103692.

248. Okumura, M.; Shimamoto, S.; Nakanishi, T.; Yoshida, Y.; Konogami, T.; Maeda, S.; et al. Effects of positively charged redox molecules on disulfide-coupled protein folding. FEBS Lett. 2012, 586(21), 3926–3930.

249. Macchione, G.; Paz, J.L. de; Nieto, P.M. Synthesis of hyaluronic acid oligosaccharides and exploration of a fluorous-assisted approach. Carbohydr. Res. 2014, 394, 17–25.

250. Larkin, M.A.; Blackshields, G.; Brown, N.P.; Chenna, R.; McGettigan, P.A.; McWilliam, H.; et al. Clustal W and Clustal X version 2.0. Bioinformatics 2007, 23(21), 2947–2948.

251. Jacobson, M.P.; Pincus, D.L.; Rapp, C.S.; Day, T.J.; Honig, B.; Shaw, D.E.; et al. A hierarchical approach to all-atom protein loop prediction. Proteins 2004, 55(2), 351–367.

252. Schrödinger, N.Y., LLC Schrödinger release 2018-4: MacroModel; http://gts.sourceforge.net/: 2018.

253. Halgren, T.A.; Murphy, R.B.; Friesner, R.A.; Beard, H.S.; Frye, L.L.; Pollard, W.T.; et al. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. J. Med. Chem. 2004, 47(7), 1750–1759.

254. Bubb, W.A. NMR spectroscopy in the study of carbohydrates: Characterizing the structural complexity. Concepts in Magnetic Resonance Wiley: 2003, 19A(1), 1–19 10.1002/cmr.a.10080.

255. Claridge, T. High-Resolution NMR Techniques in Organic Chem-

istry; Elsevier: 2016, 10.1016/c2015-0-04654-8.

256. Gonnet, G.H.; Cohen, M.A.; Benner, S.A. Exhaustive matching of the entire protein sequence database. Science 1992, 256(5062), 1443–1445.

257. Nayeem, A.; Sitkoff, D.; Krystek, S. A comparative study of available software for high-accuracy homology modeling: from sequence alignments to structural models. Protein Sci. 2006, 15(4), 808–824.

258. Rossi, K.A.; Weigelt, C.A.; Nayeem, A.; Krystek, S.R. Loopholes and missing links in protein modeling. Protein Sci. 2007, 16(9), 1999–2012.

259. Colley, K.; Varki, A.; Kinoshita, T. Cellular organization of glycosylation. In Essentials of glycobiology; n.d.

260. Mrázek, H.; Weignerová, L.; Bojarová, P.; Novák, P.; Vaněk, O.; Bezouška, K. Carbohydrate synthesis and biosynthesis technologies for cracking of the glycan code: Recent advances. Biotechnology Advances 2013, 31(1), 17–37 https://doi.org/10.1016/j.biotechadv.2012.03.008.

261. Varki, A.; Gagneux, P. Biological functions of glycans. In Essentials of glycobiology; 3rd ed. Cold Spring Harbor Laboratory Press: 2017.

262. Fernández-Tejada, A.; Cañada, F.J.; Jiménez-Barbero, J. Recent developments in synthetic carbohydrate-based diagnostics, vaccines, and therapeutics. Chemistry - A European Journal Wiley: 2015, 21(30), 10616–10628 10.1002/chem.201500831.

263. Mishra, S.; Upadhaya, K.; Mishra, K.B.; Shukla, A.K.; Tripathi, R.P.; Tiwari, V.K. Chapter 10 - carbohydrate-based therapeutics: A frontier in drug discovery and development. In Atta-ur-Rahman, Ed. Elsevier: 2016, 307–361 https://doi.org/10.1016/B978-0-444-63601-0.00010-7.

264. Sunasee, R.; Adokoh, C.K.; Darkwa, J.; Narain, R. Therapeutic potential of carbohydrate-based polymeric and nanoparticle systems. Expert Opinion on Drug Delivery Informa Healthcare: 2014, 11(6),

867–884 10.1517/17425247.2014.902048.

265. Seeberger, P.H.; Finney, N.; Rabuka, D.; Bertozzi, C.R. Chemical and enzymatic synthesis of glycans and glycoconjugates. In Essentials of glycobiology; 3rd ed. Cold Spring Harbor Laboratory Press: 2017.

