Novel Saturation Transfer Difference (STD) NMR approaches to understand biologically relevant protein-carbohydrate interactions

Serena Monaco

Faculty of Science

School of Pharmacy

Thesis submitted to the School of Pharmacy, University of East Anglia, in fulfilment of the requirement for the degree of Doctor of Philosophy

Norwich, April 2018

To Florent,

the most beautiful discoveries are the ones we are making together

Acknowledgements

I will start these acknowledgements going straight to the point: I have had the biggest luck a young researcher in my field could have. I have learnt Saturation Transfer Difference NMR from the hands of one of its main developer, my supervisor, Jesùs Angulo, sitting at the spectrometer together, day after day, when I was the only (and very proud!) member of the Angulo's Lab. My gratitude for this is immense, but this is not the only reason for thanking you, Jesùs: you have been more than a supervisor, you have been a mentor and an example of enthusiasm, perseverance and passion as few others. Thanks for your scientific and personal support through these years, for your positivity and empathy, for teaching me the concept of "tolerance to frustration" and the respect for my limits (while still pushing them further), lessons often disregarded in academy. Overall, you made my life *possible*, this made the difference.

I am deeply thankful to Yaroslav Khimyak, my second supervisor. Thanks for your constant support, the fruitful and inspiring scientific conversations we had along the years, but also for your endless culture and the care you put in widening our horizons and keeping us together as a beautiful and cheerful group.

A very special thank goes to Nathalie Juge and Louise Tailford at the Quadram Institute of Biotechnology. The chance to collaborate with you has been a real honour and the best intellectual fun for me. Thanks for your trust since the very early days of my PhD, and for your availability at any time. This has been so precious to me, and I really hope we can keep this going in the years to come.

I am very grateful to the School of Pharmacy, the funding body of my 3 years here; and with it, I would like to thank Colin Macdonald for his help with the NMR instrumentation and Chris Morris, for giving me the chance of few years of enriching demonstration in the Physical Pharmacy module and for being always available (especially in these last few important months). A big thank also goes to the students I had the chance to supervise during my PhD: Aditya, Pani, Sham, Sarah, Ian, Tom and Agne. Thanks also to Joanne, Sharron, Kate and John for helping with every other aspect of the academic life.

The NMR group has always been a big family for me, and there's a list of people I want to thank. First of all, Sam Walpole, thanks for being on my side since October '15, and thanks for all your support every day since then. I think of you as my main partner in this academic adventure, and I feel so lucky I could work along such a brilliant mind and join

our efforts to produce some exciting science: it has been a pleasure and an honour. Susana Ramalhete and Karol Nartowski, made the days in the lab a constant source of enjoyment and mutual support. In the toughest days, your presence around made the whole difference. Thanks for all the personal and scientific support since day-one until today, throughout the distance: thanks for being *real* friends. The lab days would have not been the same without Franco, Francesc, Lucka, Martyna, Fabiana, Francesca, Cholpon, Elise, Remy and Teresa. It is so tough to endure so many goodbyes, but it has been truly enriching sharing a piece of the road together. Luckily some people stay, and I want to deeply thank Valeria, for being the most joyful and supportive mate, Vani and Juan Carlos for the enlightening "post-doc input" and for their precious help with this thesis, Alex, Matthew and Dani for making the NMR group as special as it is.

I feel lucky to have found so many special people in Norwich, but I would like to add some personal acknowledgements, to the friends whom I painfully left somewhere else in Europe to start this beautiful experience. Thanks to Nando, Carlo and Marina for keeping being my Danish family, thanks to Lilli and Desi for being the corner of a big triangle across Europe together, for showing me that some things will never change. Clementine and Godot, thanks for being the closest to my soul, and the best friends I could ever whish for. A huge thank goes to my family is in UK: thanks to my uncle Matteo, for being there anytime and to my awesome brother, Ivan, for being "home" when I needed the most, I feel so lucky we got re-united in this corner of the world.

These three years have been a challenging trip, but it was worth the happiness and satisfaction I found through all my PhD and in my life in Norwich. This chance would never be real without the inspirations and the encouragement to discover which I had from my parents, two physicists, who only asked for one thing in my whole life: that I was happy. The sparkle that I saw in the eyes of my dad every night he came back from the lab since I was a child certainly played a hidden role in the choices I have made so far, and the multidisciplinary and curious mind of my mum, who always keeps learning and expanding her knowledge, was and is a great inspiration to never stop discovering.

From the deepest point of my heart, my biggest thank goes to whom arrived serendipitously, who changed life and moved to a new country, to be by my side in this PhD adventure: my husband-to-be, Florent. Thanks for all the things we are learning from each other, for sharing success and defeat with me, for believing in me, for supporting my dreams, for the beautiful life we are choosing to build together.

Abstract

Saturation transfer difference (STD) NMR spectroscopy is a powerful NMR technique extensively used to obtain epitope maps of ligands binding to biologically relevant protein receptors. This allows to reveal semi-quantitative structural details of the interaction, which is key to direct lead optimization efforts in drug discovery. However, it does not give information about the nature of the amino acids surrounding the ligand in the binding pocket.

In this thesis, the main effort has been put to develop two novel implementations of STD NMR, aimed at elucidating the surroundings of the ligand (*i.e.*, the amino acids lining the binding pocket, or an adjacent bound ligand) in biologically relevant complexes. First, we report the development of the novel "*DiffErential EPitope mapping STD NMR*" (DEEP-STD NMR), a method producing differential epitope maps through i) differential frequency and/or ii) differential solvent (D_2O/H_2O) STD NMR experiments. These two approaches provide complementary information on the architecture of the binding pocket. The second novel method we propose is "*Inter-ligand STD NMR*" (IL-STD NMR), which relies on *on-ligand* differential frequency STD NMR to detect contacts between ligands bound to adjacent sites of a receptor.

These novel STD NMR methodologies, in combination with traditional STD NMR and computational tools, have been applied to the study of two systems: the interactions of Cholera Toxin subunit B (CTB) with a set of promising inhibitors; and the interactions of an intramolecular *trans*-sialidase from *Ruminococcus gnavus*, a gut microbiota symbiont, with a set of mucin-related sialylated ligands. In the first study, we discovered the existence of a hitherto unknown binding subsite in the GM1 binding site of CTB. In the second study, we provided the first 3D molecular model of a Michaelis complex for an IT-sialidase.

In both cases, we demonstrate that our newly developed approaches increase the level of resolution of STD NMR, widening its potential to impact the field of ligand design for biologically relevant receptors.

Publications

- Monaco, S.; Tailford, L. E.; Juge, N.; Angulo, J., Differential Epitope Mapping by STD NMR Spectroscopy To Reveal the Nature of Protein–Ligand Contacts. *Angewandte Chemie International Edition* **2017**, 56, 15289-93.
- Owen, D. C.*; Tailford, L. E.*; Monaco, S.; Šuligoj T.; Vaux, L.; Lallement, R.; Khedri, Z.; Yu, H.; Lecointe, K.; Walshaw, J.; Tribolo, S.; Horrex, M.; Bell, A.; Chen, X.; Taylor, G. L.; Varki, A.; Angulo, J.; Juge, N., Unravelling the specificity and mechanism of sialic acid recognition by the gut symbiont *Ruminococcus gnavus*. *Nature Communications* **2017**, 8, 2196.
- Walpole, S.; Monaco, S.; Nepravishta, R.; Angulo, J., STD NMR as a technique for ligand screening and structural studies. Submitted.
- Park, J. B.; Kim, Y. H.; Yoo, Y.; Kim, J.; Jun, S.; El Qaidi, S.; Walpole, S.; Monaco, S.; Wu, M.; Hays, M. P.; Hurtado-Guerrero, R.; Angulo, J.; Hardwidge, P. H.; Shin, J.; Cho, H., Structural basis for arginine glycosylation of host proteins by bacterial effectors. Submitted.
- Monaco, S.; Walpole, S.; Ramos-Soriano, J.; Moreno Vargas, A. J.; Carmona, A. T.; Robina, I.; Angulo, J., Inhibition of Cholera Toxin B: Discovery of a Cryptic Subsite at the GM1 Binding Pocket by Novel Inter-Ligand STD NMR approach (IL-STDs) combined with Molecular Dynamics. Manuscript in phase of submission.
- Monaco, S.*; Tailford, L. E.*; Bell, A.; Gao, Z.; Walshaw, J.; Latousakis, D.; Owen, D. C.; Chen, X.; Withers, S.; Angulo, J.; Juge, N. Mechanism of action of the intramolecular *trans*-sialidase from the gut symbiont *Ruminococcus gnavus*. Manuscript in phase of preparation.
- Monaco, S.; Walpole, S.; Ramos-Soriano, J.; Carmona, A. T.; Robina, I.; Angulo, J., Exploring the three-subsite GM1 binding site of CTB towards three-finger inhibitors: a DEEP-STD NMR and computational study. Manuscript in phase of preparation.

*These authors contributed equally to this work.

Presentations at conferences

- "Cholera toxin inhibition: discovery of a novel binding subsite in the GM1 binding pocket", oral presentation at the Royal Society of Chemistry Carbohydrate Group 50th Anniversary Meeting, Warwick, November 2016.
- "Inter Ligand-STD - And the discovery of a novel binding subsite for cholera toxin inhibition", oral presentation and poster at the NMR DG PG Meeting, Glasgow, June 2017.
- "Cholera toxin inhibition: discovery of a novel binding subsite in the GM1 binding pocket", oral presentation at EUROCARB 2017, Barcelona, July 2017.
- "Differential Epitope Mapping STD NMR (DEEP-STD NMR) to reveal the nature of protein-ligand contacts", SMASH 2017, Baveno, September 2017.

Poster awards

- "A 3-dimensional view of protein-carbohydrate interactions by STD NMR: characterisation of the ligand specificity of an intramolecular *trans*-sialidase from the human gut microbiota": Best Poster Award for Next Generation Presentation Therapeutics at the School of Pharmacy Research Colloquium 2016, UEA.
- "DiffErential EPitope Mapping by STD NMR spectroscopy (DEEP-STD NMR) to assess protein-ligand pharmacophores": Best poster Award for Healthy Aging at the School of Pharmacy Research Colloquium 2017, UEA.

Abbreviations and acronyms

3'SL	3'sialyllactose
3NPG	3-*nitro*-phenyl-α-D-galactoside
6'SL	6'sialyllactose
AMBER	Assisted Model Building with Energy Refinement
Benz	Benzyl
BSA	Bovine Serum Albumine
CAZy	Carbohydrates-Active enZymes database
CBM	Carbohydrate Binding Module
CORCEMA-ST	COmplete Relaxation and Conformational Exchange MAtrix for Saturation Transfer
CPMG	Carr-Purcell-Meiboom-Gill sequence
CSP	Chemical Shift Mapping
CT	Cholera Toxin
CTA	Cholera Toxin subunit A
CTB	Cholera Toxin subunit B
DEEP-STD	DiffErential EPitope mapping Saturation Transfer Difference
DMSO	Dimethyl sulphoxide
DQF-COSY	Double Quantum Filter COrrelation SpectroscopY
FBDD	Fragment Based Drug Discovery
FID	Free Induction Decay
Fur	Furoate
Gal	Galactose
GH	Glycoside hydrolase
GIT	Gastro-Intestinal Tract
Glc	Glucose

GLIDE	Grid-Based Ligand Docking with Energetics
GM1	Monosialotetrahexosylganglioside
HREMD	Hamiltonian Replica Exchange Molecular Dynamics
HSQC	Heteronuclear Single Quantum Coherence
IBD	Inflammatory Bowel Disease
ILOE	Inter-Ligand NOESY
IL-STD	Inter-Ligand saturation transfer difference
ITC	Isothermal Titration Calorimetry
IT-sialidase	Intramolecular *Trans*-sialidase
MCMM	Monte Carlo Multiple Minimum
MD	Molecular Dynamics
NMR	Nuclear Magnetic Resonance
Neu5Ac	*N*-Acetylneuraminic acid
NOE	Nuclear Overhauser Effect
NOESY	Nuclear Overhauser Effect SpectroscopY
OPSL-AA	Optimised Potensial for Liquid Simulations for All Atoms
PBS	Phosphate-Buffered Saline
PDB	Protein Data Bank
PHF	Polyhydroxyalkylfuroate
ppm	part per million
PTM	Post-Translational Modifications
rf	radio-frequency
STD	Saturation Transfer Difference
TOCSY	TOtal Correlated SpectroscopY
TPPI	Time-Proportional Phase Incrementation
Tr-NOESY	Exchange transferred-NOESY
XRD	X-Ray Diffraction
ΔDEEP-STD	DEEP-STD factor
ΔIL-STD	IL-STD factor

Table of Contents

Chapter 1

Introduction and objectives

1.1 Carbohydrates as biological and pharmaceutical ligands

At the beginning of the millennium, the result of the Human Genome Project (HGP) left the scientific world astonished: the entire human genome is composed of only about 26,500 genes[1]. Previous predictions estimated this number to be ranging from 50,000 to 140,000 genes, also considering that *S. cerevisiae* (baker yeast) and *D. melanogaster* (fruit fly) have respectively 6,000 and 16,000 genes[2,3]. The existing biological diversity is then encoded in such a small number of genes. To explain this, the authors of the HGP reasoned that "the finding that the human genome contains fewer genes than previously predicted might be compensated for by the combinatorial diversity generated at the level of post-translational modification of proteins"[1]. Among the post-translational modifications, glycosylation is the second most frequent one, together with acetylation and after phosphorylation[4]. Glycosylation is the process of attaching an oligosaccharide, *i.e.* a glycan, to oxygen or nitrogen atoms of specific protein residues[5]. In this section, we will briefly introduce carbohydrates as a complex and charming class of biological molecules and functional ligands; also, we will explain why deepening our knowledge on protein-carbohydrate interactions can lead us to understand many physiological and pathological events in the human body, and to appreciate their beautiful complexity.

1.1.1 From monosaccharides to complex glycans: the chemical perspective

Carbohydrates are chemically defined as polyhydroxy ketones, aldehydes, alcohols and carboxylic acids. For most of them, the molar ratio of 1:2:1 between carbon, hydrogen and oxygen holds; hence, they have been historically defined as "hydrates of carbon", *i.e.* "carbohydrates". They are prime biological substances, metabolized as mono-, oligo- and polysaccharides[6]. In physiological conditions, monosaccharides form cyclic hemiacetals and hemiketals by intramolecular nucleophilic addition of one of their hydroxyl groups to the carbonyl function. The equilibrium is shifted towards the 5- or 6-

membered heterocycles (called "furanose" and "pyranose" for their similarity to furan and pyran, respectively).

Depending on the relative designation for the configuration of the stereogenic centre further from the carbonylic group, the stereochemistry of a sugar is designed by the use of prefixes "D-" or "L-", of historical derivation (a "D-" and an "L-" sugars are enantiomers with respect to all the stereogenic centres of the molecule)[7]. Importantly, in solution, the rings are found in equilibrium between two epimeric forms. In D-sugars, these forms are called α if the hydroxyl on C1 is in axial position and β if it is in equatorial position (Figure 1.1a,b). The carbon C1 is also called "anomeric" carbon, due to its variable stereochemistry.

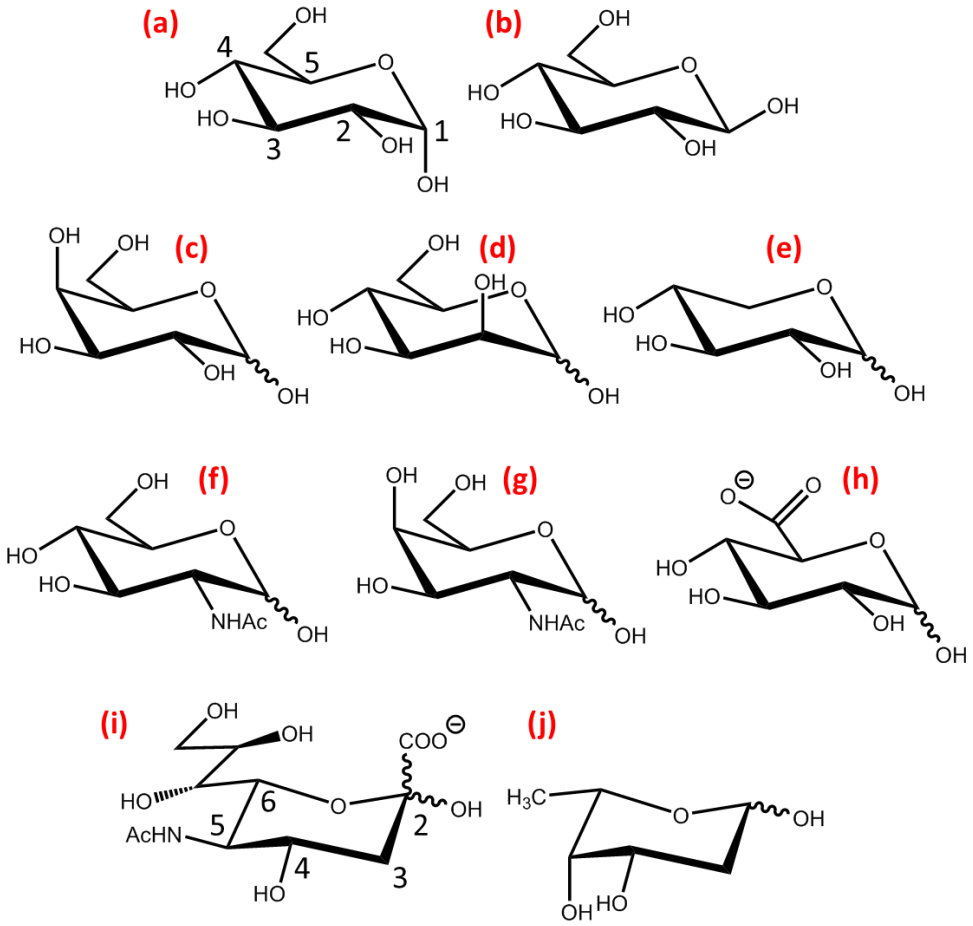

Figure 1.1. The 9 fundamental monosaccharides found in vertebrates. **(a)** α-D-glucose (α-Glc), **(b)** β- D-glucose (β-Glc), **(c)** D-Galactose (Gal), **(d)** D-Mannose (Man), **(e)** D-Xylose (Xyl), **(f)** *N*-acetyl-D-glucosamine (GlcNAc), **(g)** *N*-acetyl-D-galactosamine (GalNAc), **(h)** D-Glucuronic acid (GlcUA), **(i)** Sialic acid (Neu5Ac) and **(j)** D-Fucose (Fuc). In **(a)** and **(b)**, the two anomeric configurations of Glucose are shown. For all the other sugars, the curly bonds indicate either

one or the other configuration. Carbon nomenclature is given for Glucose and applies from **(a)** to **(h)**, and for Sialic acid, and applies for **(i)** and **(j)**.

The fundamental monosaccharides in vertebrates are 9 (Figure 1.1)[8], and they are mainly found as D-sugars, in the same way as the fundamental amino acids are 21 (mainly found in their "L-" form), and the fundamental nucleotides are 5. Glucose is the simplest of the fundamental monosaccharides and the other 8 can be metabolised starting from it. Enzymes called glycosyl-transferases can form glycosidic linkages between two or more monosaccharides. Due to the presence of two adjacent oxygens, the anomeric carbon is susceptible to nucleophilic attack from any hydroxyl groups[6]. If the attacking nucleophile is situated on a second monosaccharide unit, the new molecule is defined as a disaccharide[7]. The ring whose anomeric position is engaged in the glycosidic bond is called the non-reducing end, and the ring with a free anomeric hydroxyl is called the reducing end (Figure 1.2). It is easy to expand this scenario to the formation of consecutive glycosylic bonds yielding trisaccharides, oligosaccharides and eventually polysaccharides.

The presence of nucleophilic donors on several carbons of each monosaccharide means that the regiochemistry of the linkage can vary enormously and the opportunity for branching is very conspicuous[8]. Here it lies a substantial difference between carbohydrates and the two other main classes of bio-molecules: starting from the 9 basic aldo-pyranoses, the possible tetrasaccharides are 10 million; whereas, the possible tetramer formed by the 4 DNA nucleotides are 400, and 200,000 are the possible tetra-peptides which we can obtain from the 21 fundamental amino acids (lipids have been excluded from this discussion). In this combinatorial calculation, we have not considered the fact that, additionally, glycosidic linkages are not only defined by the regiochemistry (*e.g.*, (1-2), (1-3), (1-4) linkages, *etc.*), but also by the stereochemistry at the non-reducing ring (e.g., α(1-2), β(1-2), α(1-3), β(1-3), linkages, *etc.*) . Thus, if a galactose is transferred onto the carbon C4 of a glucose, there are two possible outcomes, depending on whether the galactose is in α or in β configuration. In the first case, we will have the α-Gal(1-4)Glc disaccharide and in the second case we will have the β-Gal(1-4)Glc disaccharide, which we commonly call lactose. Additionally, the glycosidic bond can be highly flexible, allowing rotation around the two dihedral angles ϕ and ψ[7], defined in Figure 1.2. This means that linear, but even more branched, oligosaccharides

have the potential to explore a wide conformational space. The potential of glycans of assuming several "shapes" in solution adds a further layer of information-coding and it has been referred to as "the third dimension of the sugar code" [9].

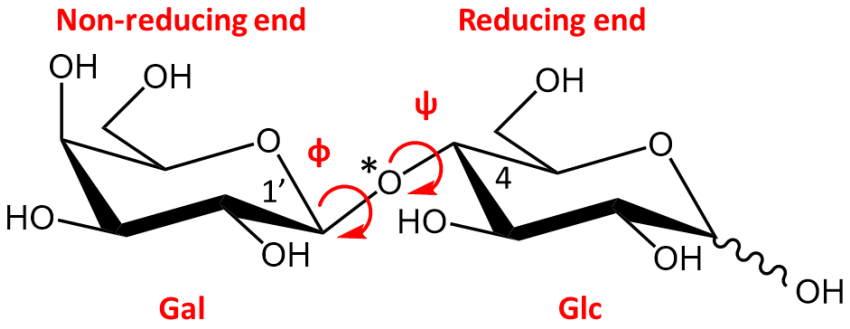

Figure 1.2. Example of a glycosidic bond in the disaccharide β-Gal(1-4)Glc (Lactose). The reducing and non-reducing ends are indicated, and the red curved arrows define the two dihedral angles: ϕ (O5'-C1'-O*-C4) and ψ (C1'-O*-C4-C3)[10].

The elements of glycan structural variability encountered so far are:

i) Nature of the monomer units (*i.e.* composition);

ii) Regiochemistry of the linkage;

iii) Stereochemistry of the linkage (at the anomeric positions);

iv) Conformational flexibility (of the glycosidic linkage).

The enormous number of possible structural variation is self-explanatory on the potential of information-storage of the totality of glycan structures in living organisms, which is what we call the "glycome".

1.1.2 The glycan code: exploring the social life of cells

Having described the elements of structural complexity of glycans from a chemical point of view, we now want to highlight their fundamental role in biological contexts after post-translational modification (PTM) takes place.

Biologically, glycans decorating proteins post-translationally modified can be divided into *N*- and *O*-glycans, depending on whether they are attached to an asparagine (Asn) or a threonine/serine (Thr, Ser), respectively[8]. *N*-glycans are biosynthesised in the endoplasmic reticulum (ER) on a lipid-like molecule termed dolichol phosphate, from which they are then transferred to the newly synthesised proteins in the lumen of the

ER. *O*-glycosylation takes place in the Golgi but is less frequent than *N*-glycosylation and less is known on how it works. Mucins are the class of glycoproteins carrying the greatest number of *O*-glycans.

Going through the Golgi to reach the cytoplasm, glycoproteins undergo further modifications to allow distribution to their final destination: to specific cells compartments, to the cell surface or to the extracellular space[10]. Glycans in PTMs can be thought of as "tags" (exposed on the membranes of cellular organelles, of the cell itself or of free glycoproteins in the extracellular matrix) carrying specific information and allowing interactions with other entities, able to recognise and "read" the tags (Figure 1.3). Carbohydrate-recognising proteins are the counterpart of this "communication". Protein-carbohydrate interactions are the processes to transduce the information delivered by PTMs. In another analogy, we can envision glycans as "keys"; then, the biological receptors, exposed on self-cells, guest-cells, viruses, or free receptors such as antibodies, toxins, carbohydrate binding modules *etc.*, are the "locks", as first postulated by Emil Fisher in 1894[11]. The "keys" fit (almost always) uniquely in their "locks" allowing the information to be de-codified to start physiological or pathological processes[12], in what has been defined as "the social life of cells"[13].

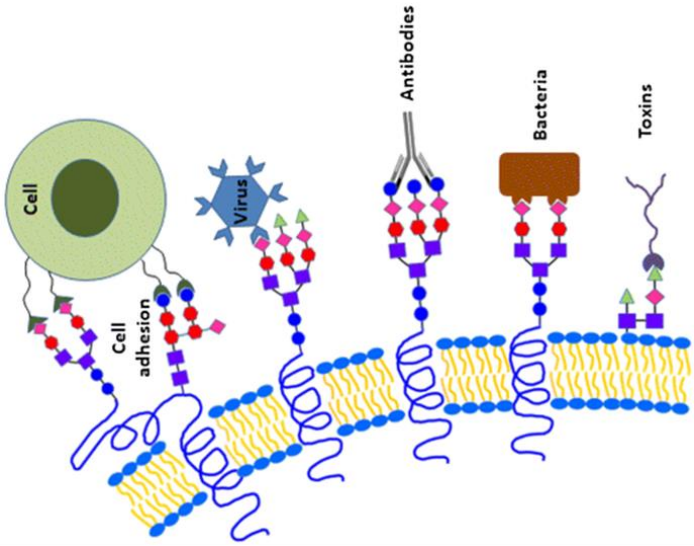

Figure 1.3. Cartoon representing five examples of biological host-guest recognition processes in which protein-carbohydrate interactions are involved. Toxins, bacteria, antibody, viruses or self-cells recognise glycans exposed on the cell surface of a host cell. Figure from 14.

1.1.3 Medicinal glycochemistry: the pharmaceutical perspective

In this thesis, we focus on two biologically relevant protein-carbohydrate interactions, involving bacterial proteins recognising glycans exposed on the intestinal lumen. In one case, the bacterial protein is the toxin from *V. cholera* (CT) which recognises a branched pentasaccharide (GM1) exposed on the epithelium of our gut. The GM1/CT interactions is one of the strongest protein-carbohydrate interactions (with a K_D in the low nM range[15]) and mediates the internalisation of the toxin in the lumen of the epithelium cells, where the enzymatically active unit of the toxin triggers a cascade disrupting the ionic channels, and leading to strong dehydration[16]. In the second system, the bacterial protein is an intramolecular *trans*-sialidase from *R. gnavus*, a commensal bacterium in our gut[17]. Sialidases are carbohydrate active enzymes cleaving sialic acid units that are frequently capping the *O*-glycans which are present in high concentration on some glycoproteins such as mucins. By cleaving, and in some cases metabolising, sialic acid units they can adapt to mucin-rich niches of the intestine and colonise them[18]. In some cases, mucin degradation is associated to severe pathologies (such as inflammatory bowel disease in the gut, or cystic fibrosis in the lungs[19]).

These are only two examples of protein-carbohydrate interactions which became of interest in medicinal studies. Among other notorious examples of non-bacterial protein-carbohydrate interactions we can mention the noxious interactions between hemagglutinin and neuraminidase expressed on the surface of the influenza virus with the sialic acid exposed on the epithelium of the host cells[20]; or the leukocyte-endothelial initiating the leukocyte recruitment during acute and chronic inflammation, mediated by selectins, a glycan-binding class of adhesin proteins[21]. Protein-carbohydrate interactions also play a fundamental role in immunology processes and another remarkable example is the definition of blood types, first distinguished by Karl Landsteiner in 1900[22]. To be blood type 0, A, B or AB simply refers to the type of glycan exposed on the blood cells of each individual, a piece of knowledge which allowed to safely perform blood transfusion already during WWII[23]. Glycochemistry is also in the spotlight of vaccinology for the development of carbohydrate-based vaccines (glycol-vaccines). This was initially based on the idea of using bacterial polysaccharides as antigens, exploiting the unique glycan structure found on diverse pathogens[24]. Starting from there, it has been possible to target the surface carbohydrates from protozoa to

cancer cells in the aim of developing always safer and more efficient vaccines[25]. Nevertheless, the main centre of interest in pharmaceutical research with respect to glycochemistry seems still to be the design and development of inhibitors of protein-carbohydrates interaction[26], which is the aim of the first of the two biological projects of this thesis (CTB). To do this in a knowledge-based fashion, a detailed understanding of how protein-carbohydrate interactions work is necessary, and we will try to cover the major features from a structural, thermodynamic and kinetic point of view in the following section.

1.2 Protein-carbohydrate interactions: structural, thermodynamics and kinetics aspects

1.2.1 Structural recognition patterns specific to protein-carbohydrate complexes

Lectins are the main family of carbohydrate-binding proteins. Their sugar-binding capacity and their lack of enzymatic activity are the main common features to the components of this class of proteins, which otherwise differ largely in their binding site architectures, ranging from shallow grooves to deep pockets[12]. There are diverse classes of lectins: legume lectins, C-type lectins, galectins, viral proteins, toxins and so on[14]. When carbohydrate-binding proteins are associated to enzymatically active domains, unlike lectins, we define them as carbohydrate-binding modules (CBMs, CAZy, http://www.cazy.org[27]). Carbohydrate-active enzymes are another family of carbohydrate-recognising proteins, performing diverse enzymatic modification to their sugar substrates, once they bind. Among the enzymatic transformation they can perform there is hydrolysis and rearrangement of glycosidic bond (glycoside hydrolases), formation of new glycosidic bonds (glycoside transferases), non-hydrolytic cleavage of glycosidic bonds (polysaccharide lyases), just to name the most populated classes (CAZy, http://www.cazy.org[27]). Sugar-recognition is a prerogative of both active and non-active classes of proteins, regardless the destiny of the sugar substrate upon binding. For this reason, it is possible to generalise several recognition elements specific to protein-carbohydrate interactions.

The slightly amphiphilic nature of carbohydrates is key to their specificity of interaction. The high density of hydroxyl groups makes the sugars suited as both H-bond donors and acceptors (Figure 1.4a,b)[12]. The most common side chains involved in the formation of

cooperative H-bonds are arginine, aspartate and asparagine[28]. In cases in which the hydroxyl groups are in a favourable geometry, they can coordinate Ca^{2+} ions inherent to the protein. Equatorial hydroxyl groups on carbons C3 and C4 are particularly suitable for Ca^{2+} complexation (this excludes galactoses), and mannose-binding C-type lectins are good examples of this recognition pattern (Figure 1.4c)[12]. In fact, most monosaccharides differ in the relative position of one or two hydroxyl groups. The different electron-density profile for a pair of epimers (*e.g.* glucose and mannose, epimers at C2, or glucose and galactose, epimers at C4) determines different directionality of H-bond (or ion complexation), making each sugar uniquely recognised in its own binding pocket at incredible levels of specificity.

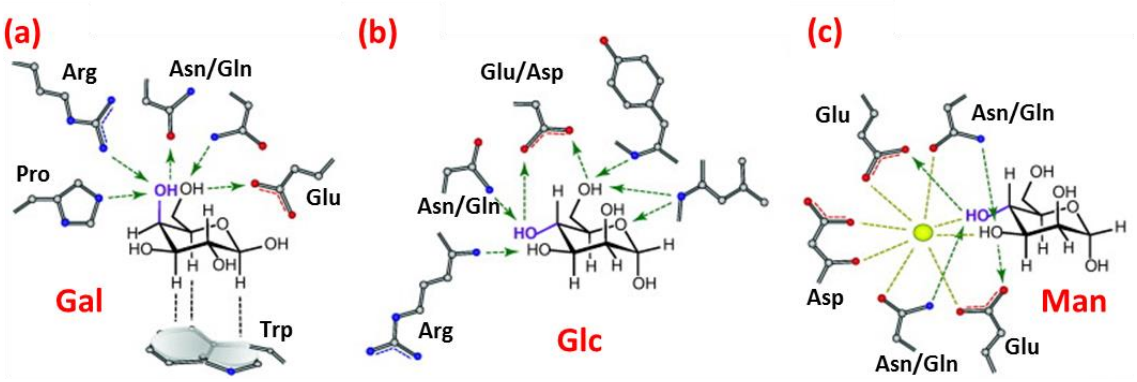

Figure 1.4. Epimer selection by three types of lectins. Arrows indicate H-bonding and point from donor to acceptor. **(a)** Galactose binding by human galectin-1 engages the axial 4-hydroxyl group in three cooperative donor/acceptor H-bonds along with the hydroxyl on C6 (OH-6), and a CH/π- stacking interaction of the H1-H3-H5 plane with the central Trp residue of the lectin site; **(b)** by contrast, the equatorial OH-4 in mannose is reliably sensed by topologically fixed bidentate H-bonding and a second H-bond, shown here for the glucose/mannose-specific leguminous lectin concanavalin A; **(c)** mannose-binding C-type serum lectin (rat collectin) involves a strategically presented Ca^{2+} ion (sphere) to probe for the presence of the di-equatorial 3- and 4-hydroxyl groups, along with two H-bonds. Figure adapted from 12.

At the same time, the slight electron-depletion resulting on the CH bond (of carbons carrying hydroxyl or carboxylic groups) enables van der Waals and stacking interactions. The most recurrent example is the CH/π-stacking taking place with the aromatic moieties of tryptophan, tyrosine and phenylalanine. This recognition pattern is particularly favourable for β-sugars (β-glucose, -galactose, -mannose, -glucuronic acid)

due to the presence of 3 axial CH (at position 1, 3 and 5), which enhances the CH contribution to the π-stacking (Figure 1.4a)[12,29].

The group of Woolfson browsed all the X-Ray structures of protein-carbohydrate complexes available until 2015, checking for the occurrence of polar, aliphatic and aromatic amino acids proximal to the carbohydrates (normalising to their occurrence in all protein sequences). They reported that aspartic acid and asparagine (Asp, Asn) occurs approximatively twice as often as expected by chance; aliphatic residues were generally disfavoured; tryptophan, tyrosine and histidine (Trp, Tyr, His) occurred 9, 4 and 3 times more frequently than as expected by chance, respectively[29].

Additionally, clusters of arginine are regularly found in sialic acid recognising binding sites, where they multi-directionally complex the carboxylic group on C2[18], as we will see in Chapter 3 and 5; the same is true for heparin-recognising binding pockets which are rich in lysine and arginine, as heparin is a polysaccharide constituted of highly sulphated disaccharides units (they are regarded as the most negatively charged class of glycans)[30].

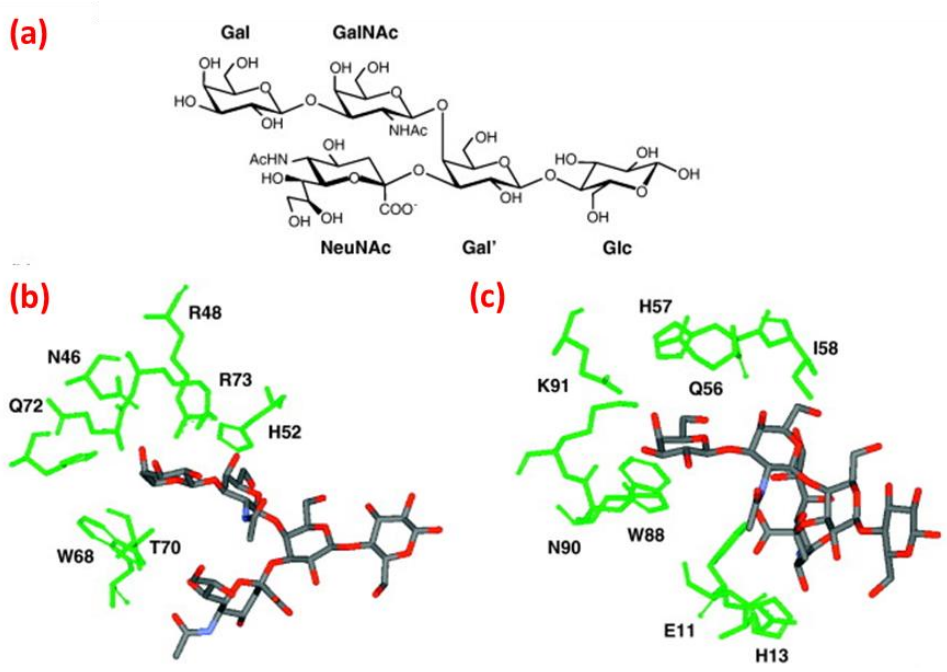

Figure 1.5. Differential conformer selection of the pentasaccharide of ganglioside GM1 by human galectin-1 and the cholera toxin B-subunit. The α2,3-sialylgalactose linkage at the branch site **(a)** can adopt three low-energy structures. Human galectin-1 **(b)** and the B-subunit of the cholera toxin **(c)** (PDB ID: 2CHB, 3CHB, respectively) bind different conformers ("keys") so that the same pentasaccharide is bioactive in two distinct shapes[31]. Figure from 12.

As we mentioned already, the conformational flexibility of the different glycosidic bonds also plays a fundamental role in protein-carbohydrate recognition. A remarkable example is found in the binding of the GM1 ganglioside to galectin-1 and to cholera toxin subunit B (CTB). GM1 (Figure 1.5a) is a branched pentasaccharide, whose non-reducing ends are a galactose and a sialic acid linked through a β-GalNAc(1-4)β-Gal unit.

GM1 free conformation has been thoroughly studied, and the pentasaccharide has been found to adopt three major conformations in solution[31]. Most interestingly, the two lectins (galectin-1 and CTB) recognise two different "shapes" of the three most populated conformations (Figure 1.5b,c). Coming back to our previous analogy, the conformational behaviour of GM1 in solution resembles that of a "flexible key", which can be recognised by different "locks", selectively.

Before concluding this discussion, it is also worth mentioning that the GM1/CTB complex has a 100-fold higher affinity than its GM1/galectin-1 counterpart (low nM *vs.* hundreds nM, respectively). This is to ascribe to the pentameric nature of CTB, which has a GM1 binding pocket on each monomer, for a total of five[32], in contrast to the homodimeric nature of galectin-1[33]. The origin of the affinity enhancement associated to multivalent binding depends on the physiognomy of the multivalent receptor. In the case of the GM1/CTB interaction, the mechanism goes through the face-to-face binding of the pentameric lectin, orienting its multiple subsites in a common direction, to a carbohydrate or glycoprotein with clustered epitopes (GM1s are highly concentrated on the epithelial cell surface)[34]. Other multivalent interactions rely on so-called "bind and jump" mechanism, when the binding sites are at opposite ends of the molecule and the receptor "moves" on the carbohydrate-rich surface[35]. However charming, this discussion falls outside the field of interest of this thesis, and in the next sub-section we will focus on the discussion of thermodynamic aspects which will be essential to analyse every kind of mechanisms of binding (protein-carbohydrate, protein-ligand, monovalent, multivalent, *etc.*).

1.2.2 The role of water and the enthalpy-entropy compensation effect

So far, we described glycans as amphiphilic units, which are at the same time able to participate in cooperative H-bond (both as donors and acceptors) and in CH/π hydrophobic stacking in their amphiphilic binding sites. When we focus on the

thermodynamics of the interactions, we must consider the solvent as the third component of the mechanism of binding: water is the medium in which the interactions take place, water molecules are often intrinsic to amphiphilic binding sites and especially, the free ligands in solutions are surrounded by several layers of water molecules.

In fact, the high density of hydroxyl groups on glycans poses interesting questions on the behaviour of these ligands when *free* in solution and surrounded by water molecules. Most interestingly, computational studies performed by the group of Vanderkooi gave an excellent answer to this question, showing how the possibility for intramolecular H-bond in monosaccharides strongly influences the structure of the water molecules constituting the first solvation layer[36]. For instance, an axial OH on C4 (OH-4) can hydrogen bond to the OH-6 forming a network which reaches the O5 (OH-4 → OH-6 → O5). Alternatively, it can interact with an equatorial OH-3 creating an OH-6 → OH-4 → OH-3 network. An axial OH-2 can engage with OH-4 and OH-6, influencing the formation of an OH-2 → OH-4 → OH-6 network. Whatever the geometry, the presence of intramolecular H-bond networks themselves is associated to *more nonpolar-like* hydration around their hydroxyl. As we already mentioned, equatorial OH-3 and OH-1 also contribute to the formation of a hydrophobic CH patch on the face of the sugar where the correspondent CH-3 and CH-1 are axial (*e.g.*, favouring the interactions with aromatic π clouds).

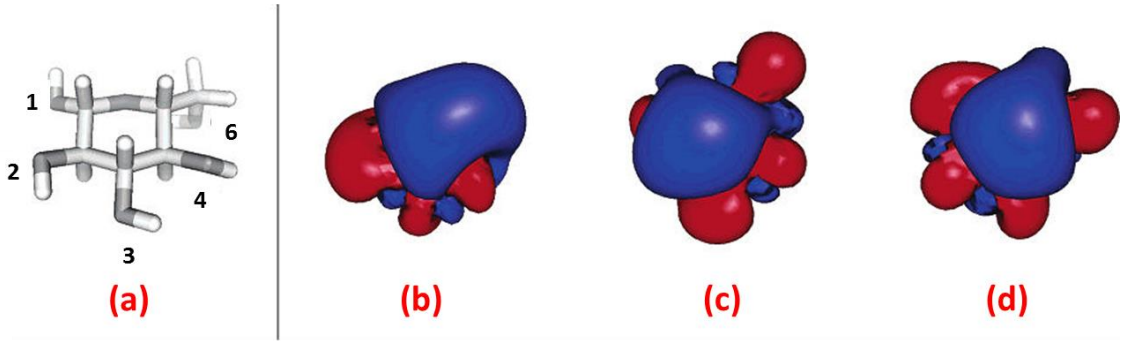

Figure 1.6. Positive (blue) and negative (red) electrostatic field iso-surface contours surrounding the hydrophobic patch of common biological sugars: β-glucose **(b)**, β-mannose **(c)**, and β-galactose **(d)**. The 3D structure of β-glucose **(a)** shows the CH groups present on the opposite face of these sugars. The figure is slightly offset from face-on view in order to show the vector orientation of these groups. Figure from 36.

The water structure surrounding this patch is relatively more polar-like than that observed when the OH-1 and OH-3 are in configurations that do not support patch formation. It is enlightening to notice that the monosaccharides common to biology are those containing both a hydrophobic patch and limited intramolecular H-bonding networks, namely, glucose, galactose, and mannose (as opposed as to allose and talose, for instance). The particular solvation properties deriving from these structural features makes these sugar units more prone to engage in molecular recognition events, arguably due to the different electrostatic potential surfaces observed on the two faces of the sugars (Figure 1.6)[36].

This brings us naturally to include into the discussion the hydrophobic effect, the tendency of nonpolar molecules (or part of molecules) to aggregate in aqueous media: a central aspect of biomolecular structure and recognition[37]. Some argue that the low occurrence of aliphatic residues in carbohydrate-recognising binding pockets is the proof that the hydrophobic effect does not play a fundamental role in protein-carbohydrate recognition[29], whereas some other may argue that already the CH/π interaction is a remarkable example of the hydrophobic effect[38]. On the other hand, the hydrophobic effect plays a key role in accommodating the methyl groups of fucose or that of the amidic group of sialic acid, *N*-acetyl-glucosamine and *N*-acetyl-galactosamine. The discussion is substantially controversial, as it is undeniable that the water molecules surrounding the apolar ligand and filling the active site of the protein are an important part of the mechanisms of binding.

The entropic contribution to binding is another focal point of this discussion. Commonly, binding implies selection of one single conformation, among the many conformations the free ligand can assume when it is free, which will fit in the (more or less tight) binding site[39]. Unavoidably, the loss of conformational freedom, plus the loss of independent freedom of rotation and translation in solution, equates to an entropic loss ($\Delta S < 0$) which usually has a negative impact on the spontaneity of the process at physiological temperature ($\Delta G = \Delta H - T\Delta S$). As a counterpart, the hydrogen bound water molecules which occupy the binding pocket of the protein are released upon binding, going back to a more "disordered" state in which they are free to tumble in the bulk solution. Clearly enough, based on what we just discussed in the previous paragraph, entropy and enthalpy go in opposite directions: the de-solvation of the free ligand and

of the apo binding pockets, which are enthalpically unfavourable, are entropically favourable; the loss of conformational freedom of the ligand, associated with the binding event, brings both: a gain in the enthalpic component (H-bond formation, hydrophobic effect, CH/π stacking), together with a loss of entropy of flexible ligands and proteins.

Many thermodynamic studies have been carried out in the attempt of rationalising: i) the unfavourable entropy of binding caused by conformational restrictions of the ligand upon binding[40]; ii) the small conformational changes throughout the protein upon ligand binding[41]; iii) and the reorganization of solvent molecules within the active site of a protein after ligand binding[42]. The combination of these evidences has resulted in the proposal of the so-called enthalpy-entropy compensation, but the topic remains controversial.

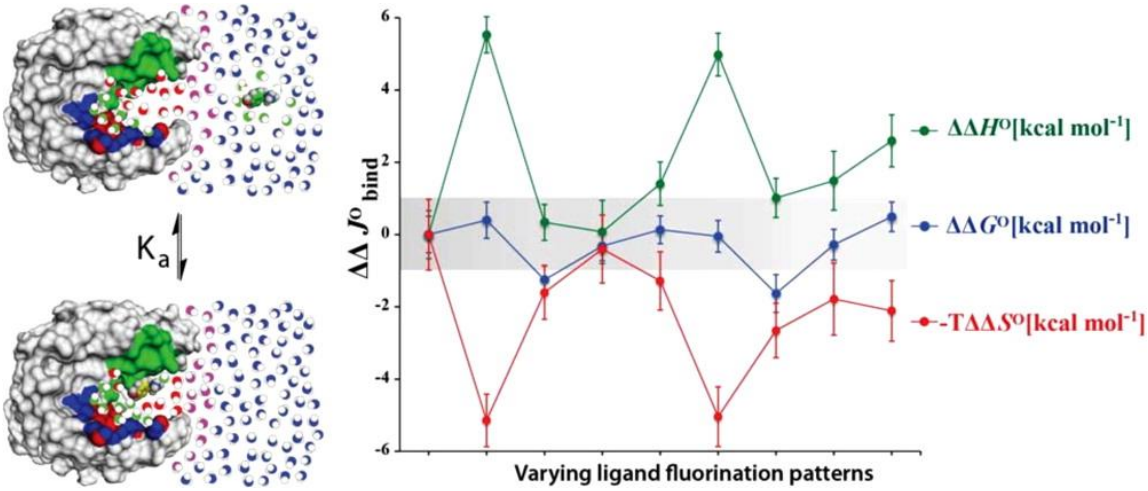

Figure 1.7. Enthalpy-entropy compensation in protein-ligand binding. Graphical abstract from 37. Diagram of the thermodynamic results for $\Delta\Delta J°_{bind}$, where J stands for Gibbs energy (G, blue), enthalpy (H, green), and entropy (S, red), respectively —obtained from ITC measurements at 298.15 K. The relative differences in the enthalpy and entropy of binding (*i.e.*, mutual H/S compensation) result in indistinguishable values of $\Delta G°_{bind}$ for 8 ligands with differential fluorinated patterns. The grey region demarcates the 95% confidence interval (*i.e.*, two standard deviations) of $\Delta\Delta G°_{bind}$ for the reference compound. The sketch of the association process is represented on the left panel.

The term was coined in 1930 and discussed ever since. One of the latest comprehensive studies on the subject, a work from the group of Whitesides, brilliantly exemplified the effect of different fluorination pattern on a library of compounds interacting with

human carbonic anhydrase[37]. They showed that changes in the structure of the ligand often lead to opposite and compensating changes in the enthalpy and entropy of binding but result in surprisingly small changes in the free energy of binding, as it appears self-explanatory in Figure 1.7.

The most cautious authors still argue that, in fact, the enthalpy-entropy compensation in water prevents to accurately dissect out measured values of ΔG into contribution from enthalpy and entropy. Whereas we insist on the relevance of dissecting ΔG into its ΔH and ΔS components (even by approximation) to discuss the mechanisms of binding, we agree that ΔG is the thermodynamic quantity most accurately accessible from measured binding constants in solution and the most relevant in calculations of equilibria. Therefore, in the next sub-section, we will present a simple model to describe protein-ligand interactions at the equilibrium and to treat them mathematically from a thermodynamics and kinetics point of view.

1.2.3 One-site model for bimolecular association in solution: thermodynamic equilibrium and kinetics of exchange

The framework that we are going to discuss does not only apply to protein-carbohydrate interactions but can describe any receptor-ligand system. The model is called "one-site model for bimolecular association in solution" (where no allosteric or cooperative effect is accounted for). Three distinct species are involved: a protein receptor (P), a ligand (L) and the complex of their association, the complex (PL).

$$P + L \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} PL$$

For the purposes of ligand-based NMR observation, the kinetics of the binding process is a determining parameter. In that context, we refer to it as the kinetics of exchange, as we observe the ligand *exchanging* between the free and the bound states. The unimolecular rate constant for the dissociation reaction (off-rate, k_{off}, in s^{-1}) is inversely proportional to the half-life time of the receptor ligand complex. The bimolecular rate constant for the direct reaction, k_{on}, measures the probability of productive encounters (*i.e.,* encounters which directly lead to the formation of the complex). For simple systems, one normally can consider k_{on} as invariable, under the main assumption that the association process of the interaction is controlled by molecular diffusion (in which

case, the best estimations give k_{on} ≈ 10^8 – 10^9 M^{-1} s^{-1}). However, if the interaction involves strong long-range forces (when electrostatic components are involved), k_{on} could be significantly larger; in contrast, it could be significantly lower if the interaction is coupled to large conformational changes.

A key parameter in any binding equilibrium is the affinity, which can be expressed in terms of the dissociation equilibrium constant:

$$K_D = \frac{[P][L]}{[PL]} = \frac{k_{off}}{k_{on}} \qquad \text{Equation 1.1}$$

where $[P]$, $[L]$ and $[PL]$ correspond to the concentration of P , L (free state) and PL (bound state) at the thermodynamic equilibrium. K_D is also given by the ratio of k_{off} over k_{on}. For systems in which we can assume k_{on} diffusion-limited (so in absence of long range interactions or of large conformational rearrangements), the K_D can be directly related to k_{off}, *i.e.*, measuring the affinity gives us direct information on the turnover of binding (or residence time in the binding pocket). For small K_D values, corresponding to strong binding, k_{off} will be small, indicating slow turnover of the ligand in the binding site, *i.e.*, a long residence time. *Vice versa*, large K_D values, corresponding to weak binding, will imply large $k_{off,}$ with a fast turnover and a short residence time.

Changing perspective, if we consider that the fraction of bound receptor is given by

$$f_B^P = \frac{[PL]}{[P]_T} = \frac{[PL]}{[P]+[PL]} \qquad \text{Equation 1.2}$$

we can then combine this equation with the definition of K_D and express the fraction of bound receptor as a function of the concentration of free ligand:

$$f_B^P = \frac{[L]}{K_D+[L]} \qquad \text{Equation 1.3}$$

This corresponds to the equation of a Langmuir isotherm for the saturation of a receptor site of interaction with the ligand (Figure 1.8).

For low ligand concentration, $[L] << K_D$, f_B^P is linearly proportional to $[L]$; on the other hand, for very large ligand concentration, $[L] >> K_D$, the receptor is fully saturated with ligand molecules and $f_B^P = 1$. If we consider the particular situation in which the

receptor is half-saturated, *i.e.* when $f_B^P = 0.5$, the Langmuir isotherm equation predicts that the free ligand concentration is $[L] = K_D$, which gives us a more practical definition of K_D, as "*the concentration of ligand in the free state that leads to half saturation of the receptor*". At this point, it is important to highlight that, in the Langmuir isotherm, $[L]$ represents the concentration of free ligand, and not the total added concentration, $[L]_T$.

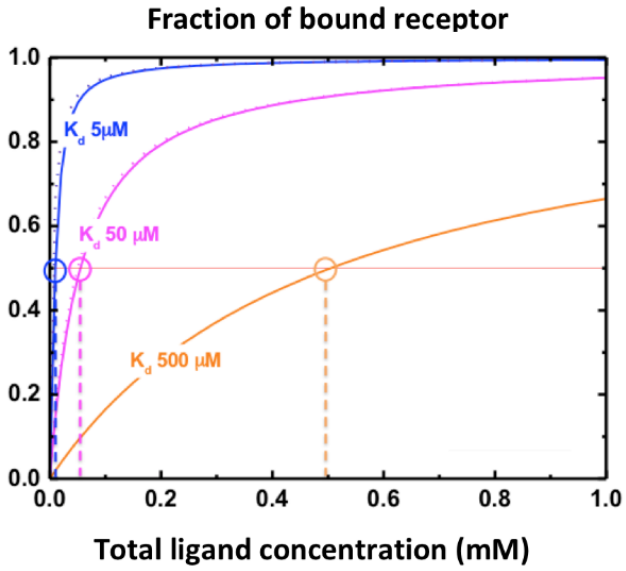

Figure 1.8. Langmuir isotherm showing the fraction of bound receptor f_B^P as a function of the total ligand concentration $[L]_T$ for three different affinities (in all cases $[P]_T$=10μM).

1.3 Theory of Solution State Nuclear Magnetic Resonance[43]

The first Nuclear Magnetic Resonance signal was detected in bulk materials by Bloch and Purcell in 1945[44,45]. As physicists, they thought to have found a precise method to measure magnetic moments of different nuclei. In the few following years, it became clear that the frequency of resonance of a given nucleus was affected by the chemical environment on the nucleus itself. The discovery of chemical shift, followed by the discovery of spin-spin coupling, disclosed the huge potential of the technique for chemical structural elucidation[46]. On these premises, the exponentially fast theoretical and technological developments of Nuclear Magnetic Resonance Spectroscopy started[47].

In 1953, the theory of the nuclear Overhauser effect (NOE) was published[48], and in the same year the theory for exchange effects in NMR spectra was postulated[49]. In a few years, while the first commercial NMR magnets started to appears, the theory was

translated in routine practice, allowing extended structural and conformational studies[50]. As the first superconductive magnets were engineered and higher and higher magnetic fields could be reached and more sophisticated pulse sequences were developed (including multi-dimensional NMR), NMR started to be applied in the field of biology and in 1983 the first small protein was assigned[51]. In parallel, great progresses were made in the field of solid state NMR, with the discovery of the magic angle spin technology, and in the field of combined imaging and spectroscopy (bringing to MRI).

Recently, the implementation of novel processing approaches, non-uniform sampling and non-Fourier signal processing methods in multidimensional NMR, enabled reducing experimental times considerably and allowed acquisition of up to 7D experiments[52], while the first 1.2 GHz magnet has been purchased by the University of Lille in France. NMR is thriving and without any doubt it can be considered as the most important and versatile of all the spectroscopies.

1.3.1 Basic concepts[53,54,55]

1.3.1.1 Spin number, spin angular momentum and magnetic moment

Atoms are composed of nuclei and electrons. While many spectroscopy techniques rely on the observation of electron energy transitions, NMR focuses on nuclear spin energy transitions. Nuclei possess several properties through which we can define them: mass, electric charge, magnetism and spin. To discriminate between NMR "active" or "silent" nuclei, the nuclear spin quantum number (I) is the most important parameter. Indeed, nuclei with $I = 0$ do not exhibit NMR properties, while for all the other nuclei we can define a spin angular momentum P, also called the nuclear spin, and defined as:

$$P = \sqrt{I(I+1)}\,\hbar \qquad \text{Equation 1.4}$$

where $\hbar$ is the reduced Plank constant ($\hbar = h/2\pi$) and $I = \frac{1}{2}, 1, \frac{3}{2}, 2,$ *etc*. Classical angular momentum is associated to a rotating object. On the contrary, P is an intrinsic property of the quantum particles. As in classical physics a charge in motion (*e.g.* associated to an angular momentum) generates a magnetic field, the intrinsic spin angular momentum leads to the generation of a magnetic moment, μ, which is proportional to P:

$$\mu = \gamma P \qquad \text{Equation 1.5}$$

where γ is the so-called gyromagnetic ratio, which shows a defining value for every different nuclide. μ and P are vectors, with a defined magnitude and orientation. We can think of the individual magnetic moment of each spin in a sample as randomly oriented in normal conditions. However, when the sample is placed in a static magnetic field B_0, the magnetic moments will show a tendency to align with the direction of the field, as little magnetic bars would do. In quantum terms, the observation of the system will lead to the detection of only a discrete number of different orientations of the spins, which is equal to $2I + 1$.

1.3.1.2 The differences in energetic state populations and the Larmor frequency

For nuclei with $I = \frac{1}{2}$ (such as ^{1}H and ^{13}C, the most commonly detected nuclei), there will be two possible orientations the spins can undertake. By convention, B_0 is always applied in the *z* direction, therefore, nuclei with $I = \frac{1}{2}$ will either align with *z* or with *–z*. We call these two energy states the α and the β state, respectively. Usually, the population of the α state (N_α) is slightly larger than the population in the β state (N_β) and the ratio between the two it increases exponentially with ΔE (therefore it depends on γ and B_0):

$$\frac{N_\alpha}{N_\beta} = e^{\frac{\Delta E}{k_B T}} \qquad \text{Equation 1.6}$$

where k_B is the Boltzmann constant and the difference in energy level between α and β, ΔE, is given by:

$$\Delta E = \hbar \gamma B_0 \qquad \text{Equation 1.7}$$

The quantum mechanics system we are trying to describe can become complex, but a very useful way to deal with it is to depict it in classical mechanics and vectorial terms, which is accurate up to a certain level. In particular, we can think of B_0 imposing a torque on the magnetic moments μ, which results in the precession of every one of the nuclear spins about the direction of the static field. The rate of the precession is expressed (in Hz) as follows:

$$\nu_0 = \frac{-\gamma B_0}{2\pi} \qquad \text{Equation 1.8}$$

In NMR, this is called the Larmor frequency of the nucleus and it also corresponds to the frequency of the electromagnetic radiation required to the nuclei to change their spin state (from α to β or *vice versa*), by acquisition of a quantum of energy ($\Delta E = h\nu$).

1.3.1.3 Bulk magnetisation and the perpendicular rf field B_1 in the rotating frame

Now, we can imagine summing up all the single μ components (precessing at the Larmor frequency and aligned with or against B_0): the result will be a so-called "bulk" magnetisation vector M_z aligned with B_0 (Figure 1.9). The magnitude of M_z will be given by the *z* projection of the sum of the magnetic moments μ. Formally, we would then write that:

$$M_z \propto N_\alpha - N_\beta \qquad \text{Equation 1.9}$$

This implies the net magnetisation on the transverse (*x*-*y*) plane to be null. As mentioned, to induce an energy transition between the α and the β states, the Larmor frequency of the nucleus of interest must be matched by the electromagnetic wave employed for excitation of the sample.

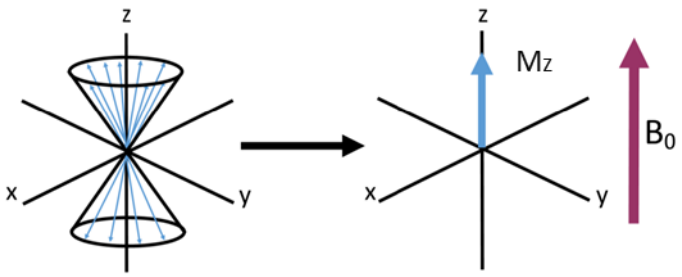

Figure 1.9. Bulk magnetisation as represented in the vector model. The single spin magnetic moments rotating at the Larmor frequency in their α and β states (pointing above and below the *x*-*y* plane, respectively) sum up to give the bulk magnetisation vector M_z, whose direction is +*z* and whose intensity is the projection of the sum of the μs on *z*. Figure adapted from 54.

Practically, as the frequencies typically associated with nuclear spin transitions fall in the range of radio waves (tens to hundreds of MHz), for typical magnets used in NMR spectroscopy, the excitation is done applying a radiofrequency *pulse* oscillating at around the Larmor frequency (passing current via a coil surrounding the sample). This produces a magnetic field, B_1, perpendicular to B_0 (the direction of B_1 can be modulated by the coil properties). To help visualising the effects of the oscillating rf pulses on the bulk magnetisation rotating along *z* in the presence of B_0, the formalism of the rotating frame of reference was introduced (Figure 1.10).

In the laboratory frame the *x*, *y*, *z* coordinates are viewed as static, while the rotating frame of reference uses coordinates *x'*, *y'*, *z'* which are rotating with the frequency of the B_1 field of the radiofrequency pulse . If the magnetic field introduced by the rf pulse oscillates in the *x* axis (*x*-pulse), it can formally be decomposed into two counter-rotating components oscillating in the transverse plane with frequencies $\pm \nu_0$ (Figure 1.10) (where ν_0 is expected to be the frequency for NMR condition to be satisfied). In the rotating frame of reference, one of the rf field vectors is then static and the second rotates with $2\nu_0$ in opposite (-) direction. Being this now far from the Larmor frequency, it can be ignored. Therefore, in the rotating frame, time dependency of the rf field is removed and both M_z and B_1 are static and perpendicular to each other.

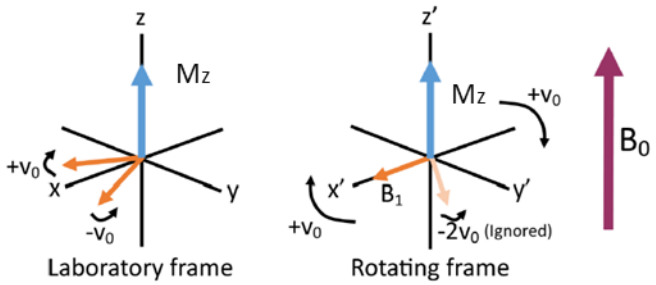

Figure 1.10. Representation of laboratory and rotating frame of reference. Figure adapted from 54.

1.3.1.4 Chemical shift and the fundamental pulse-acquire NMR experiment

We already argued that the resonance frequency of a given nucleus depends on its gyromagnetic ratio (γ) and applied external field B_0, and we mentioned that the local environment of the nuclei in the molecule (namely surrounding electrons of *any* neighbouring nucleus) has a large effect too. Electrons circulating in their orbitals can be seen as currents passing through wires and producing small (local) magnetic fields (B') in the opposite direction to B_0. Hence, the effective (local) magnetic field (B) felt by the nucleus is slightly smaller than the applied external magnetic field (B_0) and is given by the equation:

$$B = B_0 - B' = B_0(1 - \rho) \qquad \text{Equation 1.10}$$

where ρ is the shielding constant.

The electron density surrounding the atoms in any molecule is typically rather complex and depends on several factors including formation of hydrogen bonds, presence of electronegative nuclei in close proximity, unpaired electrons or through-space interactions within molecular clusters. Experimentally, the chemical shift (δ) is measured as a frequency difference of the analysed nuclei (say resonating at ν_i) with respect to one reference frequency. If we imagine the reference frequency to be on-resonance with the rotating frame (ν_0), the other sets of nuclei can either move slower or faster, and consequently will have a smaller or larger δ. The chemical shift of each set of nuclei is defined in the ppm (part per million) scale as:

$$\delta_i = 10^6 \frac{\nu_i - \nu_0}{\nu_0} \qquad \text{Equation 1.11}$$

and it is independent from the strength of the field B_0.

We have now all the tools to describe the fundamental pulse-acquire experiment, consisting in turning on an rf irradiation of given amplitude for a given time. As already B_0 did on μ (inducing precession), the rf field imposes a torque on M_z rotating it perpendicularly to the direction of the B_1. This drives the bulk magnetisation vector away from its equilibrium position (from the *z*-axis towards the *x*-*y* plane, according to the "motor rule"). The 90° pulse, bringing the magnetisation on to the *x*-*y* plane, equalises the spin populations of the α and β states, that is, it makes M_z null. A 180° pulse, instead, which will bring M_z to *–z,* does invert the populations. All these elements (and many more) are useful to manipulate the magnetisation in pulse sequences way more complicated than the pulse-acquire. For that purpose, it is also important to know that if the rf field is applied for a time t_p, the flip or nutation angle of the pulse (θ) through which the magnetisation is rotated is found as:

$$\theta = \frac{\gamma}{2\pi} B_1 t_p \text{ degrees} \qquad \text{Equation 1.12}$$

Far from its equilibrium state by any θ angle, the bulk magnetisation will tend to recover its *+z* position by precessing around *z*. As we have seen, different sets of nuclei will precess at different frequencies, each of them inducing a weak oscillating voltage in the coil (now acting as a receiver coil rather than a transmitter coil). This small signal will be amplified and processed (Fourier transformed from the detected time domain to the

frequency domain) and the result will be a frequency spectrum containing as many "peaks" as many are the sets of nuclei experimenting different electronic environment (*i.e.* having different chemical shifts).

We have explained very briefly the use of rf pulses to perturb the equilibrium between the energy states of the nuclear spins (generating sets of signals on the receiver coil, and allowing to detect an NMR spectrum of the sample). Of deepest interest for the techniques used and developed in the present thesis, though, are the complicated processes of re-establishing the equilibrium state *after perturbation*. These processes are called relaxation, and to them we dedicate the following sub-section.

1.3.2 Relaxation in one- and two-spin systems[54,55]

Understanding relaxation is relevant from many points of view. In the first place, we need to know how much time the nuclei in our sample take to come back to equilibrium after a given pulse, to allow full recovery of the bulk magnetisation on the *z* axis, before pulsing again, *i.e.*, before acquiring another scan (acquiring and accumulating more than one scan is the norm in NMR, to increase the *S/N* ratio). Nuclear relaxation is stimulated by a suitable fluctuating magnetic field to induce the necessary spin transitions that eventually will lead to re-establishing the equilibrium populations. These fluctuations act locally (as opposed as to the rf pulse, which is coherently applied to the *whole* sample) and they are modulated by the random molecular motions existing in the surrounding environment (mainly diffusional rotations). Therefore, understanding relaxation can also unveil information on the motional regimes in the sample. Among the mechanisms which can induce relaxation, the most relevant for us is the process of magnetic dipole-dipole interactions, which, in particular, is responsible for inducing the so-called "cross relaxation" process, which gives rise to one of the most relevant observables in NMR spectroscopy, the Nuclear Overhauser Effect (NOE). NOE allows to probe intermolecular distances and most (if not all) of the NMR techniques investigating intermolecular interactions are based on it.

1.3.2.1 Spectral density function and correlation time

To understand better the role of random fluctuations in relaxation processes and NOE, it is fundamental to introduce the concept of spectral density function and the related

notion of correlation time. Random fluctuations in the sample generate molecular motions with a distribution of frequencies that can be expressed by the so-called correlation function. The correlation function with respect to time (τ) for an isotropic rotational diffusion of a hard sphere is:

$$G(\tau) = \overline{B_{loc}^2} e^{-\frac{|\tau|}{\tau_c}} \qquad \text{Equation 1.13}$$

where B_{loc} is the local field, $|\tau|$ is the modulus of the time and τ_c is the *correlation time*, dictating the rate of decay of the correlation function with time. By definition, τ_c is the time a molecule takes to rotate through one radian and it is a key NMR parameter because it enables to quantify molecular tumbling motions. It is affected by molecular size (or better by radial volume), solvent viscosity and temperature. Generally, it is in the order of picoseconds for small molecules in non-viscous solvents at room temperature (in the fast motion or extreme narrowing limit), and in the order of nanoseconds for macromolecules in aqueous solutions (in the slow motion or spin diffusion limit).

The exponential part of the correlation function is independent from the source of the local fields; it is therefore possible to simplify the equation by defining a reduced correlation function, $g(\tau)$, which is independent of the size of the local fields:

$$g(\tau) = e^{-\frac{|\tau|}{\tau_c}} \qquad \text{Equation 1.14}$$

The Fourier transform (FT) of the correlation function is called the *spectral density function*, $G(\tau) \overset{FT}{\rightarrow} J(\omega)$, and translates the distribution of molecular motions at the angular frequency expressed in radian ($\omega = 2\pi\nu$):

$$J(\omega) = \overline{B_{loc}^2} \frac{2\tau_c}{1+\omega^2\tau_c^2} \qquad \text{Equation 1.15}$$

The reduced spectra density, $j(\omega)$, does not depend on the size of local fields and it is expressed as:

$$j(\omega) = \frac{2\tau_c}{1+\omega^2\tau_c^2} \qquad \text{Equation 1.16}$$

In both forms, what is crucial to highlight is that the amount of motion at the Larmor frequency is simply found by evaluating $J(\omega)$ at $\omega = \omega_0$, as shown in Figure 1.11. The expressions also show that the dependence of the spectral density for a hard sphere

with frequency follows a Lorentzian function and highlights its dependence from the correlation time.

In other words, as fluctuations lead to generation of pulsing local magnetic fields in the surroundings of the spins, the spectral density function represents the probability of finding a fluctuating magnetic component at any given frequency because of the motion. Therefore, the area under the three curves is equal (and does not dependent from τ_c).

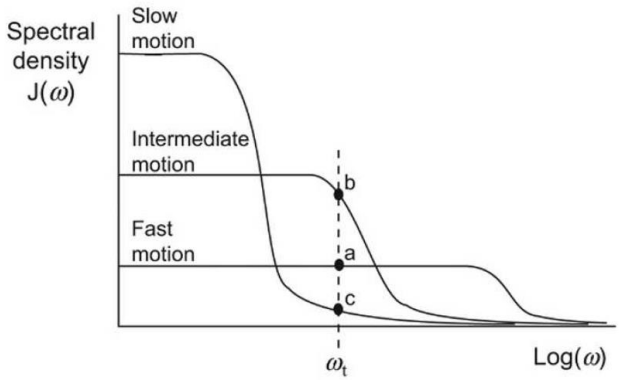

Figure 1.11. Spectral density function $J(\omega)$ as a function of the frequency of molecular motions for different correlation times (from small τ_c for fast tumbling molecules to large τ_c for slow tumbling molecules). ω_τ represents the frequency of spin transition. Figure from 54.

Fast tumbling molecules (small molecules, non-viscous solvents or elevated temperatures) can probe a wide range of motional frequencies and therefore show a wide spectral density distribution varying with ω (fast motion in Figure 1.11). In contrast with the above, slowly tumbling molecules (large molecules, or medium-sized molecules in viscous solvents or low temperatures) can only experience lower frequencies of motion, having a narrower but more intense distribution of frequencies (slow motion in Figure 1.11). Hence, the spectral density function is sharply peaked around $\omega = 0$.

1.3.2.2 Longitudinal relaxation (T_1)

Hence, we can go back to the pathways of the magnetisation in our sample. We have already presented the rf pulse as an event perturbing the spin populations at equilibrium of the α and β states (N_α and N_β). In these terms, for the equilibrium to be restored *transitions* must take place, bringing the nuclei back to their original energy state (α to β, or β to α). We can start from the assumption that the rate of this transition process

is proportional to N_α, and it is a first order process with a proportionality constant W (Figure 1.12).

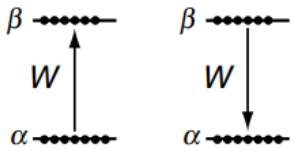

Figure 1.12. A state transition from α to β decreases the α population, a state transition from β to α increased the α population.

The rate of change of N_α and N_β can be written as:

$$\frac{dN_\alpha}{dt} = -W(N_\alpha - N_\alpha^0) + W(N_\beta - N_\beta^0), \quad \frac{dN_\beta}{dt} = +W(N_\alpha - N_\alpha^0) - W(N_\beta - N_\beta^0)$$

Equation 1.17

where $N_\alpha - N_\alpha^0$ and $N_\beta - N_\beta^0$ indicate the deviation from the population in the α and β states at the equilibrium (N^0). Based on Equation 1.9, $M_z \propto N_\alpha - N_\beta$, we can write the dependency of M_z with the time as:

$$\frac{dM_z}{dt} = \frac{d(N_\alpha - N_\beta)}{dt} = \frac{dN_\alpha}{dt} - \frac{dN_\beta}{dt}$$

Equation 1.18

Combining Equation 1.17 and Equation 1.18, we find that:

$$\frac{dM_z}{dt} = \frac{-R_z\,(M_z - M_z^0)}{dt}$$

Equation 1.19

where $R_z = 2W$ and $M_z^0 = N_\alpha^0 - N_\beta^0$. When $M_z = M_z^0$, the system is at the equilibrium and nothing happens, but when M_z deviates from the equilibrium, the rate of change is proportional to the deviation. As for any other first order kinetics, solving the differential equation above shows that *z*-magnetization returns from $M_z(0)$ to M_z^0 following an exponential law. The time constant of the exponential is $1/R_z$ (also called $1/R_1$), and this is often referred to as T_1, the longitudinal or spin-lattice relaxation time. It is important to specify that T_1 is the "time constant" of the kinetics of longitudinal relaxation and not the effective time required for recovery. The term "spin-lattice" used to describe longitudinal relaxation refers to the fact that the energy is lost by the spins and transferred in the form of heat to the surroundings (the rigid "lattice" when we talk about crystalline solids, but the term has become of widespread use for any kind of

sample). The thermal energy involved is too small as to affect the temperature of the sample.

As we mentioned already, the transfer of energy from the spin to the surrounding is not spontaneous, but it is mediated by the chaotic tumbling of molecules (reason for which, in solid samples, longitudinal relaxation can take up to minutes). Longitudinal relaxation only occurs when these molecular motions create a time-dependent magnetic field fluctuating around the Larmor frequency. Since T_1 depends on the probability of components oscillating at the appropriate frequency, the curves in Figure 1.11 predicts how the longitudinal relaxation rate vary with correlation time. In Figure 1.11, ω_τ represents the frequency of spin transition. For a fast tumbling molecule, the spectral density would be most likely low (as many rotational frequencies are possible), and the relaxation will be slow (point "a" in Figure 1.11). Relaxation can fasten if the distribution changes towards a more intermediate motional regime, where the spectral density at ω_τ would be higher (point "b" in Figure 1.11). The relaxation is again very slow for slow motion when the tumbling rate falls below ω_τ. The maximum relaxation rate occurs when $\omega_0 \tau_0 = 1$, which corresponds to $J(\omega_0) = \omega_0^{-1}$.

1.3.2.3 Transverse relaxation (T_2)

At the same time as the longitudinal z-magnetisation returns to its equilibrium state through longitudinal relaxation, a second phenomenon is occurring: the decay of the transverse magnetisation to zero. This is brought about by mutual swapping of energy between spins, meaning simultaneous transition of one spin from α to β, while another spin transits from β to α, what we call a flip-flop process. This process does not contribute to restoration of the bulk magnetisation to the *z* plane, as the overall population does not change, but it contributes to the loss of net magnetisation in the transverse plane. Just at time zero after the 90° pulse, the bulk magnetisation in the *x,y*-plane results from a coherent superposition of magnetic moments. The flip-flop process, brought about by the slightly different magnetic field that each spin experiences tumbling around, destroys that *coherence* of the bulk magnetisation, bringing it to zero (precessing spins are dephased). This process occurs until equilibrium (*x,y* magnetisation equals zero) is established with a rate characterised by the transverse relaxation time constant T_2. Dephasing of spins is due then to magnetic field inhomogeneities, which are given by two factors: either intrinsic non-uniformity of the sample (un-optimised

shimming, turbidity of the sample, imperfection of the glass, *etc.*), or local magnetic field inhomogeneities arising from intra- and intermolecular interactions in the sample. The relaxation time constant for these two sources combined is designated T_2*:

$$\frac{1}{T_2^*} = \frac{1}{T_2} + \frac{1}{T_2(\Delta B_0)} \qquad \text{Equation 1.20}$$

where T_2 refers to the contribution from local relaxation processes and T_2 (ΔB_0) refers to that from field inhomogeneity.

As when the longitudinal magnetisation is restored there is no transverse magnetisation to be destroyed on the *x*-*y* plane, it is clear that T_2 can never be longer than T_1 (in fact, longitudinal relaxation always contributes to dephasing of the spins in the transverse plane, whereas the opposite is not true). While T_1 has a minimum for intermediate motion and then increases again for slow tumbling motion, T_2 decreases steadily as the correlation time increases, as shown in Figure 1.13. Here, we do not enter the details of the effect of T_2 on the width of the peaks (smaller T_2 are associated with broader signals), but in Sub-section 1.3.5, we will see more in detail how this effect weights differently on protein-based and ligand-based methods for the investigation of intermolecular interactions.

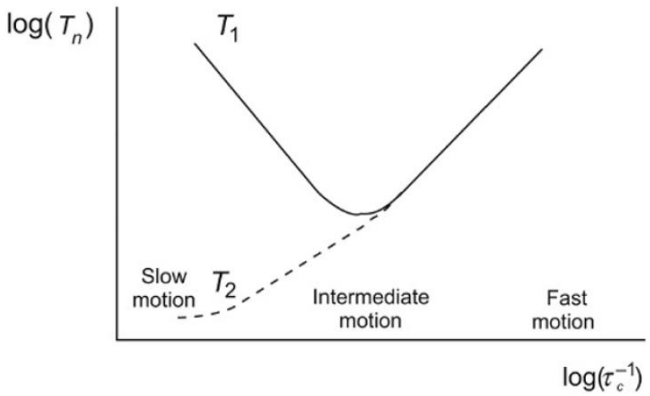

Figure 1.13. Motional dependence of longitudinal (T_1) and transverse (T_2) relaxation times. Figure from 54.

1.3.2.4 Longitudinal cross relaxation

In the same way as we treated the rate of changes in populations N_α and N_β to derive the rate of longitudinal relaxation R_z of a single spin returning to equilibrium after an rf pulse, we can derive relaxation rates of more complex processes. For instance, a sample containing two sets of spins, *I* and *S*. They will have between them four energy levels,

which can be labelled according to the spin states of the two spins, as shown in Figure 1.14.

In such systems, there are six possible pathways for relaxation, when we consider all the 4 energy levels. W_1^I and W_1^S are the four allowed spin transitions that we already encountered, and which will be denoted as "single quantum transitions". Transitions between the $\alpha\alpha$ and the $\beta\beta$ states are denoted as W_2^{IS} and are "double quantum transitions" and W_0^{IS} is the "zero quantum transition" going from $\alpha\beta$ to $\beta\alpha$, simultaneously. Double and zero quantum transitions are forbidden for nuclei with spin ½ (in spectroscopy, the *selections rules* allow only transition with $\Delta I = I(initial\ state) - I(final\ state) = \pm 1$).

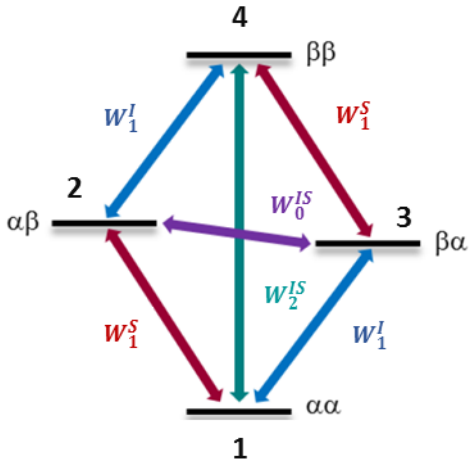

Figure 1.14. The six possible transitions in a two-spin system under the effect of an external magnetic field. Figure adapted from 55.

As we have done before, we can write the rate equation for the flow of population for each level (which we now call N_1 to N_4, rather than $N_{\alpha,\beta}$). For level 1, for example, we would write:

$$\frac{dN_1}{dt} = -W_1^S N_1 - W_1^I N_1 - W_2^{IS} N_1 + W_1^S N_2 + W_1^I N_3 + W_2^{IS} N_4 \quad \text{Equation 1.22}$$

The same equation can be written for the other three energy levels. In the two-spin system in analysis, solving the differential equations for the population of each energy level as we have done for the longitudinal relaxation of a single spin system in Equation 1.17 does not give us the required answers. Instead, the use of the product operator formalism is required. Although we have not included this approach in this dissertation,

we can define the two operators I_z and S_z, which plainly represent the magnetisations of spins *I* and *S* on the *z*-axis, respectively, which is what we want to follow from a practical point of view. The differential equations solving dI_z/dt and dS_z/dt are called the Solomon equations. To simplify them, we can write the two Solomon equations of main interest for us in the following form:

$$\frac{d(I_z - I_z^0)}{dt} = -R_{zI}(I_z - I_z^0) - \sigma_{IS}\,(S_z - S_z^0) \qquad \text{Equation 1.23}$$

$$\frac{d(S_z - S_z^0)}{dt} = -R_{zS}(S_z - S_z^0) - \sigma_{IS}\,(I_z - I_z^0) \qquad \text{Equation 1.24}$$

The combined rate constants are defined as:

$$R_{zI} = W_0^{IS} + 2W_1^I + W_2^{IS} \qquad \text{Equation 1.25}$$

$$R_{zS} = W_0^{IS} + 2W_1^S + W_2^{IS} \qquad \text{Equation 1.26}$$

$$\sigma_{IS} = W_2^{IS} - W_0^{IS} \qquad \text{Equation 1.27}$$

where R_{zI} and R_{zS} are the auto-relaxation (longitudinal relaxation) rate constants, and σ_{IS} is the rate at which the magnetisation is transferred from spin to spin by relaxation processes and is termed longitudinal cross relaxation rate constant. The main results springing from the Solomon equations is the definition of the concept of cross relaxation itself, showing that in a two-spin system the rate of recovery of *I* and *S* depends on each other (*if* the cross relaxation is not null).

The corresponding transition probabilities are:

$$W_1^S = \frac{3}{40} b^2 j(\omega_{0S}) \qquad \text{Equation 1.28}$$

$$W_1^I = \frac{3}{40} b^2 j(\omega_{0I}) \qquad \text{Equation 1.29}$$

$$W_0^{IS} = \frac{1}{20} b^2 j(\omega_{0I} - \omega_{0S}) = \frac{1}{20} b^2 j(0) \qquad \text{Equation 1.30}$$

$$W_2^{IS} = \frac{3}{10} b^2 j(\omega_{0I} + \omega_{0S}) = \frac{3}{10} b^2 j(2\omega_{0I}) \qquad \text{Equation 1.31}$$

Where b is the dipolar constant:

$$b = \frac{\mu_0\, \gamma_I\, \gamma_S\, \hbar}{4\pi r^3} \qquad \text{Equation 1.32}$$

with γ being the gyromagnetic ratio, μ_0 is the permeability of vacuum and r is the inter-nuclear distance. We assumed that the spins of interest are of the same type (*e.g.*, ^{1}H), with comparable Larmor frequencies. Hence, homonuclear cross relaxation in Equation 1.27 can be re-written as:

$$\sigma_{IS} = W_2^{IS} - W_0^{IS} = \frac{3}{10} b^2 j(2\omega_{0I}) - \frac{1}{20} b^2 j(0) \qquad \text{Equation 1.33}$$

Since double and zero quantum transitions are not allowed by the conventional selection rules, these are not observable, but they can strongly influence the observable NMR intensities associated to the transitions of the spins involved in the process of cross relaxation. In the following sub-section, we will analyse the importance of cross relaxation in two-spin system with different motional regimes and understand its relationship with the Nuclear Overhauser Effect.

1.3.3 Nuclear Overhauser Effect (NOE): from Steady state NOE to Transient NOE[54,55,56]

The nuclear Overhauser effect (NOE) itself is a manifestation of cross relaxation, stimulated by dipolar coupling. If we imagine a little magnetic field associated to each nucleus in the sample, it is easy to imagine that the nuclei will affect each other, each generating its own field and sensing the others. The prerogative for cross relaxation to take place (so for σ_{IS} to be non-zero), is that the set of nuclei *I* and *S* are close to each other, as from the previous equations we have seen that σ_{IS} is inversely proportional to the sixth power of the distance between the nuclei. The NOE's dependence on the distance between sets of nuclei is what makes this technique essential in structural studies, as we will show.

1.3.3.1 NOE enhancement in the slow and fast motion regime

In the two-spin homo-nuclear system which we are considering, we can envisage the situation in which we apply a continuous selective rf pulse on a frequency corresponding to the single transition of one of the nuclei (say *S*), saturating the populations across energy levels 1-2 and 3-4 in Figure 1.14. In this new condition, the allowed relaxation pathways are W_2^{IS} and W_0^{IS}, as the single quantum transitions of spin *S* are saturated (Figure 1.15). W_2^{IS} and W_0^{IS} are not single quantum transitions, so that they cannot be detected, but their effects are observable on the signals of spin *I* in a ^{1}H 1D spectrum acquired under the selective saturation of spin *S*.

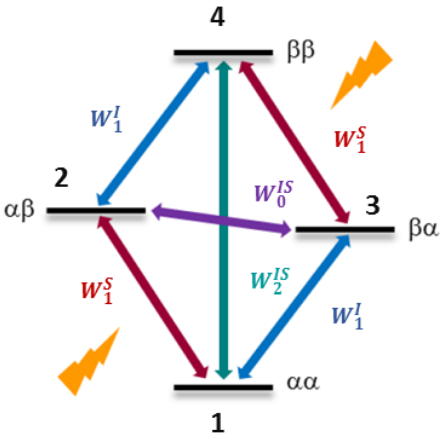

Figure 1.15. Energy levels for a two-spin system upon saturation of spin *S*. Figure adapted from 55.

This results from a change in the spin *I* magnetisation that will manifest as a modification of the NMR signal intensity of *I*. This modification resulting from selective saturation of the spin S transitions is termed NOE enhancement (Figure 1.16).

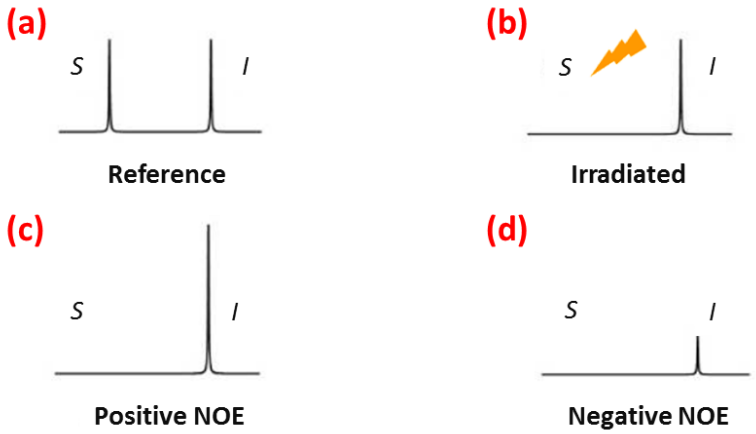

Figure 1.16. Examples of the effects of cross relaxation on 1D ^{1}H NMR spectra of spins *I* and *S* **(a)** upon selective saturation of spin *S*. When spin *S* is irradiated **(b)**, the *S* populations are equalised, and we do not observe any signal (in the irradiated spectrum the situation after instantaneous saturation is depicted). If cross relaxation between *S* and *I* is taking place, the peak intensity of spin *I* is affected. When relaxation takes place via W_2 processes, a positive enhancement is observed **(c)**; when relaxation takes place via W_0 processes, a negative enhancement is observed **(d)**. Figure adapted from [54].

When spin *S* is irradiated, the *S* populations are equalised, and we do not observe any signal for that spin. At the same time, the peak intensity of spin *I* is affected due to cross relaxation with spin *S*. When cross relaxation takes place preferentially via W_2 processes,

a positive enhancement is observed; in contrast, if the dominant mechanism for cross relaxation involves W_0 processes, a negative enhancement is observed (Figure 1.16c,d).

The NOE enhancement is indicated as η, and its sign and magnitude are proportional to the cross relaxation rate σ_{IS}:

$$\eta \propto \sigma_{IS} = W_2^{IS} - W_0^{IS} \qquad \text{Equation 1.34}$$

The magnitude of the NOE on spin *I* varies according to how the NOE experiment is performed (and it can depend on other variables), but the sign of the enhancement will be dictated by which process is more favourable between W_2^{IS} and W_0^{IS}, for the system in analysis.

We have already introduced the definition for fast motion or extreme narrowing limit, defined as $\omega_0\tau_c \ll 1$ and characterised by very short correlation times; and the slow motion or spin diffusion limit, when $\omega_0\tau_c \gg 1$ and characterised by very large correlation times.

Thus, in the extreme narrowing limit,

$$j(2\omega_0) = 2\tau_c = j(0) \qquad \text{Equation 1.35}$$

and Equation 1.33 simplifies to

$$\sigma_{IS} = \frac{1}{2}b^2\tau_c \qquad \text{Equation 1.36}$$

Under this condition, the sign of σ_{IS} is positive since $W_{2IS} > W_{0IS}$, resulting in a positive NOE enhancement.

In the spin diffusion limit,

$$j(2\omega_0) = \frac{2}{{\omega_0}^2\tau_c} = \frac{j(0)}{{\omega_0}^2{\tau_c}^2} \qquad \text{Equation 1.37}$$

and Equation 1.33 simplifies to

$$\sigma_{IS} = -\frac{1}{10}b^2\tau_c \qquad \text{Equation 1.38}$$

as $j(2\omega_0)$ is negligible relative to $j(0)$. Under this condition, the sign of σ_{IS} is negative since $\tau_c > 0$, resulting in a negative NOE enhancement.

In concluding this discussion, it is important to stress that if r is too large (usually above 5 Å), b will tend to 0, σ_{IS} will be null and there will not be NOE effect whatsoever. That is, in essence, the relevance of NOE experiments for determination of 3D structures of molecules: it is observable only for short spin-spin distances, reporting through-space connectivity associated to the 3D architecture of the molecule.

1.3.3.2 Steady state NOE

The experiment schematised in Figure 1.16, in which spin *S* was saturated selectively by a weak continuous wave (CW) rf irradiation, is called "steady state" NOE (ss-NOE). We define a long low power continuous wave rf as a "saturating pulse", which ensures full saturation of the populations of spin S, and hence results in the disappearance of the peak corresponding to *S*, and in an enhancement in the peak corresponding to *I*, proportional to σ_{IS}.

The pulse sequence to perform this experiment is shown in Figure 1.17: in a first experiment, a continuous wave is on spin *S* for a time long enough to allow homogeneous saturation across the sample (long rectangle in Figure 1.17); as soon as the CW is switched off, a hard 90° pulse is given (short and tall rectangle in Figure 1.17), and acquisition follows.

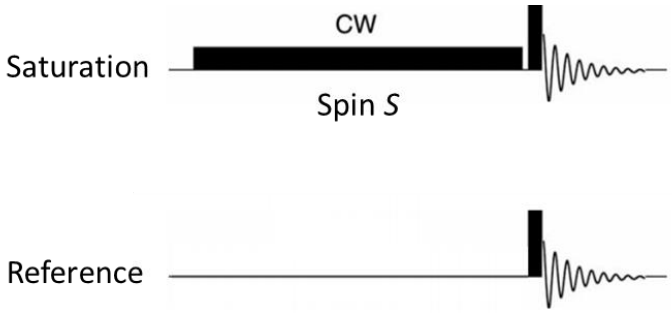

Figure 1.17. Pulse sequence for steady state (SS) NOE experiments. The saturation on spin *S* is obtained using a continuous wave rf irradiation, immediately followed by a hard 90° pulse and acquisition. The reference spectrum, in which the intensities are unperturbed, is a simple pulse-acquisition experiment. The difference spectrum is obtained by subtraction of the saturated spectrum from the reference spectrum. Figure adapted from 55.

Treating this with the product operator formalism, $S_z = 0$, *i.e.* the steady state has been reached as an effect of the CW (when this is on for long enough). This means that I_z is

not changing either ($dI_z/dt = 0$). Including these two conditions, we can rewrite Equation 1.23 as it follows:

$$\left.\frac{d(I_z - I_z^0)}{dt}\right|_{SS} = -R_{zI}\left(I_{z,SS} - I_z^0\right) - \sigma_{IS}\ (0 - S_z^0) = 0 \qquad \text{Equation 1.39}$$

therefore:

$$I_{z,SS} = \frac{\sigma_{IS}}{R_{zI}} S_z^0 + I_z^0 \qquad \text{Equation 1.40}$$

Hence, the enhancement η is given by:

$$\eta_{SS} = \frac{I_{z,SS} - I_{z,ref}}{I_{z,ref}} = \frac{\sigma_{IS}}{R_{zI}} \frac{S_z^0}{I_z^0} \qquad \text{Equation 1.41}$$

where $I_{z,ref}$ is the magnetisation of spin *I* in the reference spectrum. The element to stress here is that the enhancement on spin *I*, when spin *S* is saturated to the steady state, does not only depend on σ_{IS} alone, but the longitudinal relaxation of *I* itself, R_{zI}, plays a role too.

1.3.3.3 Transient NOE

Equation 1.41 implies that the enhancement measured on spin *I*, in a steady state experiment saturating spin *S*, is not only proportional to r_{IS}^{-6}, so it can only give a qualitative measure of the distances. This is not the case for the transient NOE experiment, which relies on a shaped and selective 180° pulse on *S* to invert its population (rather than equalising it).

The pulse sequence to perform this experiment is shown in Figure 1.18. Again, in the first experiment, inversion of population is achieved by a shaped short pulse; a mixing time, τ_m, is then allowed for the magnetisation to evolve, before a hard 90° pulse is given, and acquisition follows.

This means that, after the pulse (time 0), $I_z(0) = I_z^0$ and $S_z\ (0) = -S_z^0$. Dealing with a transient process means that in this case the differential equations for dI_z/dt and dS_z/dt must be solved. If we want to know what happens to spin *I* when the population on spin *S* is inverted, we must solve Equation 1.23 again:

$$\frac{dI_z}{dt} = -R_{zI}(I_z(t) - I_z^0) - \sigma_{IS}\ (S_z(t) - S_z^0) \qquad \text{Equation 1.42}$$

where the dependence of the magnetisation of both spins with the time has been highlighted, as we are dealing with a transient phenomenon. In the initial rate approximation (in the first instance after the shaped pulse, *i.e.*, for very short mixing times), we can re-write:

$$\left.\frac{dI_z}{dt}\right|_{init} = -R_{zI}(I_z^0 - I_z^0) - \sigma_{IS}\,(-S_z^0 - S_z^0) = 2\sigma_{IS}S_z^0 \qquad \text{Equation 1.43}$$

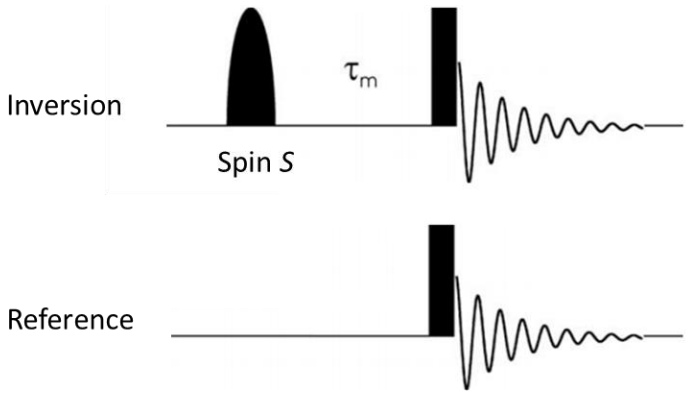

Figure 1.18. Pulse sequence for transient NOE experiments. The inversion on spin *S* is obtained using a shaped 180° pulse, followed by a mixing time, τ_m, a hard 90° pulse and acquisition. The reference spectrum, in which the intensities are unperturbed, is a simple pulse-acquisition experiment. Figure adapted from 55.

It is easy to integrate this equation for the length of the mixing time (that is from time 0 to τ_m), to find that:

$$I_z(\tau_m) = 2\sigma_{IS}\tau_m S_z^0 + I_z^0 \qquad \text{Equation 1.44}$$

where the NOE enhancement is given by:

$$\eta(\tau_m) = \frac{I_z(\tau_m) - I_{z,ref}}{I_{z,ref}} = \frac{2\sigma_{IS}\tau_m S_z^0}{I_z^0} \qquad \text{Equation 1.45}$$

Unlike the steady state experiment, the behaviour of spin *I* does not depend on its longitudinal relaxation properties, but only on the cross relaxation rate and on the mixing time, which we can set experimentally. This make transient NOE an excellent tool for correlating NOE enhancements coming from transient NOE experiments with inter-nuclear distances between the two spins based on the inverse proportionality of σ_{IS} and r_{IS}^6.

1.3.3.4 Truncated driven NOE (TOE)

If we apply short shaped selective pulses in a train, increasing the overall length of the irradiation period, we can measure the rate at which the steady state is reached; but in the first points of this experiment we are still in transient NOE condition. We call this approach the truncated driven NOE, abbreviated to TOE. As we will see in more detail in Chapter 2, Section 2.1, TOE is the key element for Saturation Transfer Difference NMR (STD NMR), which relies on saturation on the nuclei of a large molecule receptor (generally a protein) leading to intermolecular NOEs with the closest nuclei of the ligand. Whereas the NOE enhancement at the steady state is strongly dependent on the relaxation properties of the nuclei (in this case the ligand nuclei, which is what we observe), the rate at which the steady state is reached is affected by inter-nuclear distances. Before entering into the details concerning STD NMR, which is the core technique of the different systems studied in this thesis, in the following sub-section, we will introduce the concept of spin diffusion and give a brief overview on the wide range of NOE- or relaxation-based methods to study intermolecular interactions by NMR.

1.3.4 NOE in multi-spin systems: spin diffusion[54,56]

NOE is a direct consequence of the contributions to relaxation of the double- and zero-quantum transitions (W_2^{IS} and W_0^{IS}) processes, while any mechanism inducing single quantum transitions (W_1s broadly) can be considered as a competing mechanism reducing the observed NOE enhancement. The discussion so far has covered the relaxation effects in one- and two-spin systems. It goes without saying that real systems are never so simple, and many are the factors influencing the longitudinal relaxation behaviour of a molecule in a complex system. This goes from the small contribution of other relaxation effects (chemical shift anisotropy, quadrupolar interactions, *etc.*), to the presence of paramagnetic species in solution, such as dissolved O_2. We group together all these terms and call them ρ_I^*, or the "leakage" of the system.

ρ_I^* is not the only factor which can affect the detection of NOE enhancement, and a key role is played be the relayed-NOE: the so-called spin diffusion. This arises from the presence of nuclei *N*, neighbouring *S* and *I*, and themselves in position to cross-relax with *S* and *I*, with the ultimate effect of diluting the NOE.