266. Schmaltz, R.M.; Hanson, S.R.; Wong, C.-H. Enzymes in the synthesis of glycoconjugates. Chemical Reviews American Chemical Society (ACS): 2011, 111(7), 4259–4307 10.1021/cr200113w.

267. O'Neill, E.C.; Field, R.A. Enzymatic synthesis using glycoside phosphorylases. Carbohydrate Research Elsevier BV: 2015, 403, 23–37 10.1016/j.carres.2014.06.010.

268. HAMURA, K.; SABURI, W.; ABE, S.; MORIMOTO, N.; TAGUCHI, H.; MORI, H.; et al. Enzymatic characteristics of cellobiose phosphorylase fromRuminococcus albusNE1 and kinetic mechanism of unusual substrate inhibition in reverse phosphorolysis. Bioscience, Biotechnology, and Biochemistry Informa UK Limited: 2012, 76(4), 812–818 10.1271/bbb.110954.

269. KINO, K.; SATAKE, R.; MORIMATSU, T.; KURATSU, S.; SHIMIZU, Y.; SATO, M.; et al. A new method of synthesis of alkyl β-glycosides using sucrose as sugar donor. Bioscience, Biotechnology, and Biochemistry Informa UK Limited: 2008, 72(9), 2415–2417 10.1271/bbb.80097.

270. Tran, H.G.; Desmet, T.; Saerens, K.; Waegeman, H.; Vandekerckhove, S.; D'hooghe, M.; et al. Biocatalytic production of novel glycolipids with cellodextrin phosphorylase. Bioresource Technology Elsevier BV: 2012, 115, 84–87 10.1016/j.biortech.2011.09.085.

271. Ernst, B.; Magnani, J.L. From carbohydrate leads to glycomimetic drugs. Nat Rev Drug Discov 2009, 8(8), 661–677.

272. O'Neill, E.C.; Field, R.A. Enzymatic synthesis using glycoside phosphorylases. Carbohydr. Res. 2015, 403, 23–37.

273. Pergolizzi, G.; Kuhaudomlarp, S.; Kalita, E.; Field, R.A. Glycan Phosphorylases in Multi-Enzyme Synthetic Processes. Protein Pept.

Lett. 2017, 24(8), 696–709.

274. Kuhaudomlarp, S.; Walpole, S.; Stevenson, C.E.M.; Nepogodiev, S.A.; Lawson, D.M.; Angulo, J.; et al. Unravelling the specificity of laminaribiose phosphorylase from paenibacillus sp. YM-1 towards donor substrates glucose/mannose 1-phosphate by using x-ray crystallography and saturation transfer difference NMR spectroscopy. ChemBioChem Wiley: 2018, 10.1002/cbic.201800260.

275. Nakajima, M.; Tanaka, N.; Furukawa, N.; Nihira, T.; Kodutsumi, Y.; Takahashi, Y.; et al. Mechanistic insight into the substrate specificity of 1, 2-β-oligoglucan phosphorylase from lachnoclostridium phytofermentans. Scientific Reports Springer Nature: 2017, 7(1) 10.1038/srep42671.

276. Olsson, M.H.; S?ndergaard, C.R.; Rostkowski, M.; Jensen, J.H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. J Chem Theory Comput 2011, 7(2), 525–537.

277. Jayalakshmi, V.; Krishna, N.R. Complete relaxation and conformational exchange matrix (CORCEMA) analysis of intermolecular saturation transfer effects in reversibly forming ligand-receptor complexes. J. Magn. Reson. 2002, 155(1), 106–118.

278. Han, B.; Liu, Y.; Ginzinger, S.W.; Wishart, D.S. SHIFTX2: Significantly improved protein chemical shift prediction. Journal of Biomolecular NMR Springer Nature: 2011, 50(1), 43–57 10.1007/s10858-011-9478-4.

279. Bewley, C.A.; Shahzad-ul-Hussan, S. Characterizing carbohydrate-protein interactions by nuclear magnetic resonance spectroscopy. Biopolymers 2013, 99(10), 796–806.

280. Broeker, N.K.; Andres, D.; Kang, Y.; Gohlke, U.; Schmidt, A.; Kunstmann, S.; et al. Complex carbohydrate recognition by proteins: Fundamental insights from bacteriophage cell adhesion systems. Perspectives in Science Elsevier BV: 2017, 11, 45–52 10.1016/j.pisc.2016.10.001.