Mathematically, the NOE enhancement on *I*, upon irradiation of *S,* is affected by the presence of a third set of nuclei *N* in the following way:

$$\eta_I\{S\} = \eta_{I,max}\ \frac{r_{IS}^{-6}}{r_{IS}^{-6}+r_{IN}^{-6}} \qquad \text{Equation 1.46}$$

Therefore, the closest *N* is to *I*, the least is the chance to measure distances accurately. But diluting the NOE enhancement is not the only thing that a neighbouring nucleus can do; in fact, if *N* is close enough to *I*, it can even sense the effect of the irradiation on *S* through *I*. This is what we call a relayed-NOE. For small molecules, for which direct NOE is positive (*S*→*I*), the relayed NOE (*I*→*N*), is negative. This is because a decrease in the population difference of *I* from the NOE between *S* and *I* results in an increase of the difference in population between *I* and *N*. If the direct effect is positive, the indirect effect will be, in turn, negative. The same is not true for large molecules, for which both the direct and the indirect NOE are negative. Therefore, the presence of relayed NOE in the structural study of fast tumbling molecule is not a dramatic problem, as the negative sign of the relayed NOEs makes it easier to distinguish direct effect from indirect effects. For slow tumbling molecules, the problem is more conspicuous, as both direct and indirect effects exhibit a negative enhancement. Additionally, for larger, slow tumbling molecules the NOE grows more rapidly because of the long correlation time (the same reason why the NOE for large molecules are larger in absolute values than the NOE for small molecules). In such systems, extended saturation times selective to one spin can end up being spread throughout the (large) molecule until all the nuclei experience roughly the same saturation level. This extreme manifestation of relayed-NOE is called *spin diffusion* (noticeably, the limit of slow tumbling is also referred to as spin diffusion limit). In fact, this terminology has been extended to indirect NOE in general, whereas its use in general contexts is controversial.

1.3.5 NMR approaches to investigate intermolecular interactions

The NMR techniques to study intermolecular interactions divide into i) ligand-based, whether we put ourselves in the conditions of observing the response of a small molecule in the presence of its receptor; or ii) receptor-based techniques, observing the receptor itself. The two are schematised in Figure 1.19.

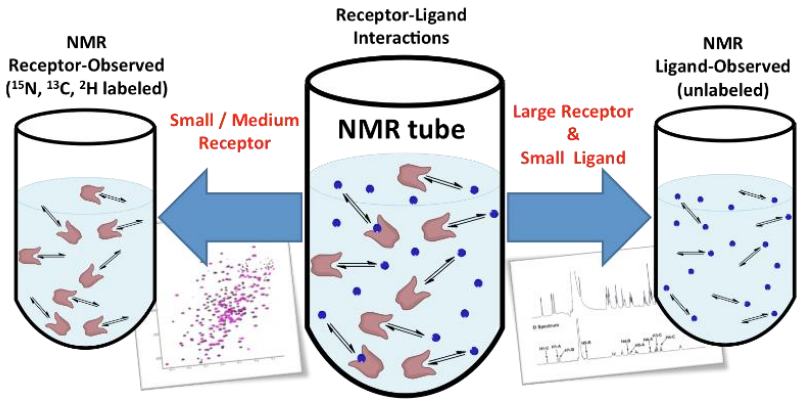

Figure 1.19. Cartoon schematising the two approaches to receptor ligand interactions: ligand-based and receptor-based point of view.

1.3.5.1 Ligand-based methods for intermolecular interactions investigation[57]

Ligand-based NMR methods imply observing intermolecular systems composed of a small molecule binding to a much larger receptor, which we monitor through NMR experiments focusing *only* on the signals from the small molecule (in the free state). These ligands are characterized by large relaxation time constants T_1 and T_2, vanishing or weakly positive NOE enhancements (short correlation times), and large translational diffusion coefficients. Bound compounds share the NMR relaxation properties of the receptor, and have short T_2, large and negative NOE enhancements with efficient spin-diffusion, and slower molecular diffusion. These significant differences make possible the detection of the interactions, as, when bound, the ligand will transiently assume the properties of the receptor, and it will behave differently in the free and the bound state. For the kinetics reasons which we have seen in Sub-section 1.2.3, large excess of ligand over the receptor is required for saturating the binding site of the protein receptor when dealing with weak binders ($[L]_T \gg [P]_T$ and $[L]_T \gg K_D$). In these conditions, though, $f_F^L \gg f_B^L$. For this reason, in ligand-based techniques, the difference in the property under observation between the free and the bound state, which we generally call Q_F and Q_B should be such that $f_B^L Q_B$ is significant relative to $f_F^L Q_F$. One advantage of this is the fact that in a sample containing ligand in large excess to the receptor, the signals of the small molecules will be predominant, and the protein envelope will be on the baseline.

The plainest example of how the properties of the receptor can be transferred to the ligand and detected in a 1D ^{1}H NMR spectrum is the T_2 broadening experiment. Short

T_2 is associated to broad signals, whereas long T_2 is associated to sharp signals. In a solution containing a large receptor with two molecules, say one binder and one non-binder, conditions can be optimised to determine, just from the simple 1D ^{1}H NMR spectrum which of the two molecules is the ligand, as only its signals will transiently be affected by a T_2 typical of the large receptor and will be much broader than the signals of the molecule which does not interact with the receptor (Figure 1.20a).

The same mechanism can be exploited in a more elegant way applying the Carr-Purcell-Meiboom-Gill sequence (CPMG), also called T_2-filter experiment. This relies on the repetition of spin-echo building blocks, applied after the usual 90° pulse. The spin-echo building block consists of a 180° pulse, anticipated and followed by a fixed interval τ. As we know, the 180° pulse inverts the spin populations, therefore the *x,y*-magnetisations which were defocused according to their different chemical shifts during the first period τ, will be inverted and then refocused after the second interval τ. This will only be true for the magnetisation which has not been lost already by loss of coherence in the *x*-*y* plane due to intrinsic transverse relaxation (T_2). Hence, in a CPMG experiment, after several spin-echo building blocks, the transverse magnetisation of molecules with short T_2 will quickly disappear, whereas that from molecules with long T_2 will survive. This is very useful for removing the interfering signals of proteins in samples to study protein-small molecule interactions, clearing up the spectra of the large molecule, and showing only the small molecule signals (this is called a T_2-filter). But, what is more relevant, it can be applied to detect protein-ligand binding. Thus, applying a CPMG sequence to the one-receptor-and-two-ligand system described before, only the signal of the non-binder will survive the spin-echo loop, while the signals of the molecules interacting with the protein will disappear completely (as well as any visible protein signal)[58].

Introducing a paramagnetic element in the protein-ligand system and inducing paramagnetic relaxation is another possibility to detect intermolecular interactions. The transverse relaxation rate constants, R_2, are proportional to the product of the squares of gyromagnetic ratios of the two interacting spins. Considering that the electron gyromagnetic ratio is 658 times higher than that of the proton, the presence of one unpaired electron in one molecule, will induce a relaxation effect onto a nearby proton significantly more efficient than other proton located at that position. A first approach to study intermolecular interactions is to employ a first ligand of the receptor carrying a

"spin label", *i.e.*, a functional chemical group containing an unpaired electron (*e.g.*, the radical TEMPO). We then study the interactions of secondary ligands binding in an adjacent binding pocket, in the presence of saturating amounts of the first spin-labelled ligand. The unpaired electron will act as a transverse magnetization sink for all molecules that bind in the secondary binding site, an effect that will strongly depend on the distance between the adjacent site and the spin-label of the first ligand. If we carry out T_2-filter experiments on this system, the ligand signals will be dramatically reduced due to the R_2 relaxation enhanced by the presence of the spin-label[59].

The CPMG and the paramagnetic relaxation experiments, together with ^{19}F relaxation experiments for fluorinated compounds, are quick methods relying on the large difference between the relaxation properties of the large and the small molecule to detect binding. In drug discovery, these are routinely used to screen large libraries obtaining an immediate answer about the potential to bind.

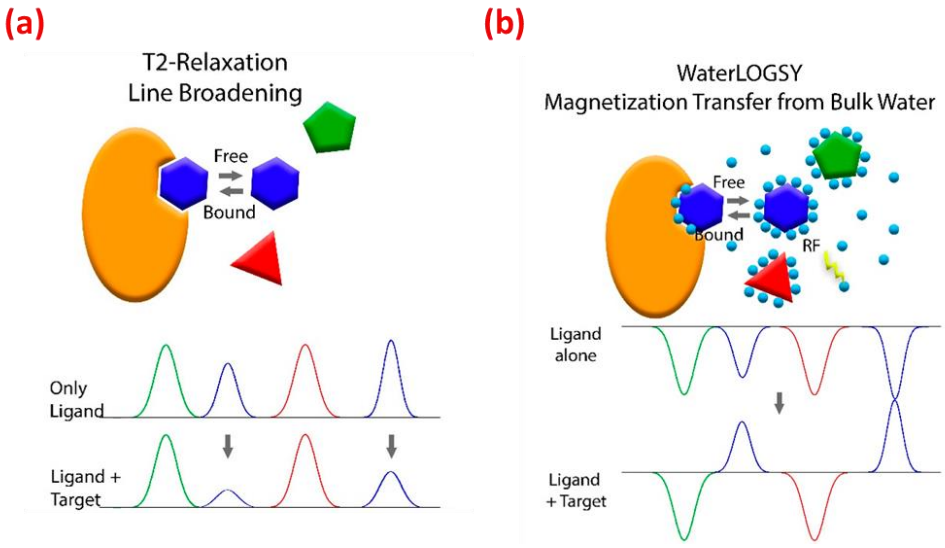

Figure 1.20. Cartoon representation of **(a)** the T_2-filter experiment and **(b)** the WaterLOGSY experiment in the situation where multiple fragments (represented by geometric shapes) and one receptor (represented by a yellow oval) are in solution. The schematic representation of the appearance of the spectra in the presence and the absence of the receptor is shown below each experiment. Figure from 60.

Another very widespread technique in pharmaceutical research is WaterLOGSY, which relies on the transfer of magnetisation by intermolecular NOE between the ligand and the bulk or bound water. This experiment can be considered as a variant of the STD NMR

experiment, in which we employ the water molecules present at the receptor-ligand interface to produce the perturbation in the magnetization of the protein and eventually of the ligand protons.[20] These water molecules show characteristic features of large-sized molecules, as they are bound to the macromolecular receptor, and hence give rise to strong negative NOEs with both the protons at the surface of the protein and the ligand protons in the bound state. In contrast, water molecules in the solvation shell of the ligand in the free state show characteristic features of small or medium-sized molecules (small τ_C) and hence give rise to small positive NOEs with the ligand protons. In practice, we use steady state NOE-like pulse sequences in which we selectively saturate the water signal (light water must then be used). In this experiment, the ligand signals are affected by strong negative NOEs with both, water and receptor protons, and hence have opposite sign to the signals of the non-binding molecules in the sample (Figure 1.20b)[61].

The ligand-based methods described so far, albeit reliable and very easy to run, cannot give any information on the binding mode of the potential ligands. We have to resort to STD NMR and transfer-NOESY (tr-NOESY) to know more on the epitope of binding of a small molecule and on its bound conformation, respectively. Being the studies proposed in this thesis strictly structural, STD NMR and tr-NOESY are the key techniques used, therefore Section 2.1 of Chapter 2 will be devoted to them.

1.3.5.2 Receptor-based methods for intermolecular interactions investigation[62]

Ligand-based techniques have on their side the ease of employment, the lack of upper limit in the size of the receptor, and the fact that no previous information on the receptor is required upon analysis (no need of assignment and therefore no need of isotopic labelling).

In turn, this means that it is not possible to know what is going on at the receptor level and to establish directly the location of binding on the surface of the receptor (competition experiments are required). Also, interactions with slow kinetics (strong affinity) are hard to detect with ligand-based methods, and small receptors do not guarantee reliable results.

Therefore, receptor-based methods are highly complementary and can answer many questions that ligand-based methods cannot. To this purpose, the chemical shift

assignment of the protein is almost always required. Protein assignment requires ^{13}C and/or ^{15}N isotopic labelling (as well as ^{2}H in some cases) and poses an upper limit on the size of the protein (fast transverse relaxation precludes detection of signals from very large proteins). Additionally, for large proteins, signal overlapping becomes too conspicuous for the protein to be resolved. In this regard, advanced multi-dimensional experiments, together with powerful non-uniform sampling techniques, which help reducing the experimental times, and higher magnetic fields are pushing this limit higher and higher.

Once the chemical shifts of each residue (or most of them) are known, and the structure of the protein is determined, changes in chemical shifts of the amide protons of the receptor can be easily followed upon ligand addition (*e.g.* in a titration experiment) by carrying out simple $^{1}H,^{15}N$-HSQC experiments. This technique is called chemical shift perturbation (Figure 1.21) and allows to identify the binding site and/or any allosteric effect, if present.

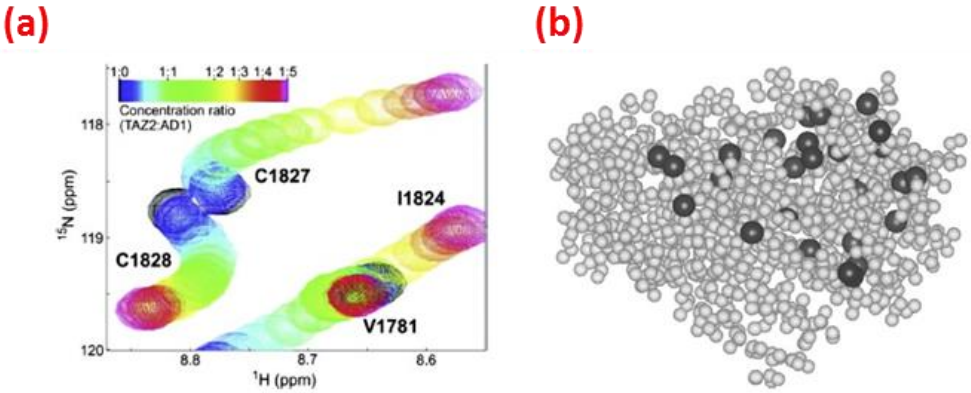

Figure 1.21. General example of a CSP experiment. **(a)** Detail of a ^{1}H-^{15}N HSQC experiment showing significant variations of peak positions after successive additions of the ligand. **(b)** Map of the residues that determine the interaction interface (amino acids suffering large perturbations in chemical shifts). Figure from 63.

For protein-protein complexes, a selective subunit isotopic labelling can be used to remove the signals of the non-isotopically enriched subunits from the spectra, and again follow the chemical shift perturbation of the non-deuterated protein during the interaction. For high molecular weight proteins, the non-labelled protein can be per-deuterated (*i.e.*, all the the non-labile ^{1}H spins are substituted by ^{2}H spins), to reduce the broadening of the labelled protein under observation. A more efficient methodology

consists in selectively introducing protons in a per-deuterated protein. Particularly effective is the introduction of protonated methyl groups. Very recently, there has been an increased interest in ^{13}C direct observation NMR methods, allowing the development of new schemes for specific ^{13}C labelling, allowing selective visualisation of given residues in the protein.

Other methods to characterise the interfaces between species are: intermolecular, inter-subunit or inter-domain NOEs and paramagnetic relaxation enhancements (PREs). Subsequently, distances and relative orientations of the subunits can be restrained by using NOEs, residual dipolar couplings (RDCs), and pseudo-contact chemical shifts (PCSs).

1.4 Motivation and objectives of this thesis

1.4.1 Why high-field NMR in combination to computational tools is an ideal tool for protein-carbohydrate investigation

In the previous sections we have set the theoretical framework of this thesis, introducing the world of glycobiology and intermolecular interactions, on one hand, and the field of Nuclear Magnetic Resonance, with an eye on the phenomenon of NOE and its extensive application in structural studies, on the other hand.

Before stating the aims of the current work, it is worth emphasising the vast potential of NMR to study biologically relevant protein-ligand interactions. More than any other class of ligands, carbohydrates are complex and flexible entities with a high degree of stereochemistry which make them interesting subjects for NMR observation. Nevertheless, the chemical similarity of carbohydrate monomers, and of the protons within each ring, poses the problem of chemical shift overlapping: the proton chemical shift dispersion of carbohydrates is generally low, with chemical shifts concentrated between 3.2 ppm and 4.1 ppm for the ring protons H2 to H6, and between 4.4 ppm and 5.2 ppm for the anomeric ones[64]. Thus, to study carbohydrates at atomic detail, high-field spectrometers are necessary. The 800 MHz Bruker spectrometer available at the School of Pharmacy at UEA is an optimal tool to investigate the protein-carbohydrate interactions.

A proficient level of expertise in the field of computational analysis is also required to model the 3D structure of the interactions in solution, on the bases of the structural

experimental data. Molecular docking is a good starting point for providing 3D models for protein-ligand interactions and the suite Maestro Schrodinger has been reported to be the most efficient tool to model flexible carbohydrates in relatively shallow binding pockets[65] (still, Molecular Dynamics is always the method of choice to account for the full flexibility of the system). As we will see in the following chapter, validation of STD NMR data against the simulated 3D models is necessary, and the best available tool for this is a program called CORCEMA-ST (COmplete Relaxation and Conformational Exchange Matrix for Saturation Transfer), a MatLab code released in 2002 by the group of Rama Krishna[66].

In our research group, the coexistence of these three fields of expertise, the availability of the 800 MHz Bruker spectrometer and the Maestro suite, as well as the expertise in the use of the CORCEMA-ST script (under licence) provided the perfect environment for undertaking the development of novel STD NMR approaches for the structural investigation of a number of biologically relevant protein-carbohydrate interactions at atomic detail.

In this environment and over the three years of this doctoral project, it has been very frequent to be in touch with research groups inside and outside the UK and to take small or large part into several collaborations. Only the two biggest (and most successful) works of biological relevance are included in this thesis: namely, a drug discovery project on the structural study of cholera toxin inhibitors; and the fundamental structural investigation of an intramolecular *trans*-sialidase from the gut microbiota. Still, the fact that many other side projects kept coming and going was an exceptional training and inspiration. This allowed us to study and observe many different biological systems from any natural kingdom, what constantly encouraged us to experiment further.

1.4.2 Our initial intentions of novel STD NMR methodology development

Investigating such a wide range of protein-ligand interactions was a strong stimulus to experiment with STD NMR, as we were pushed to tailor the technique to the features of every different system.

The initial aim of the present thesis was to expand the potential of Saturation Transfer Difference NMR, exploiting some novel conceptual ideas.

Specifically, we wanted to answer the following questions:

1. It was known that different irradiation frequencies slightly affect the outcome of the STD NMR results. Can these differences be exploited? Can the inhomogeneity of spin diffusion help tracking the different pathways for direct and indirect saturation transfer, to get information on the architecture of the binding pocket?
2. Does the presence of protons from water affect the saturation transfer from the polar residues containing exchangeable protons?
3. Can we extract further information from STD NMR experiments with direct irradiation on the proton frequencies of the ligand? Can we observe intramolecular NOEs across the bound ligand on STD NMR experiments? Can we observe inter-ligand NOEs on STD NMR experiments, in a system containing two ligands bound to adjacent subsites?
4. In the case of positively answering those questions, can we provide a standardised protocol for the scientific community to implement our findings in their research?

The answers to these questions are reported in Chapter 3.

1.4.3 Cholera toxin inhibition: investigation of a novel class of GM1 antagonists

The first of the biological investigations undertaken involved the structural elucidation of the binding of a promising lead and its fragments to the GM1 binding subsite of the Cholera Toxin subunit B (CTB), coming from the research group of Inmaculada Robina (University of Seville).

In the wake of a long history of drug discovery studies aimed at designing GM1 antagonist to serve as CTB inhibitors and prevent the onset of the cholera infection, the particularity of the CTB binders investigated by us was their limited carbohydrate nature, making them more "drug-like" than many of the ligands proposed before.

Namely, the main lead was based on a scaffold containing a thio-galactose and a polyhydroxyalkylfuroate-aromatic moiety, designed to bind in the GM1 binding pocket, to the two well-characterised galactose and sialic acid subsites, respectively. In a

previous work, on the bases of the STD NMR binding epitopes of the three ligands, a *qualitative* bidentate binding mode was postulated[67].

Hence, the main question we wanted to answer was:

1. Can we determine the binding mode of the main lead and its fragments, quantitatively? Does the binding actually involve the two known binding subsites?

Our finding that the polyhydroxy moiety did not occupy the sialic acid binding subsite, but a novel groove adjacent to it, opened new questions:

2. Do the sialic acid and the novel sub-sites exclude each other, or do they coexist? Can they both be occupied at the same time?
3. Where is the source of the specificity of the novel class of ligands (the polyhydroxyalkylfuroate binding to the novel binding sub-site or the galactose binding to the galactose subsite)?
4. What is the impact of structural variations of the scaffold on the affinity of binding? How can penalties associated to the flexibility of the polyhydroxy chain be reduced?

The answers to these questions are reported in Chapter 4.

1.4.4 Investigating the specificity of an IT-sialidase from gut microbiota: structural study on the binding mode of syaloglycans

The research group of Nathalie Juge at the Quadram Institute of Biotechnology is strongly focused on the investigation of the gut symbiont *Ruminococcus gnavus* (*R. gnavus*), a mucin degrader with the ability of binding, hydrolysing and metabolising the sialic acid capping the glycans exposed on the mucus. The recently discovered link between *R. gnavus* and the inflammatory bowel disease brought renewed attention to this organism[68].

In particular, we have studied a sialidase from *R. gnavus*, *Rg*NanH. The enzyme consists of i) a carbohydrate binding module recognising sialic acid (CBM40), and ii) an enzymatically active domain (GH33), which converts the α2/3 sialic acid domain capping

the mucins into a tricyclic derivative of sialic acid itself: 2,7-anhydro-Neu5Ac. For this feature of its enzymatic domain, *Rg*NanH is defined as an intramolecular *trans*-sialidase (IT-sialidase), a class of enzymes currently containing three members in total.

We undertook two sub-projects, focusing on the enzymatic domain and on the carbohydrate binding module, with the aim of investigating the domains in terms of affinity, specificity and mechanism of recognition towards a library of α2/3 and α2/6 sialoglycans.

For GH33, our target questions were:

1. Which are the main elements of the ligands for molecular recognition?
2. Does GH33 select α2/3 sialoglygans over α2/6 s sialoglygans?

The finding that GH33 binds to both 3'-sialyllactose (3'SL) and 6'-sialyllactose (6'SL), despite the fact that only 3'SL is the hydrolysable substrate, opened more questions:

3. Do 3'SL and 6'SL bind to the same subsite? Which are their relative affinities?
4. What is the 3D structure of the complexes?
5. Which are the key elements of the interactions?
6. How to explain the specificity of the reaction for α2/3 sugars?
7. How to explain the prevalence of the intramolecular *trans*-reaction over the intermolecular *trans*-reaction?

For CBM40, our target questions were:

1. Which are the main elements of the ligands for molecular recognition?
2. Which is the minimum sugar entity recognised by CBM40?
3. What is the selectivity of CBM40 towards sialoglycan linkages?
4. How do sugar decorations (*e.g.*, the presence of a glycolylic group in the place of the acetamide group on the sialic acid, the *N*-acetylation of the galactose and glucose moieties) affect the binding?

The answers to these questions are reported in Chapter 5.

Chapter 2

Techniques and tools

2.1 Saturation Transfer Difference NMR and Exchange-transferred-NOE

2.1.1 Effect of chemical exchange on chemical shift[54]

Rarely, if ever, we can consider the system in our samples static from an NMR point of view. Some phenomena, such as the internal bond vibrations for the molecules in our tube (in the frequency range of 10^{13}-10^{14} Hz), are too fast to be noticeable in the NMR time scale. Still, many interesting dynamic processes are in a suitable time scale for NMR observation in the frequency range between 10 and 10^{10} Hz[69] and we call them *chemical exchange* processes. These include restricted bond rotation, group interconversion or tautomerism, as examples of intramolecular processes of great interest for organic and structural chemists. In our spotlight are the intermolecular chemical exchange processes involving ligand-protein binding events.

Any process of chemical exchange can be written as an equilibrium between two states A and B:

$$A \underset{k_{ex}}{\overset{k_{ex}}{\rightleftharpoons}} B$$

As already mentioned in Chapter 1, Sub-section 1.2.3, the biomolecular association in solution between a protein (P) and a ligand (L) to form a protein-ligand complex (PL) can be written as:

$$P + L \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} PL$$

In the context of ligand-based NMR techniques in which we have framed our work, we can formalise the chemical exchange process as:

$$L_F \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} L_B$$

Where L_F and L_B are the free and bound ligand, respectively.

In many cases, the chemical exchange will influence the appearance of the spectra. This will be most obvious when the chemical shift of one or more protons of the ligands is affected by the chemical exchange (in this case the appearance of the spectra will depend on the kinetics of the exchange). For simplicity, let's imagine a two-site exchange between equal populations, for example an amide with a methyl group sitting on the amidic nitrogen (Figure 2.1a). The semi-rigidity of the amidic bond is such that the methyl group can rotate from a *cis* to a *trans* configuration relative to the amidic oxygen, undergoing the equilibrium shown in Figure 2.1a. In each configuration, the methyl group would experience a different chemical environment and therefore have different chemical shift. If the rate of exchange between the *cis* and the *trans* configuration is slow enough in the chemical shift time scale ($k_{ex} << \Delta\nu$), each species (*A* and *B*) will have a lifetime long enough and we will be able to clearly distinguish two sharp separated peaks (one coming from the methyl group in *A* and one coming from the methyl group in *B*, Figure 2.1b left).

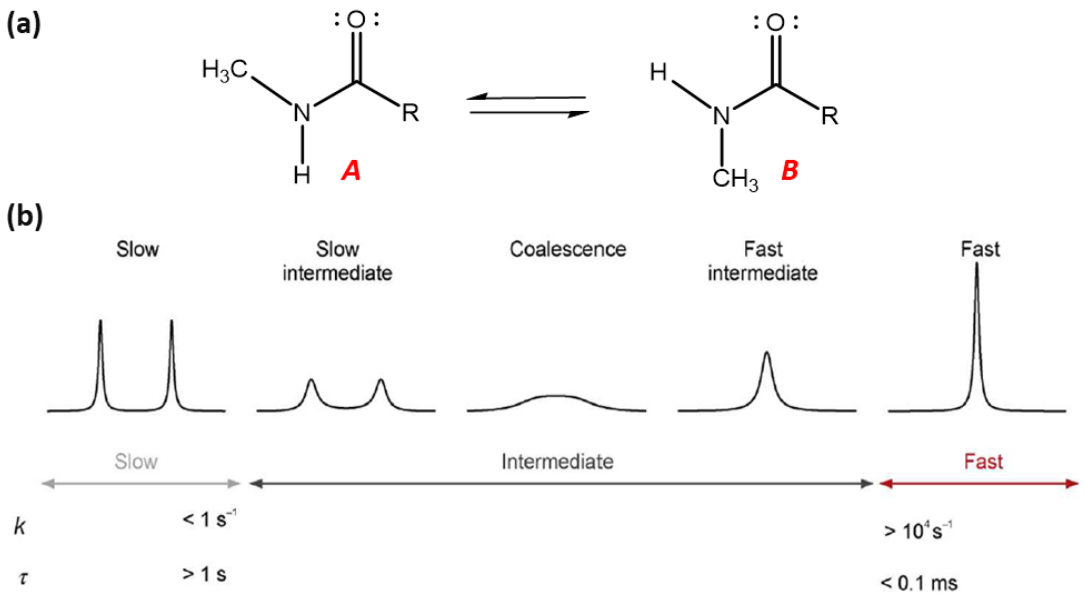

Figure 2.1. **(a)** Restricted rotation around the amidic bond. The methyl group is in chemical exchange between a *cis* and a *trans* configuration relative to the amidic oxygen. **(b)** Schematic classification of the dynamic exchange regimes observed in NMR spectra, with approximate rates (k) or life time (τ) associated with these regimes. Figure adapted from 54.

If we started to steadily increase the temperature of the system, *i.e.* to steadily increase the exchange rate of the process, we will see the peaks "moving" toward each other. Eventually, when the chemical exchange rate is so fast relative to the chemical shift time

scale (k_{ex} >> $\Delta\nu$) that the two species cannot be distinguished anymore, we will observe one single sharp peak, resulting from an "average" between the two species (Figure 2.1b right).

The transition between the slow and fast motion regimes is called *intermediate motion* and goes through broadening of the two peaks until a point, called *coalescence point*, in which the signals "flatten out" seeming to disappear ($k_{ex} \approx \Delta\nu$). As the exchange rate increases, one broad signal at the average chemical shift among the two starts to be visible and becomes sharper as the kinetics fastens (k_{ex} >> $\Delta\nu$).

The broadening effect arises as the same spin (*e.g.* the methyl protons in Figure 2.1a) will experience two different precession frequencies while they transition from the *A* to be *B* form and *vice versa*. In the intermediate regime, the precession frequency changes within the scale of the acquisition time, leading to a net dephasing of the bulk magnetisation vector (the exchange does not happen coherently for all the molecules at the same time). This will lead to short FIDs (faster loss of net magnetisation) which are associated to broad signals. If the exchange rate is very fast, little net dephasing of transverse magnetisation can occur and the FID records the average frequency between the two species, producing a sharp single signal.

2.1.2 Effect of chemical exchange on other NMR parameters: slow and fast motion[54]

When the system under study is either in the slow or in the fast regime of chemical exchange, we will obtain sharp signals (two in slow exchange conditions and one in fast exchange conditions) which we can analyse to obtain useful information on the dynamics of the system. For instance, in Figure 2.1, the presence of one or two signals alone informs us on the rate of exchange around the semi-rigid bond. But the chemical shift is not the only NMR parameter influenced by chemical exchange processes (albeit it is the most promptly observed). Other NMR parameters, such as longitudinal or transverse relaxation rate constants, scalar couplings, and residual dipolar couplings are sensitive to various types of chemical exchange and can assume different values when experiencing transitions between different environments.

In the case of ligand-receptor binding processes, in particular, the ligand transiently adopts NMR parameters typical of the much larger receptor (we will generally refer to

any of these parameters as Q), and, at the same time transiently perturbs the binding site microenvironment in the receptor. In both cases, $Q_F \neq Q_B$, *i.e.* $\Delta Q \neq 0$ (where the subscripts *F* and *B* refer to the parameters in the free and the bound state, respectively). Therefore, monitoring ΔQ upon binding can provide precious information on the nature of the process. When we consider both ν and Q for a given system, three scenarios are possible[69]:

1) $k_{ex} << \Delta Q$ and $k_{ex} << \Delta \nu$;
2) $\Delta \nu >> k_{ex} >> \Delta Q$;
3) $k_{ex} >> \Delta Q$ and $k_{ex} >> \Delta \nu$.

In case 1), the exchange is slow relative to both time scales, therefore we observe two resonances and can measure both Q_F and Q_B. In case 2), the exchange is slow relative to the chemical shift time scale and fast relative to Q, so we observe two signals on which we can measure an average observed Q, as follows:

$$Q_{obs} = f_F^L Q_F^L + f_B^L Q_B^L \qquad \text{Equation 2.1}$$

Where f_F^L and f_B^L are the fractions of free and bound ligand, respectively. In case 3), the exchange is fast relative to both time scales, and we have a single signal on which we can equally measure the average observed Q_{obs}, as in Equation 2.1 (the same three cases and equations apply to any chemical exchange process, involving *A* and *B* rather than Q_F and Q_B).

2.1.3 Saturation transfer by chemical exchange. Saturation Transfer Difference (STD) NMR[70]

For any system under chemical exchange in which we can see relatively sharp NMR signals in the spectrum from the molecule/s involved in the exchange, we can apply *saturation transfer* methods. Intuitively, this involves a "transfer" of magnetisation from a state *A* to another state *B* (*e.g.*, from bound to free state, in our protein-ligand interaction system), and *vice versa*. This is implemented by perturbing (or saturating) the resonance of given protons in one state and observing how it is transferred to the other state. For the magnetisation to be transferred, the transfer process must occur at a faster or at least comparable rate relative to the spin relaxation; else, the effect of the spin perturbation is lost before the exchange is completed (*i.e.*, it must be true that ΔR_1

<< k_{ex}, where ΔR_1 is the difference between the R_1s of the ligand in the free and the bound state).

Let us now focus again on the ligand-receptor interaction, as observed from the ligand perspective. Also, let us imagine that the exchange rate is fast relative to the chemical shift exchange rate and relative to both relaxation time scales (k_{ex}>>$\Delta\nu$ and k_{ex}>>ΔR_1, ΔR_2).

These are the optimal conditions for ligand-based NMR methods for characterising molecular interactions. It is worth emphasising again that for such fast kinetics, large excesses of ligands are necessary to saturate the protein sites. Due to the fast exchange and to the large excess of ligand, which means that the chemical exchange is taking place with highly unequal populations, the observed signals of the ligand will mostly correspond to those of the free form, with the very weak broad protein signals on the baseline. That means that, under these conditions, the bound state is not directly accessible (not observable), and hence, it would not be possible to directly perturb the ligand signals selectively in the bound state, with the aim of transferring that perturbation to the free state. Indeed, in STD NMR we resort to protein saturation, followed by one first step of intermolecular saturation transfer (by NOE), and then, the chemical exchange will eventually carry that saturation from the bound to the free state of the ligand. For a large protein, selective saturation of a single amino acid resonance will result in a rapid spread of the magnetisation over the entire macromolecule. This is the spin diffusion effect which we described in Chapter 1, Sub-section 1.3.4. When a ligand binds to the saturated receptor, it will receive part of this saturation by intermolecular NOE. The assumption that ΔR_1 << k_{ex} means that the ligand will have time go back in the bulk solution before losing the transferred magnetisation. As NOE is negative slow-tumbling molecules, the sign of the NOE acquired by the ligand upon contacting the saturated protein will be negative too, resulting in a decrease of the ligand signals upon protein irradiation[70a]. The amount of saturation transferred to the binder depends on the macromolecular mobility, the lifetime of the complex (saturation transfer is only detected for complexes with dissociation constant K_D in the range of 10^{-3} to 10^{-8} M) and the geometry of binding (binding mode, which depends on the bioactive conformation of the ligand, and the intrinsic architecture of the binding pocket), as we

will see in more detail in the following sub-sections dedicated to the practical aspects of STD NMR.

2.1.4 Technical aspects of STD NMR and pulse sequence[70]

Based on saturation transfer from the protein to ligands in fast chemical exchange, the technical implementation of Saturation Transfer Difference NMR (STD NMR) is very simple. On a sample containing the receptor (large molecule with MW > 15 kDa) at low concentration (10^{-5} to 10^{-6} M) and a large molar excess (1:10 up to 1:1000) of a pool of small test compounds, two experiments are recorded (Figure 2.2 top): i) 1D ^{1}H-NMR under conditions of thermal equilibrium, the so-called reference spectrum or *off-resonance* experiment, and ii) a second 1D ^{1}H-NMR experiment, the *on-resonance* experiment, in which some protons of the receptor are selectively irradiated with low power radiofrequency during a certain period (saturation time) that can be varied typically in the range of several seconds.

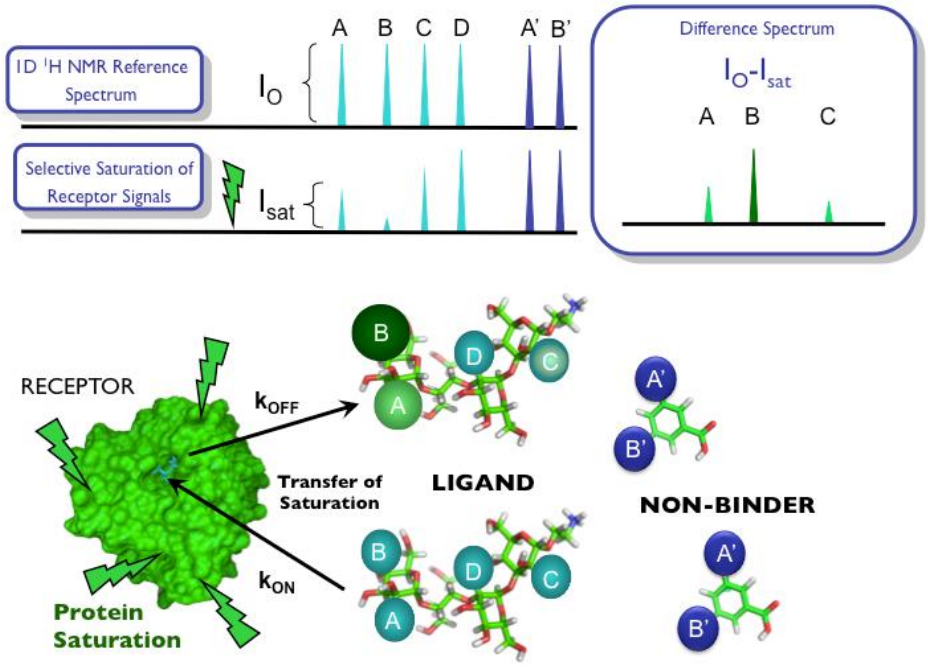

Figure 2.2. Illustration of the 1D STD NMR experiment (top) applied to a sample containing a protein receptor in the presence of a mixture of two small test compounds in molar excess (bottom). Two 1D NMR spectra are recorded: i) a standard 1D ^{1}H NMR (*off-resonance* or *reference* spectrum), and ii) a 1D ^{1}H NMR spectrum with selective saturation of receptor proton signals. Saturation of protein proton polarisation can be transferred to a small molecule by intermolecular NOE if protein-ligand ^{1}H-^{1}H distances are short enough (< 4-5 Å). Thus, saturation transfer will only take place to molecules that transiently bind to the receptor. The difference

spectrum ($I_0 - I_{sat}$, top right) will show only signals corresponding to the ligand, while signals of non-binders cancel out. Figure from 70c. See also Figure 3.1.

During the on-resonance experiment, the existence of non-equilibrium magnetisation (saturation) on ^{1}H nuclei of the receptor leads to intermolecular NOEs with ^{1}H nuclei of the ligand, in the bound state, as the ligand most likely will have at least some of their ^{1}H nuclei at very short distances (r < 4-5 Å) from those ^{1}H nuclei of the receptor.

This process results in the macroscopic detection of (transferred) saturation on the ligand NMR signals (I_{sat}). Since signals are reduced in the corresponding off-resonance experiment (I_0), subtraction of both experiments ($I_0 - I_{sat}$) leads to positive difference signals allowing the identification of the binding compounds. The intensities of non-binding molecules remain the same in both spectra and the difference spectrum cancels them out (dark blue signals in Figure 2.2).

As we can see in the pulse program (Figure 2.3), the key element is the saturating irradiation selective to the protein. This could be achieved by a low power continous wave saturating pulse, as in the pulse sequence of the steady state NOE (as in Figure 1.16), but to avoid undesirable spillover of the saturating field, this is replaced by a train of Gaussian-shaped selective pulses of 50 ms each (the filled half oval in Figure 2.3). The power is set to cover a bandwidth of about 100 Hz, while the number of pulses in the train (n) is adjusted to obtain the desired length of total saturation time.

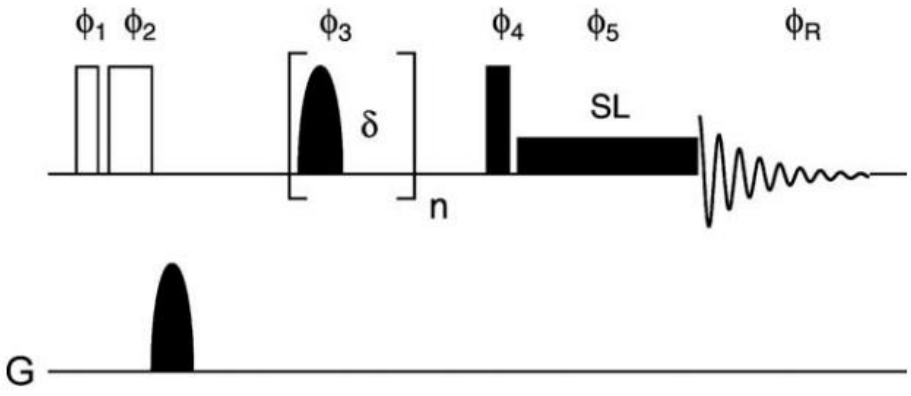

Figure 2.3. Pulse sequence for STD NMR . The narrow and wide-open bars correspond to trim pulses with a length of 2.5 and 5 ms, respectively, performed at the power level for spin-lock. The narrow-filled bar is a 90° hard pulse. A series of selective pulses (50 ms duration, 4 ms inter-pulse delay, frequency switched for on-resonance and off-resonance spectrum) is applied for saturation. The first two pulses as well as the gradient are applied to destroy residual

magnetization. Phase cycling: ϕ1=x; ϕ2=y; ϕ3=x; ϕ4= ϕR=x, 2(−x), x, y, 2(−y), y; ϕ5=4(y), 4(−x), 4(−y), 4(x). Figure from 71.

Hence, a hard 90° pulse (the filled bar with phase ϕ_4) is applied, followed by a spin lock purge pulse (usually between 10 and 40 ms), to remove the broad protein signals. Any other unwanted residual magnetisation can be cancelled by using trim pulses with a length of 2.5 and 5 ms (the empty bars with phase ϕ_1 and ϕ_2, respectively), performed at a power level for spin-lock. The use of a gradient in correspondence of the trim pulses can help this process[71].

Due to the variable length of the saturating Gaussian train (which can go from 0.1 s to 20 s, depending on the system and the purpose), the STD NMR experiment can be regarded as a truncated driven NOE (TOE) experiment, as described in Chapter 1, Sub-section 1.3.3. The usual saturation time for STD NMR in screening experiments is 2-3 s, as a compromise between sensitivity and reasonable experimental time.

2.1.5 Ligand epitope mapping by STD NMR[70c, 72]

The potential of STD NMR goes well beyond simple screening, as not all the signals of the ligand in a STD NMR spectrum show the same amount of saturation. Theoretically, the magnetization transferred from receptor to ligand protons by intermolecular NOE in the bound state depends on the inverse of the sixth power of their distances[56]. Thus, the observed differences in saturation transferred among the different ligand protons indicates different spatial proximities between the distinct parts of the ligand molecule and the protein in the bound state. Qualitatively, stronger ligand STD intensities reflect closer inter-proton distances between the corresponding ligand protons and the receptor surface in the bound state.

This means that we can *map* the protons of the ligand closer to the protein surface, and get semi-quantitative structural information about the binding, by comparing the relative STD intensities of the different protons of the ligand at a given saturation time. This is done by normalizing all the measured STD factors ($I_0 - I_{sat}/I_0$) against the most intense one, which is arbitrarily assumed to be 100%. Thus, the resulting percentage STD values represent a map of the ligand–protein contacts delineating which ligand moieties are key for recognition by the receptor. This is what we call *ligand epitope mapping*, an example of which is shown in the left panel of Figure 2.4.

In this process, the choice of the saturation time is crucial. The STD intensity of any given proton of a ligand typically increases at increasing protein saturation time, until reaching a plateau, the STD_{max}, which depends on the relaxation properties of the ligand (for fast-tumbling small molecules we usually expect around 5-6 s). The plateau is affected by the longitudinal relaxation (R_1) of the ligand in the free state. This means that R_1 plays a more relevant role the closer we get to the STD_{max}. It is worth considering that, ideally, the group epitope mapping information should not depend on the chosen saturation time. However, significantly different R_z relaxation rates of the different protons of the ligand can produce artefacts in the epitope definition.

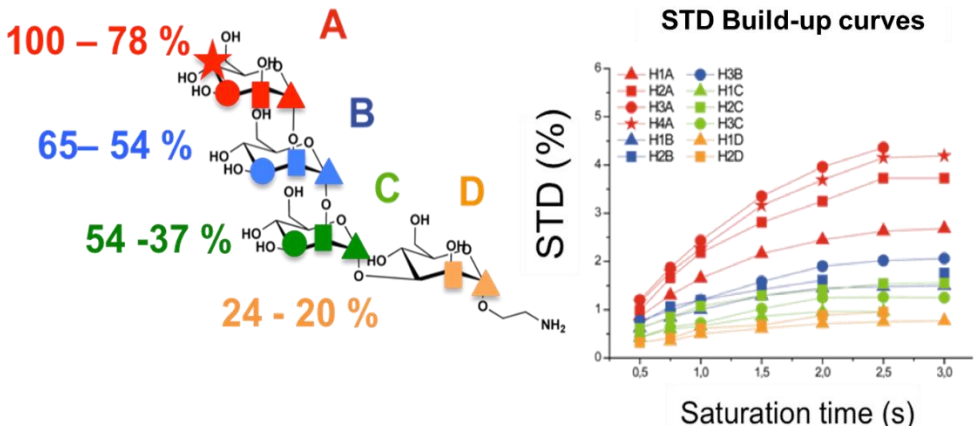

Figure 2.4. Ligand epitope mapping by STD NMR. Left: group epitope mapping of a tetramannoside binding to the human anti-HIV-1 antibody 2G12 by normalization of the STD initial slopes from the experimental curves (right). In this example, the distribution of saturation transfer indicates binding through the non-reducing mannose residue (mannose A). Figure adapted from 70c.

Specifically, protons with slower R_1 relaxation will accumulate more efficiently saturation in solution. Thus, their relative STD intensity will be enhanced in comparison to other ligand protons, and, thus, their proximity to the binding pocket may be largely overestimated at long saturation times. Since all these artefacts are consequence of differences in the ability to accumulate saturation in the free state, a way to cancel them out is to derive STD intensities for a saturation times tending to zero, when virtually no accumulation of saturated ligand takes place. To that aim, Mayer and James[72] proposed measuring STD intensities at increasing saturation times, as opposed as to a single saturation time experiment, and fitting the experimental build-up curves to the mono-exponential function:

$$STD(t_{sat}) = STD_{max}(1 - e^{k_{sat} * t_{sat}})$$ Equation 2.2

where $STD(t_{sat})$ is the observed STD intensity at a given saturation time, t_{sat} , STD_{max} is the asymptotic maximum of the build-up curve, and k_{sat} is a rate constant that measures the speed of STD build-up. An example of the resulting build-up curves is reported in the right panel of Figure 2.4. $k_{sat}$and STD_{max} are derived by least-squares fitting, and the initial slope of the curve is obtained as:

$$dSTD(t_{sat})/dt_{sat}|_{t_{sat} \to 0} = STD_{max}\ k_{sat}$$ Equation 2.3.

The group epitope mapping of the ligand is obtained by normalisation of the relative values of the set of initial slopes obtained, relatively to an arbitrarily chosen proton of the set, usually the one with highest initial slope.

2.1.6 Exchange-transferred-NOESY: bio-active conformation of the ligand[54,56]

STD NMR is not the only ligand-based technique with the potential of elucidating structurally the binding mode in protein-ligand interactions. Exchange-transferred NOESY (tr-NOESY) is a 2D NMR technique enabling us to study the conformation of the bound ligand in solution. From the technical point of view, tr-NOESY is purely a NOESY experiment applied to a sample containing a receptor and ligand in an appropriate ratio (which relates to the affinity of the interaction). The basic pulse program for any NOESY experiment is shown in Figure 2.5.

So far, we have not entered the details of 2D experiments. Simplistically, any 2D experiment consists of the following steps: preparation, evolution, mixing and detection. In the pulse sequence in Figure 2.5, the preparation is the 90° hard pulse with phase ϕ_1 and the evolution time is the time interval t_1. The evolution time is the key element here, as it allows to generate the second dimension of the spectra. Let us imagine, for a sample containing a single signal, that t_1 is set to 0; then, the sum of the first two 90° hard pulse is a hard 180° pulse, which would simply invert the magnetisation from *z* to *-z*.

The net transverse magnetisation would be null, and no signal would be detected. If we started to increment the length of t_1, we would let the magnetisation vector evolve by a given angle, so that, when we pulse again with phase ϕ_2, a small signal would be detected. As we increase t_1 by regular intervals, the detected signal will increase, until

reaching a maximum when the magnetisation vector had time to evolve by a 90° angle; for larger t_1 the magnetisation would then decrease, pass through a null, then through a negative minimum, and so on according to the sinus, until the signal entirely decays due to relaxation.

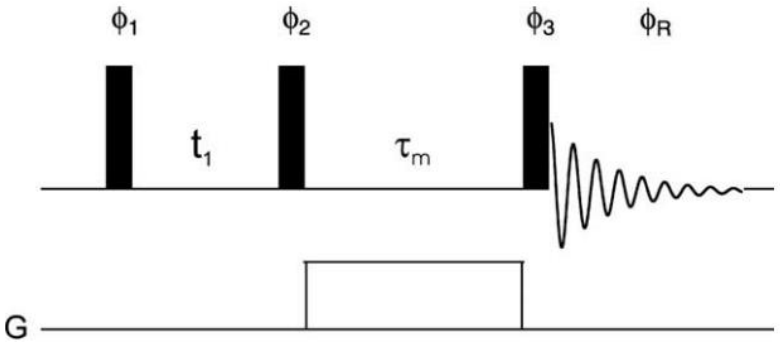

Figure 2.5. Pulse sequence for the 2D NOESY experiment. Narrow filled bars correspond to 90° hard pulses. Phase cycling: ϕ1=x, −x (incremented according to time-proportional phase incrementation (TPPI)); ϕ2= ϕ3=x; ϕR=x, −x. Figure from 71.

The resulting set of FIDs constitutes a pseudo-FID itself, which defines the evolution of magnetisations in what we call the indirect dimension (as it is indirectly generated). A 2D Fourier transform of this data, on the one-signal sample imagined, would yield a signal with the same frequency in both dimensions. The situation becomes more interesting when more than one set of protons is present in the sample, as we can observe correlations between the signals (or peaks) in some fashion (*e.g.*, through bond or through space), leading to the appearance of the so-called cross peaks. In the case of NOESY, the cross peaks indicate correlation through space, meaning that the two protons linked by that correlation are spatially close.

Thus, the resulting spectra will show *diagonal peaks* with identical frequency in both dimensions (large blue dots in the spectra in Figure 2.6), and *cross peaks* featuring the two frequencies of the correlating protons (small dots in the scheme in Figure 2.6). It is important to highlight that cross peaks only exist when the cross relaxation σ_{IS} is not null. Therefore, it is important to choose mixing times such that the cross relaxation of the nuclei of interest is at its maximum. As we have seen profusely, the sign of σ_{IS} depends on the correlation time of the molecules under investigation. Thus, the sign of the cross peaks will be the same or the opposite relative to the diagonal peaks (assumed to be positive), depending on the correlation time of the analytes, and their intensity

will be modulated as a function of the mixing time τ_m of the experiment (among other things), as shown in Figure 2.7.

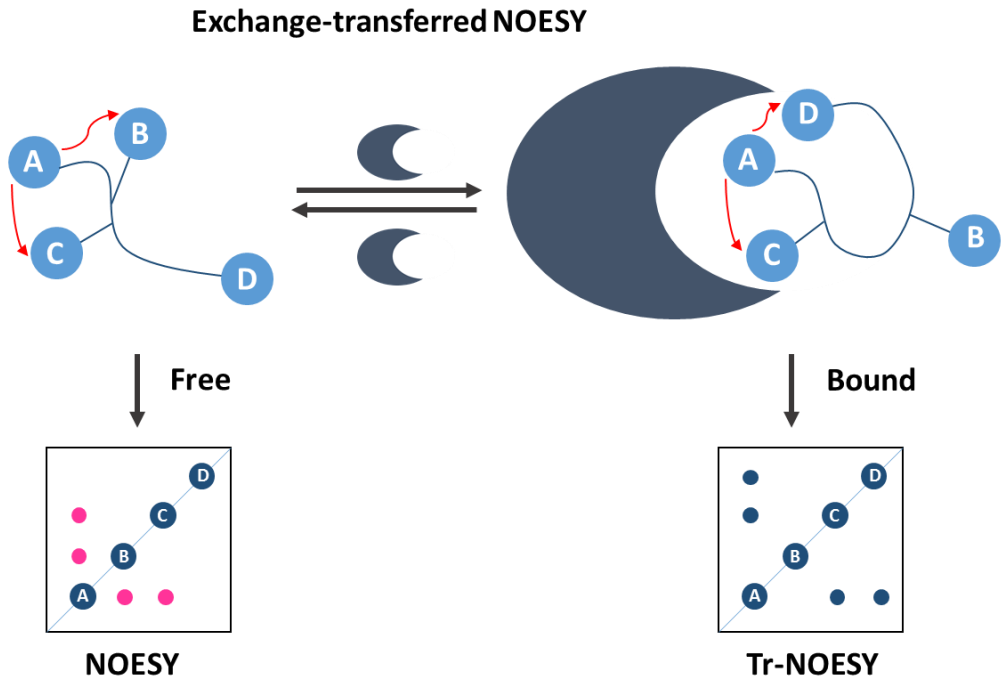

Figure 2.6. Schematic representation of the exchange-transferred NOESY. The existence of receptor-ligand interaction can be proved by the change in sign of the NOESY cross peaks comparing samples in the absence and in the presence of the receptor.

From Chapter 1, Sub-sections 1.3.2 and 1.3.3, we are also aware that low molecular weight molecules typically show small NOEs and reach their maximum in NOESY experiments at mixing times in the range of seconds. In contrast, large molecules (*e.g.*, proteins) show strong NOEs and reach their maximum magnitude for much shorter mixing times (in the range of tens or hundreds of milliseconds).

Exchange-transferred NOE experiments take advantage of the fact that a small ligand in the presence of a high molecular weight receptor will also show negative NOEs if the kinetics of the binding is within the fast exchange condition, and the ligand-to-protein ratio is carefully selected[73].

In this experiment, the chemical exchange can transfer to the free state the perturbations in the populations of the protons of the ligand due to the intra-ligand longitudinal ^{1}H-^{1}H cross relaxation processes in the bound state. These processes lead to strong negative NOEs in the bound state that can be transferred to the bulk solution

if the kinetics is fast enough. In the free state, those populations (bound state NOEs) survive long enough (due to the smaller R_Z of small molecules) as to produce an accumulation in the bulk solution, allowing the observation of strong NOEs in the NMR signals of the free ligand (Figure 2.6).

Under these conditions, the observed longitudinal cross relaxation rate or NOE (σ_{Tr-NOE}) is given by the population-weighted average of the NOEs in the free and the bound state:

$$\sigma_{Tr-NOE} = f_F^L \sigma_F^{NOE} + f_B^L \sigma_B^{NOE} \qquad \text{Equation 2.4}$$

where f_F^L and f_B^L are the fraction of free and bound ligand, and σ_F^{NOE} and σ_B^{NOE}, the NOEs in the corresponding free and bound states. To detect NOEs from the bound state, we must assure that $|f_B^L \sigma_B^{NOE}| \gg |f_F^L \sigma_F^{NOE}|$. If the ligand : protein ratio is too high, $f_F^L \gg f_B^L$, we will enhance the properties of the free state, detecting the NOEs of the small molecule (σ_F^{NOE}). In fact, the ligand-to-protein ratios for transferred NOE observation are within a much narrower range than in the case of STD NMR experiments (typically between 10:1 to 50:1).

From the observation of transferred NOEs, we can determine the bioactive conformation of the ligand in the binding site of the receptor. Nevertheless, those protons of the receptor close to the ligand in the bound state can produce relay NOEs that can affect the intra-ligand NOEs. This phenomenon is called protein mediated spin diffusion, and to avoid its effect on structural calculations we must resort again to initial slopes analysis, or, if feasible, the use of a per-deuterated receptor[74].

2.2 Molecular modelling

2.2.1 Molecular Mechanics[75]

Molecular modelling is a broad field ranging from orbital-based quantum mechanical *ab initio* calculations to predict small system interactions (*e.g.*, in the water dimer) or more ordered interactions (*e.g.*, lattice and solid states studies) to empirical and semi-empirical methods. Indeed, most of the interesting systems in chemistry and biology are too big to be considered purely through quantum mechanics. To simplify the problem, some assumptions and approximations should be made.

In molecular mechanics, the essential assumption is the Born-Oppenheimer approximation, stating that: *the motions of the nuclei and of the electrons can be treated separately*. Due to the minuscule size of the electrons relative to the nuclei, indeed, their motion can be neglected, and the molecular system can be regarded as an ensemble of spherical masses, bringing the problem back to a classical mechanics one. The masses are given coordinates and interact with each other through processes of bond stretching, dihedrals distortion, angles opening and closing according to classical mechanical rules. The method based on these assumptions is indeed called "force field" approach and has huge applications, despite the apparent oversimplification on which it is based.

In fact, the potential energy U is expressed as a function of the internal coordinates $\vec{R}$ of the *N* particles constituting the system (Equation 2.5). The nature of each particle and bond must be specified and the parameters relative to them, *e.g.*, the atoms mass and volume. The constants k_b, k_θ and k_χ, as well as the bonds and angles reference values b_0 and θ_0 in Equation 2.5, are extracted from experimental data (hence these methods are called semi-empirical). In solving this equation, penalties are associated with distortion from these *reference* values. To find the optimal molecular geometry, the values of the coordinates of each atom are searched by minimizing the force field equation, which follows in its simplest form:

$$U(\vec{R}) = \sum_{bonds} k_b\,(b - b_0)^2 + \sum_{angles} k_\theta\,(\theta - \theta_0)^2 + \sum_{dihedrals} k_\chi\,(1 + \cos(\eta\chi - \delta)) + \sum_{Coul\ i,j} \frac{q_i q_j}{\varepsilon\, r_{ij}} + \sum_{LJ\ i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} + \frac{B_{ij}}{r_{ij}^{6}} \right)$$

Equation 2.5

where the contribution of bond stretching, angles and dihedrals re-arrangement are considered. In the second line of the equation, the non-bond terms are also taken into account, namely the electrostatic interactions ($Coul\ i,j$) and the Lennard-Jones ($LJ\ i,j$) repulsion-dispersion potential energy terms, where q_i and q_j are the charges of the interacting particles, ε is the permittivity of the material, A_{ij} and B_{ij} are the Lennard-Jones coefficients and r_{ij} is the distance between the particles.

2.2.2 Protein-ligand docking calculations[76,77]

Based on the force field theory, molecular docking is a computational procedure that aims at predicting the preferred orientation and conformation of a ligand bound to its target protein (protein-ligand docking). Each docking program operates slightly differently, but they share common features that enable them to search for the locations on the protein surface that lead to favourable interactions with the ligand, sample the conformational space of the ligand and compute the interaction energy between the protein and ligand.

In classic protein-ligand docking, the receptor is considered as a rigid body, while the ligand can be set free to rotate, translate and change conformation (or it can be treated as a rigid body only able to rotate and translate). More computationally demanding techniques allow some degree of flexibility of the receptor as well, and they are referred to as *induced fit* docking techniques.

In both cases, the practical steps to follow to perform docking calculations are:

- Protein preparation and ligand conformational minimisation;
- Grid generation;
- Docking calculations;
- Scoring function.

Here, we will shortly describe them one by one.

- Protein preparation and ligand conformational minimisation

Most often, the receptor structure comes from X-ray crystallography data, but it can be derived from NMR studies or homology modelling. Thus, the protein structure need to be checked and prepared, to ensure that all the atoms and bonds are correctly specified, that no hydrogen atom is missing, and that all the residues in the binding pocket have unique orientation (in some cases, amino acid side chains can present several conformations in the X-ray structure). At this point, it is also possible to select the number and type of water molecules to keep in the binding pocket, and finally a round of minimisation is performed.

For the ligand, a round of conformational search minimisations is frequently performed. This allows to start from a conformation which is a low relative minimum, representative of the free molecule in solution (keeping in mind that the ligand conformation will be explored during the docking steps as well).

- Grid generation

The *space* of the receptor in which one wants to *fit* the ligand must be defined, by generating a grid. Its main feature are its centre and its shape and dimension. To centre the grid, one generally uses the reference ligand in the binding pocket, but it is possible to centre the grid in any position, simply by choosing the residues of interest.

The dimension and the shape of the grid are arbitrary: by default, the grid has cubic shape, but the three sides can be set to have different length. In Glide, one can define an inner and an outer grid: the inner grid will enclose the centre of the ligand (defined as the midpoint of the longest segment which can be drawn between any two atoms of the molecule), while the outer grid will contain the whole ligand, as shown in Figure 2.7. The size of the grid strongly affects the outcome of the docking calculations: larger grids will allow to probe unusual and asymmetric binding pocket, exploring a wider space; on the other hand, smaller grids will limit the number of results, excluding odd poses and saving calculation time (therefore, ideally, the smaller the better).

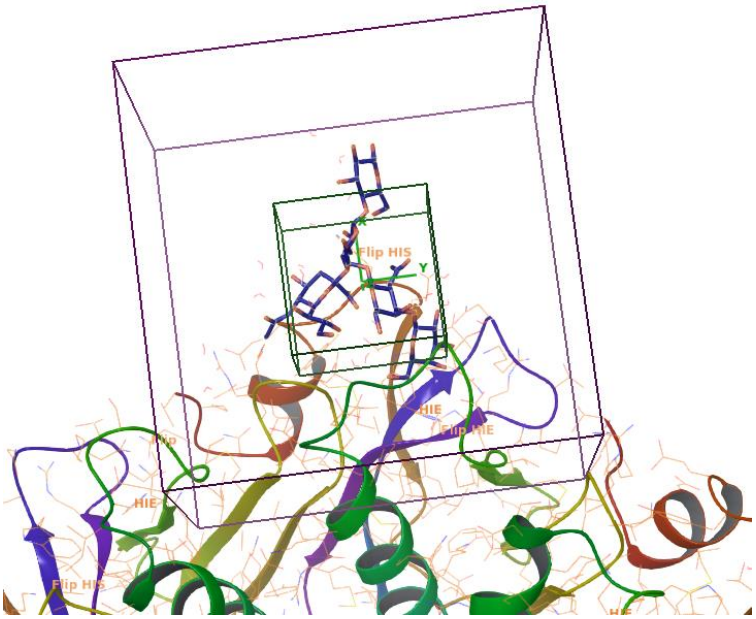

Figure 2.7. Example of grid generation in the case of cholera toxin subunit B with its natural ligand GM1 (from PDB ID: 3CHB). The inner grid is lined in green and the outer grid is lined in purple.

- Docking calculations

The main objectives of the actual calculations are: i) to define the best *location* for the ligand on the surface (within the grid), ii) to sample the conformational space of the ligand and iii) to compute the energy relative to the interaction, giving a higher score to the most favourable ones. To this purpose, different minimization and filtering steps may be implemented in the docking protocol from software to software. Generally, the starting point is a conformational search on the ligand, from which the lowest energy conformations are then minimized in the field of the receptor, using a force field which can be chosen (usually the Optimised Potential for Liquid Simulations for All Atoms, shortened to OPLS-AA). Finally, the resulting poses are subjected to a Monte Carlo procedure that examines torsional minima, before providing the final solution. This must be considered as a "funnel" process, where the number of poses decreases stepwise.

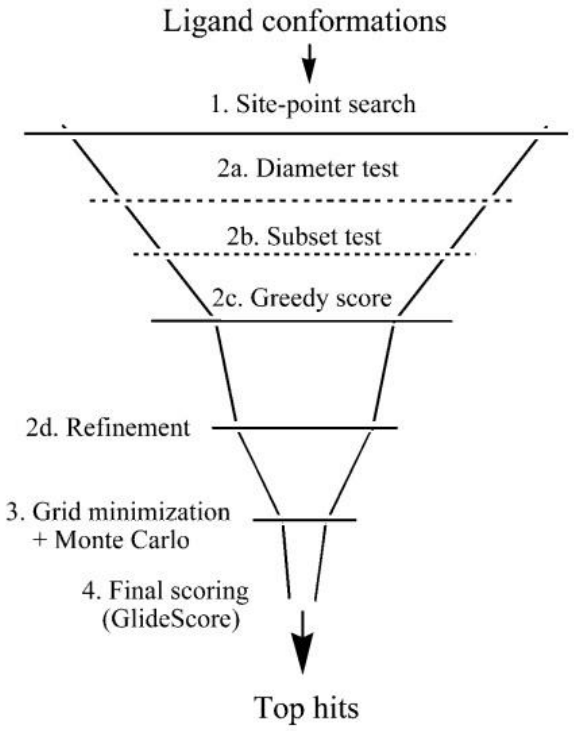

Figure 2.8. Glide docking "funnel" showing the protocol followed to generate docked poses. Figure from 76.

The algorithm shown in Figure 2.8 refers to the Glide docking method, which is the main programme used in the current project, but the general procedure is similar among the different software available on the market (*e.g.*, DOCK, FlexX, AutoDock, AutoDock Vina, *etc.*). An interesting study published by the group of Imberty reported a comparison of

three software (DOCK, AutoDock and Glide) to predict the binding mode of flexible carbohydrates in shallow lectins binding site. They found that Glide is the most successful to account for the flexibility of the oligosaccharides and to represent the complexity of their binding modes[65].

- Scoring functions

The scoring of the poses after the calculation is crucial to rank the solutions according to energy and to compare one another. Ideally, one would require assessing the relative free energy of binding (ΔG) of the ligand-receptor interactions generated, which is a computationally demanding task. Thus, empirical scoring function approximating ΔG have been developed to enable ranking[77]. They consider mainly solvation, ionic interactions and hydrogen bonds to yield reliable approximations. Several functions are available, and they serve diverse needs. "Glide gscore" in Glide, for example, is particularly suitable to compare binding of different ligands to one receptor (*e.g.*, to define the best binder among many); on the other hand, "Glide emodel", weighting more significantly the force fields component such as electrostatic and van der Waals energies, is more convenient to compare conformations of the same ligand to its receptor.

2.3 Model validation. COmplete Relaxation and Conformational Exchange MAtrix approach: CORCEMA-ST[66,78]

A quantitative assessment of the goodness of a 3D molecular model of a protein-ligand complex in relation to the experimental STD NMR data requires a tool able to theoretically predict STD intensities from the Cartesian atomic coordinates of the 3D molecular model. The theoretical framework for this quantitative validation of STD NMR experiments is CORCEMA-ST (Complete Relaxation and Conformational Exchange Matrix for Saturation Transfer), developed by Rama Krishna in 2002. Written as a MatLab code, CORCEMA-ST can predict the STD intensities for the protons of a ligand at different saturation times, upon input of the Cartesian co-ordinates of the structure of the ligand-receptor complex obtained by any experimental or computational method (*e.g.*, X-ray crystallography, NMR, docking simulations) (Figure 2.9).

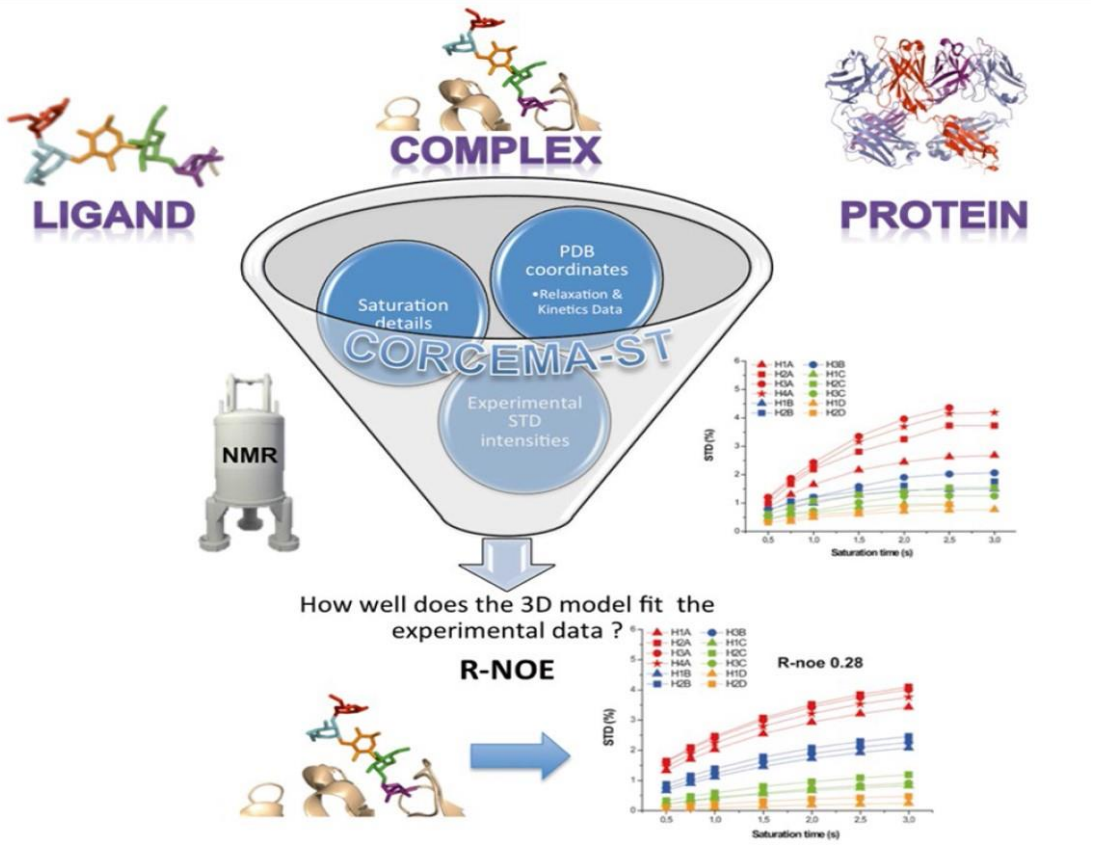

Figure 2.9. Cartoon representation of the CORCEMA-ST protocol. Figure from 78.

CORCEMA-ST calculates the observable magnetization transferred in the protein-ligand NOE occurring when the protons of the protein are irradiated to saturation. This is achieved by solving the matrix differential equation which describes the intensity of magnetization (I_z) as a function of the saturation time (t_{sat}), including the kinetics of the magnetization transfer (motion of magnetisation Q) combined with the kinetics of chemical exchange (K) and that of relaxation R_1 (R). The matrix equation is of the form:

$$I_z(t_{sat}) = I_z(0) + \frac{[1-e^{-(R+K)*t_{sat}}]\,Q}{R+K} \qquad \text{Equation 2.6}$$

Hence, the transferred magnetisation is translated into STD factors and build-up curves can be plotted for the different ligand protons and compared with the experimental ones to prove whether the 3D model fits with the STD NMR data.

The goodness of the fit between the molecular model and the experimental NMR data can be quantified by the so-called NOE R-factor:

$$NOE\ R-Factor = \sqrt{\frac{\sum(STD_{\ exp,i} - STD_{\ calc,i})^2}{\sum(STD_{\ exp,i})^2}} \qquad \text{Equation 2.7}$$

where $STD_{exp,i}$ and $STD_{cal,i}$ are the experimental and calculated STD intensities of each proton i, respectively. By its definition, the NOE R-factor can assume any positive value, with lower values corresponding to better fit between experimental and theoretical data. Thus, different theoretical structural models derived, *e.g.*, from different docking runs, can be ranked according to how well they explain the experimental STD NMR data. NOE R-factors below 0.3 are regarded as satisfactory to validate the 3D structure[79].

The preparatory steps to run CORCEMA-ST calculations are:

- preparation of the files containing the structure of the complex, ligand, and protein;
- preparation of a text file containing the NMR chemical shifts (simulated from ShiftX (www.shiftx2.ca) or experimentally available) for all the protein protons;
- definition of the parameters relative to the system in analysis in the MatLab script.

To limit the size of the matrix on which to make calculations, one defines a cut-off around the ligand. 5 Å is usually the lowest limit to be used, but 8-10 Å give more accurate results. To account for the kinetics of the chemical exchange, the equilibrium constant of the interaction must be entered; the concentration of protein and ligand must also be specified in order to calculate the distribution of the species in solution. To represent the spin diffusion and relaxation effect, it is important to specify the correlation times of the free ligand and of the complex. Relaxation can be also affected by non-specific relaxing agents present in the sample, such as traces of paramagnetic elements in solution: this is described by the parameter ρ (expressed in Hz), for which higher values corresponds to the stronger relaxation effect on all the nuclei in solution (resulting in a lower overall intensities of the signals). These parameters are used to simulate the T_1 values for the relaxation of the ligand's protons, which can also be input manually, if they are experimentally available[80]. Finally, the experimental conditions relative to the spectrometer and to the acquisition must be specified (*i.e.*, the field strength of the instrument, the frequency range of irradiation and the saturation times used).

Chapter 3

Novel Saturation Transfer Difference (STD) NMR approaches: DEEP-STD NMR and IL-STD NMR

Parts of the findings from this Chapter are published as:

Monaco, S.; Tailford, L. E.; Juge, N.; Angulo, J., Differential Epitope Mapping by STD NMR Spectroscopy To Reveal the Nature of Protein–Ligand Contacts. *Angewandte Chemie International Edition* **2017**, 56, 15289-93.

3.1 Introduction

3.1.1 From bulk magnetisation to NOE in protein-ligand systems

As discussed in Chapter 1, Sub-section 1.3, we can think of macroscopic magnetisation as resulting from the difference in population (N) between $2I+1$ states, which generates a bulk magnetisation vector M_z aligned with B_0, *i.e.* along the *z* axis. The acquisition of NMR experiments relies on a perturbing event (application of a radiofrequency) misaligning the bulk magnetisation from its position of equilibrium. Thereafter, recovery of the equilibrium conditions occurs, which is a universal driving force in all natural systems. In the case of nuclear spin perturbation, the *rate* of this recovery (dM_z/dt), *i.e. relaxation*, is of paramount importance. Relaxation is favoured by physical phenomena inducing transitions. Among them, if two nuclei called *I* and *S* are close enough through space, then *cross relaxation* -induced by *dipole-dipole interactions*- is particularly of interest for us, as i) its rate is dictated by the motional regime of rotational diffusion, and ii) the kinetics of cross relaxation, or the rate constant of cross-relaxation σ_{IS} is (inversely) dependent to the distance between the interesting nuclei. As already derived from the Solomon equations (in Chapter 1, Sub-section 1.3.3, Equation 1.32 and 33):

$$\sigma_{IS} = b^2 \frac{3}{10} j(2\omega_0) - b^2 \frac{1}{20} j(0) \qquad \text{Equation 3.1}$$

where:

$$b = \frac{\mu_0 \gamma_I \gamma_S \hbar}{4\pi r^3} \qquad \text{Equation 3.2}$$

and τ_C is the correlation time, γ is the gyromagnetic ratio of each nuclei, μ_0 is the permeability of vacuum and *r* is the inter-nuclear distance. In fact, cross relaxation is responsible for the Nuclear Overhauser Effect (NOE)[55]. Detecting NOEs is crucial for structural studies by NMR, allowing complicated assignments of natural products, distance measurements (if reference distances are available) and hence obtaining constraints definitions for protein structural calculations. Moreover, it plays an irreplaceable role in the investigation of chemical exchange processes, such as the protein-ligand recognition processes which are our objects of study. We already described our protein-ligand system in exchange as follow:

$$P + L \leftrightharpoons PL \qquad \text{Equation 3.3}$$

Which, from the ligand perspective rewrites as:

$$L_f \leftrightharpoons L_b \qquad \text{Equation 3.4}$$

To observe the ligand in its bound state, we must look at NMR variables which vary dramatically when the ligand interacts with the protein. It should be clear at this point that the correlation time τ_C is an ideal variable for this purpose. Therefore, a profound understanding of the NOE phenomena, the process/es leading to "spin diffusion" and their modulation in systems under fast exchange conditions is essential to be able to optimise ligand-based techniques and to generate new developments. Based on this theoretical knowledge, the work presented in this chapter aims at expanding the already brilliant potentialities of Saturation Transfer Difference (STD) NMR, beyond its current limits.

3.1.2 The process of protein-ligand saturation transfer and the long-standing question of protein spin diffusion homogeneity

We can define the STD NMR experiment as a truncated driven NOE (TOE) experiment[71]. It thus relies on the generation of non-equilibrium magnetisation (selective irradiation) on a small group of receptor nuclei, followed by transfer of non-equilibrium population all over the nuclear spins of the protein, and eventually leading to intermolecular NOEs with the closest nuclei of the ligand in the bound state[70a]. For the purpose of this

discussion, we can outline the STD NMR experiment in four processes at the atomic level, as cartooned in Figure 3.1: i) a selective shaped pulse of variable length is applied to a given set of protein protons to perturb the equilibrium, long enough as to induce saturation; ii) by dipole-dipole direct NOE, the irradiated protons transfer magnetisation to the adjacent protons (which can belong to the protein or to the ligand); iii) by dipole-dipole relayed NOE (spin diffusion), the protein nuclei transfer magnetisation to the whole network of nuclei in an intramolecular process; iv) if a small molecule binds to the surface of the protein, the magnetisation is transferred to the interacting ligand nuclei via an intermolecular process. Quantification of the relative distribution of saturation transferred among the different ligand protons yields the binding epitope mapping of the ligand, which allows to identify the main contacts of the ligand with the protein residues in the complex[70b].

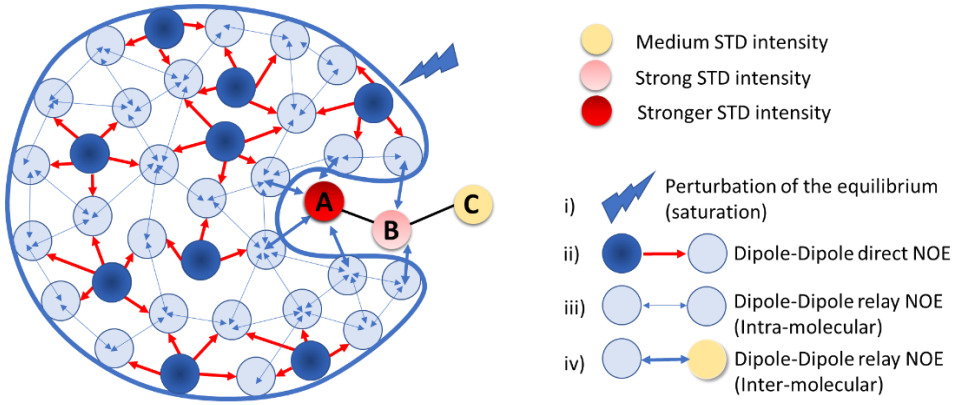

Figure 3.1. Cartoon of the sequence of events taking place during an STD NMR experiment. The outcome is the binding epitope mapping of the ligand constituted by the three different moieties A-B-C, reporting on its binding mode.

However, this approach fails to provide distinctive information about the ligand location or orientation in the bound state; and on the presence of ligands bound in adjacent subsites. Additional experiments are required for these aims: to characterize the location of binding, competitive STD NMR experiments with "spy" ligands of known binding location must be carried out[81,82]; while to identify the orientation of binding in relation to other bound fragments, ILOE tr-NOESY experiments must be performed[83,84].

Importantly, STD relies on the efficacy of spin diffusion to homogeneously spread the saturation over the receptor, but this is not always ensured and, indeed, differences in ligand binding epitope mappings can be observed if STDs are acquired at different

saturating frequencies[66]. Here, we will try to show the fact that small inhomogeneities in saturation, under appropriate experimental setup, can deliver more specific information on the intermolecular interactions than just simply the ligand epitope, but it can reveal the nature of the protein amino acids *where* the saturation comes from.

3.1.3 Protein spin diffusion inhomogeneity: can it be exploited to get structural information?

Our idea is to exploit the differences in ligand epitope mappings arising from acquiring sets of two experiments under different conditions (*e.g.*, different irradiation frequencies or different solvent composition). The differences in the resulting epitopes, albeit expected to be small, can be magnified and mapped into the ligand to provide additional layers of structural information on the architecture of the binding site compared to the standard STD NMR experiments. We then devised two independent approaches:

- *Approach 1:* Exploring the nature of the protein residues around the ligand in the bound state (lining the binding pocket). Elucidating their nature should facilitate the determination of the orientation of the ligand relative to them (Figure 3.2), if the 3D structure of the binding pocket is known;
- *Approach 2*: Exploring the presence of different bound ligands in proximity to each other, in the cases of ternary protein-ligands complexes. This approach should facilitate elucidating the relative orientation of the ligands in the ternary complex (Figure 3.3).

Approach 1, which we call DiffErential EPitope mapping STD NMR (DEEP-STD NMR), consists of two implementations a) differential irradiation frequencies and b) differential solvent composition. These are based on the following hypotheses, respectively:

- *Hypothesis A: Ligand protons will receive larger amounts of saturation if the protein protons in their proximity are "directly irradiated" instead of being "relayed-NOE" saturated* (Figure 3.2a).
- *Hypothesis B: Ligand protons close to slow-exchanging protein exchangeable protons will receive less saturation if the latter are exchanged to deuterium, i.e.*

running the experiments in D_2O, instead of proton, i.e. running them in H_2O (Figure 3.2b).

Implementation of these two hypotheses is expected to provide distinctive information on the nature of the protein protons (aromatic, polar, or apolar) the ligand is in contact with in the bound state, allowing to get some information on the pharmacophore of the protein, as we discuss more in detail in Section 3.3. It is worth to highlight that, with this approach, we might get access to information on the pharmacophore of the protein using a single ligand experiment, in contrast with the considerable number of active molecules that are required for classic ligand-based determination of a protein pharmacophore[85,86].

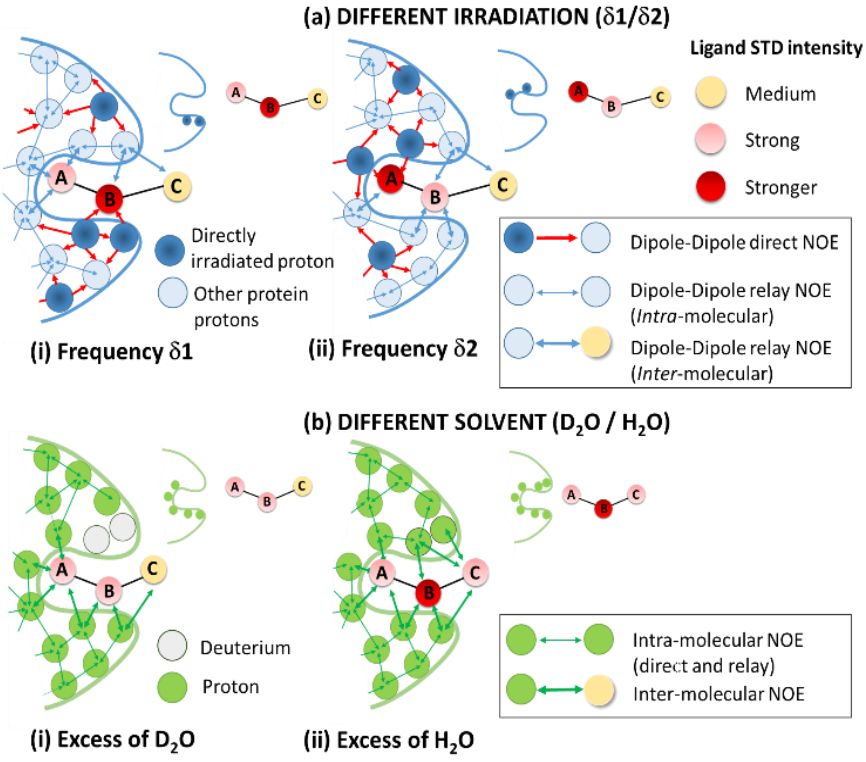

Figure 3.2. Cartoon representing the two implementations of the DEEP-STD NMR protocol. **(a)** Different irradiation frequencies: **(i)** STD NMR with selective irradiation (δ1) on protein protons close to ligand proton B. **(ii)** STD NMR with selective irradiation (δ2) on protein protons close to ligand proton A; the distinct binding epitopes of the ligand are sketched in the free state. **(b)** Different solvent composition: **(i)** STD NMR experiment in D_2O. **(ii)** STD NMR experiment in H_2O. Figure adapted from 87.

Approach 2, which we call Inter-Ligand STD NMR (IL-STD NMR), is a variation of the differential irradiation DEEP-STD NMR approach applicable to ternary complexes (one protein and two or more ligands), where ligand protons and protein protons are directly irradiated alternatively (*on-ligand vs. off-ligand irradiation*). We postulate the following:

- *Hypothesis C: In a ternary complex composed of one protein and two ligands X and Y, protons of ligand X close to protons of ligand Y will receive larger saturation when adjacent Y (and protein) protons are directly irradiated (or vice-versa), compared to when only the protein is irradiated* (Figure 3.3).

It must be considered that the magnetization can be transferred from one ligand to the other either directly (intermolecular through space direct NOE) or through an intermediate protein proton placed close to both (intermolecular relayed NOE, or protein-mediated spin diffusion). Nevertheless, in both cases, the observation of IL-STD necessarily originates from proximity in the space between those protons from different ligands. In this way, detection of IL-STDs is expected to deliver information facilitating the determination of the relative orientation of ligands bound in adjacent binding subsite.

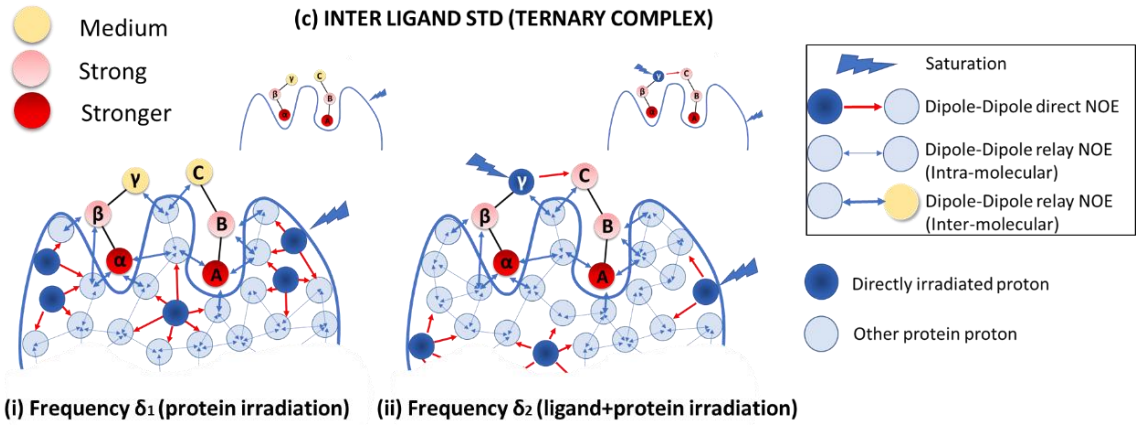

Figure 3.3. Cartoon representing the IL-STD NMR approach. **(i)** STD NMR with selective irradiation (δ1) on protein protons. **(ii)** STD NMR with selective irradiation (δ2) on ligand proton γ (close to proton C of the adjacent fragment) as well as on protein protons.

This is because an STD increase on a given proton (or set of protons) of ligand X when a given proton (or set of protons) of ligand Y is directly irradiated (or *vice-versa*) would arise when these "regions" of the two ligands are closer in space than any other region of the molecules. This is discussed in detail in Section 3.4.

3.1.4 Model systems: brief introduction

To validate our hypotheses, we chose three biologically relevant protein-ligand model systems for which high resolution 3D structures of the complexes from X-ray diffraction data were available. The two DEEP-STD NMR approaches (differential irradiation and differential solvent composition) have been applied to:

i) The complex between 2,7-anhydro-Neu5Ac and GH33, the glycosyl hydrolase domain of the intramolecular (IT) *trans*-sialidase from *Ruminococcus gnavus* (IT-sialidase from *Rg*NanH, PDB ID: 4X4A)[17]. The compound 2,7-anhydro-Neu5Ac is a Neu5Ac derivative, resulting from the intramolecular trans-glycosilating activity of GH33. So far, only three IT-sialidases have been discovered, and their role of mucin-degraders make them a subject of great interest in gut microbiota fundamental research.

ii) The complex between 3-nitrophenyl-α-D-galactopyranoside (3NPG) and Cholera Toxin Subunit B (CTB, PDB ID: 1EEI). 3NPG is a well-known CTB inhibitor that occupies the galactose binding subsite of the GM1 pocket, mimicking the galactose of the natural binder[88]. As one of the first and simplest CTB ligands developed, it is frequently chosen as a reference (and/or starting material) in CTB inhibitor drug design projects[89,90,91,92].

The IL-STD approach has been tested on:

iii) The complex of Naproxen with Bovine Serum Albumin (BSA, PDB ID: 4OR0). The physiological importance of Serum Albumin and its remarkable ability of binding endogenous and exogenous compounds determined the usefulness of this protein as a model for the study of ligand protein interactions[93]. In complex with Naproxen, a non-steroidal anti-inflammatory, and analogous small molecules have been used for optimisation of many experimental techniques, like ligand-based NMR experiments[94,95]. BSA has three Naproxen binding sites, two of which adjacent to each other and with affinities compatible with strong STD detection[96].

3.2 Structural information from protein spin diffusion inhomogeneity - Preliminary data

3.2.1 Binding epitope anomalies and "on-resonance scanning"

Serendipitously, during our studies of the interactions of the Carbohydrate Binding Module domain of *Rg*NanH (CBM40) with Neu5Acα6Lac (6'SL, Figure 3.4a,b), we obtained the first results that indicated that relevant structural information can indeed be encrypted in the spin diffusion inhomogeneity. CBMs are small carbohydrate recognising domains associated to glycosidases. We were analysing the ligand specificity of CBM40 (21 kDa) through STD NMR investigation of a library of 15 sialic acid-containing ligands, leading to the structural elucidation of key features of their binding to CBM40 (whole study reported in Chapter 5, Section 5.3)[97]. The binding pocket of CBM40 is quite shallow (Figure 3.4b) and the main element of recognition for all the ligands is the non-reducing sialic acid capping (Figure 3.4a). The binding of CBM40 to 6'SL (and to all the other ligands studied) can be considered weak (transient), with K_D in the mM range, as quantified by ITC.

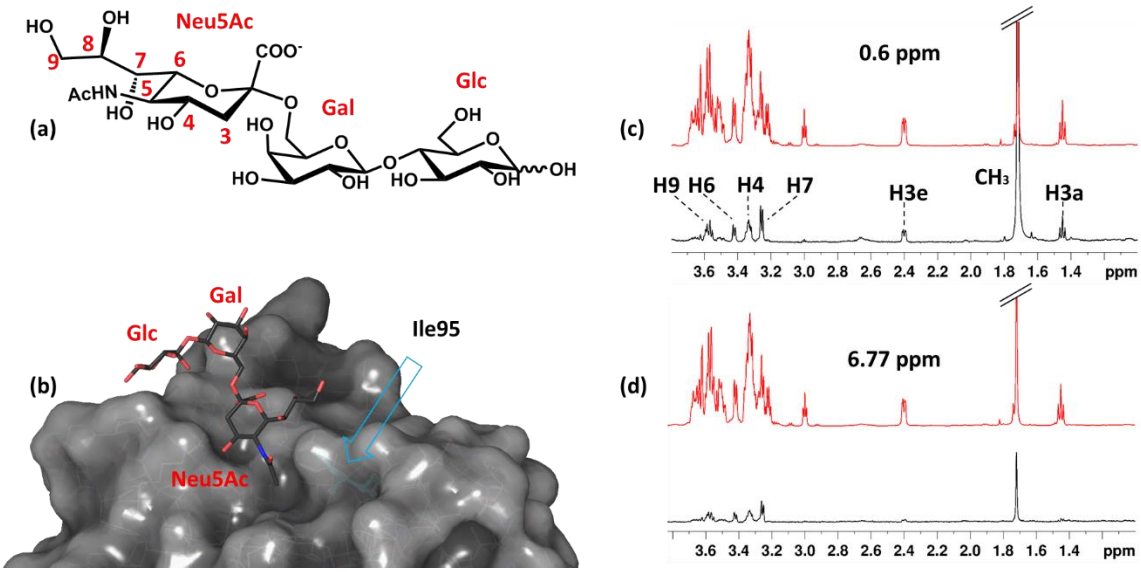

Figure 3.4. (a) Chemical structure of Neu5Acα6Lac (6'SL). **(b)** Crystal structure of the complex 6'SL/CBM40 (PDB ID: 6ER4[97]): the binding site is shallow and the Neu5Ac is the main recognition element. Ile95 is coloured in turquoise and indicated by an arrow (see Figure 3.5c for more details). **(c)** STD NMR experiment at 0.60 ppm and 2 s saturation time: reference spectra (x 1) in red and difference spectra (x 4) in black. **(d)** STD NMR experiment at 6.77 ppm and 2 s saturation time: same colouring scheme and scaling as **(c)**. The signals coming from the sialic acid (main

element of recognition) are labelled. The relative ratio between the signal from the methyl group and the rest of the sialic acid signals is visibly higher when irradiating at 0.60 ppm.

The binding of 6'SL to CBM40 could be detected by STD NMR, with STD (%) signals for the sialic acid protons in the range of 10% when CBM40 was irradiated at a frequency of 0.60 ppm (where most of aliphatic resonances are found) and 2 s saturation time. STD signals from the lactose moiety where weaker as this part of the ligand is loosely bound. Interestingly, the methyl group of the sialic acid moiety showed a significantly higher STD % (60%, Figure 3.4c). The over 5-fold difference observed between the strongest STD signals of the methyl group and the signal of the rest of the protons lead us to name this effect the "STD incremental effect". However, when the same sample was irradiated at 6.77 ppm (in the aromatic region of the protein) the result was a stark drop in the STD intensity of the methyl group, relative to the rest of the signals (Figure 3.4d). The same behaviour of the STD intensity of the methyl group in STD NMR experiments at two different frequencies (0.60 ppm and 6.77 ppm) was observed for all the other sialic-acid containing binders (as reported in Chapter 5, Section 5.3).

To exclude the likelihood of direct ligand irradiation during irradiation at the aliphatic region, *i.e.* at 0.60 ppm (the methyl group resonates at 1.74 ppm), STD NMR experiments at 2 s saturation time were performed at increasing irradiation frequencies with 0.20 ppm increments. This meant scanning the spectral region from -0.20 ppm to 1.40 ppm, in what we call "on-resonance scanning" (Figure 3.5a). The STD intensities for the methyl protons and for H6 of the sialic acid (as a reporter signal) were recorded, showing a very interesting trend: after an increase of the STD signal of the methyl group from 60% to 80% moving from 0.60 ppm to 0.80 ppm, there was a drop between 1.00 ppm and 1.20 ppm. The signal then slightly increased at 1.40 ppm, where we were only 0.40 ppm away from the methyl resonance frequency itself and direct irradiation was certainly playing a role. On the contrary, the control peak H6 stayed almost steady along the spectral window (Figure 3.5a). This suggests that the Neu5Ac methyl group reaches a maximum STD intensity between 0.80 ppm and 1.00 ppm.

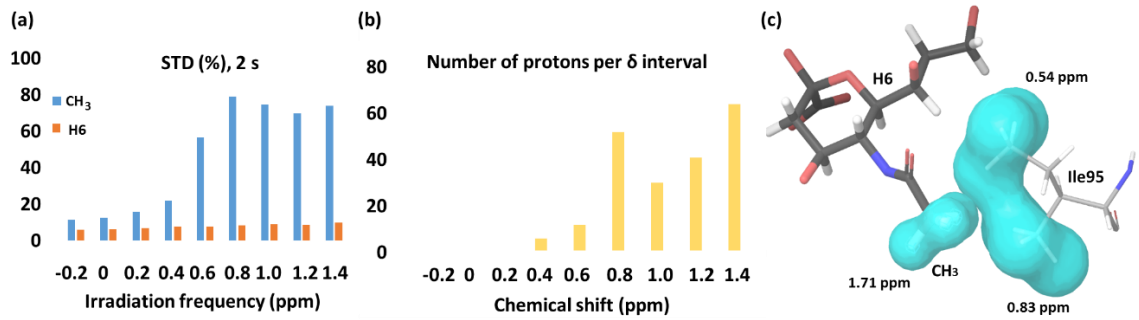

Figure 3.5. (a) "On-resonance scanning" from -0.2 ppm to 1.4 ppm for the CBM40/6'SL complex. For each experiment, the absolute STD intensity is recorded at 2 s saturation time. **(b)** Predicted distribution of protein protons resonating at each 0.20 ppm interval, as obtained from ShiftX2 on the crystal structure of CBM40 at the experimental conditions. **(c)** Detail of the crystal structure showing the relative orientation of the sialic acid of 6'SL and Ile95 (PDB ID: 6ER4[97]). The turquoise surface encloses the directly irradiated protein protons and the ligand methyl group object of the analysis.

We identified the two methyl groups of Ile95 at the binding pocket of CBM40 as having predicted resonance frequencies (by ShiftX2) at 0.84 ppm and 0.58 ppm. Ile95 is pointing the two methyl protons towards the ligand methyl group in the bound state (Figure 3.5c), and thus, potentially direct irradiation of those Ile95 protons along the scanning were responsible for the behaviour of the Neu5Ac methyl group.

This data helps ruling out that the "STD incremental effect" observed for the methyl group of Neu5Ac is given by direct irradiation (further evidence to this argument is provided in Chapter 5, Sub-section 5.3.2). More importantly, this data gave ground to formulate *Hypothesis A*: arguably, as the isoleucine methyl protons are being directly irradiated, thus they can transfer an enlarged amount of saturation to the Neu5Ac's methyl group, very close in space in the bound state, in comparison to other frequencies, hence boosting its STD intensity (Figure 3.5c). In this case, central irradiation of the isoleucine seemed to be achieved at 0.80 ppm in agreement with the ShiftX2 prediction (at 0.60 ppm the effect is also appreciable).

To our best knowledge, the data shown in Figure 3.5a is the first reported evidence that selective irradiation of residues lining the binding pocket and contacting the ligand can help identify which part of the ligand is closer to those amino acids.

This was a preliminary result of paramount importance for us, as it paved the way to carry out differential frequency irradiation experiments, which, applied to the optimal systems, lead to the formulation of DiffErential EPitope mapping STD NMR (DEEP-STD NMR), object of Section 3.3.