281. Hidaka, M.; Kitaoka, M.; Hayashi, K.; Wakagi, T.; Shoun, H.; Fushinobu, S. Structural dissection of the reaction mechanism of cellobiose phosphorylase. Biochemical Journal Portland Press Ltd.: 2006, 398(1), 37–43 10.1042/bj20060274.

282. HONDA, Y.; KITAOKA, M.; HAYASHI, K. Reaction mechanism of chitobiose phosphorylase from vibrio proteolyticus: Identification of family 36 glycosyltransferase in vibrio. Biochemical Journal Portland Press Ltd.: 2004, 377(1), 225–232 10.1042/bj20031171.

283. Coburn, B.; Grassl, G.A.; Finlay, B.B. Salmonella, the host and disease: a brief review. Immunol. Cell Biol. 2007, 85(2), 112–118.

284. Eng, S.-K.; Pusparajah, P.; Mutalib, N.-S.A.; Ser, H.-L.; Chan, K.-G.; Lee, L.-H. Salmonella: A review on pathogenesis, epidemiology and antibiotic resistance. Frontiers in Life Science Informa UK Limited: 2015, 8(3), 284–293 10.1080/21553769.2015.1051243.

285. Crump, J.A.; Sjolund-Karlsson, M.; Gordon, M.A.; Parry, C.M. Epidemiology, Clinical Presentation, Laboratory Diagnosis, Antimicrobial Resistance, and Antimicrobial Management of Invasive Salmonella Infections. Clin. Microbiol. Rev. 2015, 28(4), 901–937.

286. Garai, P.; Gnanadhas, D.P.; Chakravortty, D. Salmonella enterica serovars Typhimurium and Typhi as model organisms: revealing paradigm of host-pathogen interactions. Virulence 2012, 3(4), 377–388.

287. Fabrega, A.; Vila, J. Salmonella enterica serovar Typhimurium skills to succeed in the host: virulence and regulation. Clin. Microbiol. Rev. 2013, 26(2), 308–341.

288. Lathrop, S.K.; Binder, K.A.; Starr, T.; Cooper, K.G.; Chong, A.; Carmody, A.B.; et al. Replication of Salmonella enterica Serovar Typhimurium in Human Monocyte-Derived Macrophages. Infect. Immun. 2015, 83(7), 2661–2671.

289. Richardson, L.A. How Salmonella survives the macrophage's acid attack. PLoS Biol. 2015, 13(4), e1002117.

290. Cota, I.; Sanchez-Romero, M.A.; Hernandez, S.B.; Pucciarelli, M.G.; Garcia-Del Portillo, F.; Casadesus, J. Epigenetic Control of Salmonella enterica O-Antigen Chain Length: A Tradeoff between Virulence and Bacteriophage Resistance. PLoS Genet. 2015, 11(11), e1005667.

291. Velge, P.; Wiedemann, A.; Rosselin, M.; Abed, N.; Boumart, Z.; Chausse, A.M.; et al. Multiplicity of Salmonella entry mechanisms, a new paradigm for Salmonella pathogenesis. Microbiologyopen 2012, 1(3), 243–258.

292. Boumart, Z.; Velge, P.; Wiedemann, A. Multiple invasion mechanisms and different intracellular Behaviors: a new vision of Salmonella-host cell interaction. FEMS Microbiol. Lett. 2014, 361(1), 1–7.

293. Coburn, B.; Sekirov, I.; Finlay, B.B. Type III secretion systems and disease. Clin. Microbiol. Rev. 2007, 20(4), 535–549.

294. Zhang, K.; Riba, A.; Nietschke, M.; Torow, N.; Repnik, U.; Pütz, A.; et al. Minimal SPI1-t3ss effector requirement for salmonella enterocyte invasion and intracellular proliferation in vivo. PLOS Pathogens A.J. Baumler, Ed. Public Library of Science (PLoS): 2018, 14(3), e1006925 10.1371/journal.ppat.1006925.

295. Zhou, D.; Mooseker, M.S.; Galan, J.E. Role of the S. typhimurium actin-binding protein SipA in bacterial internalization. Science 1999, 283(5410), 2092–2095.

296. Hayward, R.D.; Koronakis, V. Direct nucleation and bundling of actin by the SipC protein of invasive Salmonella. EMBO J. 1999, 18(18), 4926–4934.