3.2.2 Further considerations on the "STD incremental effect" observed on the 6'SL/CBM40 complex

The "STD incremental effect" observed on the methyl group of 6'SL as bound to CBM40 can be considered as an anomalous STD outcome. To fully justify such a large effect, a few additional considerations seem necessary before moving to more standard situations (where 5-fold STD differences between the first and the second strong signal are rarely observed). First, it must be considered that CBM40 is a relatively small protein (21 kDa) whose 1D ^{1}H NMR peaks look not as broad as larger-sized proteins, due to its shorter τ_c (estimated to be *circa* 20 ns). For comparison, Figure 3.6 shows the 1D ^{1}H NMR spectra of a larger protein (GH33, τ_c estimated to 50 ns, Figure 3.6a), whose peaks look more like a so-called "protein envelope" (Figure 3.6b). Arguably, sharper protein peaks allow more intense direct irradiation and stronger saturation transfer, magnifying the effect of inhomogeneous spin diffusion for small proteins as compared as large proteins, as sketched in Figure 3.6c,d. The chemical shift dispersion of the protein protons under the effect of T_2 broadening have already been reported by Kang *et al.* to affect strongly the STD outcome (for a large antibody they had to correct CORCEMA-ST calculations to account for line broadening)[98]. Furthermore, the low affinity, the shallow geometry of the subsite and most likely a very fast kinetics of interaction intensify the intensity gap between the saturation received by the methyl group, which is the only one buried in the hydrophobic pocket (containing Ile95, Tyr116 and Tyr210), and the other Neu5Ac protons. Low affinity typically involves short residence time, which could explain the fact that the rest of the Neu5Ac protons receive a considerably lower saturation. This effect is partially cancelled out when irradiation at 6.77 ppm is performed, where a more homogeneous protein saturation is achieved, so there are no "hot spots" in the binding pocket close to ligand protons (like the "hot" methyl groups of Ile95 when irradiating at 0.60 ppm).

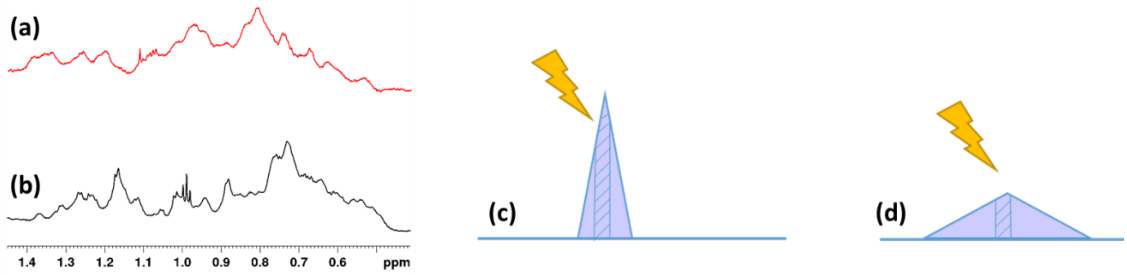

Figure 3.6. Left panel: ^{1}H NMR spectral region (0.10 ppm-1.40 ppm) of the proteins GH33 (52 kDa) **(a)** and CBM40 (21 kDa) **(b)** Fourier transformed with the same line broadening factor of the window function. Due to its smaller size (and τ_c), the peaks of CBM40 look sharper and less "envelope-like". **Right panel**: cartoon representing the T_2 broadening effect on the protein signals. Sharp signals receive much more magnetization when primarily irradiated (and thus transfer it more efficiently to the close ligand protons). Ideally isolating a single protein signal, irradiation pulse of the same width will saturate a much higher percentage of the sharp signal **(c)** than on the broad one **(d)**.

3.3 DiffErential EPitope Mapping STD NMR (DEEP-STD NMR)

3.3.1 Significance and potential of Differential Epitope Mapping by STD NMR

The potential of the preliminary data discussed above is that, in favourable cases, we can tailor the STD NMR experiments to take advantage of the inhomogeneity of the spin diffusion process within the protein, *i.e.* we can select which types of protein protons will be *directly irradiated*. Thus, the differences in binding epitopes will highlight those parts of the ligand contacting a given type of irradiated protein residues (*Hypothesis A*, Figure 3.2a).

Additionally, as pointed out in *Hypothesis B*, another source of minor differences in epitope maps is the solvent. In D_2O, the polar side chains of amino acid in the binding pocket have their exchangeable protons replaced by deuterium, which is inefficient for transferring saturation[56]. In contrast, in H_2O, these protons would be available to receive saturation through intra-protein spin diffusion, and hence be able contribute to an additional transfer of saturation to ligand protons, compared to the STD NMR experiments carried out in D_2O[99]. The efficiency of this process depends on exchange rate of those polar protons with bulk water, with slow exchanging protons in the binding pocket being expected to produce the largest variations[99]. We can exploit this effect to detect clusters of slow exchanging residues (arginine, in particular, as we will explain in

detail in Sub-section 3.3.6) in close contact to the bound ligand, as increases in relative STD intensities of protons located in one part of the ligand, when the STD NMR experiment is carried out in water, will indicate contacts with protons in slow exchange with water.

To detect and quantify these differences we resort to experimentally determining a "*Differential Epitope Map*" of a ligand under two differential conditions (which might be either two different frequencies, or two different solvents, D_2O or H_2O). As we will describe in the following subsection, this is a map of the differences in relative ligand STDs (*i.e.*, epitope differences) from each pair of STD NMR experiments reflecting contacts to specific types of amino acids.

Potentially, if the 3D structure of the protein is unknown, this approach would spread light on the nature *and* architecture of the binding site, allowing to identify possible "patches" (aliphatic, aromatic, polar, *etc.*) within the pocket. In cases for which the 3D structure of the protein is known, this approach could provide the orientation of the ligand in the binding pocket and/or help discriminating where a ligand binds between two alternative binding sites.

3.3.2 Definition of DEEP-STD factor (ΔDEEP-STD) and protocol for acquisition and processing DEEP-STD NMR experiments

The DEEP-STD NMR protocol then consists of running pairs of STD NMR experiments under two different experimental conditions (Figure 3.2), and quantifying their differences in relative STDs. Each pair comprises ***experiment-1*** (*exp1*) and ***experiment-2*** (*exp2*), performed under differential conditions. The "*Differential Epitope Map*" of a ligand under two experimental conditions (*exp1* and *exp2*) is determined through the DEEP-STD factor ($\Delta\mathrm{STD}_i$) for each proton *i* as:

$$\Delta\mathrm{DEEP}-\mathrm{STD}_i = \frac{STD_{\ exp1,i}}{STD_{\ exp2,i}} - \frac{1}{n}\sum_i^n \left(\frac{STD_{\ exp1,i}}{STD_{\ exp2,i}}\right) \qquad \text{Equation 3.5}$$

The first step in the analysis of DEEP-STD NMR data consists in determining which of the two experimental conditions, *exp1* or *exp2*, produced stronger STD intensities (comparing their total sum of STD values). The experiment with the largest total sum

will be called *exp1*. Next, the ratio of STD intensities $\left(\frac{STD\ \%_{exp1,i}}{STD\ \%_{exp2,i}}\right)$ is calculated for each proton of the ligand. These ratios report not only on differences in the epitopes but also on the different global level of protein saturation achieved in both experiments (*e.g.,* protein saturation on the aromatics leads to a reduced global saturation, as the number of aromatic residues is normally much lower than the number of aliphatic ones). For that reason, to obtain the differential epitope, the intrinsic differences in protein saturation must be removed. To that aim, the average ratio of STDs over all protons must be calculated $\left[\frac{1}{n}\sum_i^n \left(\frac{STD\ \%_{exp1,i}}{STD\ \%_{exp2,i}}\right)\right]$. To remove the contribution from differences in saturation level, and to establish a "zero" in the scale for analysis of the differential epitope, this average factor must be subtracted to each of the individual STD ratios. This assures that the $\Delta\mathrm{DEEP}-\mathrm{STD}_i$s reveal only those differences arising from the distinct types of amino acids hit by the saturating radiofrequency. If the change in conditions is affecting the residues lining the binding pocket, we will observe positive values for protons experiencing a relative increase in STD values in *exp1* and negative values for protons experiencing a relative increase in STD values in *exp2*. It is important to highlight that the increase/decrease of STD values is "relative", as it comes from the comparison of the binding epitopes under the two conditions, as expressed in Equation 3.5, and not from the comparison of absolute STD values.

3.3.3 *Differential irradiation* DEEP-STD NMR experiments

We first tested the protocol by analysing the effect of different irradiation frequencies on the binding of 2,7-anhydro-Neu5Ac to GH33[17]. We ran two STD NMR experiments irradiating (0.5 s) at 0.60 ppm (*exp1*) and 6.55 ppm (*exp2*). These frequencies are known to be centred in the aliphatic and aromatic protein proton spectral regions, respectively[100]. It is worth noticing that the selection of frequencies to irradiate different types of protein protons can be based on either known NMR spectral properties of the protein (if chemical shifts are assigned), on NMR databases (*e.g.*, BMRB[101]), or on predictions if a 3D model of the protein is available (as in this study, Table 3.3[101]).

The DEEP-STD factors (0.60 ppm/6.55 ppm) are shown in Figure 3.7a. Positive DEEP-STD factors report relative STD increases when irradiation was at 0.60 ppm (aliphatics); negative ones indicate increases when irradiation was at 6.55 ppm (aromatics). The

resulting differential epitope map is shown in Figure 3.7b. The results clearly show how different protons of the ligand occupy distinct areas of the GH33 binding pocket lined by either aliphatic or aromatic residues. The positive ΔDEEP-STD for CH_3 and H3a suggest vicinity to aliphatic side chains, whereas negative ΔDEEP-STDs for H8, H9 and H9' suggest vicinity to aromatic protons.

The DEEP-STD NMR results shown in Figure 3.7a are in excellent agreement with the published crystal structure of the complex between 2,7-anhydro-sialic acid and GH33 (Figure 3.7c), where the ligand sits between aliphatic (Ile258, Ile338, and Val502) and opposite aromatic patches (Tyr667 and Trp698). The ligand protons CH_3 and H3a point towards the aliphatic residues, while H8, H9, and H9' are projected towards the aromatic side chains. Protons H3e, H4, H5, H6 and H7 sit in between these two regions in agreement with their negligible ΔDEEP-STD factors.

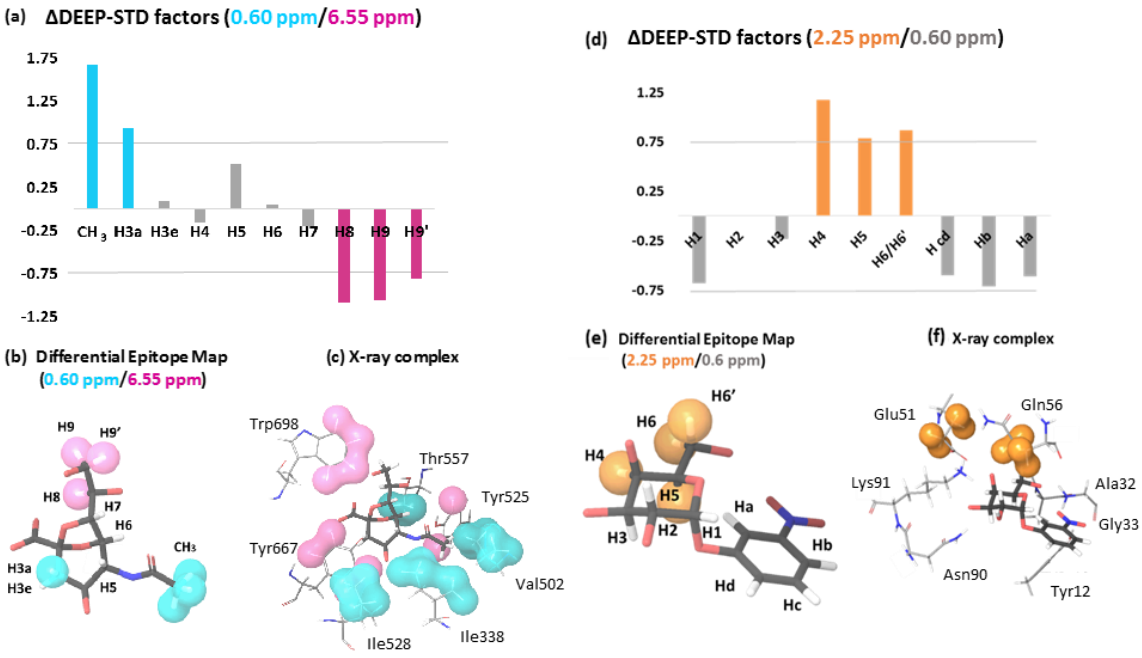

Figure 3.7. **Left panel**: Differential Epitope Mapping (0.60 ppm/6.55 ppm) of 2,7-anhydro-Neu5Ac in complex with GH33. **(a)** ΔDEEP-STD histogram: positive ΔDEEP-STD (above the limit of +0.75) after aliphatic irradiation (0.60 ppm) are in cyan, and negative ΔDEEP-STD (below -0.75) after aromatic irradiation (6.55 ppm) are in magenta. **(b)** Differential Epitope: ΔDEEP-STD map of the ligand. Cyan surfaces highlights ligand contacts with aliphatic side chains; magenta, contacts with aromatic side chains. The ligand polar protons have been omitted. **(c)** Crystal structure of the complex (PDB ID: 4X4A[17]). Protein protons are coloured as a function of their NMR frequencies: directly irradiated at 0.60 ppm (aliphatics) in cyan and directly irradiated at

6.55 ppm (aromatics) in magenta (Table 3.3). **Right panel**: differential Epitope Mapping (2.25 ppm/0.60 ppm) of 3-nitrophenyl-α-D-galactopyranoside (3NPG) in complex with Cholera Toxin Subunit B (CTB). **(d)** ΔDEEP-STD histogram**:** protons with positive ΔDEEP-STD (above the limit of +0.75) after irradiation at 2.25 ppm are in orange. **(e)** Differential Epitope**:** ΔDEEP-STD map of the ligand. Orange surfaces indicate ligand contacts with protein side chains directly irradiated at 2.25 ppm. The ligand polar protons have been omitted. **(f)** Crystal structure of the complex (PDB ID: 1EEI[88]). Protein protons directly irradiated at 2.25 ppm are enclosed in orange surface. For both systems, protons were added using Schrodinger software[102]. Comparison of **(b)** and **(c)** highlights the excellent match of the differential epitope map of the ligand with the distribution of the residues in the binding pocket. Figure adapted from 87.

Further, we applied the novel protocol to the complex of CTB, a larger receptor (65 kDa), with 3NPG[88]. Unfortunately, the ligand contains an aromatic moiety, which precludes protein irradiation in this spectral region. However, in DEEP-STD NMR, it is possible to select other groups of protein protons for irradiation, providing that they are in spectral regions devoid of ligand signals. For CTB, we targeted protein resonances at 2.25 ppm, where no ligand protons showed signals, and protein signals were available for saturation. We predicted the chemical shifts of protons of CTB within 4 Å of the ligand in the X-ray structure, and the results indicated Glu51 and Gln56 protons as the ones likely to be directly irradiated (Figure 3.7f, and Table 3.3). Experiments conducted with differential frequencies (2.25 ppm/0.60 ppm), resulted in positive ΔSTD values for protons H4, H5, H6 and H6' on the galactose, indicating an increase in relative STDs when irradiating at 2.25 ppm (Figure 3.7d). In contrast, negligible ΔSTD factors were observed for H1, H2, H3, and the aromatic protons at the opposite end of the molecule.

The differential epitope map of 3NPG (Figure 3.7e) was found to be in perfect agreement with the published crystal structure of the complex between 3NPG and CTB (Figure 3.7f),[10] in which the galactose ring area of H4 to H6 is surrounded by the side chains of Glu51 and Gln56. In contrast, H1, H2, H3 and aromatic carbons are pointing far from those side chains in the binding pocket (Figure 3.7f). These results confirm that it is possible to identify the nature of the ligand-receptor contacts by means of differential protein irradiation. The spectra and the raw and processed data are shown in Figure 3.8, Tables 3.1 and 3.2.

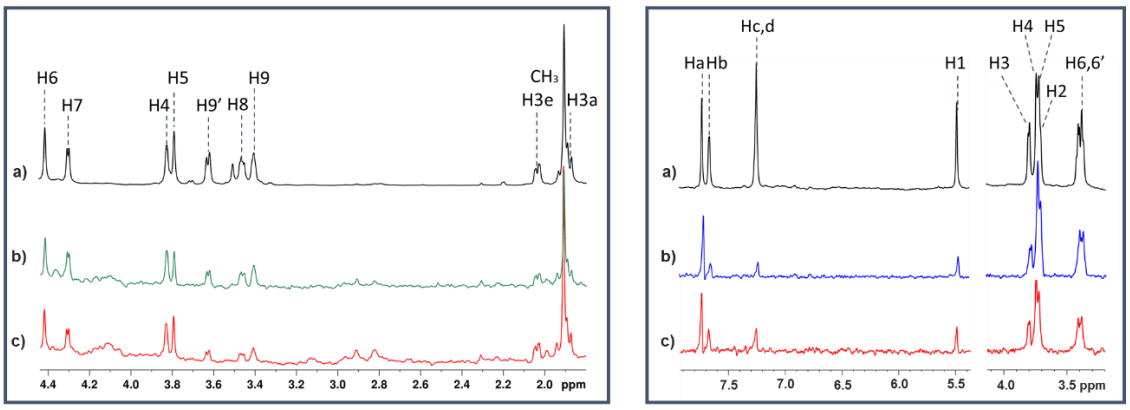

Figure 3.8. Left panel: 1 mM of 2,7-anhydro-Neu5Ac in the presence of 50 μM GH33 in deuterated tris-d_{11} buffer, 293 K. **(a)** Reference spectra (x 1); **(b)** STD NMR spectrum with on-resonance irradiation at 6.55 ppm (x 64); **(c)** STD NMR spectrum with on-resonance irradiation at 0.60 ppm (x 32). **Right panel:** 1 mM of 3NPG in the presence of 10 μM CTB (50 μM binding pockets) in deuterated Phosphate Buffer Saline (PBS buffer), 278 K. **(a)** Reference spectra (x 1); **(b)** STD NMR spectrum with on-resonance irradiation at 2.25 ppm (x 8); **(c)** STD NMR spectrum with on-resonance irradiation at 0.60 ppm (x 16). Figure from 87.

It is worth mentioning here that, which ΔSTD values should be considered significant will depend on the sizes of the STD factors for the protein-ligand system under study. Based on the results of the systems analysed here, we experimentally determined that ΔSTD greater than 0.75 in magnitude were significant.

Proton	^1H δ (ppm)[a]	STD % 0.60 ppm	STD % 6.55 ppm	Ratio STD 0.60 ppm/6.55 ppm	ΔDEEP-STD
CH3	1.90	3.66	0.94	3.89	1.68
H3a	1.89	2.87	0.91	3.15	0.94
H3e	2.04	2.42	1.05	2.30	0.09
H4	3.82	2.84	1.38	2.06	-0.16
H5	3.79	2.58	0.94	2.74	0.53
H6	4.41	2.68	1.18	2.27	0.06
H7	4.30	2.61	1.29	2.02	-0.19
H8	3.41	1.41	1.24	1.14	-1.08
H9	3.46	1.03	0.89	1.16	-1.06
H9'	3.63	1.15	0.82	1.40	-0.81
		Sum	**Sum**	**STD average**	
		23.25	10.64	2.21	

Table 3.1. DEEP-STD NMR using *Differential Irradiation (0.60 ppm/6.55 ppm)* of the GH33/2,7-anhydro-Neu5Ac complex. [a] Spectra for assignment acquired at 293 K. Table from 87.

Proton	^{1}H δ (ppm)[a]	STD % 2.25 ppm	STD % 0.60 ppm	Ratio STD 2.25 ppm/0.60 ppm	ΔDEEP-STD
H1	5.50	6.73	3.66	1.84	-0.68
H2	3.73	12.51	4.95	2.53	0.01
H3	3.81	10.81	4.72	2.29	-0.23
H4	3.75	21.82	5.9	3.70	1.18
H5	3.73	15.21	4.6	3.31	0.78
H6/H6'	3.40	19.79	5.84	3.39	0.87
H cd	7.27	2.81	1.46	1.92	-0.60
Hb	7.69	5.24	2.9	1.81	-0.71
Ha	7.75	12.21	6.38	1.91	-0.61
		Sum	**Sum**	**STD average**	
		107.13	40.41	2.52	

Table 3.2. DEEP-STD NMR using *Differential Irradiation (2.25 ppm/0.60 ppm)* CTB/3NPG. [a] Spectra for assignment acquired at 278 K. Table from 87.

3.3.4 Selection of irradiation frequencies and identification of directly irradiated residues in the binding pocket

For proteins of known 3D structure, directly irradiated protons can be easily identified in the binding pocket if their chemical shift assignments are known. In many cases, however, the protein chemical shifts are not available. Then, statistical averages of chemical shifts of the residues present in the binding pocket (from existing NMR databases) can be used.

GH33 Binding Pocket Residues				CTB Binding Pocket Residues			
258	ILE	HA	4.11	12	TYR	HA	4.63
		HB	2.01			HB2	2.73
		HD1	0.49			HB3	2.90
		HG2	0.98			HD1	7.07
		HG12	1.48			HD2	7.01
		HG13	1.29			HE1	6.8
338	ILE	HA	4.71			HE2	6.83
		HB	1.69	32	ALA	H	8.29
		HD1	0.60			HA	4.16
		HG2	0.84			HB	1.37
		HG12	1.25			HB2	1.73
		HG13	1.14			HB3	1.51
502	VAL	HA	4.70	33	GLY	HA	3.89
		HB	1.74			HA2	3.87
		HG1	0.70			HA3	3.88
		HG2	0.36	51	GLU	HA	4.35
557	THR	HA	3.62			HB2	2.02
		HB	4.10			HB3	2.08
		HG2	0.97			HG2	2.28
525	TYR	HA	5.98			HG3	2.30
		HB2	2.95	56	GLN	HA	4.05
		HB3	2.84			HB2	1.81
		HD1	7.02			HB3	1.93
		HD2	6.93			HE21	7.08
		HE1	6.64			HE22	6.87
		HE2	6.64			HG2	1.94
677	TYR	HA	4.66			HG3	2.02
		HB2	3.05	90	ASN	HA	4.83
		HB3	2.76			HB2	3.02
		HD1	6.82			HB3	2.82
		HD2	6.83			HD21	7.44
		HE1	6.68			HD22	6.97
		HE2	6.63	91	LYS	HA	4.62
698	TRP	HA	4.47			HB2	1.85
		HB2	3.18			HB3	1.72
		HB3	3.01			HD2	1.62
		HD1	6.63			HD3	1.62
		HE1	9.89			HE2	2.96
		HE3	7.25			HE3	2.94

Table 3.3. ^{1}H chemical shifts of amino acid residues in the binding pocket (within 4 Å from the ligand) of GH33 (left) and CTB (right) as simulated by Shiftx2 (http://www.shiftx2.ca). Resonances within 0.4 ppm from the directly irradiated frequencies (*i.e.*, 0.60 ppm and 6.55 ppm in the case of the differential irradiation DEEP-STD studies of the GH33/2,7-anhydro-Neu5Ac

complex, and 2.25 ppm in the case of the CTB/3NPG complex,) are highlighted in the same colour scale as in Figure 3.7. Table from 87.

Alternatively, the 3D structure of the protein can be used to predict the chemical shifts using existing software. Here, we selected the irradiation frequencies based on predictions of chemical shifts using the available 3D structures of the proteins by using ShiftX2 (http://www.shiftx2.ca[100]), as we report in Table 3.3.

Additionally, the chemical shifts of selected protons on the proteins were also checked against average chemical shift histograms from the Biological Magnetic Resonance Data Bank (http://www.bmrb.wisc.edu). All molecular graphics were generated with Schrodinger Maestro 11 version 2016-4. Hydrogen atoms were added to the ligand and to the protein using the *Protein Preparation Wizard* and *LigPrep*, and protonated residues within 4 Å from the ligand where selected as shown in Figure 3.7.

3.3.5 *Differential solvent* D_2O/H_2O DEEP-STD NMR

We next studied 2,7-anhydro-Neu5Ac in complex with GH33 under two different solvent conditions: D_2O (*exp1*), or H_2O (*exp2*). The irradiation frequency was set at 0.60 ppm. In H_2O, the large pool of solvent protons acts as a magnetization sink [103], resulting in a global saturation lower than in D_2O. For this reason, we set *exp1* in Equation 3.5 to be the experiment in D_2O.

Thus, negative DEEP-STD factors correspond to ligand protons with reduced relative STDs in D_2O, that is, being adjacent to slow exchanging polar residues in the bound state. We recorded negative DEEP-STD factors from differential D_2O/H_2O experiments at protons H3a, H3e, H9, and H9' (Figure 3.9a). Here again, the DEEP-STD NMR results were in line with the crystal structure (Figure 3.9c), where the ligand protons H3a, H3e, H9, and H9' point towards a highly polar patch in the GH33 binding pocket (Arg257, Arg276, Arg575, and Arg637).

This is in excellent agreement with the known slow exchanging behaviour of the η protons of arginine residues in H_2O[99]. Interestingly, the ligand methyl group showed a positive DEEP-STD factor (Figure 3.9a), which is explained by the presence of a fast exchanging hydroxyl group (Y525) close in space to the methyl groups of V502 (Figure 3.7c and 3.11)[99].

Finally, we conducted DEEP-STD NMR experiments with differential D_2O/H_2O conditions on the 3NPG/CTB complex. The ΔDEEP-STDs of 3NPG were negligible, and no differential epitope was obtained, indicating that changing solvent did not significantly affect the STD pattern (Figure 3.9d). This is in agreement with the lack of slow exchanging polar residues in the CTB binding pocket (Figure 3.9f)[88]. This suggests that when no differential epitope is obtained after a change from D_2O to H_2O, the protein binding pocket is likely to lack slow exchanging polar side chains (*e.g.* Arg)[99]. The spectra and the raw and processed data are shown in Figure 3.10, Tables 3.4 and 3.5.

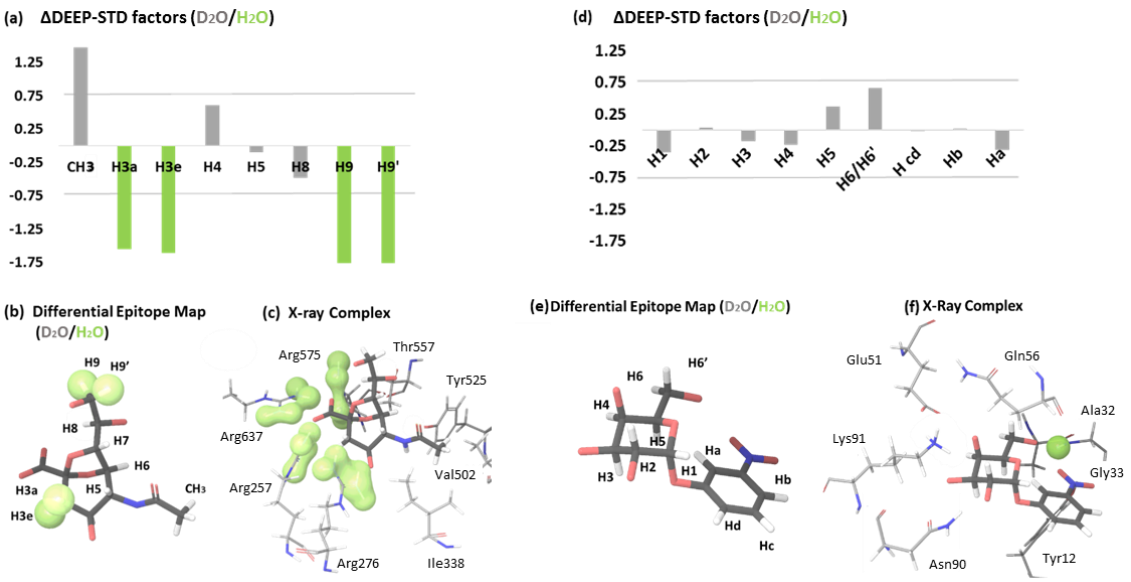

Figure 3.9. Left panel: differential Epitope Mapping (D_2O/H_2O) of 2,7-anhydro-Neu5Ac in complex with GH33. **(a)** DEEP-STD histogram**:** protons with DEEP-STD factor < -0.75 are in green. Protons H6 and H7 were excluded from the analysis, due to their proximity to the water peak, and the use of solvent suppression. **(b)** Differential Epitope**:** DEEP-STD map of the ligand. Green surfaces indicate ligand contacts with the protein side chains carrying slow exchanging protons. **(c)** Crystal structure of the complex (PDB ID: 4X4A[17]). The slow exchangeable protons in the binding pocket are enclosed in a green surface. **Right panel**: differential Epitope Mapping (D_2O/H_2O) of 3NPG in complex with CTB. **(d)** DEEP-STD histogram: all the protons showed values close to 0, indicating that none of the residues in the binding pocket is affected by the change in conditions. **(e)** Differential Epitope: no DEEP-STD map of the ligand was obtained. **(f)** Crystal structure of the complex (PDB ID: 1EEI[88]). The slow exchangeable protons in the binding pocket are enclosed in a green surface. Figure adapted from 87.

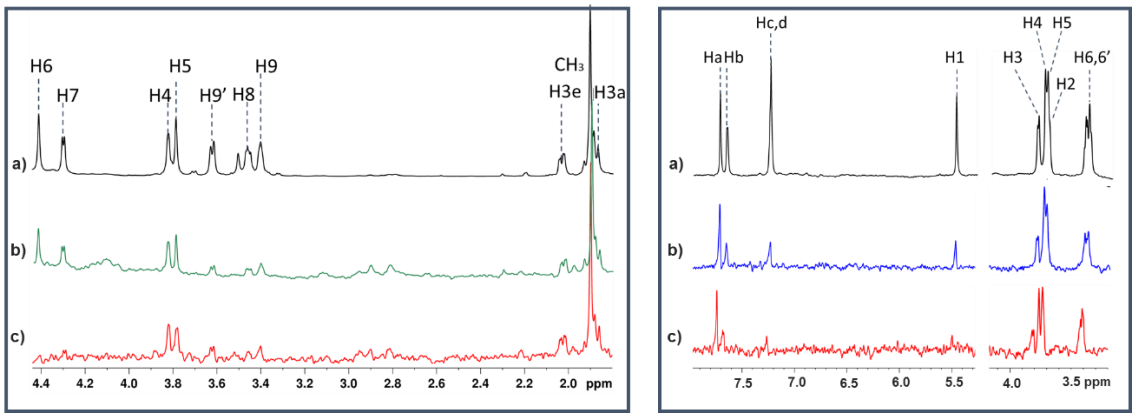

Figure 3.10. Left panel: 1 mM of 2,7-anhydro-Neu5Ac in the presence of 50 µM GH33 50 uM in PBS buffer, 298 K. **(a)** Reference spectra (x 1); **(b)** STD NMR spectrum in D_2O (on-resonance at 0.60 ppm; off-resonance at 40 ppm) (x 32); **(c)** STD NMR spectrum in H_2O:D_2O 90:10 (x 128). **Right panel:** 1 mM of 3NPG in the presence of 10 µM CTB (50 µM binding pockets), 278 K. **(a)** Reference spectra (x 1); **(b)** STD NMR spectrum in D_2O (on-resonance at 0.60 ppm; off-resonance at 40 ppm) (x 16); **(c)** STD NMR spectrum in in H_2O:D_2O 90:10 (x 32). Figure from 87.

Proton[a]	STD % D_2O	STD % H_2O	Ratio STD D_2O/H_2O	ΔDEEP-STD
CH3	3.66	0.51	7.18	1.75
H3a	2.87	0.69	4.16	-1.26
H3e	2.42	0.59	4.10	-1.32
H4	2.84	0.45	6.31	0.89
H5	2.58	0.46	5.61	0.19
H8	1.41	0.27	5.22	-0.20
H9	1.03	0.3	3.43	-1.99
H9'	1.15	0.31	3.71	-1.71
	Sum	**Sum**	**STD ratio average**	
	17.96	3.65	5.42	

Table 3.3. DEEP-STD NMR using *Differential Solvent (D_2O/H_2O)* of the 2,7-anhydro-Neu5Ac/GH33 complex. Table from 87. [a] For assignment see Table 3.1.

Proton[a]	STD % D_2O	STD % H_2O	Ratio STD D_2O/H_2O	ΔDEEP-STD
H1	3.66	1.4	2.61	-0.62
H2	4.95	1.51	3.28	0.05
H3	4.72	1.51	3.13	-0.10
H4	5.9	1.83	3.22	-0.01
H5	4.6	1.24	3.71	0.48
H6/H6'	5.84	1.6	3.65	0.42
H cd	1.46	0.51	2.86	-0.37
Hb	2.9	1.1	2.64	-0.59
Ha	6.38	2.02	3.16	-0.07
	Sum	**Sum**	**STD ratio average**	
	40.41	12.72	3.14	

Table 3.5. DEEP-STD NMR using *Differential Solvent (D_2O/H_2O)* of the 3NPG/CTB complex. Table from 87. [a] For assignment see Table 3.2.

3.3.6 The exchange rate of polar protons and Differential D_2O/H_2O Epitope Mapping

In H_2O, the exchange rate of exchangeable polar protons on amino acid side chains of the protein in the ligand binding pocket has a strong influence on the ability of these protons (and their adjacent non-exchangeable ones) to transfer magnetization to a bound ligand. In H_2O, *slow exchanging* protons will contribute to transfer the magnetization from the binding site to the ligand. In contrast, in the case of *fast exchanging protons*, their exchange with the bulk water will be too fast for them to take part efficiently in the protein-ligand saturation transfer process. In addition, the magnetization of non-exchangeable protein protons in the binding pocket close in space to fast exchanging protein protons will be lost in the bulk water due to an efficient exchange-mediated magnetization leakage (Figure 3.11)[99].

Based on these premises, the presence of slow exchanging protons in the binding pocket will enhance the STD response of ligands protons in close contact with them, if the experiment is performed in H_2O as opposed to D_2O (a red "+" in Figure 3.11). It is worth noting that the enhancement of STD signal is "relative"; that is, after comparison of the "binding epitopes" (relative normalized STDs) under the two conditions (H_2O and D_2O), and not by comparing the absolute STD values (Equation 3.5).

The presence of "isolated" fast exchanging protons will not have a significant effect on the STD of the ligand protons when increasing the percentage of protonated sites over deuterated ones (in light water they will exchange fast and they will be "invisible" in terms of magnetization transfer). However, as mentioned above, the presence of fast exchanging protons near to non-exchangeable protein protons leads to leakage of the magnetization of the latter.

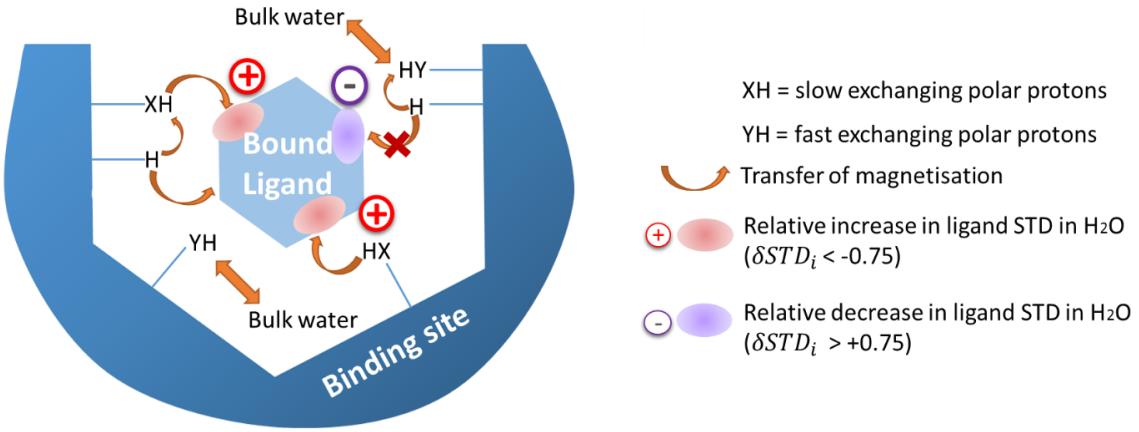

Figure 3.11. Cartoon representing the different pathways for saturation transfe r from slow and fast exchanging polar protons of the binding pocket to the bound ligand in H_2O. **Slow exchanging protein polar protons** contribute to additional STD intensities on the ligand (**plus signs**). **Fast exchanging protons** either do not contribute or can lead to ligand STD reduction due to exchange mediated leakage (**minus sign**). This figure can be considered as a modified version of Figure 1 in [99], adapted to explain DEEP-STD NMR experiments with differential solvent conditions (D_2O/H_2O). Figure from 87.

What is more, if the latter is/are in close contact with ligand protons (X'H-H in Figure 3.11), the result will be a relative reduction of the STD intensities of the closest ligand protons (a purple "-" in Figure 3.11). This explains the positive DEEP-STD factor D_2O/H_2O observed for the methyl group of 2,7-anhydro-Neu5Ac (>+0.75) in our study. In the X-ray structure (PDB ID: 4X4A), the methyl group of V502, which is contributing

significantly to the transfer of magnetization in the STD NMR experiments (Figure 3.9), points toward the hydroxyl group of Y525, which is known to have a very fast kinetics of exchange[99]. This induces an efficient exchange-mediated leak of magnetization from V502 to the bulk water, having the overall effect of decreasing the STD intensity on the methyl group in H_2O (when compared to the same sample in D_2O).

3.3.7 CORCEMA-ST validation of the DEEP-STD data for 2,7-anhydro-Neu5Ac/GH33

To validate the DEEP-STD NMR approach, we tried to reproduce the differential epitope mapping at 0.5 s for the complex 2,7-anhydro-Neu5Ac with GH33, running the CORCEMA-ST calculations simulating the two approaches experimentally followed: i) differential frequency STD (0.60 ppm/6.55ppm), and ii) differential solvent (D_2O/H_2O). Even when no optimization of the parameters needed for the calculations was pursued, the theoretical results produced a pattern of positive and negative ΔDEEP-STDs (Figure 3.12) very similar to what observed experimentally (Figure 3.7a and 3.9a).

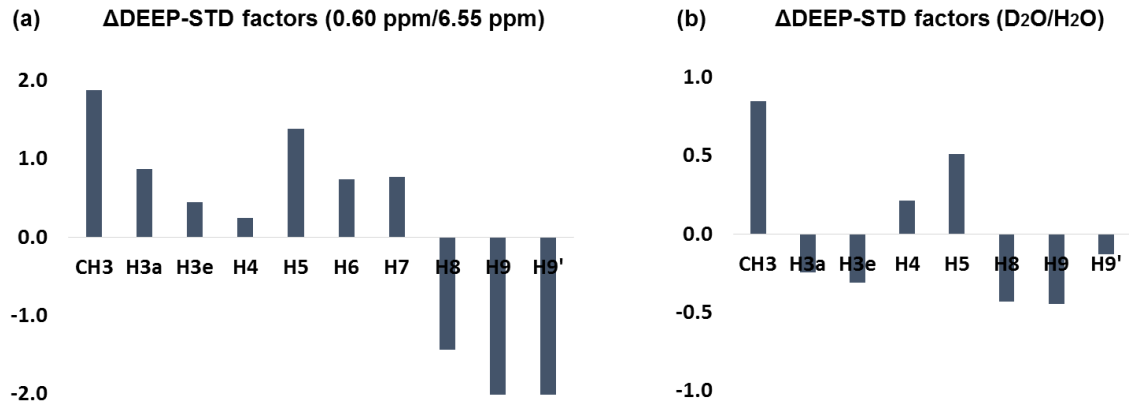

Figure 3.12. ΔDEEP-STD histograms for 2,7-anhydro-Neu5Ac in complex with GH33 calculated on the STD (%) at 0.5 s predicted by CORCEMA-ST: **(a)** differential irradiation ΔDEEP-STD (0.60 ppm/6.55 ppm) and **(b)** differential solvent ΔDEEP-STD (D_2O/H_2O).

These theoretical results further strongly support the proposed DEEP-STD NMR approach, highlighting how depending on the architecture and chemical nature of the amino acids at the binding pocket, as well as depending on the protonation of polar side chains, the transfer of saturation to the ligand is different, and we can pick up those differences.

3.4 Inter-Ligand STD NMR (IL-STD)

3.4.1 Significance of the IL-STD approach and its potential limitations in comparison to the ILOE approach

In Sections 3.2 and 3.3, we demonstrated the validity of *Hypotheses A* and *B* and established the DEEP-STD NMR protocol, as a tool to investigate the nature of the intermolecular contacts in protein-ligand complexes, getting information otherwise inaccessible by traditional STD NMR. The technique relies on harnessing the inhomogeneity of spin diffusion *on the receptor* to "localise" the source of strongest magnetisation from the protein.

Based on *Hypothesis C*, Inter-Ligand STD NMR (IL-STD NMR) can be thought as a reintroduction of the intra- and inter-ligand NOEs in ternary protein-ligand systems (Figure 3.3). The idea is to "directly" irradiate ligand proton(s) and to measure the effect on the neighbouring-ligand protons. Whereas changes in saturation of the "self" protons (of the irradiated molecule) are not structurally useful, variations in saturation of the protons of the other ligand in the ternary complex can reflect spatial vicinity of the two ligands (which here can as well be thought as "fragments", to use a common terminology in drug discovery). Notably, this is the same information available from tr-NOESY experiments carried out on the ternary complexes, an elegant approach developed by the group of Pellecchia and named "ILOE" (Inter-Ligand NOE) [84]. Compared to ILOE, IL-STD NMR is less time-consuming (the long experimental time of tr-NOESY and the complication arising from the acquisition of 2D experiments make ILOE the least popular among ligand based NMR methods in pharmaceutical companies[104]). The disadvantage is the risk of direct irradiation in the spectra, intrinsic to the act of irradiating a single proton within a crowded spectrum.

IL-STD NMR is a valuable tool to validate 3D molecular models of ternary complexes (e.g. generated by multiple ligand docking), as it provides information on how the two fragments are located relative to each other. On the other hand, if the 3D structure of the ternary complex is unknown, the experiment must be repeated for each proton of each ligand to obtain thorough information (as one could not know where to irradiate *a priori*). In both cases, to mitigate the risk of direct irradiation, each set of experiments

must be repeated for each ligand in the absence of the other (*i.e.* on the binary complex). This allows to assess the genuine inter-ligand nature of the relative STD increments observed upon irradiation of one of the two ligands (such controls must also be run in the ILOE approach, to verify that the inter-ligand cross peak does not come from protein-protein or protein-ligand contacts).

3.4.2 Definition of IL-STD factor (ΔIL-STD) and data processing protocol

Like differential irradiation DEEP-STD NMR, IL-STD NMR relies on acquiring pairs of STD NMR experiments at two irradiation frequencies and quantifying their differences in relative STD intensities. We call these frequencies *on-ligand*, when we irradiate on resonance of ligands protons (keeping well in mind that we are at the same time hitting the protein too), and *off-ligand*, when we irradiate on the protein solely.

To obtain positive values for the protons receiving inter-ligand saturation, here the "**experiment-1**" will always be the *on-ligand* and the "**experiment-2**" will always be the *off-ligand*. Thus, the value of interest for the IL-STD data processing is $\frac{STD\ _{on-ligand}}{STD\ _{off-ligand}}$.

The unavoidable presence of some effects of direct irradiation, intra- and inter-ligand, leads to values of $\frac{STD\ _{on-ligand,i}}{STD\ _{off-ligand,i}}$ much larger and non-homogeneous than what observed for the calculation of the DEEP-STD factor. This makes the term $\left[\frac{1}{n}\sum_i^n\left(\frac{STD\ \%_{exp1,i}}{STD\ \%_{exp2,i}}\right)\right]$ not relevant for the IL-STD data processing. This is because it would strongly depend on the number of directly irradiated protons, which varies from situation to situation, therefore subtracting this value to each $\frac{STD\ _{on-ligand}}{STD\ _{off-ligand}}$ would not be generally informative. The term $\left[\frac{1}{n}\sum_i^n\left(\frac{STD\ \%_{exp1,i}}{STD\ \%_{exp2,i}}\right)\right]$ was introduced in the ΔDEEP-STD calculation to remove the contribution from differences in saturation level, and to establish a "zero", as the final aim was the definition of a differential epitope. Changing irradiation from *on* to *off* causes differences in saturation level which are not possible to attenuate.

On the other hand, to detect inter-ligand contacts, our aim is merely to distinguish relative *STD increments* arising from the irradiation of a neighbouring ligand proton. Based on this argument, for a ternary complex formed of a protein and two ligands **X**

and **Y**, we can define an IL-STD factor for each proton *i* of **X** when a given proton of **Y** is irradiated (or *vice-versa*) as:

$$\Delta\mathrm{IL}-\mathrm{STD}_{i,ternary}=\frac{STD_{\ on-ligand,i}}{STD_{\ off-ligand,i}}-N \qquad \text{Equation 3.6}$$

Where *N* is an entire number, defined as the composite function mode of the population of $\left\lfloor\frac{STD_{\ on-ligand,i}}{STD_{\ off-ligand,i}}\right\rfloor$ ($\lfloor x\rfloor$ or *floor function* of *x* is the integer part of any number *x*). If two modes are present, we recommend choosing the lowest value. *N* will in most cases be equal to 1, unless there are factors which increase the $\frac{STD_{\ on-ligand}}{STD_{\ off-ligand}}$ to higher values, as in the example reported in this section. This allows to establish a baseline against which assessing the increase or decrease of each proton (which is similar to what one would do equalising by-eye the intensity of the difference spectra, *on-ligand* and *off-ligand*, against the protons not affected by direct irradiation). Most importantly, each $\Delta\mathrm{IL}-\mathrm{STD}_{i,ternary}$ must be compared to the correspondent $\Delta\mathrm{IL}-\mathrm{STD}_{i,binary}$ calculated under the same *on-ligand* and *off-ligand* irradiation conditions but in the absence of the irradiated ligand (**X** if we are analysing **Y** or *vice-versa*). The protons affected by direct irradiation (or intra-ligand effects) will show a positive $\Delta\mathrm{IL}-\mathrm{STD}_{i}$ in both experiments (ternary and binary); the protons which will have a positive $\Delta\mathrm{IL}-\mathrm{STD}_{i,ternary}$ and a negative $\Delta\mathrm{IL}-\mathrm{STD}_{i,binary}$ can safely be considered at NOE distance from the directly irradiated proton of the adjacent ligand.

3.4.3 BSA/Naproxen as a "pseudo-ternary" complex of the type X-X'

Ternary fragment complexes of a protein and two fragment **X** and **Y** are very common in fragment-based drug design (FBDD) projects in the developing stage. Obtaining such confidential systems from a pharma company or a FBDD group involves complications, in terms of IP issues, therefore we tested the approach on the complex of commercially available BSA with Naproxen, which we consider as a "pseudo-ternary" complex, for the purposes of our proof of concept. Naproxen occupies three well characterised binding sites of BSA: "drug site 2", "drug site 1" and "fatty acid site 6". The latter two being adjacent to each other (Figure 3.13a)[96,105].

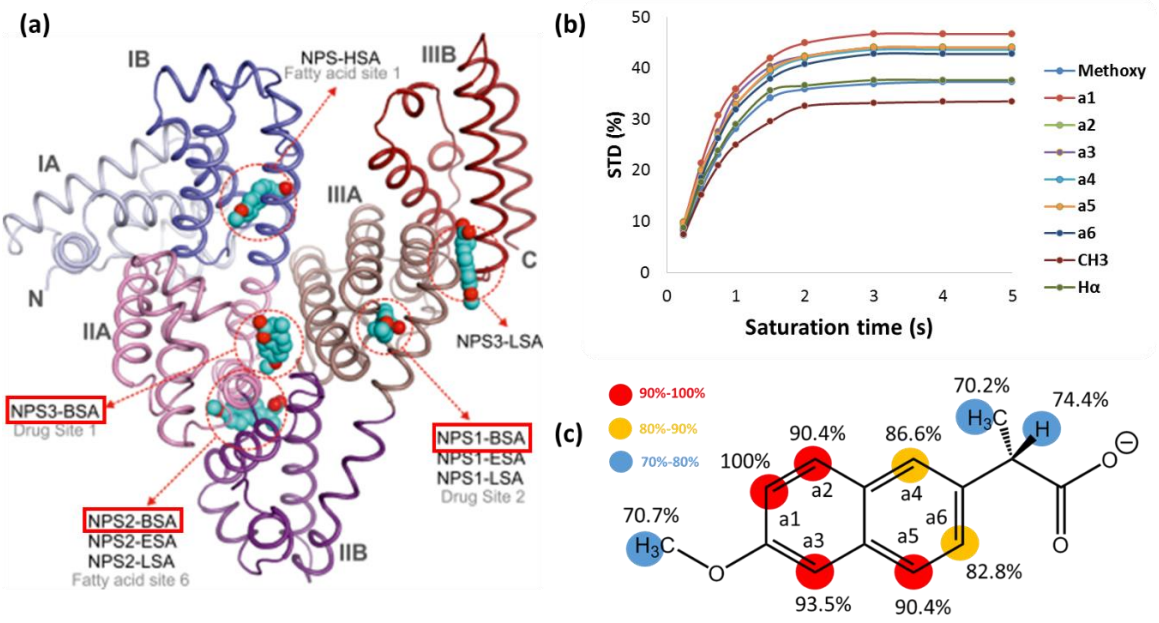

Figure 3.13. **(a)** Distribution of Naproxen (NPS) binding sites in bovine (BSA), equine (ESA), leporine (LSA) and human (HSA) serum albumin complexes, from [96]. The three BSA binding sites of relevance for us are highlighted in red. **(b)** STD NMR build-up curves for Naproxen in complex with BSA. **(c)** STD NMR binding epitope mapping.

Defining binding affinities to the three sites is controversial, "drug site 2" is reported to be the one with strongest affinities and many K_D values for this system have been published, ranging from 0.034 µm to 0.82 µm[94,106,107].

STD analysis of this complex shows very high STD intensities and very steep build up curves, already plateauing at 2 s (Figure 3.13b). This can be ascribed to the high binding affinities, the peculiar relaxation properties of Naproxen and the strong contribution of rebinding (from the three subsites). Compatibly, the binding epitope map is rather homogeneous with the lowest value of relative STD intensities on the three non-aromatic sets of protons (the methoxy group, the proton in α and the methyl in β to the carboxylic group, between 70% and 74%) and the strongest value on a1, the aromatic proton in *ortho* to the methoxy group (Figure 3.13c). It is important to underline that the observed STD signals arise from the combination of Naproxen binding in the three different sites, most likely with different kinetics and affinities. This makes very cumbersome to accurately interpret the STD NMR binding epitope. Nevertheless, based on the observed epitope pattern, we can speculate that the major contribution for the observed STD intensities comes from the Naproxen occupying the fatty acid site 6 (Naproxen NPS2 in Figure 3.13a). Naproxen NPS2 seems the ligand most

homogeneously buried in the protein, and the one which leaves the non-aromatic ends slightly more solvent exposed, while the binding mode of Naproxen NPS1 and NPS3 looks less compatible with the observed binding epitope.

On the other hand, if drug site 2 is the one with highest affinity, and the lowest K_D measured is 0.034 µM, it makes the interaction of Naproxen NPS1 in that site too strong to be detected by STD NMR. K_Ds in the range of units of µM, instead, are compatible with the very strong STD signals observed. On these assumptions, we can consider the BSA/Naproxen system a "pseudo-ternary" complex formed of a protein and two ligands **X** and **X'**.

3.4.4 IL-STD NMR on Naproxen allows to confirm the binding of Naproxen NPS2 and NPS3 to adjacent BSA subsites

We irradiated the methyl group in α to the carboxylic group (1.46 ppm) and observed the response on the methoxy group (3.71 ppm), as from the crystal structure the methyl group of NPS3 is remarkably close to the methoxy group of NPS2. We acquired two build-up curves (0.1 s to 2 s) irradiating at 1.464 ppm (*on-ligand*) and 0.60 ppm (*off-ligand*). The spectra at 1 s saturation time are shown in Figure 3.14.

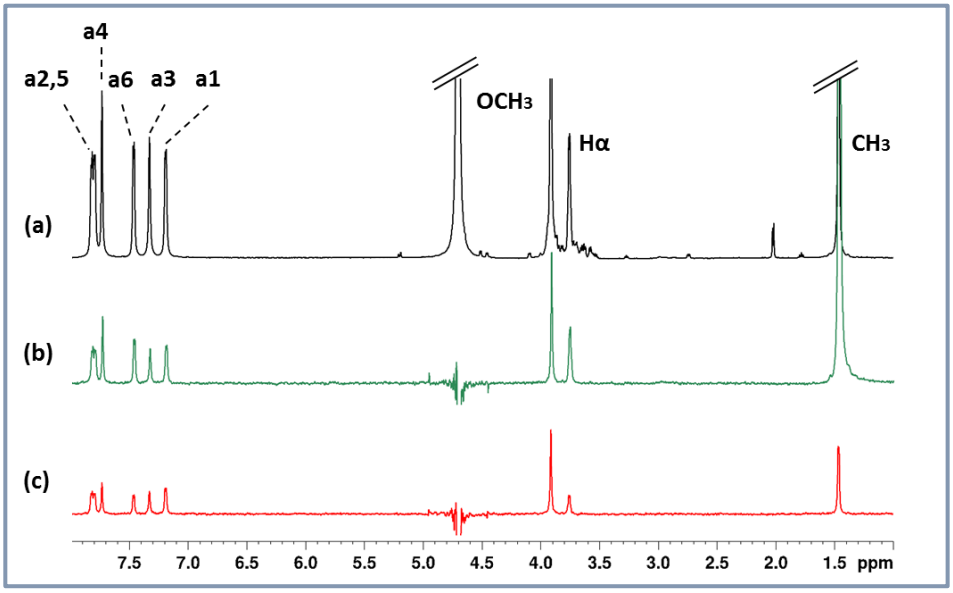

Figure 3.14. 4 mM of Naproxen in the presence of 50 µM BSA in deuterated PBS buffer, pH 7.4, 303 K, saturation time 1 s. **(a)** Reference spectra (x 1); **(b)** STD NMR spectrum with on-resonance irradiation at 1.46 ppm (x 32); **(c)** STD NMR spectrum with irradiation at 0.60 ppm (x 32).

The $\Delta\mathrm{IL}-\mathrm{STD}_{i,ternary}$ were calculated using the initial slopes (non-normalised) extrapolated from the build-up curves at the two frequencies ($\frac{initial\ slope\ _{on-ligand,i}}{initiatl\ slope\ _{off-ligand,i}}$), rather than on the STD values at a given saturation time ($\frac{STD\ _{on-ligand,i}}{STD\ _{off-ligand,i}}$) (Figure 3.15a). This looked necessary for such steep build up curves. Not surprisingly, a very similar pattern of ΔIL-STDs was obtained when the calculations were repeated with the STD intensities at 0.1 s, as we discuss in the following sub-section. The raw and processed data are found in Table 3.6.

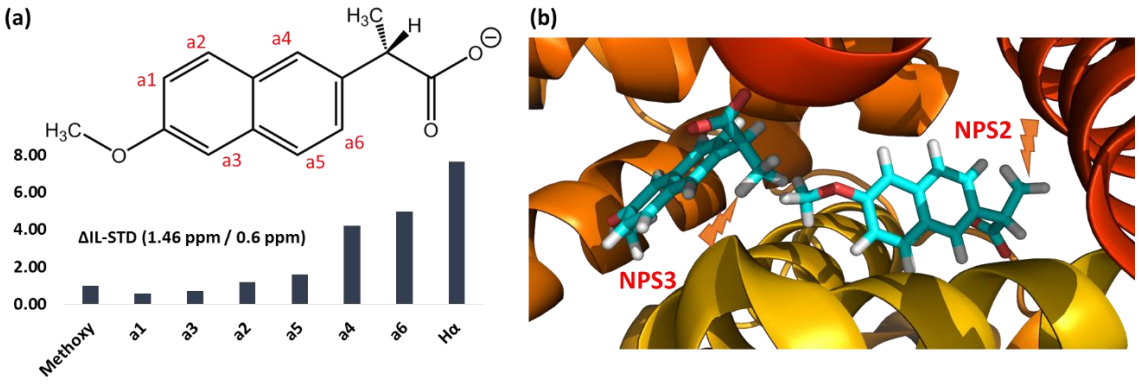

Figure 3.15. (a) ΔIL-STD histogram for the system Naproxen/BSA calculated using the initial slopes of the build-up curves acquired at 1.46 ppm (*on-ligand*, methyl group irradiation) and 0.60 ppm (*off-ligand*). Intra-ligand NOE is the main effect, with ΔIL-STDs decreasing from the methyl group towards the other end of the molecule: the trend is broken at the ΔIL-STD of the methoxy group, which is larger than the ΔIL-STD of a1 and a3. This reports on the proximity of the two bound Naproxen NPS2 and NPS3, as observed in **(b)** the crystal structure of the complex (PDB ID: 4OR0).

For a **X-X'** ternary complex, running the experiment on the binary system is not possible. Nevertheless, the ΔIL-STD patterns obtained were self-consistent and in good agreement with the crystal structure (Figure 3.14b). Naproxen is a small and rigid molecule, and it is clear from the ΔIL-STD (Figure 3.14a) that the main effect of irradiating on-ligand is an increase of the neighbouring aromatic groups due to intra-ligand NOE.

The IL-STD factors decrease along the molecule going from the carboxylic to the methoxy end (Figure 3.14a). Remarkably, the IL-STD factor on the methoxy group is higher than the one for a1, a3 and a2. Considering the rigidity of the molecule, this can solely be explained by the inter-ligand transfer of magnetisation from Naproxen NPS3

to Naproxen NPS2 (also confirming our assumption of observing Naproxen NPS2 mainly).

Due to the multiple-binding modes, the steep build-up curves and the "pseudo-ternary" **X-X'** nature of the complex, the BSA/Naproxen complex was a very hard system on which to detect IL-STD signals. The fact that we do see them, gives us ground to propose that IL-STD NMR can be detected and has the potential to be useful for assessing the relative orientation of two bound ligands, or of two binding subsites (when one of the two is known) as we report in Chapter 4, Section 4.2 on a **X-Y** ternary complex.

Proton	1H δ (ppm) [a]	Initial slope 1.46 ppm (s^{-1})	Initial slope 0.60 ppm (s^{-1})	Ratio initial slope 1.46 ppm/0.60 ppm	ΔIL-STD
OCH3	3.71	43.27	4.83	8.96	0.96
a1	6.98	59.32	6.93	8.57	0.57
a3	7.13	56.97	6.22	9.16	1.16
a2	7.61	48.78	5.61	8.70	0.70
a4	7.53	64.33	5.27	12.20	4.20
a5	7.59	54.59	5.69	9.60	1.60
a6	7.24	61.55	4.76	12.94	4.94
Hα	3.55	71.29	4.56	15.64	7.64
				N	
				8	

Table 3.6. ΔIL-STD calculated using the initial slopes of the STD NMR build-up curves (1.46 ppm/0.60 ppm) for the Naproxen/BSA complex. [a] Spectra for assignment acquired at 298 K.

3.4.5 Selection of the saturation time for ΔDEEP-STD and ΔIL-STD studies

The interpretation of the data obtained on the Naproxen/BSA system triggered some reflection on using a single saturation time, as opposed to a whole build-up curve, to calculate IL-STD factors. The IL-STD factors (1.46 ppm/0.60 ppm) of Naproxen NPS2 at three different saturation times are reported in Figure 3.16. It is interesting to observe how the information present at 0.1 s gets lost at increasing saturation times; and how

the absolute intensities of ΔIL-STD become lower at longer saturation times. This "flattening" is due to the role of longitudinal relaxation. Not surprisingly, the pattern observed at 0.1 s is very similar the one observed when the ΔIL-STD are calculated using the initial slopes obtained from the build-up curves in Figure 3.14a (the initial slope extrapolates the STD (%) when the saturation time approaches zero).

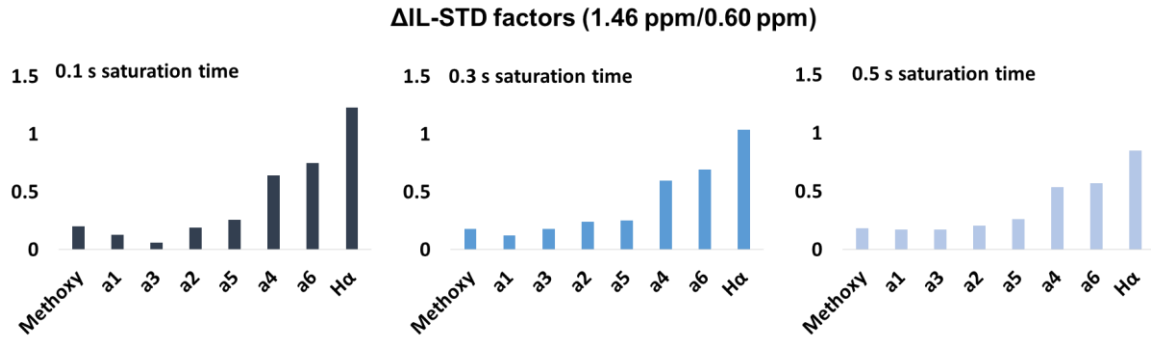

Figure 3.16. ΔIL-STD factors (1.46 ppm/0.60 ppm) for Naproxen (NPS2) in complex with BSA at increasing saturation times (0.1 s, 0.3 s, 0.5s). The pattern observed at 0.1 s (very similar to the one obtained with the initial slopes in Figure 3.14a) is lost as the saturation time increases (at 0.5 s the IL-STD factors for protons Methoxy, a1, a3 and a2 are similar).

In Figure 3.17, the same comparison is drawn for differential irradiation DEEP-STD study on 2,7-anhydro-Neu5Ac/GH33 from Figure 3.7a, left panel. In this case, the stark differences observed at 0.5 s disappear at 2 s, where all the ΔDEEP-STD are below |0.5|.

It is interesting to observe that, at longer saturation times, the pattern of the ΔDEEP-STD map is retained ($H3_a$ and CH_3 are still the highest positive values and H8, H9 and H9' still the lowest negative values), but the overall map is "flattened" around 0.

STD NMR can be considered as a truncated NOE (TOE) experiment, where the *rate* at which the steady state is reached carries the information on spatial proximity. This means that the ligand longitudinal relaxation (R_1) starts playing a role the closer we get to the steady state (at longer saturation times). In the binding epitope mapping approach, eliminating the R_1 contribution is crucial to retain the spatial information. STD_{max} (reached at the steady state, when the curve plateaus) is dependent on both cross relaxation (σ_{IS}) and longitudinal relaxation (Equation 1.40, in Chapter 1). For this reason, the binding epitope of ligands containing protons with different relaxation properties could be misleading if calculated at a single time point too close to the

STD_{max}[70c,108]. Therefore, for determining STD NMR binding epitope mappings, relying on the initial slope is unavoidable, albeit time-consuming to acquire and process.

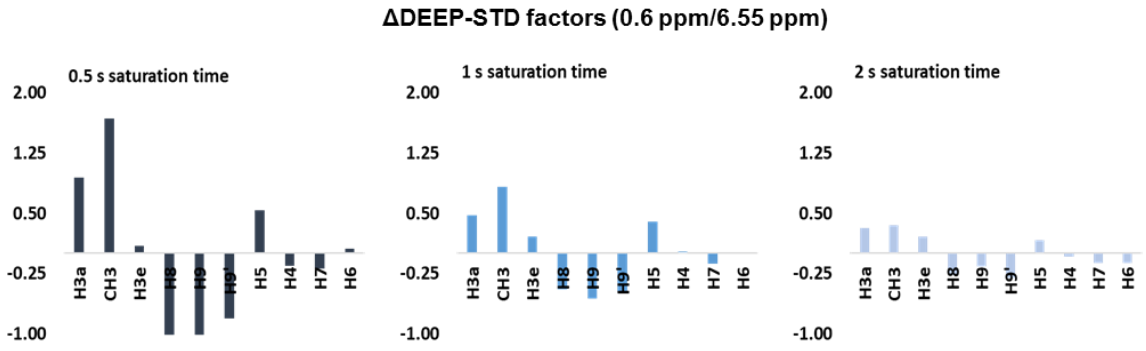

Figure 3.17. ΔDEEP-STD factors (0.60 ppm/6.55 ppm) for 2,7-anhydro-Neu5Ac in complex with GH33 at increasing saturation times (0.5 s, 1 s, 2 s). As observed for the IL-STD data (Figure 3.16), the starker differences among the protons are observed at the shorter saturation times.

For the DEEP- and IL-STD NMR, comparison of a single saturation time at two different frequencies is shown to be reliable and allows to keep the protocol simple enough (STD NMR build-up curves can always be run in particularly problematic systems or in case of doubt). The condition for the choice of the saturation time is that it is short enough in the relaxation time scale of the system in analysis (far from the plateau of the build-up curve). Thus, the choice of the saturation time is system-dependent. Seemingly, analysis at long saturation times can cause loss of information (flatter histograms), not its mismatching. In defining the saturation time at which to perform the analysis, it is also fundamental to consider the absolute intensities of the signals for each complex. If the signal intensities are too small, the experimental error could invalidate the approach (*e.g.*, for most protein-ligand systems, the STD signals at 0.1 s would be within the noise). For very low affinity systems, or in situations where many experiment needs to be performed, one can consider increasing the saturation time to 1 s or 2 s, upon making sure that the information is not lost.

Based on this, to extract the maximal information coming from spin diffusion inhomogeneity using DEEP-STD and IL-STD approaches, we recommend to always choose the shortest saturation time that the system and the situation allow.

3.5 Conclusions

In this chapter, we proposed two novel and versatile approaches to exploit inhomogeneity of spin diffusion to obtain information on the molecular space surrounding the ligand (Figure 3.18).

Firstly, the information coming from IL-STD NMR (spatial vicinity of two adjacently bound fragments in a binding pocket) could already be obtained by other ligand-based techniques (tr-NOESY under ILOE conditions). Thus, we propose IL-STD NMR rather as a proof of concept that inter- (and intra-) ligand STD can be observed: to the best of our knowledge, this had never been reported before. For favourable cases, IL-STD NMR looks like a promising and time-efficient technique to help the process of knowledge-based drug design. Together with ILOE, it allows to explore the orientation of binding sub-sites and bound fragments in known and unknown binding pockets, or to validate ternary complex models in fundamental research projects, as we report in Chapter 4, Section 4.2.

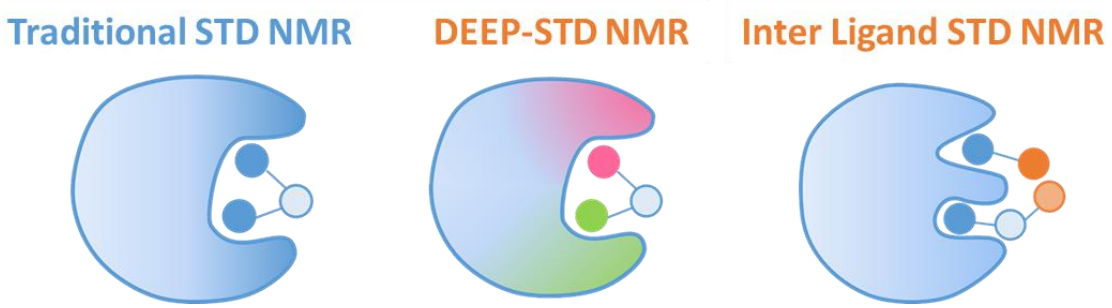

Figure 3.18. Cartoon schematising the information obtained on protein-ligand interactions by traditional STD NMR and by the two approaches proposed by us (DEEP-STD NMR and IL-STD NMR).

Secondly and more importantly, we propose DEEP-STD NMR as an original approach, which gives access to a layer of structural information previously inaccessible through ligand-based NMR techniques: exploring the nature of the amino acid residues surrounding the ligand in the binding site, and hence assessing the pharmacophore of the protein. We designed a simple and robust protocol relying on differential irradiation or differential deuteration of the polar residues in the binding site; and we envision that DEEP-STD NMR, in combination with classical STD NMR, will become a popular approach to characterize intermolecular interactions revealing the nature of the protein-ligand contacts. Given the amphiphilic nature of many ligand-receptor interactions (both

synthetic and natural ligands), the approach is advantageous both in fragment-based drug discovery and in fundamental biological research.

If the 3D structure of the protein is known, the protocol allows for the first time to gain information from STD NMR on the orientation of the ligand. We have demonstrated this here for small ligands, which are more difficult to orient than larger ones. DEEP-STD NMR combined with the 3D structure of the protein is expected to define the orientation of larger ligands in the bound state even more easily. As we will show in Chapter 4 and 5, if different ligands of a well characterized protein are studied, DEEP-STD NMR will provide different fingerprints to readily obtain their different binding orientations (Section 5.2). From there, the technique can be expanded by performing "on-resonance scanning" and detailing even more in depth the nature of the contacts (Section 4.3).

Noticeably, DEEP-STD NMR is applicable even when the 3D structure of the protein is unknown; in this case, the nature *and* architecture of the binding site could be unveiled, as the technique allows to identify possible "patches" (aliphatic, aromatic, polar, *etc.*) within the unknown pocket.

The possibility of detecting "polar patches" is especially appealing in the investigation of protein-carbohydrate interactions, as carbohydrate-recognising binding pockets are often lined with arginine, histidine and asparagine[109] (promoting cooperative H-bonds, bidentate interactions and other stabilizing contacts[12]). GH33, among the many sialidases, is a remarkable example, with its arginine triad complexing the carboxylate group to keep the sialic acid in place[12]; but it is also true for heparin-binding proteins and other targets[110].

IL-STD NMR and DEEP-STD NMR should be regarded as versatile tools providing structural information at the atomic level on the binding pocket environment in which the ligand resides.

3.6 Materials and Methods

3.6.1 Chemicals, ligands and proteins

Compounds 3-*nitro*-phenyl α-D-galactopyranoside (3NPG) and Naproxen > 98.8%, cholera toxin subunit B (CTB), bovine serum albumin (BSA), deuterium oxide (99.9% ^{2}H),

tris-(hydroxymethyl-d_3)amino-d_2-methane (Tris-d_{11}, 98% ^2H), disodium hydrogen phosphate (Na_2HPO_4), potassium dihydrogen phosphate (KH_2PO_4), sodium chloride (NaCl) and potassium chloride (KCl) were purchased from Sigma. Sugar Neu5Acα6Lac (6'SL) was purchased by Glycom. Sugar 2,7-anhydro-Neu5Ac was prepared from our collaborators in the group of Nathalie Juge (Quadram Institute) as previously described in [111]. [Briefly, the compound was enzymatically synthetised by incubating 4-Methylumbelliferyl-Neu5Ac with *Rg*NanH in ammonium formate buffer (100 mM) at 37°C, pH 6.5 overnight. The reaction was terminated by the addition of an equal volume of ethanol and the precipitate removed by centrifugation (4000 *g* for 20 min). The mixture was evaporated to dryness and the 2,7-anhydro-Neu5Ac was purified by Folch partitioning.]

CBM40 and GH33 were also expressed by our collaborators at the Quadram Institute as well. CBM40 was expressed in *Escherichia coli* BL21 Rosetta (DE3) (Novagen) with a six-histidine tag and purified on a nickel-Sepharose column, as described in [97]. GH33 was expressed in *Escherichia coli* BL21 cells (New England Biolabs, Boston, USA) and purified by immobilized metal anion chromatography (IMAC) as described[17], where this domain is denoted "*Rg*NanH NI-domain".

3.6.2 NMR measurements and processing

^1H and ^{13}C resonance assignment for the ligands was performed on the bases of 1D ^1H, 2D ^1H-^1H DQF-COSY, ^1H-^{13}C HSQC and ^1H-^1H NOESY experiments run on the free ligands in deuterium oxide. For STD NMR experiments, all the samples consisted of 50 µM in binding sites (CBM40 and GH33 and BSA have a single binding site per protein, whereas CTB is a pentamer with five identical binding sites) and 1 mM ligand (except for Naproxen which was at 4 mM). 6'SL/CBM40 sample was analysed at 288 K and 2,7-anhydro-Neu5Ac/GH33 samples were analysed at 293 K, both buffered in 10 mM Tris-d_{11} pH 7.8 and 100 mM NaCl; 3NPG/CTB samples were analysed at 278 K and Naproxen/BSA sample was analysed at 303 K, both buffered in 10 mM PBS buffer, 137 mM NaCl and 2.7 mM KCl pH 7.4. For the experiments in light water, buffers and ligands were freeze dried and dissolved in Milli Q and the H_2O:D_2O ratio adjusted to 90:10 (10% D_2O for locking purposes). An STD pulse sequence that included 2.5 ms and 5 ms trim pulses and a 3 ms spoil gradient was used. Saturation was achieved by applying a train

of 50 ms Gaussian pulses (0.40 mW) on the f2 channel, at the variable frequencies indicated in the text (on-resonance experiments) and 40 ppm (off-resonance experiments). The broad protein signals were removed using a 40 ms spinlock ($T1\rho$) filter (stddiff.3). When required, water suppression was achieved by excitation sculpting with gradient (stddiffesgp.3). Saturation times ranging from 0.1 s to 2 s as indicated in the text and 5 s relaxation delay were used. 6'SL/CBM40 experiments were run with 256 scans; DEEP-STD experiments (2,7-anhydro-Neu5Ac/GH33 and 3NPG/CTB) were run with 256 scans for samples in D_2O and 512 scans for samples with H_2O:D_2O 90:10; Naproxen/BSA were run with 32 scans. All the experiments were recorded at ^1H frequency of 800.23 MHz on a Bruker Avance III spectrometer equipped with a 5-mm probe TXI 800 MHz H-C/N-D-05 Z BTO. For Naproxen/BSA, the build-up curves (0.1 s to 2 s) for each proton were fitted mathematically to a mono-exponential equation (y=a*[1-exp(b*x)]), from which the initial slopes (a*b) were obtained. The binding epitope mapping was obtained by dividing the initial slopes by the one proton a1, to which an arbitrary value of 100% was assigned.

3.6.3 CORCEMA-ST[66]

For CORCEMA-ST calculation in Sub-section 3.3.7, the cut-off distance around the binding pocket was 8 Å. The concentration of ligand and protein used was respectively 2 mM and 45 µM. k_{on} was set to 1×10^{-8} M^{-1} s^{-1}. The bound ligand correlation time was 50 ns, whereas the free correlation time was 1 ns. The equilibrium constant used was 2000 M^{-1} (K_D = 0.5 mM). The ρ leak used was 0.4 Hz, except for the simulation of the experiment in H_2O, where the ρ leak was put at 6 Hz. The irradiation frequency was set to the range 0.1 ppm to 1.1 ppm to simulate the conditions of irradiation at 0.60 ppm; and to the range 5.1 ppm to 7.0 ppm to simulate the conditions of irradiation at 6.55 ppm. To simulate the presence of H_2O as a solvent, the "i4" ("indata11") was set to 0 (1 is for measurement in D_2O, 0 for measurement in H_2O). In these conditions, the STD intensities were simulated for a saturation time of 0.5 s.

Chapter 4

New avenues for Cholera Toxin B (CTB) inhibition

4.1 Introduction

4.1.1 Cholera: onset and treatment of a neglected tropical disease

Cholera is a bacterial infection classified within the group of "neglected tropical diseases"[112]. It is considered endemic in 50 countries, but it manifests as epidemic in association with natural disasters and emergency situations characterised by poor hygienic conditions[113]. The World Health Organization estimates 3-5 million people affected by cholera and 100.000 to 130.000 deaths, every year[114]. The infection is spread by ingestion of the bacteria *Vibrio cholera O1* or *O139*[112]. Once the microorganism reaches the small intestine and colonises it, it secretes an enterotoxin called cholera toxin (CT), which triggers disruption of the ion channels of the epithelial cells of the guest. This leads to severe water and salt loss, resulting in strong de-hydration and death[115].

At the molecular level, the ion channel disruption pathway is initiated by the interaction of cholera toxin with the glycosphingolipid ganglioside-monosialylated GM1 exposed on the membranes of the mucosal cells of the hosts intestines. GM1 is a branched pentasaccharide of structure β-Gal(1-3)β-GalNAc(1-4)-[αNeuAc(2-3)-]β-Gal(1-4)Glc (Figure 4.1a). CT belongs to the lectin family and is an 84 kDa AB_5 protein, composed of a monomeric A and a pentameric B_5 subunit (Figure 4.1b)[116]. CTA is the enzymatically active domain responsible for the toxic effects; CTB is responsible for complexing GM1 in a multivalent fashion and mediating entry into the cell[115].

The mechanism of penetration of CT through the membrane to reach the Golgi apparatus has been extensively investigated through a range of methods but some parts of the mechanisms are still unclear[117,118,119]. CTA triggers a signalling cascade that disrupts the normal biochemical pathway of the regulatory enzyme adenyl cyclase (AC). The outcome is the continuous activation of AC, and thus, the uninterrupted hydrolysis of ATP to AMP. The high concentration of AMP up-regulates the chloride channel

proteins in the lumen; therefore, an excess of Cl^- ions is expelled. To maintain the electrochemical balance, an equal amount of Na^+ is excreted, followed by water due to osmosis, thus causing severe de-hydration[115].

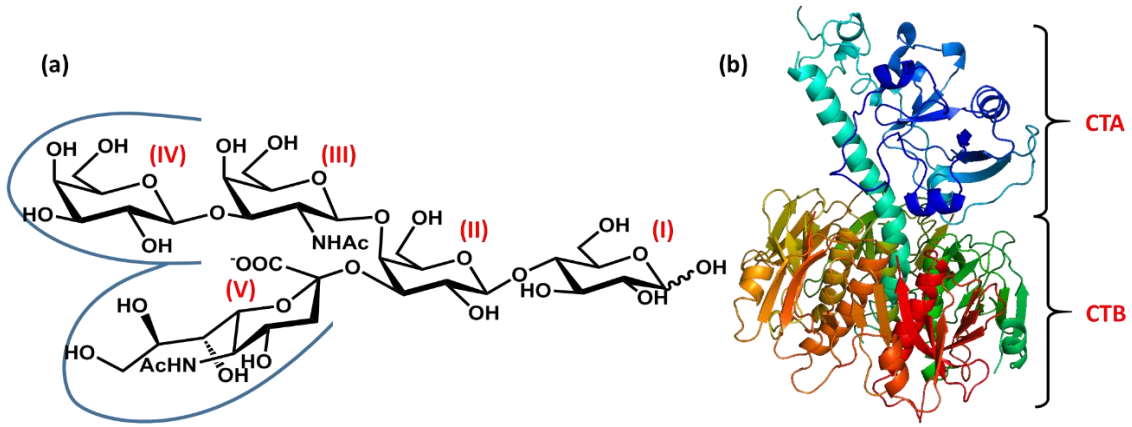

Figure 4.1. (a) Structure of the GM1 pentasaccharide with residues labelled in red; the main recognition elements for binding to CTB are enclosed in blue curves. **(b)** Crystal structure of CT (PDB ID: 1XTC[120]): CTB pentamer in green to yellow and CTA in cyan and blue, as shown by the curly brackets.

The first treatment for cholera consist of re-hydration of the infected individuals with clean water. This becomes very challenging when cholera occurs in areas with contaminated water sources. For severe cases, antibiotics are used, mainly tetracyclines and doxycyclines[112]; however, antibiotic resistance is a growing concern and multidrug resistant strains of *Vibrio cholera* have been observed worldwide[121,122,123]. This evidence has boosted i) research towards vaccine development (the oral administration of CTB vaccine is widespread and there are many other options available on the market[114]); and ii) the design of CTB antagonists, to inhibit the formation of the GM1/CTB complex.

4.1.2 Dissecting the Cholera Toxin B/GM1 interaction

The present work pursues the optimisation of inhibitors of cholera toxin internalisation, by impeding the molecular recognition of GM1 by the cholera toxin subunit B (CTB). Remarkably, the GM1/CTB interaction is one of the strongest observed for protein-carbohydrate complexes, with a K_D in the nM range[124]. Thus, a profound knowledge of the structural features that govern the molecular recognition between CTB and GM1 is key to the design of strong inhibitors. High-resolution crystallography studies [120,125,126], and one thorough thermodynamic analysis[127] of the native GM1/CTB complex are the

germinal works which allowed to unravel the mechanism of molecular recognition. The X-ray structure of the GM1/CTB complex (PDB ID: 3CHB[126]) shows that the main elements of recognition of GM1 are the non-reducing galactose (IV) and sialic acid (V) residues (Figure 4.1 left and 4.2), so that the binding mode can be considered as bi-dentate.

The research group of Homans investigated the GM1/CTB interaction in solution by isothermal titration calorimetry (ITC), detailing the thermodynamic contribution of each monosaccharide of GM1[127]. The study showed that residues (III), (IV) and (V) contribute for 17%, 39% and 43%, respectively, of the buried surface area of the intermolecular interaction. On the other hand, residues (IV) and (V) account for 54% and 44% of the intrinsic binding energy, respectively, thus indicating that the galactose binding subsite is more specific and shielded from the solvent than the sialic acid one.

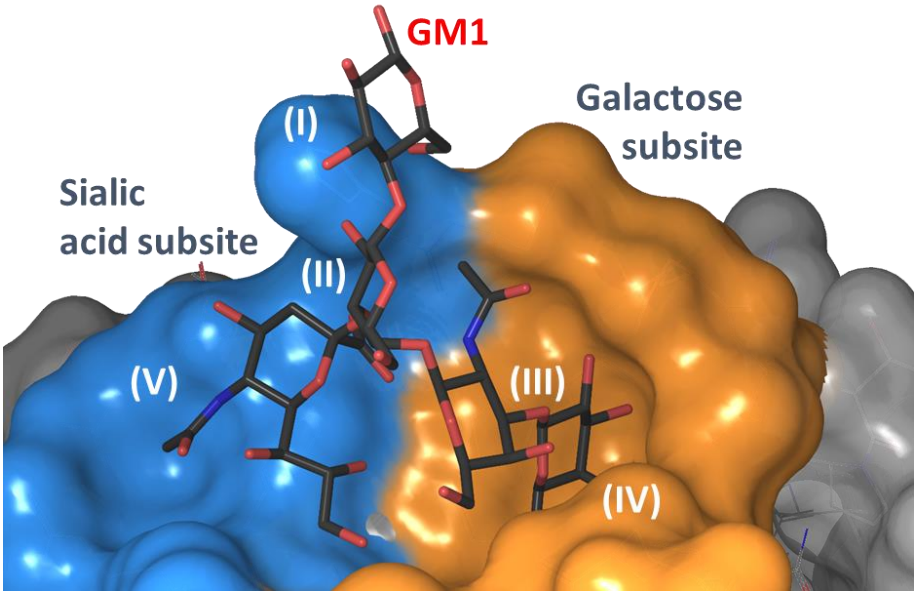

Figure 4.2. 3D representation of the binding site of GM1 bound to CTB (PDB ID: 3CHB[126]). CTB is shown as a surface coloured in blue and orange to identify the sialic acid and galactose subsites, respectively. GM1 is shown as sticks and labelled as in Figure 4.1a. The protons of GM1 are omitted for clarity.

Two valid arguments to account for the high monovalent affinity of the interaction are[127]:

- the conformational pre-organization of the pentasaccharide (*i.e.*, free conformation is similar to the bound conformation);

- the key-residues for recognition, galactose (IV) and sialic acid (V), being spaced by residues (II) and (III), which prevents any corresponding couple of fragments of GM1 from achieving interaction of comparable efficiency (i.*e.*, preventing multiple binding modes).

Based on these fundamental works, the research groups of Jimenez-Barbero and Bernardi used NMR (tr-NOESY and ROESY) and computational techniques to analyse the binding of GM1 analogues to CTB[128].

Firstly, they demonstrated that substituting GM1 residues (II) and (I) with cyclohexanediol does not affect its binding affinity. Hence, they synthetised a small library of glycomimetcs, in couple of stereoisomers (Figure 4.3), in which the sialic acid (V) of GM1 was replaced with an α-decorated carboxylic group, meant to mimic the less Neu5Ac specific interaction.

Ligand 5 (*R* configuration) was found to be 10 times more active than its stereoisomer **Ligand 4** (K_Ds 190 µM and 1000 µM, respectively). Hence, **Ligands 6** and **7** were synthesised in the *R* configuration and gave binding affinities to CTB of 45 µM and 10 µM, respectively.

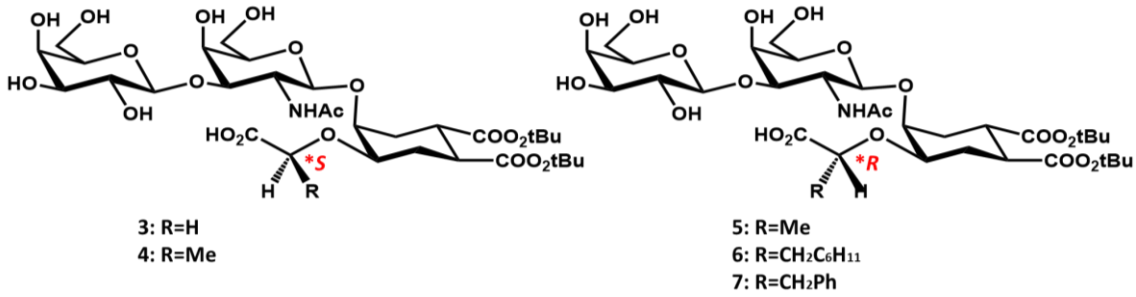

Figure 4.3: Synthetic glycomimetics based on the scaffold of a non-reducing β-Gal(1-3)β-GalNAc O-glycosilated to a substituted cyclohexanediol, from the work of Jimenez-Barbero and Bernardi[128].

Tr-NOESY experiments showed that, while **Ligand 6** rearranged upon binding, the free and bound conformations of **Ligand 7** was identical. Specifically, the non-reducing galactose and aromatic ring of **Ligand 7** seemed to interact through a stabilizing stacking where the phenyl ring is sandwiched between the GlcNAc and the protein surface, as shown by MD and tr-NOESY. **Ligand 7** inspired many of the following GM1 antagonists which we will cover in the following sub-section, and the observation of a

"conformational lock" within the ligand marked a milestone in the research on GM1 glycomimetics[128].

4.1.3 Overview of CTB inhibitors: from cyclohexanediol scaffolds to polyvalent ligands

The group of Bernardi built on the cyclohexanediol scaffold modifying systematically different fragments of the structure without succeeding to improve the affinity relative to **Ligand 7**[129,130]. Thereafter, they designed a new class of ligands based on a galactose linked to a sialic acid moiety through a non-hydrolysable spacer, in the attempt to reproduce the bidentate binding mode of GM1. While bioavailability increased due to insertion of the non-hydrolysable linker, the affinity decreased (K_D in the low mM range). Among them, **Ligands 58** and **59** in Figure 4.4 showed the best affinities[90].

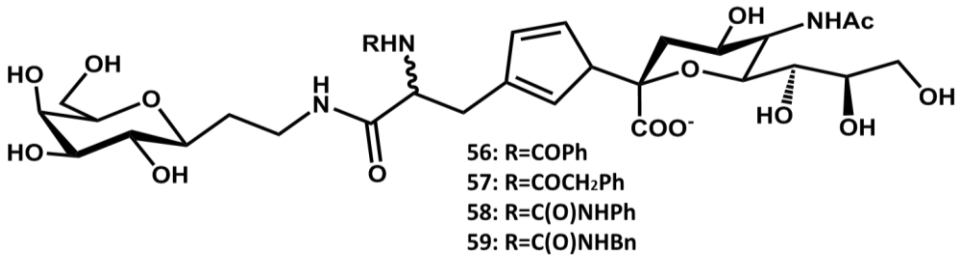

Figure 4.4. Scaffold of the non-hydrolysable ligands, from the work of Bernardi[90].

Pieters and co-workers conjugated a lactose moiety to aryl residues through a thiourea (**Ligands 23** and **24** in Figure 4.5)[131]: the galactose served as an anchor to the galactose binding subsite, and the phenyl groups was conceived to occupy the sialic acid subsite. The strongest binder of this library showed a K_D of 23 µM by weak affinity chromatography, while no 3D structural information was provided on the binding of this library in the study.

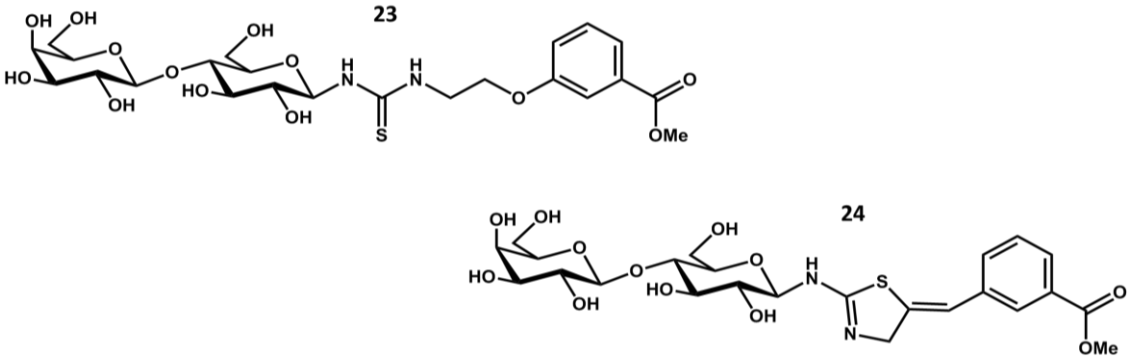

Figure 4.5. The lactose-based ligands, form the work of Pieters[131].

A great research effort has also been put in the field of multivalent binding, which is known to enhance the affinity by 100- to 1000-fold relative to monovalent binding[28]. Jointly, the groups of Bernardi and Pieters proposed a class of glycodendrimeric ligands composed of 3,5-d-(2-aminoethoxy)-benzoic acid repeating units, substituted with lactose moieties (K_D=10 µM). Conjugating the dendrimer core with **Ligand 5** (Figure 4.3) via a spacer yielded a K_D of 0.5 µM, showing the validity of the approach[132]. Other solutions resorted to α-cyclodextrins[133] or poly-(N-acryloylmorpholine) polymers as bivalent ligands (that is, consisting of two galactose moieties connected through a linker)[134]. The group of Gibson specifically focused on the investigation of galactose-functional polymers. First they showed how the length of the polymeric linker is a key element for binding to such a deep binding site (most lectins present much shallower clefts of interaction)[135]. More recently, they introduced a secondary binding motif on the linker, branched relative to the original galactose one (Figure 4.6a). Whereas the installed "side chains" were simple organic residues, based on benzylic groups decorated with halogens in different patterns (Figure 4.6b), they showed that their presence increased the efficiency of binding relative to the previous galactose-only-functional polymers[136].

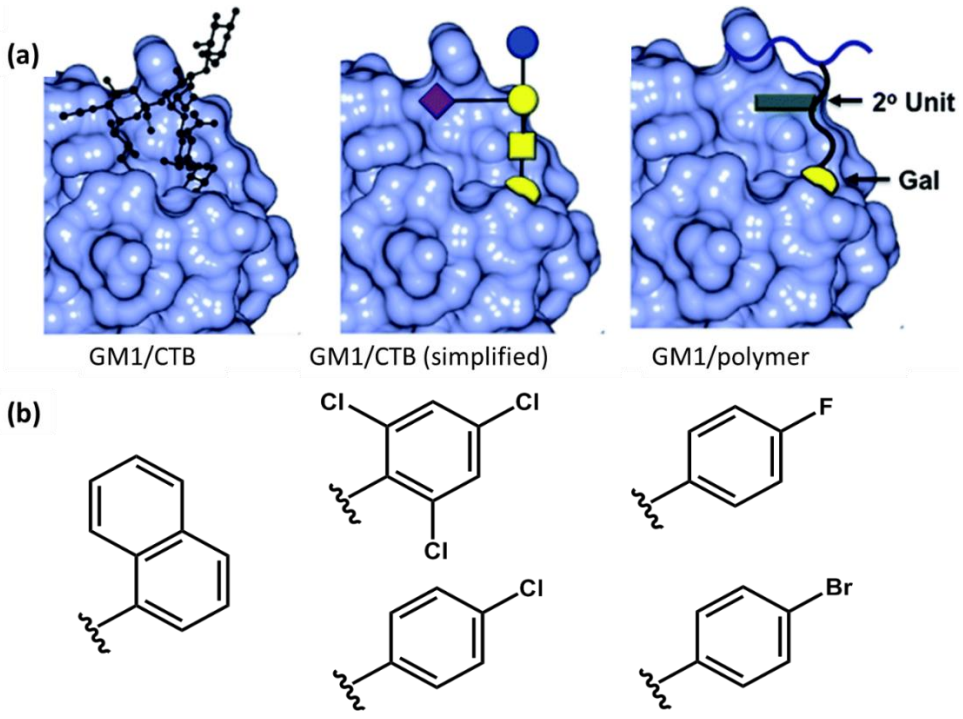

Figure 4.6. **(a)** The glycopolymers with secondary branched binding motifs approach proposed by the group of Gibson; **(b)** some of the small organic residues used as the second unit in the study. Figure from 136.

In this study, the authors highlighted two very important points: i) resorting to glycan branching, trying to mimic the binding mode of GM1 as we know it from the available structural biology information, achieved improved results; ii) complicated glycan-synthesis is not the only way to reach powerful lectins inhibition.

The lowest K_D reported so far, with affinity in the picomolar range, was achieved by means of an elegant approach using an engineered neoglyco-CTB equipped with five GM1 units attached through linkers to each monomer of the engineered CTB (Figure 4.7). This design allowed to match the geometrical requirements of the wild type toxin and deactivate it in multivalent fashion with optimal spacing[137].

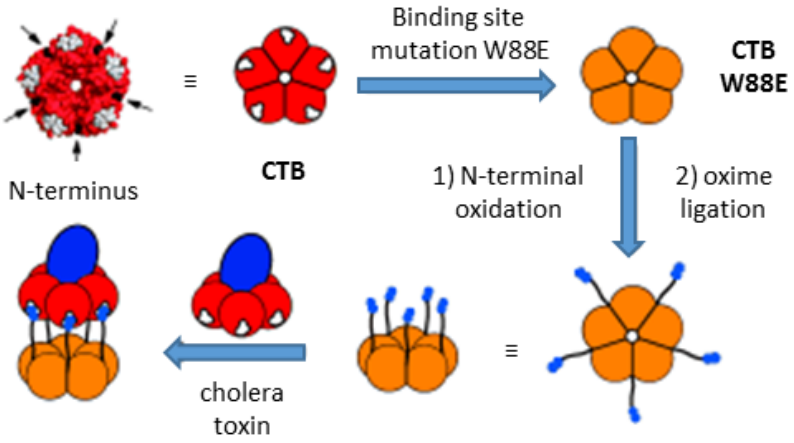

Figure 4.7. Engineered CTB W88E (in orange), carrying five GM1s (blues points) to bind wild-type CT (in red). Figure from 137.

4.1.4 CTB ligands based on polyhydroxyalkylfuroate scaffolds[92]

Drawing from the knowledge acquired in the previous decade of research on GM1-mimicking molecules, Robina and co-workers, our collaborators at the University of Seville (Spain), designed four new generations of compounds as potential inhibitors of CTB[92]. The novel underlying idea was to promote bi-dentate monovalent binding while reducing the number of carbohydrate units in the antagonists, to ease synthetic efforts.

The rationale behind this scaffold (Figure 4.8a) is the presence of a β-thio-galactose: this, on one hand, anchors the ligand into the specific galactose subsite and, on the other hand, prevents hydrolysis when compared to an O-glycosidic linkage. The thio-galactose is linked to a polyhydroxyalkylfuroate chain (PHF chain), meant to mimic sialic acid and to occupy its subsite. The pharmacophores are there connected by a non-hydrolysable click chemistry spacer, synthetically convenient and chemically stable. The polyhydroxy

chain, together with the triazole linker, helps increasing the solubility and offers possibilities for further decoration of the chain or assembly of multivalent binders.

The affinity of the four libraries towards CTB were assessed by weak affinity chromatography (WAC), using a dedicated approach developed by Ohlson and co-workers in 2009[91]. In this approach, 3-nitrophenyl-α-D-galactopyranoside (3NPG), a well-known CTB binder, was used as reference (K_D=1.10 mM). The results pointed at **Ligand 30** as the main lead from Generation II, with a K_D of 1.05 mM (Figure 4.8b).

Ligand 30, together with its constituent binding **Ligand 20** and **Ligand 33** (Figure 4.9), were analysed by ^{1}H STD NMR to determine the epitope mapping of their interactions to CTB, and to elucidate the role of the polyhydroxyalkylfuroate chain in the interaction.

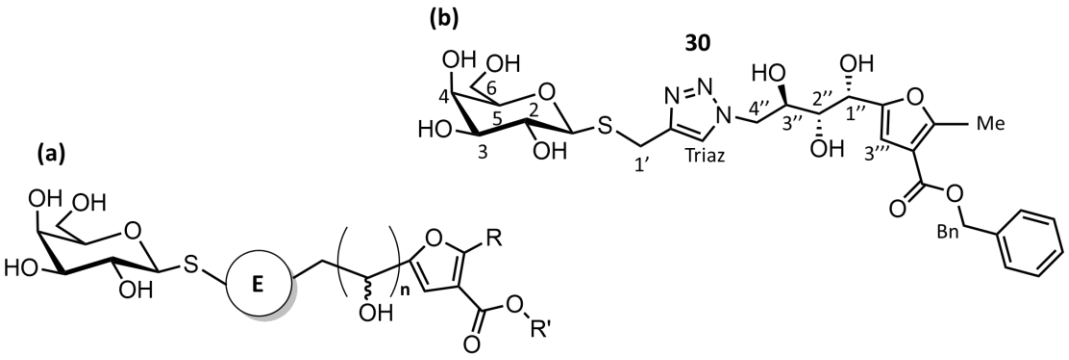

Figure 4.8. **(a)** Schematic representation of Generation II of S-linked galactosides (E=1,2,3 triazole), in 92; **(b)** Chemical structure of **Ligand 30**, the best performing CTB binder from Generation II.

The STD NMR analysis showed the thio-galactose moiety and the furoate-aromatic region of **Ligand 30** to be the closest contact points to the CTB surface. This qualitatively confirmed the hypothesis of a bi-dentate binding mode adopted by **Ligand 30**. The fact that the two fragments (**Ligand 20** and **33**) show relative STD intensity patterns very similar to that of the parent compound (Figure 4.9) supported the case of specific binding to the same binding subsites, although it did no shed light on the *location* of the subsites, particularly in the case of the polyhydroxyalkylfuroate chain.