297. Schlumberger, M.C.; Hardt, W.D. Salmonella type III secretion effectors: pulling the host cell's strings. Curr. Opin. Microbiol. 2006, 9(1), 46–54.

298. Zhou, D.; Galan, J. Salmonella entry into host cells: the work in concert of type III secreted effector proteins. Microbes Infect. 2001, 3(14-15), 1293–1298.

299. Terebiznik, M.R.; Vieira, O.V.; Marcus, S.L.; Slade, A.; Yip, C.M.; Trimble, W.S.; et al. Elimination of host cell PtdIns(4,5)P(2) by bacterial SigD promotes membrane fission during invasion by Salmonella. Nat. Cell Biol. 2002, 4(10), 766–773.

300. Fu, Y.; Galan, J.E. A salmonella protein antagonizes Rac-1 and Cdc42 to mediate host-cell recovery after bacterial invasion. Nature 1999, 401(6750), 293–297.

301. Figueira, R.; Holden, D.W. Functions of the Salmonella pathogenicity island 2 (SPI-2) type III secretion system effectors. Microbiology (Reading, Engl.) 2012, 158(Pt 5), 1147–1161.

302. Yu, X.J.; Ruiz-Albert, J.; Unsworth, K.E.; Garvis, S.; Liu, M.; Holden, D.W. SpiC is required for secretion of Salmonella Pathogenicity Island 2 type III secretion system proteins. Cell. Microbiol. 2002, 4(8), 531–540.

303. Meresse, S.; Unsworth, K.E.; Habermann, A.; Griffiths, G.; Fang, F.; Martinez-Lorenzo, M.J.; et al. Remodelling of the actin cytoskeleton is essential for replication of intravacuolar Salmonella. Cell. Microbiol. 2001, 3(8), 567–577.

304. Kuhle, V.; Abrahams, G.L.; Hensel, M. Intracellular Salmonella enterica redirect exocytic transport processes in a Salmonella pathogenicity island 2-dependent manner. Traffic 2006, 7(6), 716–730.

305. Kujat Choy, S.L.; Boyle, E.C.; Gal-Mor, O.; Goode, D.L.; Valdez, Y.; Vallance, B.A.; et al. SseK1 and SseK2 are novel translocated proteins of Salmonella enterica serovar typhimurium. Infect. Immun. 2004, 72(9), 5115–5125.

306. Yang, Z.; Soderholm, A.; Lung, T.W.; Giogha, C.; Hill, M.M.; Brown, N.F.; et al. SseK3 Is a Salmonella Effector That Binds TRIM32 and Modulates the Host's NF-ÎºB Signalling Activity. PLoS ONE 2015, 10(9), e0138529.

307. Deng, W.; Puente, J.L.; Gruenheid, S.; Li, Y.; Vallance, B.A.; Vazquez, A.; et al. Dissecting virulence: systematic and functional

analyses of a pathogenicity island. Proc. Natl. Acad. Sci. U.S.A. 2004, 101(10), 3597–3602.

308. Newton, H.J.; Pearson, J.S.; Badea, L.; Kelly, M.; Lucas, M.; Holloway, G.; et al. The type III effectors NleE and NleB from enteropathogenic E. coli and OspZ from Shigella block nuclear translocation of NF-kappaB p65. PLoS Pathog. 2010, 6(5), e1000898.

309. Gunster, R.A.; Matthews, S.A.; Holden, D.W.; Thurston, T.L. SseK1 and SseK3 Type III Secretion System Effectors Inhibit NF-ÎºB Signaling and Necroptotic Cell Death in Salmonella-Infected Macrophages. Infect. Immun. 2017, 85(3).

310. Li, S.; Zhang, L.; Yao, Q.; Li, L.; Dong, N.; Rong, J.; et al. Pathogen blocks host death receptor signalling by arginine GlcNAcylation of death domains. Nature 2013, 501(7466), 242–246.

311. Gao, X.; Wang, X.; Pham, T.H.; Feuerbacher, L.A.; Lubos, M.L.; Huang, M.; et al. NleB, a bacterial effector with glycosyltransferase activity, targets GAPDH function to inhibit NF-ÎºB activation. Cell Host Microbe 2013, 13(1), 87–99.