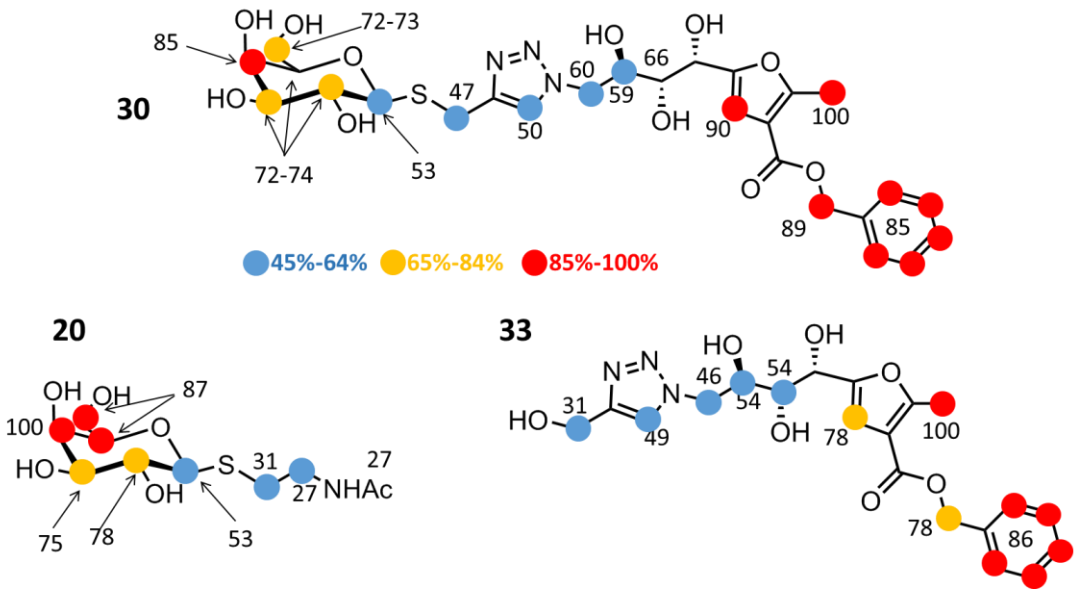

Figure 4.9. ^{1}H STD NMR binding epitope mapping of **Ligand 30**, **Ligand 20** and **Ligand 33** (data from 92). The normalised STD values from the initial slope of each proton, relative to the most intense one (for which a 100% is arbitrarily assigned), is shown next to it.

4.2 Discovery of a novel binding subsite in the GM1 binding pocket of CTB

4.2.1 Crystallisation trials of the Ligand 33/CTB complex

Based on the work of Robina and co-workers, we first attempted to define the location of binding of the polyhydroxyalkylfuroate ligands by trying to co-crystallise the **Ligand 33**/CTB complex. In the first place, the complex solution (1 : 1 ligand to protein ratio at 4% DMSO) was plated in phosphate buffer pH 7.4 at variable concentration of PEGs of variable sizes (4x4 in triplicates), incubated at 16 °C and followed over two weeks. Despite these being the favourable conditions for crystallisation of CTB in the apo form[116,88], the few crystals observed were too small for X-Ray diffraction. Therefore, 48 x 8 conditions were tried in duplicate with commercially available screening kits, namely Structure Screen 1 Eco Screen (MD1-01-ECO and MD2-02-ECO), PACT *Premier* Eco Screen (MD1-29-ECO), JCSG-*plus* Eco Screen (MD1-37-ECO) and Morpheus® (MD1-48). About ten crystals were fished after 2 weeks of incubation at 16 °C, and analysed at the Diamond Light Source synchrotron from our collaborators at the School of Biology (Dr. Andrew Hemmings and MSc Isabella Acquistapace). Unfortunately, none of the resulting structures contained the ligand and only a blurry fit of the CTB in apo form was obtained (resolution > 3 Å). This could be explained by the presence of DMSO in the drops (necessary to solubilise the ligand but probably disturbing the crystallisation process),

but also and especially by the flexible nature of the ligand and by its low affinity (1.35 mM).

4.2.2 STD NMR competition experiments: probing the fragments specificity to the CTB subsites within the GM1 binding pocket

Among ligand-based NMR techniques, STD NMR competition is a powerful approach to probe the location of binding of unknown ligands[82,81]. To this aim, STD-active probe ligands known to occupy the subsites of interest are needed. Ligand 3-*nitro*-phenyl α-D-galactopyranoside (3NPG) was chosen as a probe-ligand for the galactose sub-site (as in PDB ID: 1EEI[88]); and 3'-syalillactose (3'SL), the Neu5Ac-containing branch of GM1, was chosen as probe-ligand for the sialic acid subsite (Figure 4.10).

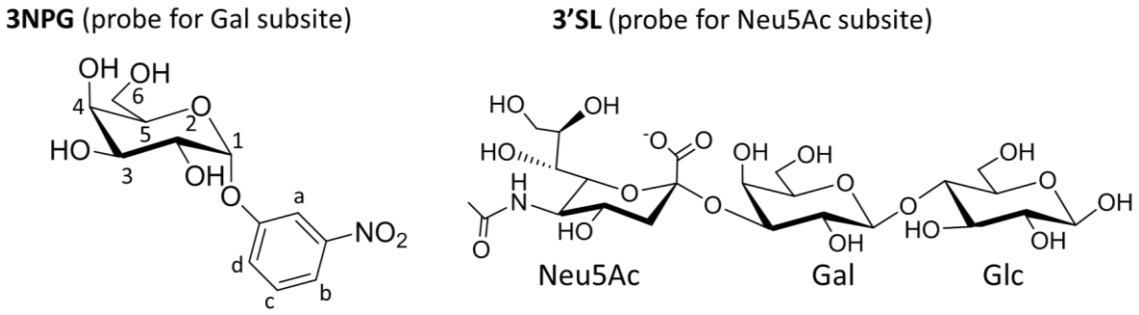

Figure 4.10. Chemical structures of 3-nitrophenyl-α-D-galactopyranoside (3NPG) and 3'-sialyllactose (3'SL). The numbering of 3NPG and the residue names of 3'SL are shown.

Four STD NMR competition experiments were carried out. The results are summarised in Table 4.1 and the spectra are shown in Figure 4.11 and in Figure A.1 of the Appendix.

Experiment	Reference Binder	Competitor	Displacement
(a)	**Ligand 30**	3NPG	Y
(b)	**Ligand 30**	**Ligand 33**	Y
(c)	**Ligand 33**	3NPG	N
(d)	**Ligand 33**	3'SL	N

Table 4.1. Table of results of the competition experiments. The column "Displacement" shows whether the competitor displaced (Y) or not (N) the reference binder.

In experiment **(a)**, **Ligand 30** was observed competing with 3NPG. When an equimolar fraction of 3NPG was added to the solution containing the **Ligand 30**/CTB complex, some displacement of **Ligand 30** was observed. In particular, the STD intensities of **Ligand 30** were halved, a result compatible with both compounds presenting a similar K_D for CTB (1.10 mM for 3NPG and 1.05 mM for **Ligand 30**). This confirms that the thio-galactose moiety of **Ligand 30** binds to the galactose subsite, previously occupied by 3NPG.

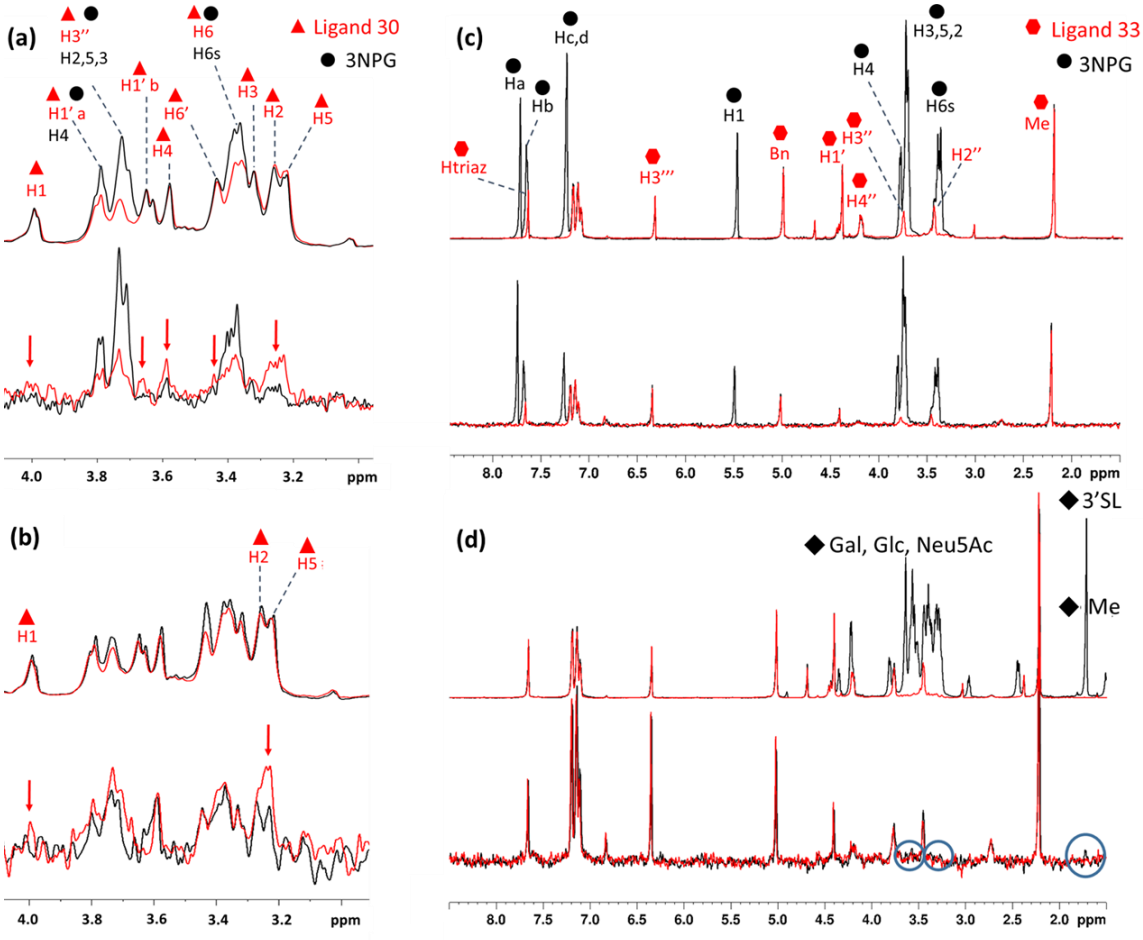

Figure 4.11. STD NMR spectra of competition experiments **(a), (b), (c)** and **(d)** (as labelled in Table 4.1). For each panel, the superposition of the reference spectra and the STD difference spectra is shown on the top and bottom, respectively. The spectra of the reference binder alone are shown in red and the spectra upon addition of the competitor are shown in black. Protons nomenclature is given in Figure 4.8b and 4.10. Red arrows on the difference spectra highlight the decrease of STD signals of the reference binders upon addition of the competitor. **(d)** The small 3'SL STD signals on the baseline of the difference spectra are indicated by a blue circle. A line broadening factor of 3 Hz was applied to the FID before FT. **Ligand 30** signals are highlighted by red triangles, 3NPG signals are highlighted by red hexagons, **Ligand 33** signals are highlighted by black circles and 3'SL signals are highlighted by black diamonds. Full spectra of competitions

(a) and **(b)**, and the STD NMR spectra of the 3'SL/CTB complex alone are reported in the Appendix (Figure A.1 and A.2).

In the competition between **Ligand 30** and **Ligand 33**, experiment **(b)**, only protons H1, H2, and H5 of ligand 30 could be monitored due to complete overlapping of the non-galactose peaks of the two ligands. The STD signals of these protons dropped upon addition of **Ligand 33** (K_D = 1.35 mM), thus showing that both Ligand 30 and Ligand 33 occupy a common subsite, most likely via the polyhydroxyalkylfuroate moiety. This could be expected based on the similar binding epitopes of the PHF region of both ligands, as reported in Figure 4.9.

In experiment **(c)**, **Ligand 33** was tested in competition with 3NPG. No competition was observed in this experiment, showing that the binding of **Ligand 33** does not involve the galactose subsite.

Finally, to investigate the binding location of the PHF moiety, competition between **Ligand 33** and 3'SL was carried out (experiment **(d)**). The ^{1}H STD intensities of **Ligand 33** were not affected upon addition of 3'SL. Despite being of low intensity, the STD signals corresponding to 3'SL can be seen in the baseline of both difference spectra (the STD NMR experiment on the complex 3'SL/CTB is shown in Figure A.2 of the Appendix). This suggests that the binding of **Ligands 30** and **33** does not involve the sialic acid subsite but may require the presence of an unknown subsite able to accommodate the PHF moiety of both ligands.

4.2.3 ILOE experiments: exploring proximity of the binding subsites

The scenario depicted by the STD NMR competition experiments did not exclude that **Ligand 30** and **Ligand 33** could compete for a subsite outside the GM1 binding pocket. In fact, the binding of **Ligand 30** to the galactose subsite is compatible with binding to a second binding pocket, located elsewhere on the protein surface, and where **Ligand 33** would bind too. To shed light on this, the proximity of the galactose subsite and the unknown subsite accommodating the PHF chain of both ligands must be probed. Following an ILOE approach[84], we ran tr-NOESY experiment on the potential ternary complex 3NPG/CTB/**Ligand 33** to explore the vicinity of the two binding subsites. As observed in experiment **(c)** in Table 4.1, 3NPG and **Ligand 33** bind independently from each other. Therefore, if they were bound in proximity to each other, *i.e.* their binding

subsites were adjacent, there would be chances of protons of both ligands getting close in space in the bound state, and then significant possibilities of observing inter-ligand cross peaks on the ILOE spectra. Expansions of the ILOE experiment are reported in Figure 4.12.

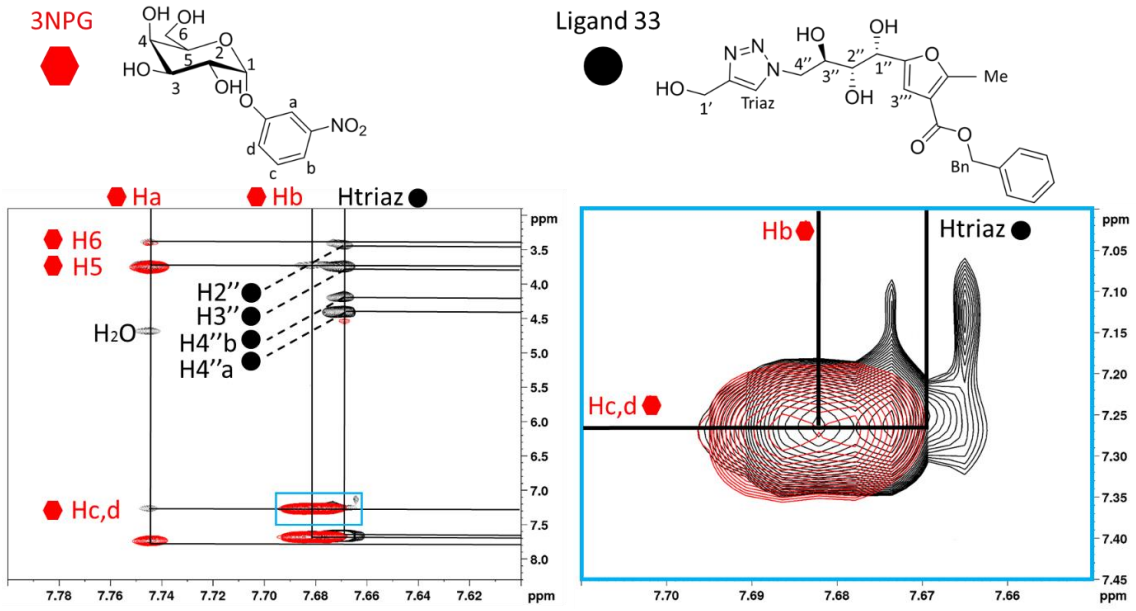

Figure 4.12. Left, expansion of the tr-NOESY experiments of the ternary complex 3NPG/CTB/**Ligand 33** (in black) and the control 3NPG/CTB (in red). Diagonal and cross peaks are assigned and the ILOE signal between Hc,d of 3NPG and Htriazole of **Ligand 33** is squared in turquoise and magnified on the right. 3NPG frequencies are highlighted by red hexagons, and **Ligand 33** frequencies are highlighted by black circles. The structure and nomenclature of the two ligands is reported here again for clarity.

In addition to the expected intra-ligand cross peaks from both ligands, we did observe a single inter-ligand cross peak between the triazole proton (Htriazole) of **Ligand 33** (7.67 ppm) and the Hc,d protons of 3NPG (positions meta and para to the nitro group, both at 7.27 ppm). As a control, the same experiment was performed on the binary 3NPG/CTB and **Ligand 33**/CTB complexes and on CTB alone (the control spectra and the full width spectra of the ternary complex are shown in the Section A.3 of the Appendix). The 7.67 ppm/7.27 ppm cross peak, which we interpret as an intermolecular contact between 3NPG and **Ligand 33**, was absent in the three control spectra. This confirmed that the cross peak genuinely comes from a close proximity between the two "fragments" **Ligand 33** and 3NPG in the bound state.

Albeit, arguably, the inter-ligand NOE could be mediated by the presence of an intermediate relay protein proton, the information encoded in the inter-ligand cross peak is that the Htriazole of **Ligand 33** and the Hc and Hd of 3NPG are closer to each other than any other pair of protons of the two molecules. This is a qualitative result of paramount importance in our study, proving:

i) that the subsite accommodating the PHF chain is adjacent to the galactose subsite (excluding unspecific interactions),

ii) that the experiment provided information on the binding orientation of **Ligand 33**, such that the triazole end is the part pointing towards the galactose subsite.

4.2.4 Molecular docking failed to provide a satisfactory model of the binding mode of the PHF chain

Translating the experimental data available from STD NMR competitions experiments and ILOE is into a quantitative binding model requires the construction of a 3D model compatible with the whole set of experimental data.

This is usually accomplished by molecular docking calculations. The available X-Ray structures of CTB bound to GM1 and 3NPG (PDB ID: 3CHB[126] and 1EEI[88]) were good starting points to model our thio-galactose ligands.

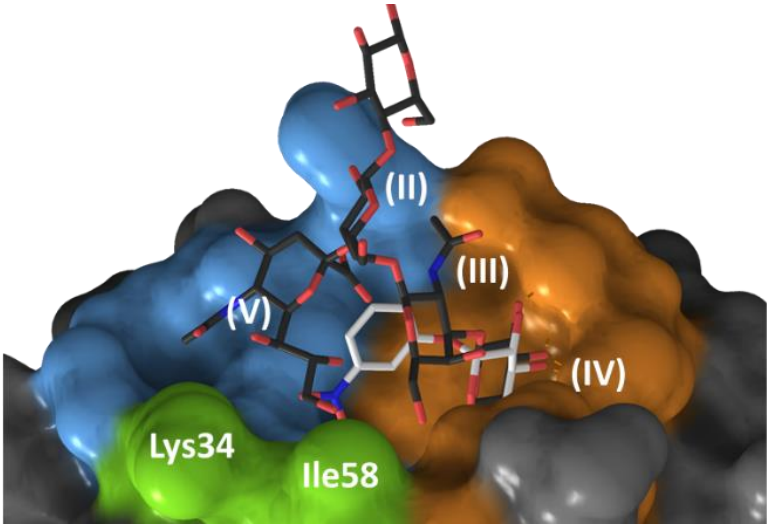

Figure 4.13. (a) Front view of GM1 (dark grey) and 3NPG (white) as bound to CTB (PDB ID: 3CHB[126] and 1EEI[88]). CTB is shown as a surface representation, with Ile58 and Lys34 coloured in green (these residues turned out to be fundamental for PHF binding, as reported in the following

sub-section). The galactose and sialic acid subsites are coloured in orange and blue, respectively. All ligand protons are omitted for clarity.

Figure 4.13 shows the front view of the GM1 binding pocket from the X-Ray structures of GM1/CTB and 3NPG/CTB, to visualise the relative orientation of the two binding subsites. It is interesting to pinpoint how the galactose of GM1 and of 3NPG fit the galactose subsite in the very same orientation, despite the different glyosidic linkage, in accordance to the specificity of the interaction.

Firstly, we carried out classic molecular docking (flexible ligand onto a rigid receptor). The coordinates of CTB from the X-Ray structure with GM1 (PDB ID: 3CHB) were used to generate the receptor grid and the docking parameters were optimised by re-docking GM1 to it with the aim of obtaining the complex as reported from the crystal structure (Figure 4.14a). With this optimised setting, docking of **Ligand 20** (Figure 4.9) showed that the thio-galactoside moiety fits in the galactose subsite with the same orientation as the galactose of GM1 and 3NPG (Figure 4.14b). The CORCEMA-ST calculations validated the model with an NOE R-factor of 0.21 (Figure 4.15).

Ligand 30 docked consistently and fit the thio-galactose in the expected subsite, as 3NPG, **Ligand 20** and GM1 (Figure 4.16a). Superposition of the galactose moiety of the lowest energy solution with the validated structure of **Ligand 20** showed good convergence for the orientation of the thio-galactosidic linkage, and the triazole of **Ligand 30** seemed to fit similarly to the acetamide group of **Ligand 20**.

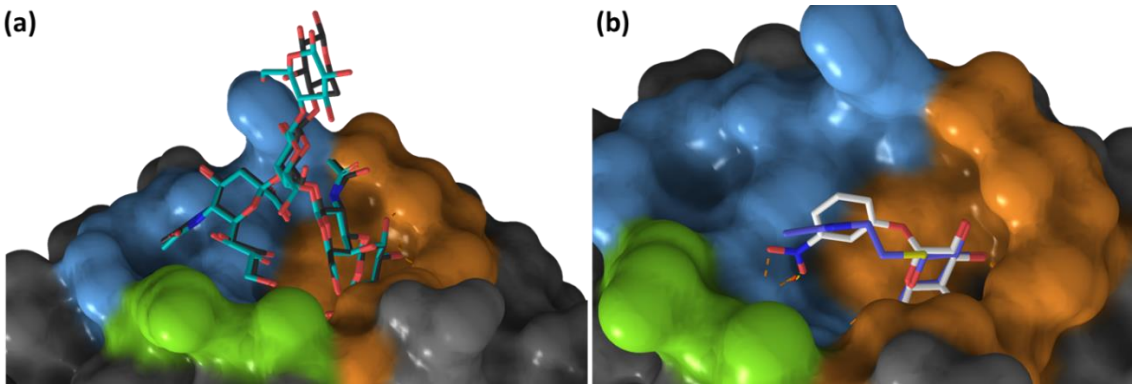

Figure 4.14. (a) Best docking solution of GM1 (cyan) in complex with CTB, superposed to the X-ray structure of GM1/CTB (grey), PDB ID: 3CHB[126]. **(b)** Best docking solution of **Ligand 20** (violet) in complex with CTB, superposed to the X-ray structure of 3NPG/CTB (white), PDB ID: 1EEI[88].

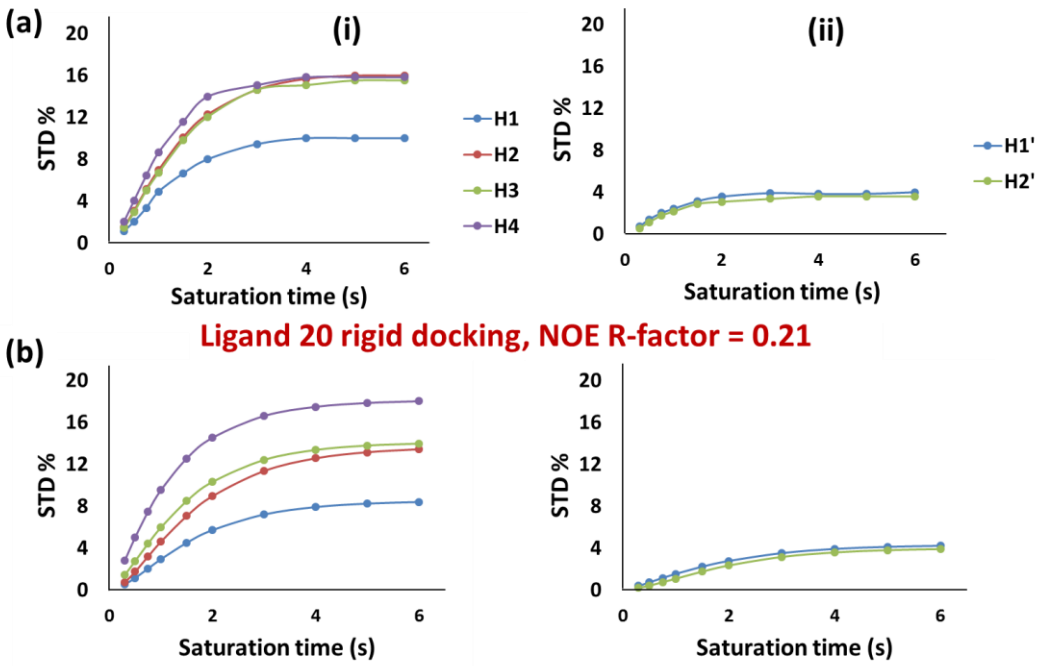

Figure 4.15. CORCEMA-ST calculations for the best docking solutions of **Ligand 20** with CTB. **(a)** Experimental STD NMR build-up curves for **Ligand 20** in complex with CTB[92]. **(b)** CORCEMA-ST derived STD NMR build-up curves for the lowest energy docking solution of the complex. For clarity, the molecule atoms are divided in two columns **(i)** and **(ii)**.

However, it should be noted that CORCEMA-ST did only validate the mode of binding of protons belonging to the thio-galactose moiety (Figure 4.17 (i, ii)). In the docking calculations, the PHF moiety failed to converge to a highly populated solution, sampling instead the whole surface around the galactose subsite and entirely hindering the sialic acid subsite, in disagreement with the competition experiments (Figure 4.16b). Accordingly, the CORCEMA-ST calculations for protons from the PHF chain to the aromatic moiety did not match the experimental build-up curves (Figure 4.17 (iii, iv)).

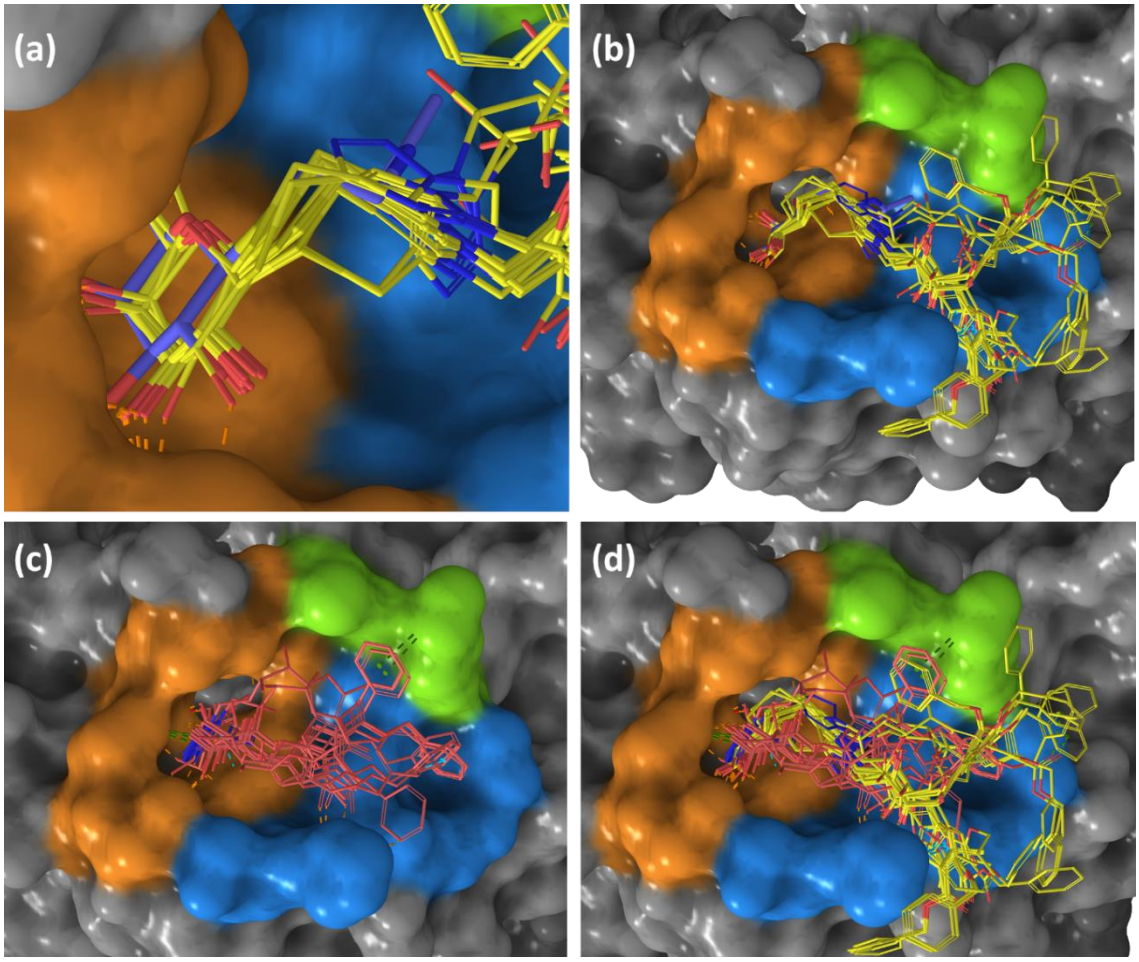

Figure 4.16. (a) Zoom in of the galactose subsite: the 15 lowest energy docking solution for **Ligand 30** (yellow) as superimposed to the lowest energy docking solution of **Ligand 20** (violet); **(b)** zoom out of **(a)** to the entire GM1 binding pocket. **(c)** The 15 lowest energy docking solution for **Ligand 33** (light red); **(d)** superimposition of **(b)** and **(c)**.

On the other hand, the docking solutions for **Ligand 33** failed to converge at all, neither among them nor with the PHF moiety of **Ligand 30** (Figure 4.16c and d). Most solutions occupied both the galactose and sialic acid subsites, and CORCEMA-ST calculations failed to validate the model (Figure 4.18).

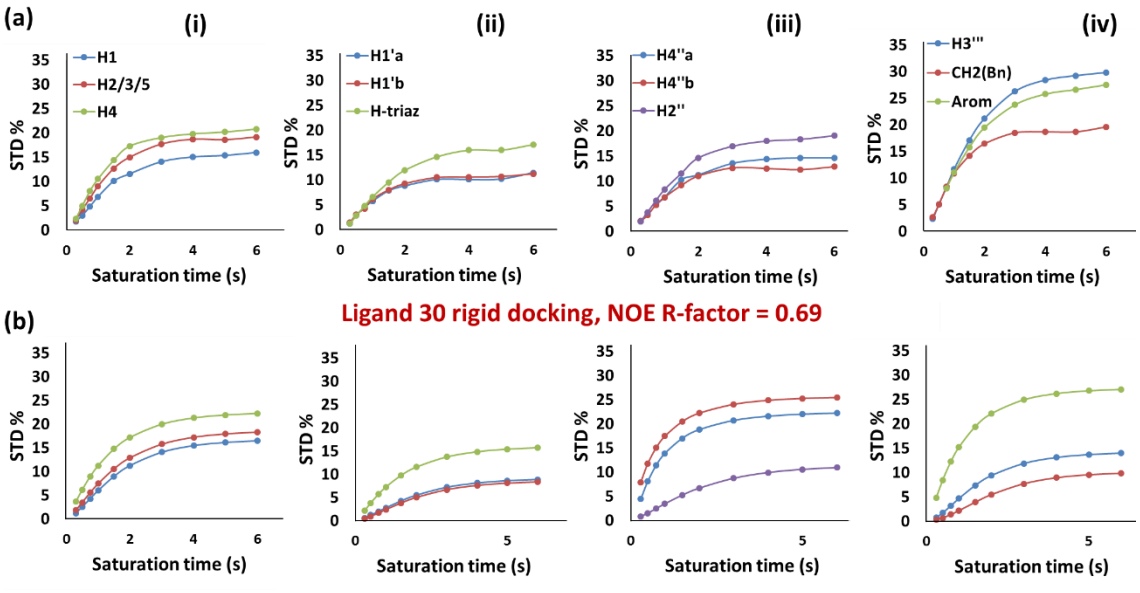

Figure 4.17. CORCEMA-ST calculations for the best docking solutions of **Ligand 30** with CTB. **(a)** Experimental STD NMR build-up curves for **Ligand 30** in complex with CTB[92]. **(b)** CORCEMA-ST derived STD NMR build-up curves for the lowest energy docking solution of the complex. For clarity, the molecule atoms are divided in four columns **(i)**, **(ii)**, **(iii)** and **(iv)**.

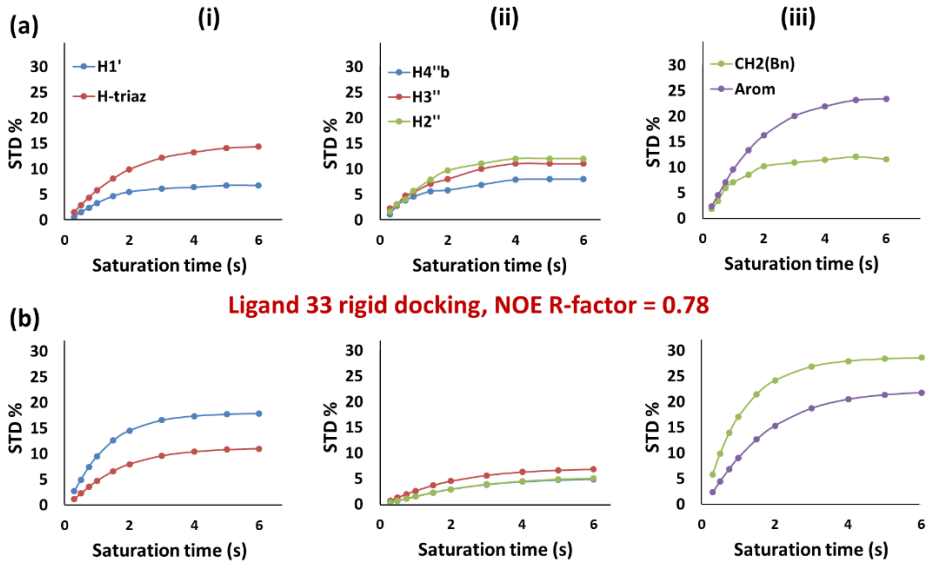

Figure 4.18. CORCEMA-ST calculations for the best docking solution of **Ligand 33** with CTB. **(a)** Experimental STD NMR build-up curves for **Ligand 33** in complex with CTB[92]. **(b)** CORCEMA-ST derived STD NMR build-up curves for the lowest energy docking solution of the complex. For clarity, the molecule atoms are divided in four columns **(i)**, **(ii)** and **(iii)**.

4.2.5 MD and CORCEMA-ST validation of a novel subsite accommodating the PHF moiety

One of the main weaknesses of considering the receptor as rigid in docking calculations is that potential conformational rearrangements of the receptor are not considered. To account for the flexibility of the ligand and the receptor, a thorough molecular dynamics (MD) of the system was performed by MSc Samuel Walpole in our group.

In the first place, to enhance sampling of the free state of the receptor, a Hamiltonian Replica Exchange Molecular Dynamics (HREMD) method was applied to the apo CTB. HRMED approaches have previously been shown to be effective in identifying druggable transient binding sites[138]. Analysis of the trajectory over all five subunits consistently revealed the presence of a transient groove close to the known galactose and sialic acid subsites. While this groove is closed in the crystal structure of CTB, opening is facilitated by rotation of the Ile58 and Lys34 sidechains (Figure 4.19a), both in the presence and in the absence of the native ligand GM1.

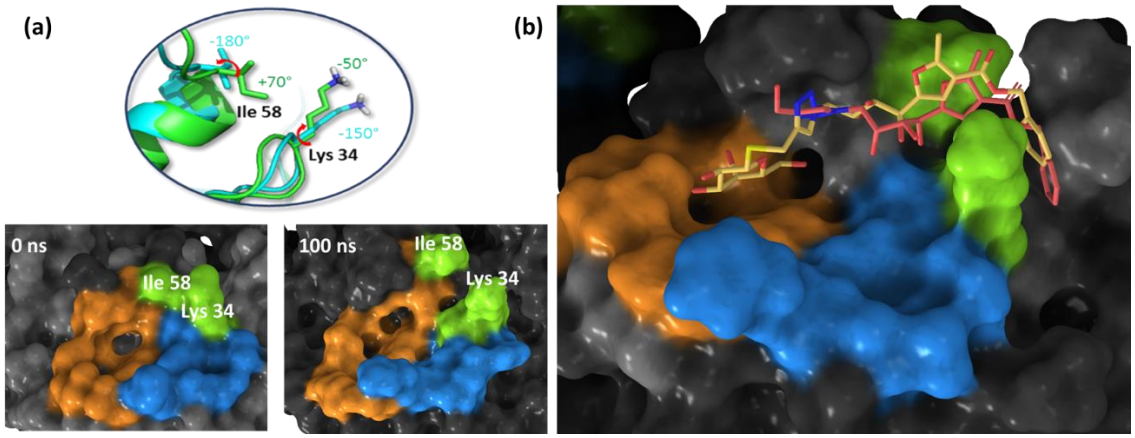

Figure 4.19. **(a)** Top: novel CTB subsite in the open (turquoise) and closed (green) conformations overlaid. The red arrows indicate the χ1 torsional rotation from closed to open states for the two residues involved (χ1 values reported for both). Bottom: top view of the closed and open conformations in surface representation, showing the initial and final frames of the HRMED simulation of apo CTB. **(b)** Representative frames of the most populated clusters of **Ligand 30** (pale yellow) and **Ligand 33** (light red) with CTB ("open" conformation). In both panels, the residues forming the galactose binding subsite are shown as orange surface. The residues constituting the sialic acid binding subsite are shown as blue surface. Ile58 and Lys34, which constitute the novel groove, are shown as green surface and labelled.

Since no large chain rearrangement is observed, the energy barrier between the "closed" and "open" conformations is expected to be very small. Furthermore,

repeating the simulation with the receptor bound to GM1 revealed that opening of this groove still takes place in the presence of the native ligand. This shows that this subsite is accessible even when the galactose and sialic acid sites are engaged, which represents a structural feature of crucial importance for further inhibitors design.

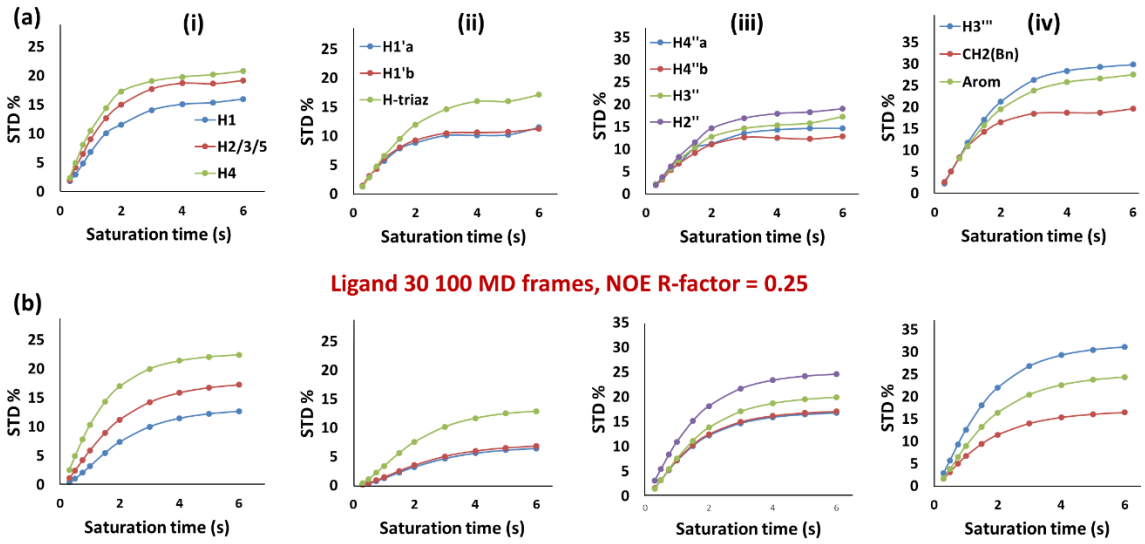

Figure 4.20. **(a)** Experimental STD NMR build-up curves for **Ligand 30** in complex with CTB[92]. **(b)** CORCEMA-ST calculated STD NMR build-up curves averaged over the 100 most representative frames from MD of **Ligand 30** with CTB.

Docking of **Ligand 30** to this "open" conformation again showed good convergence of the galactose residue to the galactose subsite, with the PHF moiety consistently occupying the novel subsite, suggesting that it acts as specific subsite for the non-carbohydrate part of the ligand. Starting from the best solution of the open conformation docking calculations, a 100 ns molecular dynamics simulation of this complex demonstrated that this conformation is dynamically stable, and the novel subsite is occupied for most of the simulation (Figure 4.19b).

The same approach was applied to **Ligand 33**, which docked consistently in the "open conformation" occupying the novel subsite, as opposed as when docked in the "closed onformation". A 100 ns dynamics simulation of the **Ligand 33** docking solution in the open conformation of CTB revealed an even greater dynamic stability when compared to the **Ligand 30**/CTB complex, with the PHF chain of **Ligand 33** exclusively occupying the novel subsite throughout the simulation. Without a galactose moiety to act as an anchor, **Ligand 33** can bury itself further into the hydrophobic binding subsite (Figure 4.19b). To validate the MD data, we performed CORCEMA-ST calculations on the 100

most populated frames from each of the MD simulations of the CTB complexes with **Ligand 30** and **Ligand 33**. In this case, the agreement with the experimental data was very satisfying, with an NOE R-factor of 0.25 and 0.31, respectively (Figure 4.20 and 4.21).

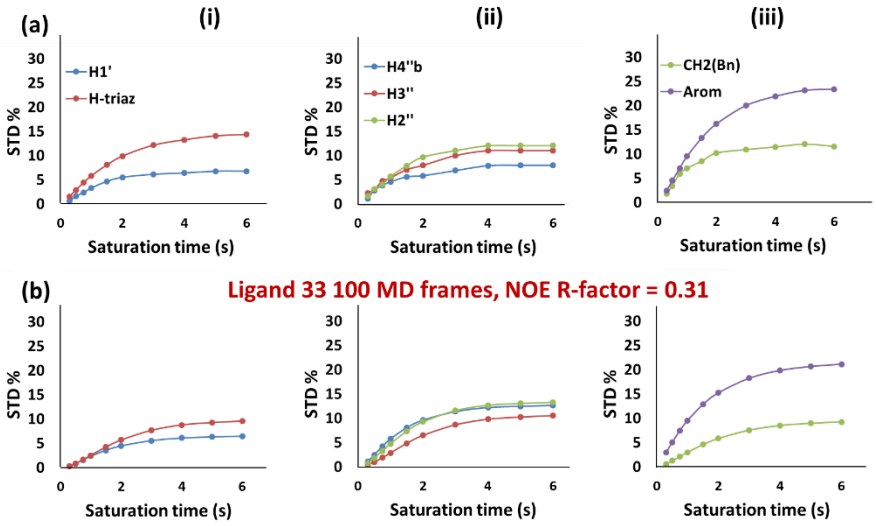

Figure 4.21. Experimental STD NMR build-up curves for **Ligand 33** in complex with CTB[92]. **(b)** CORCEMA-ST calculated STD NMR build-up curves averaged over the 100 most representative frames from MD of **Ligand 33** with CTB.

The binding mode of **Ligand 33** as validated by CORCEMA-ST in the novel binding subsite is compatible with the presence of 3NPG in the galactose binding subsite (as suggested by experiment **(c)** in Table 4.1 and by the ILOE data). To investigate this hypothesis, we superposed the structure of 3NPG/CTB available by X-Ray (PDB ID: 1EEI[88]) and the most representative frame from the MD simulations of the **Ligand 33**/CTB and we proposed it as a qualitative model for the ternary complex (Figure 4.22). In the resulting model containing the two ligands, indeed, the Htriazole of **Ligand 33** and the aromatic protons Hc and Hd of 3NPG are at a distance close enough as to be compatible with generation of inter-ligand NOE.

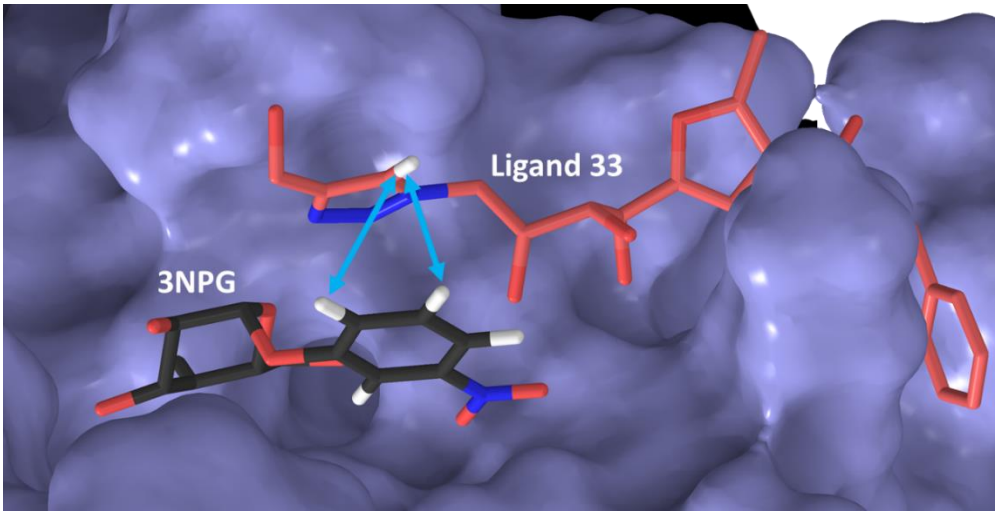

Figure 4.22. Top view of the superposition of 3NPG from PDB ID: 1EEI[88] (dark grey) with the most representative MD frame of **Ligand 33** (light red) in the CTB open binding site (homogenous surface colouring). All protons are omitted, but the Htriaz of **Ligand 33** and the aromatic protons of 3NPG, while the turquoise double-point arrows shows the inter-ligand NOE observed in the ILOE tr-NOESY in Figure 4.12.

4.2.6 Inter-Ligand-STD NMR (IL-STD NMR): an additional tool to detect adjacent binding subsites

Even without being quantitative, the model proposed for the ternary complex in Figure 4.22 agrees with the ILOE results. This gave us ground to approach the system by IL-STD NMR, as introduced in Chapter 3, Section 3.4. The ternary complex at a 1 to 40 protein to ligand ratio was analysed by STD NMR, first at the standard irradiation frequency of 0.60 ppm (aliphatic region) and then at 7.27 ppm, on resonance with the protons Hc and Hd of 3NPG.

By simple visual analysis of the STD difference spectra acquired at 0.60 ppm and 7.27 ppm, an overall decrease of the STD signals when irradiating at the aromatic region (7.27 ppm) rather than at the aliphatic region (0.60 ppm) was evident. This evidence is easily explained by the generally lower occurrence of aromatic residues relative to aliphatic residues, meaning that a lower number of directly irradiated protein protons is irradiated at 7.27 ppm than at 0.60 ppm. This translate into less overall available magnetization and lower STD intensities upon irradiation in the aromatic region. An expected increase was observed for the STD intensities of the peak coming from Hc and Hd (7.27 ppm, directly irradiated) and for the aromatic signals of **Ligand 33**, as an effect

of direct irradiation (these resonate within 0.15 ppm high-field from it, respectively 7.21 ppm, 7.16 ppm, 7.12 ppm). Also, the intensity of the Hb proton (7.69 ppm) increased when irradiating the adjacent Hc,d, due to intra-ligand NOE.

Interestingly, we observed an increase in the STD intensity of proton Htriazole of **Ligand 33** (7.67 ppm) upon irradiation of proton Hc and Hd of 3NPG (7.27 ppm) (Figure 4.23). This STD increase cannot be due to direct irradiation (the signal is 0.40 ppm apart from the source of irradiation) or intra-ligand NOE (as the directly irradiated aromatic protons of **Ligand 33** are far in space from the triazole); hence, the signal increment can only be explained by the inter-ligand transfer of magnetization between Hc and d and Htriazole, which is only possible if those protons from the two ligands established a close contact in the bound state. This is in perfect agreement with the ILOE data (Figure 4.12).

As argued in Chapter 3, Section 3.4, the risk of direct irradiation is the main limitation of this approach. For example, the inter-ligand correlation Hc and Hd/Htriazole observed by ILOE could be detected by IL-STD only via irradiation on the Hc and Hd signal of 3NPG. This is because the irradiation of Htriazole would have implied the direct irradiation of the 3PNG, due to the proximity of Htriazole to Hb of 3NPG, thus leading to spurious results.