312. Singh, D.G.; Lomako, J.; Lomako, W.M.; Whelan, W.J.; Meyer, H.E.; Serwe, M.; et al. beta-Glucosylarginine: a new glucose-protein bond in a self-glucosylating protein from sweet corn. FEBS Lett. 1995, 376(1-2), 61–64.

313. Esposito, D.; Gunster, R.A.; Martino, L.; El Omari, K.; Wagner, A.; Thurston, T.L.M.; et al. Structural basis for the glycosyltransferase activity of the Salmonella effector SseK3. J. Biol. Chem. 2018, 293(14), 5064–5078.

314. Breton, C.; Snajdrova, L.; Jeanneau, C.; Koca, J.; Imberty, A. Structures and mechanisms of glycosyltransferases. Glycobiology 2006, 16(2), 29R–37R.

315. Cui, Qiu ?; Lewis, Ian ?; Westler, Gareth ?; Anderson, M.E.; Markley, J.L. Biological Magnetic Resonance Bank: 2006, 10.13018/bmse000205.

316. Park, J.B.; Kim, Y.H.; Yoo, Y.; Kim, J.; Jun, S.H.; Cho, J.W.; et al. Structural basis for arginine glycosylation of host substrates by bacterial effector proteins. Nat Commun 2018, 9(1), 4283.

317. Pearson, J.S.; Giogha, C.; Ong, S.Y.; Kennedy, C.L.; Kelly, M.; Robinson, K.S.; et al. A type III effector antagonizes death receptor signalling during bacterial gut infection. Nature 2013, 501(7466), 247–251.

318. El Qaidi, S.; Chen, K.; Halim, A.; Siukstaite, L.; Rueter, C.; Hurtado-Guerrero, R.; et al. NleB/SseK effectors from Citrobacter rodentium, Escherichia coli, and Salmonella enterica display distinct differences in host substrate specificity. J. Biol. Chem. 2017, 292(27), 11423–11430.

319. Gottlieb, H.E.; Kotlyar, V.; Nudelman, A. NMR chemical shifts of common laboratory solvents as trace impurities. The Journal of Organic Chemistry American Chemical Society (ACS): 1997, 62(21), 7512–7515 10.1021/jo971176v.

320. Castanar, L.; Sistare, E.; Virgili, A.; Williamson, R.T.; Parella, T. Suppression of phase and amplitude J(HH) modulations in HSQC experiments. Magn Reson Chem 2015, 53(2), 115–119.

321. Sastry, G.M.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. J. Comput. Aided Mol. Des. 2013, 27(3), 221–234.

322. Vanquelef, E.; Simon, S.; Marquant, G.; Garcia, E.; Klimerak, G.; Delepine, J.C.; et al. R.E.D. Server: a web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments. Nucleic Acids Res. 2011, 39(Web Server issue), W511–517.

323. Salomon-Ferrer, R.; Case, D.A.; Walker, R.C. An overview of the amber biomolecular simulation package. Wiley Interdisciplinary Reviews: Computational Molecular Science Wiley: 2012, 3(2), 198–210 10.1002/wcms.1121.

324. Oxenoid, K.; Kim, H.J.; Jacob, J.; Sonnichsen, F.D.; Sanders, C.R. NMR assignments for a helical 40 kDa membrane protein. J. Am. Chem. Soc. 2004, 126(16), 5048–5049.

325. Li, X.; Krafczyk, R.; Maco?ek, J.; Li, Y.L.; Zou, Y.; Simon, B.; et al. Resolving the Î±-glycosidic linkage of arginine-rhamnosylated translation elongation factor P triggers generation of the first ArgRha specific antibody. Chem Sci 2016, 7(12), 6995–7001.

326. Tvaroska, I.; Taravel, F.R. Carbon-proton coupling constants in the conformational analysis of sugar molecules. Adv Carbohydr Chem Biochem 1995, 51, 15–61.

327. Parker, D.; Prince, A. Immunoregulatory effects of necroptosis in bacterial infections. Cytokine 2016, 88, 274–275.

328. Ahn, D.; Prince, A. Participation of Necroptosis in the Host Response to Acute Bacterial Pneumonia. J Innate Immun 2017, 9(3), 262–270.

329. Bender, L.M.; Morgan, M.J.; Thomas, L.R.; Liu, Z.G.; Thorburn, A. The adaptor protein TRADD activates distinct mechanisms of apoptosis from the nucleus and the cytoplasm. Cell Death Differ. 2005, 12(5), 473–481.