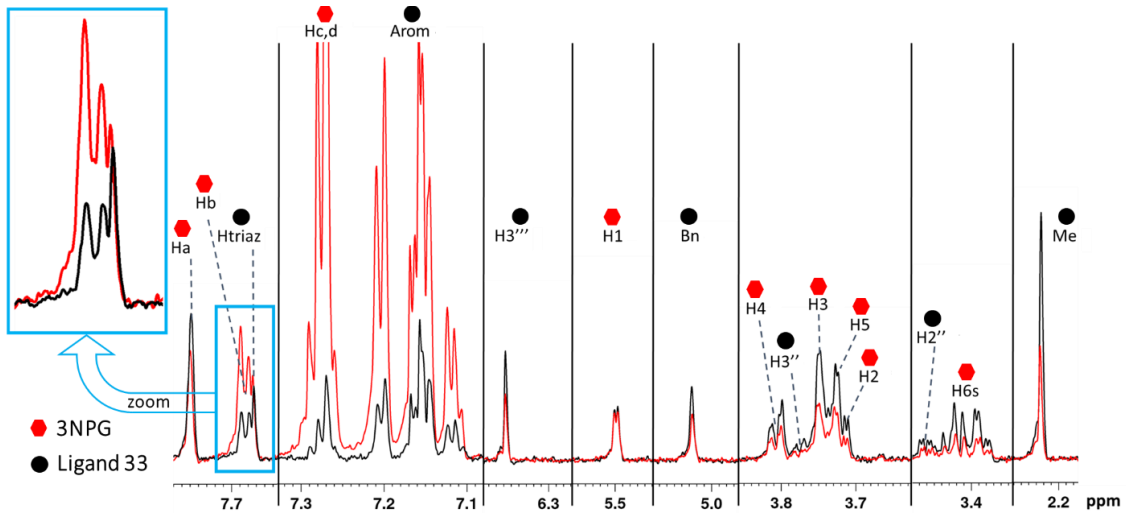

Figure 4.23. IL-STD difference spectra of the ternary complex 3NPG/CTB/**Ligand 33** with irradiation at 0.60 ppm (in black) and at 7.27 ppm, on resonance with Hc and Hd of 3NPG (in red). The spectra are zoomed on the single protons for space optimization purposes. The assignment of all peaks is given and the peak of Htriazole is squared in turqouise and magnified, showing increased intensity when irradiating at 7.27 ppm, whereas all the other protons (with

the exceptions discussed in the main text) decrease. A saturation time of 2 s and line broadening factor of 1 Hz were employed.

To exclude any possible artefacts, acquisition of control experiments under the same experimental conditions on the two binary complexes was necessary. The control experiment carried out on the binary complexes **Ligand 33**/CTB showed that, in the absence of 3NPG, the intensity of the triazole proton (as well as the rest of protons of **Ligand 33**) decreased when irradiating at 7.27 ppm, ruling out that the observed IL-STD was due to direct irradiation. An increase was observed for the aromatic protons between 7.24 and 7.04 ppm, which experience direct irradiation as indicated above (Figure 4.24a).

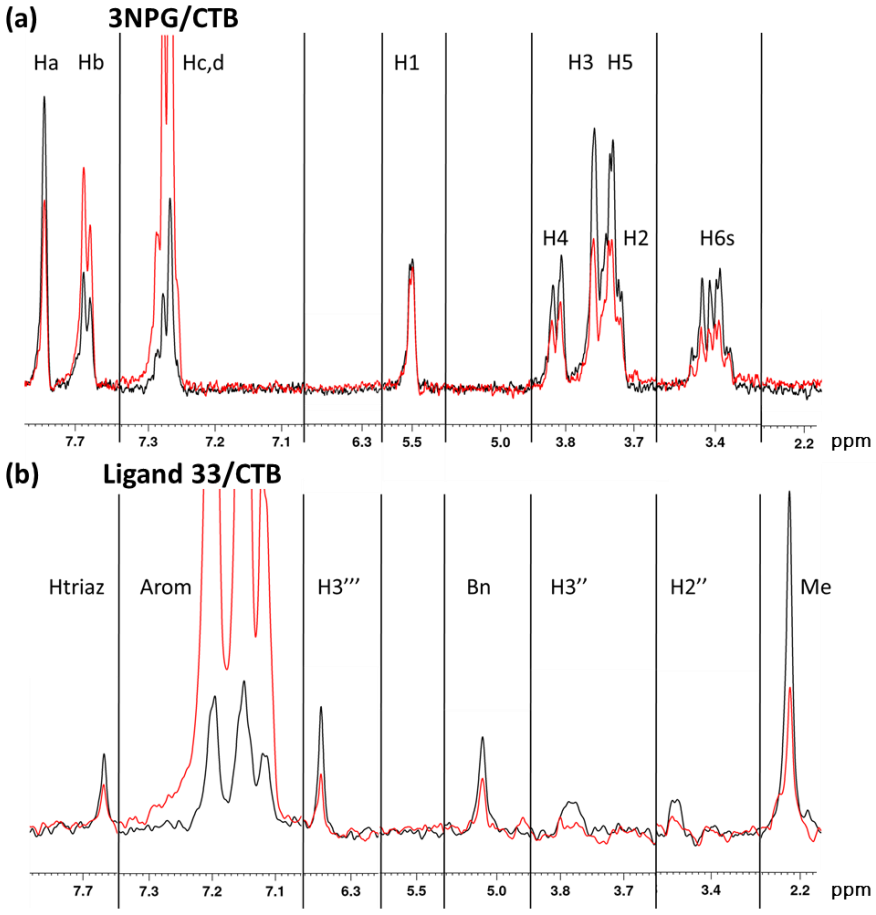

Figure 4.24. IL-STD control experiments on **(a)** 3NPG and **(b) Ligand 33** in a complex with CTB. Difference spectra with irradiation frequency of 0.60 ppm are shown in black, and difference spectra with direct irradiation at 7.27 ppm are shown in red. Same experimental conditions as for the spectra in Figure 4.23.

The control experiment on the binary complex 3NPG/CTB is shown in Figure 4.24b. Expectedly, all the STD intensities decreased when irradiated at 7.27 ppm, except for

the directly irradiated protons Hc and Hd and the Hb, which received direct intra-ligand NOE from the adjacent Hc and Hd.

Quantitatively, these data can be better analysed by determining the IL-STD factor for each proton of the ligands constituting the ternary complex, and by comparing them with the IL-STD factor obtained in the binary complexes used as control samples (Figure 4.25). The equation for IL-STD calculation is described in Sub-section 3.4.2 (Equation 3.6). In our case, *N* = 1, therefore the equation simplifies to:

$$\Delta\mathrm{IL}-\mathrm{STD}_{i,ternary} = \frac{STD_{\ on-ligand,i}}{STD_{\ off-ligand,i}} - 1$$

By definition, the IL-STD factor will only be positive for protons receiving inter-ligand saturation from the irradiated proton of the adjacent fragment (if we are able to exclude the protons receiving extra saturation by intra-ligand NOE or direct irradiation excluded from the analysis).

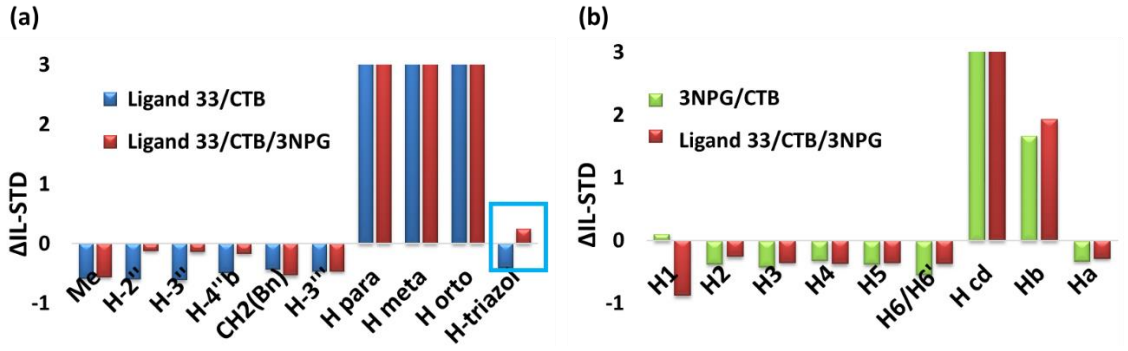

Figure 4.25. ΔIL-STD histograms of the 3NPG/CTB/**Ligand 33** ternary complex, and the two control complexes **Ligand 33**/CTB and 3NPG/CTB (and controls), represented individually for **Ligand 33 (a)** or 3NPG **(b)**. The values relative to the Htriazole of **Ligand 33** are squared in turquoise. The bar chart is cut at 3.0. Raw and processed data in Table 4.2.

In fact, protons presenting positive ΔIL-STD in both the experiment and the control are subject to either intra-ligand NOE or direct irradiation (this is easily determined by looking at the spectra), while protons presenting negative values in both the experiment and the control are too far in space from the irradiated proton(s) to be affected. By looking at the histogram representation in Figure 4.25a we can easily appreciate how the Htriazole from **Ligand 33** is the only proton with a positive ΔIL-STD only in the ternary complex experiment (and a negative ΔIL-STD in the control experiment). The IL-STD data are reported in Table 4.2.

		Ligand 33/CTB/3NPG			Ligand 33/CTB		
^{1}H δ	Atom ID	STD% off ligand	STD% on ligand	ΔIL-STD	STD% off ligand	STD% on ligand	ΔIL-STD
2.24	**Me**	11.46	5.02	-0.56	10.91	4.54	-0.58
3.48	**H-2''**	4.19	3.69	-0.12	5.62	2.28	-0.59
3.79	**H-3''**	4.95	4.26	-0.14	5.99	2.33	-0.61
4.23	**H-4''b**	2.9	2.41	-0.17	3.28	1.67	-0.49
5.03	**CH_2(Bn)**	6.97	3.29	-0.53	6.14	3.45	-0.44
6.36	**H-3'''**	10.61	5.68	-0.46	10.7	5.73	-0.46
7.12	**H para**	9.16	42	3.59	11.57	52	3.49
7.16	**H meta**	11.13	54	3.85	10.5	57	4.43
7.21	**H orto**	9.25	65	6.03	10.15	73	6.19
7.68	**Htriazol**	7.83	9.71	0.24	5.84	3.48	-0.40
					3NPG/CTB		
5.50	**H1**	12.27	1.42	-0.88	12.51	13.43	0.11
3.73	**H2**	16.6	12.27	-0.26	18.49	11.92	-0.38
3.76	**H3**	18.13	11.69	-0.36	19.79	11.69	-0.42
3.81	**H4**	16.77	10.5	-0.37	17.44	11.24	-0.32
3.73	**H5**	17.1	11.02	-0.36	18.31	11.24	-0.38
3.40	**H6/H6'**	12.27	7.68	-0.37	15.81	6.7	-0.55
7.27	**H cd**	6.9	99	13.35	7.68	94.31	11.65
7.69	**Hb**	12.63	36.96	1.93	13.27	36.25	1.68
7.75	**Ha**	21.61	15.36	-0.29	23.59	15.81	-0.34

Table 4.2. Raw STD intensities for the ternary complex (3NPG/CTB/**Ligand 33**) and the two controls. The $STD_{off\ ligand}$ at 0.60 ppm and $STD_{on\ ligand}$ at 7.27 ppm, as well as the calculated ΔIL-STD are shown. The ΔIL-STD relative to Htriazole are highlighted in red.

4.2.7 Relevance of the discovery of a novel binding subsite and the IL-STD approach from a drug design perspective

The work presented in this section can be regarded as an example of the strong potential of NMR when combined to molecular modelling approaches, in drug discovery contexts. The combination of both techniques becomes particularly precious when studying systems unsuitable for crystallisation, a recurrent situation for highly flexible and low affinity ligands. In this project, the combination of these methods allowed to establish a

milestone in cholera toxin inhibitors design: the discovery of a hitherto unknown binding subsite in the GM1 binding pocket of CTB (Figure 4.26). For the first time here, we used MD to obtain a 3D model of CTB-inhibitors as bound to the GM1 binding pocket, strongly supported by high resolution STD NMR data, as confirmed by CORCEMA-ST. Importantly, the flexibility of the ligands in solution was taken into account by simulating build-up curves for the complexes (**Ligand 30**/CTB and **Ligand 33**/CTB) over the 100 most representative frames obtained by MD simulations, with excellent outcomes. The NOE information extracted by both ILOE and IL-STD was also crucial to establish the relative orientation of the subsites.

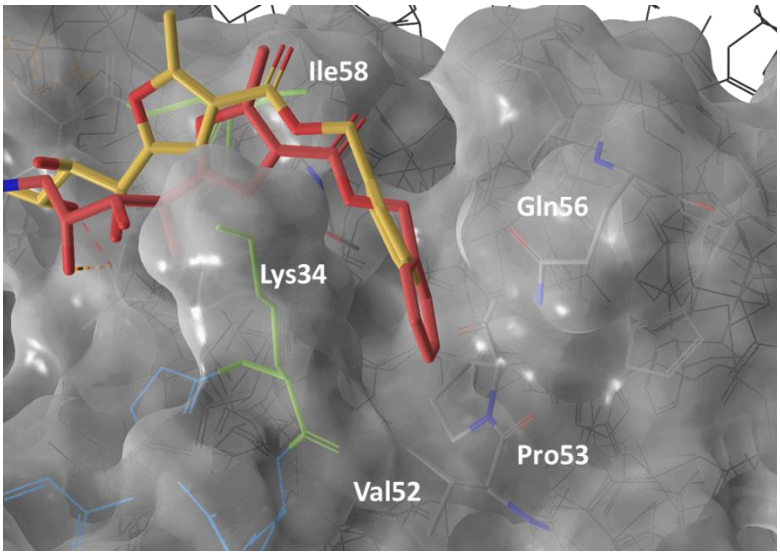

Figure 4.26. Representative frames of the most populated clusters of **Ligand 30** (pale yellow) and **Ligand 33** (light red) with CTB ("open" conformation). Zoom on the furoate-benzyl moieties: residues forming a slightly lipophilic cleft are represented in thin sticks.

Whereas the traditional Fragment Based Drug Discovery (FBDD) approach relies on combinations of simple fragments to construct effective leads, we decomposed a good-performing lead (**Ligand 30**) to understand its binding mode and explore the surroundings of the two well-characterised binding subsites, finding a new groove with good specificity for the non-carbohydrate and non-hydrolysable PHF moiety.

To the purpose of the exploration of the binding pocket, the flexibility of the ligands was an advantage (**Ligand 30** contains 10 rotatable bonds). Although, comparing the K_D of **Ligand 30** (1.05 mM), **Ligand 33** (1.35 mM) and 3NPG (1.10 mM), one can argue that the flexibility of the PHF chain provides a considerable entropic disadvantage to the binding

event. The similar K_Ds of 3NPG and **Ligand 30** suggest that the enthalpic gain due to the fit of the furoate and aromatic ring in the novel binding subsite is compensated by the negative entropy associated with the binding of the flexible ring. It is also interesting to remark that **Ligand 33**, even lacking the galactose moiety and leaving the galactose subsite free, has an almost comparable K_D to **Ligand 30**. This could be justified by the lack of "galactose-anchor", which allows **Ligand 33** to bury the aromatic moiety deeper in the slightly lipophilic novel subsite (Figure 4.26).

The furoate ring of **Ligand 33** fits tightly between Lys34 and Ile58, and its benzyl ring interacts closely with Lys34, Val156 and Pro157. On the other hand, **Ligand 30** seems to be slightly more solvent exposed, for not being able to reach the same depth of **Ligand 33**. It is interesting to point out that the residues Lys34 and Ile58 had already been noted as a lipophilic patch with the potential of interacting with hydrophobic ligands by Fan and co-workers, when they provided the X-Ray structure of 3NPG and three aromatic derivatives with CTB[88]. As in the X-Ray structure of GM1 with CTB, the two residues were modelled in the close conformation. With the HREMD approach, we clearly showed that the intuition of Fan and co-workers was well-grounded.

We believe that our findings pave the path for the design of three-finger ligands, to be tuned to an optimal degree of rigidity allowing the simultaneous occupation of the three subsites with minimal loss of entropy.

4.3 Exploring the three-subsite GM1 binding pocket of CTB: NMR structural investigation of small analogues of Ligand 30

4.3.1 A novel set of small of "small CTB ligands"

The route to three-finger CTB inhibitors design based on **Ligand 30** requires a further dissection of the thio-galactose/PHF scaffold. In this way we can: i) understand the contribution of each fragment to the binding specificity, and ii) probe the potential offered by the thio-glycosidic linkage (which can be decorated, as opposed to the O-glycosidic linkage).

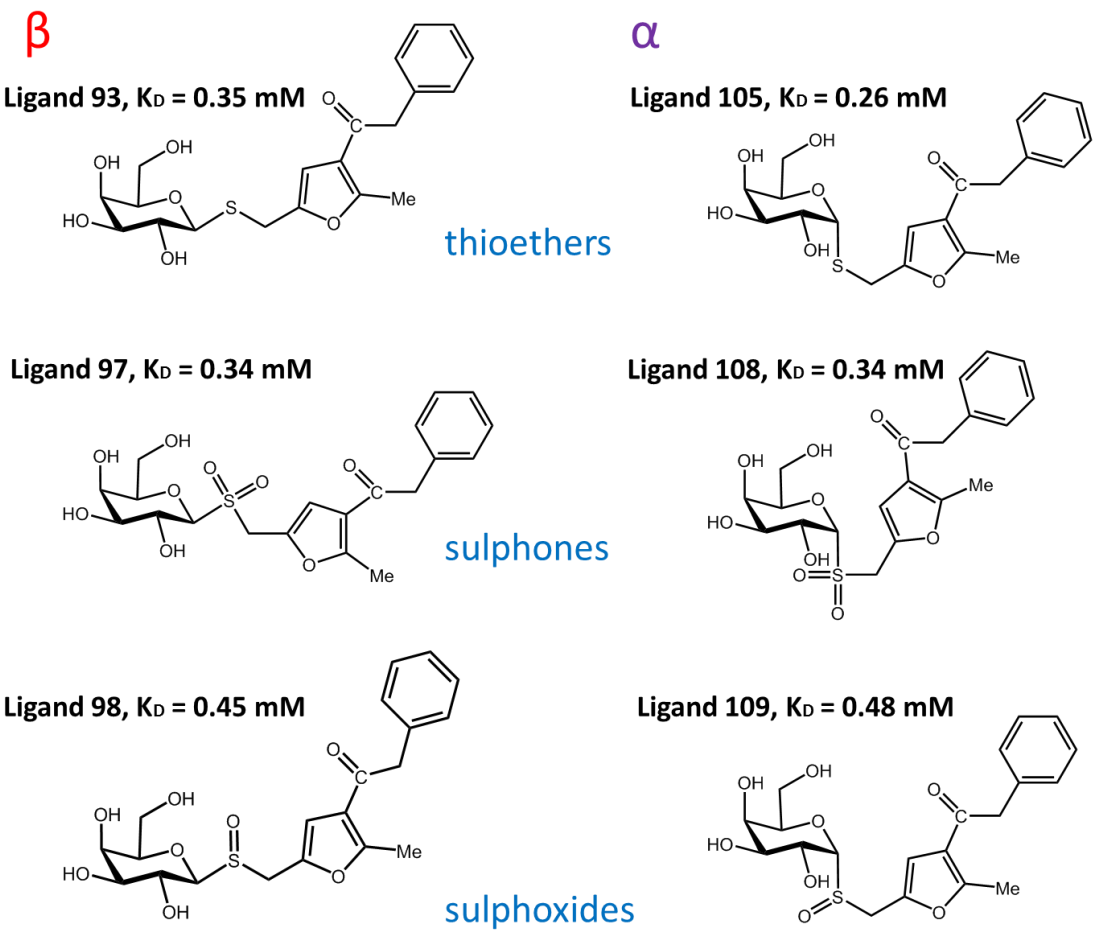

Figure 4.27. Chemical structures of the six analogues of **Ligand 30** analysed in this section. The K_D values measured by weak affinity chromatography (WAC) are given for each ligand (unpublished).

To this aim, our collaborators from the University of Seville provided us with a set of "small" **Ligand 30** analogues, that we call the "*SCJ set*", in reference to MSc Sebastian Carrion Jimenez who synthesised them. The novel compounds lack the triazole-polyhydroxy linker and have the thio-galactose residue directly connected to the furoate-benzyl moiety. The variability across the set is given by the anomeric configuration (α or β), and the oxidation state of the glycosidic sulphur (present as a thioether, sulphone or sulphoxide functional group), for a total of 6 compounds (Figure 4.27).

The ligands are named numerically as **Ligands 93**, **97**, **98** (β-thioether, -sulphone, and -sulphoxide respectively) and **Ligands 105**, **108** and **109** (α- thioether, -sulphone, and -sulphoxide respectively). The K_D values of the ligands were determined by weak affinity

chromatography (WAC) in the group of Maria Bergstrom at Linneus University (Sweden), following an approach developed by Ohlson and co-workers in 2009[91]. Interestingly, these ligands showed higher affinities (K_D values ranging from 0.35 mM to 0.55 mM), as compared to **Ligands 20**, **30** and **33**, whose K_D was above 1 mM[92].

In the *SCJ set*, the presence of the two main elements of recognition of **Ligand 30**, *i.e.*, the thio-galactose and the fuorate-benzyl moieties, linked through one single carbon atom, poses interesting questions. The smaller size of the ligands does not possibly allow them to occupy the galactose subsites and the novel binding subsite at the same time, as **Ligand 30** does. Therefore, we concentrated our efforts in finding out the subsite(s) where the smaller ligands bind to. In addition, we aimed to gain insights on the relative specificities of the two pharmacophores, in particular, whether the affinity of the furoate-benzyl moiety for the non-carbohydrate novel subsite is significant enough as to direct the ligands to bind to the novel groove rather than to the galactose subsite. To that aim, we combined STD binding epitope determination, DEEP-STD NMR, DEEP-STD on-resonance scanning, STD competition experiments and tr-NOESY experiments.

4.3.2 STD NMR binding epitope mappings: larger impact of sulphur oxidation state on the binding mode in comparison to the effect of anomeric configuration

First, we acquired the STD NMR build-up curves for the *SCJ set* in complex with CTB. The six binding epitopes looked rather different (Figure 4.28). The galactose moiety seemed to be the main recognition element across the set of ligands, with **Ligand 93** and **Ligand 97** showing strong contacts from H2 to H6 of the sugar ring, while the other four ligands have a less uniform STD pattern, where alternatively H3, H4 or H5 and H6 are the closest protons to the surface. It is also interesting to highlight that the protons sitting on the carbons adjacent to the sulphur show the weakest contact to the protein surface, while the furoate and the benzyl group always show from medium to strong contacts. Particularly, **Ligand 98** showed a binding epitope indicating that the whole ligand is intimately contacting the binding pocket surface in the bound state.

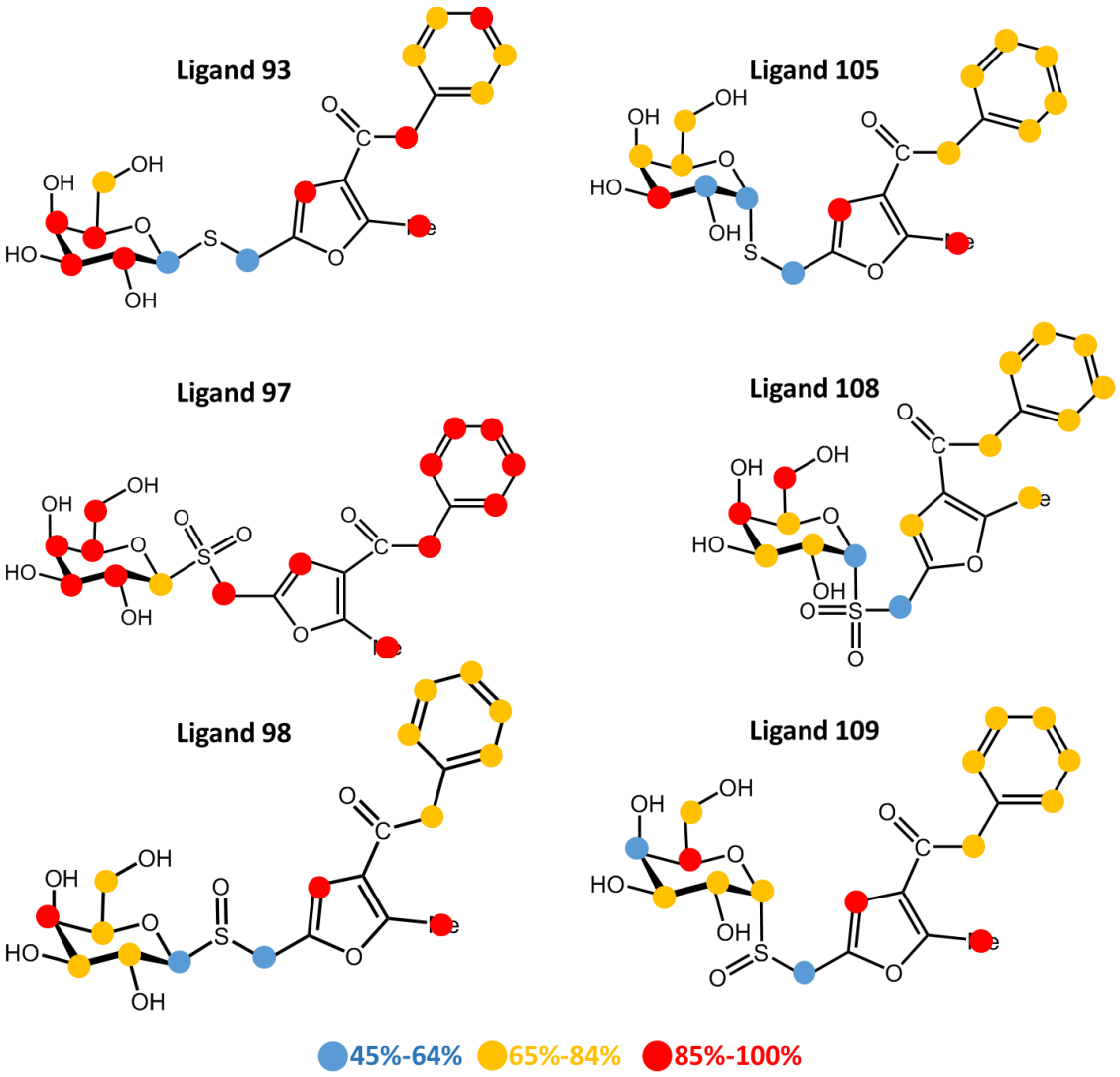

Figure 4.28. STD NMR binding epitope mapping of the "small" ligands, based on the normalised STD values from the initial slope of each proton, relative to the most intense one (for which a 100% is arbitrarily assigned). Legend indicates weak (blue), medium (yellow) and strong (red) contacts. The proton assignment of the ligands, the raw build-up curves data and the normalised STD values are reported in Section A.4 of the Appendix.

Albeit the many differences, some analogies were observed between pairs of ligands with the same sulphur oxidation state, especially between the two thioethers (**Ligand 93** and **105**) and the two sulphoxides (**Ligand 98** and **109**). Interestingly, more significant differences were observed between the two sulphones (**Ligand 97** and **108**) and among the triplets of ligands in the same anomeric configuration. This suggests that the oxidation state of the sulphur in the thio-glycosidic linkage is a key determinant of the ligand binding mode, showing a larger effect than the anomeric configuration.

4.3.3 DEEP-STD fingerprint: zoom on the galactose subsite

In Chapter 3, Section 3.3, we described the DEEP-STD approach and validated it against two complexes with high-resolution crystal structures[87]. Here, we show how we can use DEEP-STD NMR to obtain a "fingerprint" of the binding pocket, thus aiding to discriminate whether different ligands bind to the same subsite. This implies comparing the differential epitope map of a complex of known 3D structure with the differential epitope maps of unknown ligands (with the same protein and under the same differential conditions). If the unknown ligands respond to the differential conditions similarly to how the known ligand responds, *i.e.* their DEEP-STD maps are comparable, it means that they are exposed to the same residues (or the same kind of residues) and they most likely occupy the same binding pocket (or highly similar binding pockets).

Unfortunately, at this stage, the two sulphoxides, **Ligand 98** and **109**, had to be excluded from the analysis, for the following reasons. Firstly, they had the lowest affinities (K_D 0.48 mM and 0.49 mM), but, more importantly, they showed a significant degree of hydrolysis in the sample over time. As a result, some of the peaks coming from the hydrolysis overlapped with the peaks of the parent ligand, and although the sub-products did not give STDs, their presence complicated the STD analysis and interpretation.

Hence, to elucidate the location of binding of these four ligands (**Ligand 93**, **97**, **105**, **108**), we compared the differential irradiation DEEP-STD maps (2.25 ppm/0.60 ppm) with the DEEP-STD map (2.25 ppm/0.60ppm) of the 3NPG/CTB complex, of known 3D structure from X-ray diffraction. For an optimal signal-to-noise ratio and a reduced experimental time, the DEEP-STD NMR experiments for these systems were performed with 2 s saturation time (rather than 0.5 s as in the proof of concept study; the effect of the saturation time in DEEP-STD studies has been discussed in Sub-section 3.4.5). The DEEP-STD maps (2.25 ppm/0.60 ppm, Figure 4.29) showed a stark increase of the STD intensities of the galactose protons (from H3 to H6) upon irradiation at 2.25 ppm, similar to what was observed for 3NPG. The raw DEEP-STD data are reported in the Section A.5 of the Appendix.

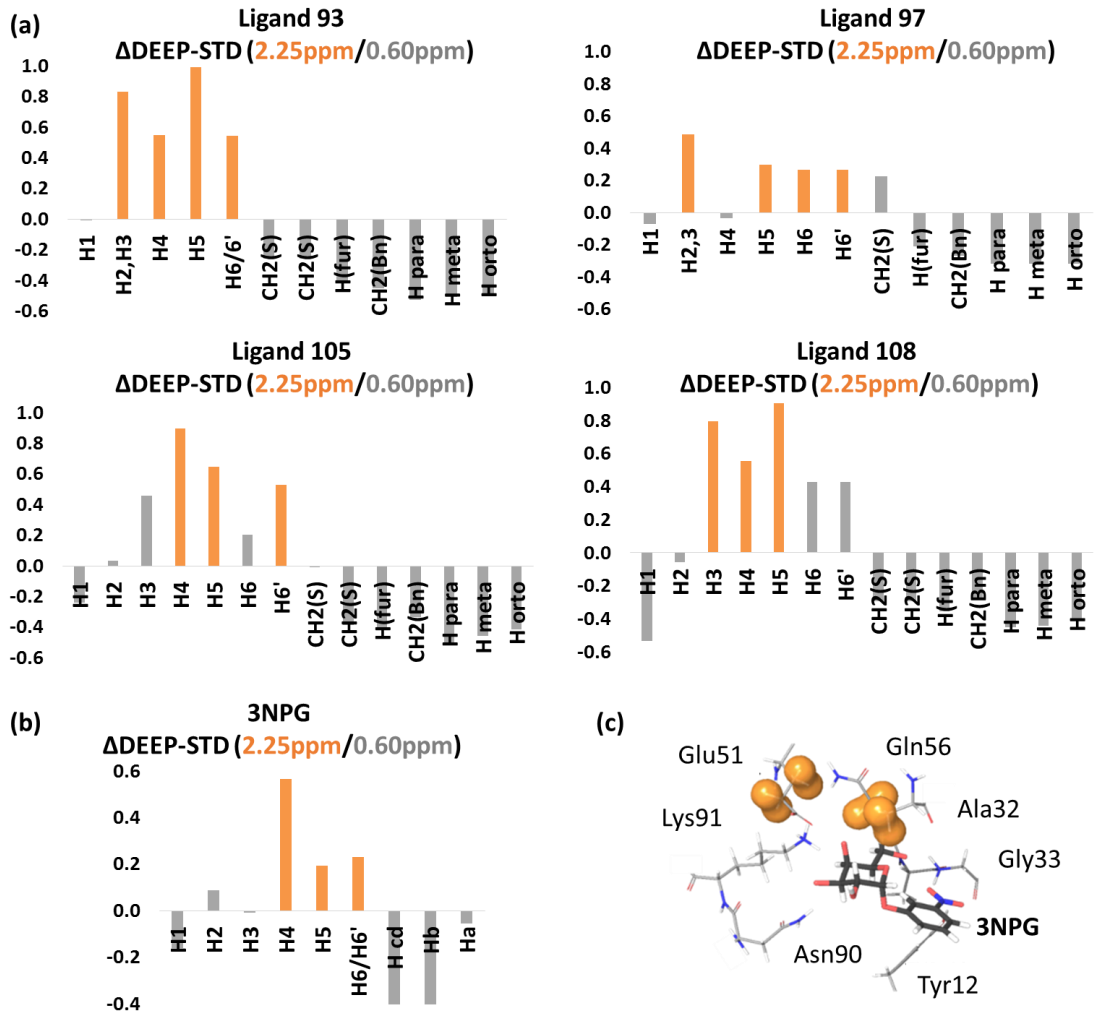

Figure 4.29. Differential Epitope Mapping (2.25 ppm/0.60 ppm) of **(a) Ligands 93**, **97**, **105**, **108** and **(b)** 3NPG in complex with CTB, obtained at 2 s saturation time. For each ligand, the three largest positive ΔDEEP-STD values are shown as orange bars. **(c)** Crystal structure of the complex 3NPG/CTB (PDB ID: 1EEI[88]). Protein protons directly irradiated at 2.25 ppm are enclosed in orange surface.

The increase in galactose STDs at 2.25 ppm is already evident by simple inspection of the resulting 1D STD NMR spectra comparing the two irradiation frequencies (Figure 4.30). As shown in Chapter 3, Section 3.3, 2.25 ppm is the predicted frequency around which the β and γ protons of Gln56 and Glu51 resonate. The proximity of Gln56 and Glu51 to H4, H5 and H6 of the galactose moiety is appreciable in the crystal structure of the 3NPG/CTB complex (PDB ID: 1EEI[88], Figure 4.29c). Overall, the positive ΔDEEP-STDs observed for the galactose protons of the four ligands confirm that the thio-galactose moiety occupies the galactose binding site.

It is interesting to think about the type of information obtained by DEEP-STD NMR in comparison with the information we would obtain by STD NMR competition. That is, the STD competition experiments of the four ligands with 3NPG could demonstrate that the small ligands compete with 3NPG. Strictly, however, this finding would purely inform us that the galactose subsite is involved in the binding of the small ligands, without clarifying whether it is the thio-galactose moiety of the small ligands which occupies the galactose subsite (although that would be expectable from structural reasons, no evidence would be available to rule out binding through the other end of the molecule). Thus, the nature of the information disclosed by the DEEP-STD NMR fingerprint approach is not accessible by either simple STD NMR competition experiments or by STD binding epitope mappings (as already shown).

It should be noted that DEEP-STD NMR experiments work on standard STD sample conditions, thus allowing to perform the experiments on the same sample used for STD binding epitope mapping (saving protein and synthetic ligands). In addition, DEEP-STD NMR can be carried out in a relatively short time (about 15 minutes for each of the ligands discussed above, using 2 s of saturation time).

It is also important to mention that irradiation at 2.25 ppm does hit the furoate methyl group directly (resonating between 2.30 and 2.40 ppm), but due to the spatial isolation of this group of protons and relative distance from the galactose ring, this does not affect the STD intensities observed for the thio-galactose protons (as shown in the correspondent "- protein" STD control experiments, run for each ligand in the same experimental conditions and in absence of the protein).

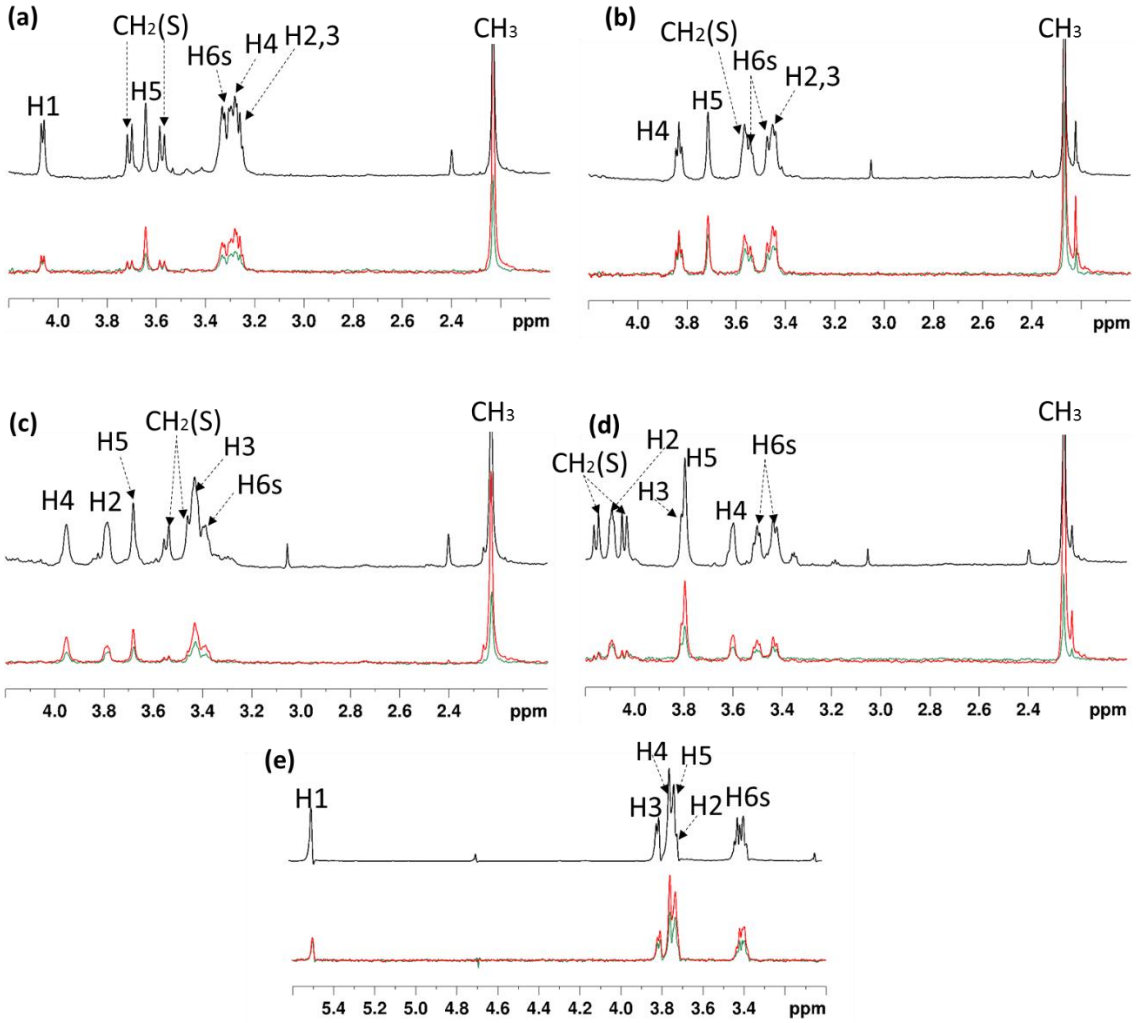

Figure 4.30. Differential Irradiation DEEP-STD NMR (2.25 ppm/0.60 ppm) spectra of **(a) Ligand 93**, **(b) Ligand 97**, **(c) Ligand 105**, **(d) Ligand 108** and **(e)** 3NPG in complex with CTB, at 2 s saturation time. In each panel, the reference spectrum (x1) is reported in black, and the difference spectra (x2) are reported in red (2.25 ppm) and green (0.60 ppm). Only the galactose spectral region is shown and the assignment of each peak is shown on top.

4.3.4 DEEP-STD "on-resonance scanning" approach: towards picking up single "amino acid-ligand" interactions

A way of getting more detailed information about the contacts and spatial proximity between ligands and protein protons is to acquire consecutive STD NMR experiments moving the irradiation frequency stepwise (increments of 0.1-0.2 ppm) around the irradiation frequency of interest (2.25 ppm in the case of 3NPG and the *SCJ set*). In Chapter 3, Section 3.2, we described this approach as "on-resonance scanning" as it consists of carrying out a set of STD NMR experiments at different *on-resonance* irradiation frequencies.

We "scanned" the irradiation frequencies from 1.45 ppm to 2.95 ppm (14 frequencies in total) and represented the STD intensities *vs.* irradiation frequency as a curve for each proton for **Ligand 93**, **97**, **105**, **108** and 3NPG (as a reference) in complex with CTB (Figure 4.31). The furoate and benzyl protons were excluded from this analysis, as the focus is here on the orientation of the thio-galactose moiety in the galactose binding subsite.

The results could be gathered within two groups showing different behaviours: curves that were rather flat (within the experimental error, *i.e.* including noise) and those which showed one or more maxima. Interestingly, all the non-flat curves showed maxima at around 1.70 ppm and/or 2.22 ppm. We interpret the frequencies at which the maxima are observed as *the central frequency of the protein protons closest to the ligands protons*. The expected average chemical shift values of Glu51 and Gln56, the two residues identified as directly irradiated at 2.25 ppm according to the predictive program Shiftx2.ca (Chapter 3, Table 3.3), are 2.17 ppm and 1.92 ppm, respectively. These values are reasonably close to our observed maxima in the on-resonance scanning curves (1.70 ppm and 2.22 ppm), supporting our interpretation.

Interestingly, in the on-resonance scanning of 3NPG, no maximum was observed at 1.70 ppm, but a maximum at 2.60 ppm was detected for protons H2, H4 and H5 (Figure 4.31e). Examining the crystal structure of the 3NPG/CTB complex (Figure 4.29c), it can be seen that Lys91 pointing towards the H2-H4 area of the sugar ring, which is a very good agreement with H2 having a maximum at 2.60 ppm but not at 2.22 ppm. According to the chemical shifts simulate with shiftx2, the Lys91 η protons resonate at around 2.95 ppm (which is not far from 2.60 ppm where the maximum is experimentally observed). The Lys91 protons were not accounted as direct irradiated protons in the DEEP-STD NMR study, as they are more than 0.40 ppm away from the irradiation frequency of 2.25 ppm. Still, the predicted chemical shifts should be regarded as an indication of the averaged chemical shift of each protein proton, while slight variations of temperature and pH can affect the validity of the prediction by some tenths of ppm. On the other hand, for protein signals with very broad linewidths (as it might be the case for a 65 kDa toroid-shaped protein, such as CTB), spurious saturation can be efficiently achieved, enlarging further the frequency range in which a given amino acid can be saturated.

On the other hand, comparing the structure of 3NPG/CTB with the on-resonance scanning in Figure 4.31e, it can be observed that the ligand protons laying outside the

galactose binding subsite showed a relatively flat on-resonance scanning curve (namely, H1, Ha, Hb and Hcd). This reinforces our interpretation of the on-resonance scanning profile representing direct contacts between ligands protons and directly irradiated protein protons. Reasonably, those ligand protons which are too far from the directly irradiated protein protons, as to efficiently receive intermolecular NOE, will not show any maxima (corresponding to those protons with ΔDEEP-STDs close to zero, in the DEEP-STD NMR approach).

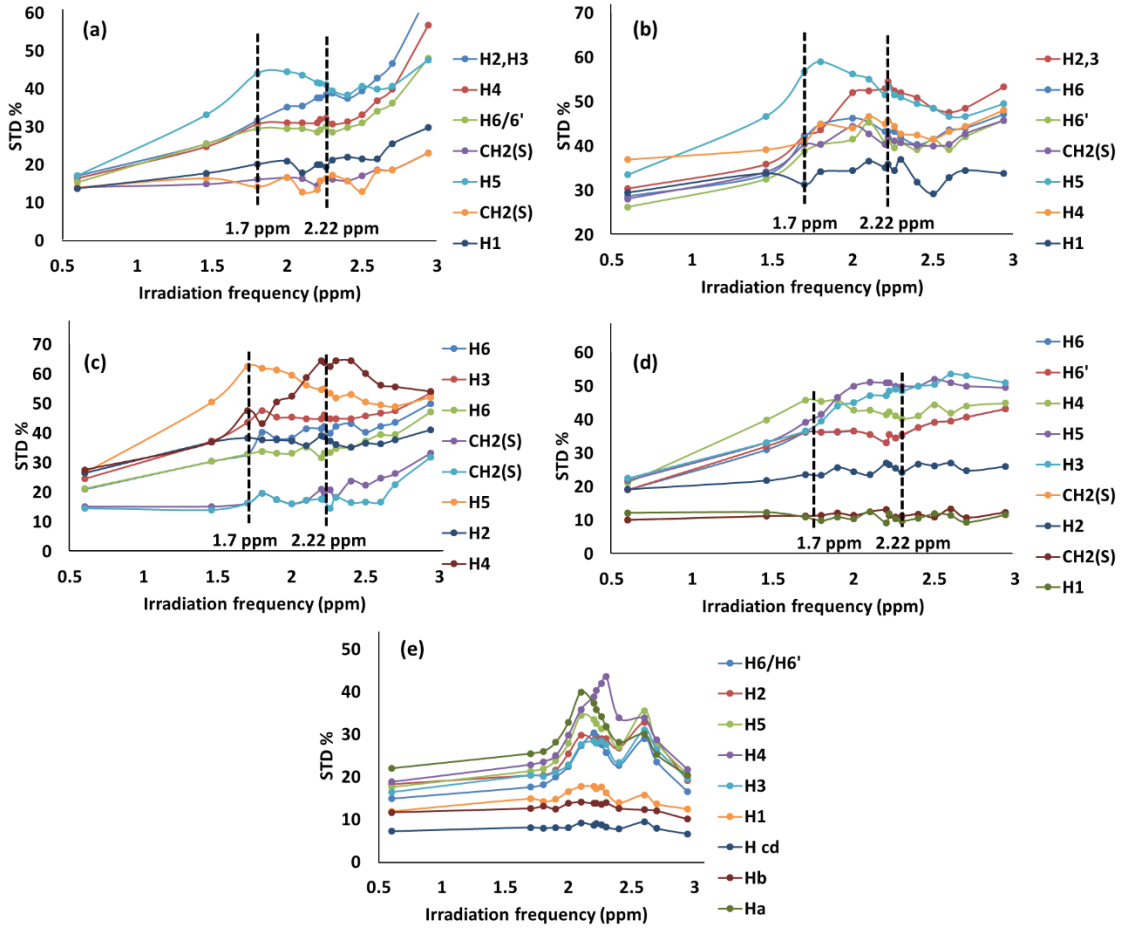

Figure 4.31. On-resonance scanning of **(a) Ligand 93**, **(b) Ligand 97**, **(c) Ligand 105**, **(d) Ligand 108** and **(e)** 3NPG in complex with CTB. The maxima in STD% represents the ligand protons closest to given residue protons.

Overall, comparing the on-resonance scanning profiles of the *SCJ set* and 3NPG indicates that the four small ligands fit, in slightly different orientations, in the galactose binding subsite, and none of them bind in the exact same orientation as 3NPG (Figure 4.31e). **Ligand 105** seems to have the most similar orientation to 3NPG (Figure 4.31c), as expected given their similar structure.

To dissipate any concerns on the possibility of having direct ligand irradiations, the same on-resonance scanning experiments were run in the absence of the CTB. For our five ligands, the STD NMR difference spectra at each of the on-resonance scanning frequency did not show any signal, except for the STD signal of the methyl group of the *SCJ* ligands, which resonates at around 2.20 ppm. The STD intensity of the methyl group in the blank experiments was 100% at 2.20 ppm, 2.22 ppm or 2.25 ppm, and it decrease down to 5% at 1.46 ppm and 2.94 ppm. What is more important for our approach, no STD signal was observed for any of the other ligand protons in the STD NMR experiments of the free ligands. Taking this into consideration, and the fact that the methyl group was excluded from the analysis, we can safely interpret the on-resonance scanning maxima as genuinely reporting on close interactions of single protein-ligand protons.

Aiming to obtain more detailed information on the specific interactions of Glc56 and Glu51 to the *SCJ set*, we run DEEP-STD NMR experiments and determined the ΔDEEP-STD values of the ligand protons differentially irradiating at the frequencies of the two STD maxima observed in Figure 4.31 (1.70 ppm and 2.22 ppm), *i.e.*, ΔDEEP-STD (1.70 ppm/0.60 ppm) and (2.22 ppm/0.60 ppm). For ease of analysis, we show their comparison on the same histogram (Figure 4.32). Positive ΔDEEP-STDs at both frequencies carry information on the relative distances between ligand protons and the single residues identified as resonating at each frequency (*i.e.*, Gln56 at 1.70 ppm and Glu51 at 2.22 ppm, in blue and green respectively in Figure 4.32).

Thus, with DEEP-STD NMR providing differential epitope mappings reported on the ligands structures second (Figure 4.32), we gain access to a second layer of structural information, beyond the classical STD NMR epitope mapping, showing those ligand protons receiving more saturation when irradiating at different frequencies. Figure 4.32 provides that "extra" layer of structural information reporting on ligand protons receiving more saturation when irradiating at 1.70 ppm (in green) or at 2.22 ppm (in blue). This allows us to identify which protons of the different ligands are pointing more towards Gln56 or Glu51, respectively, within the GM1 binding pocket of CTB.

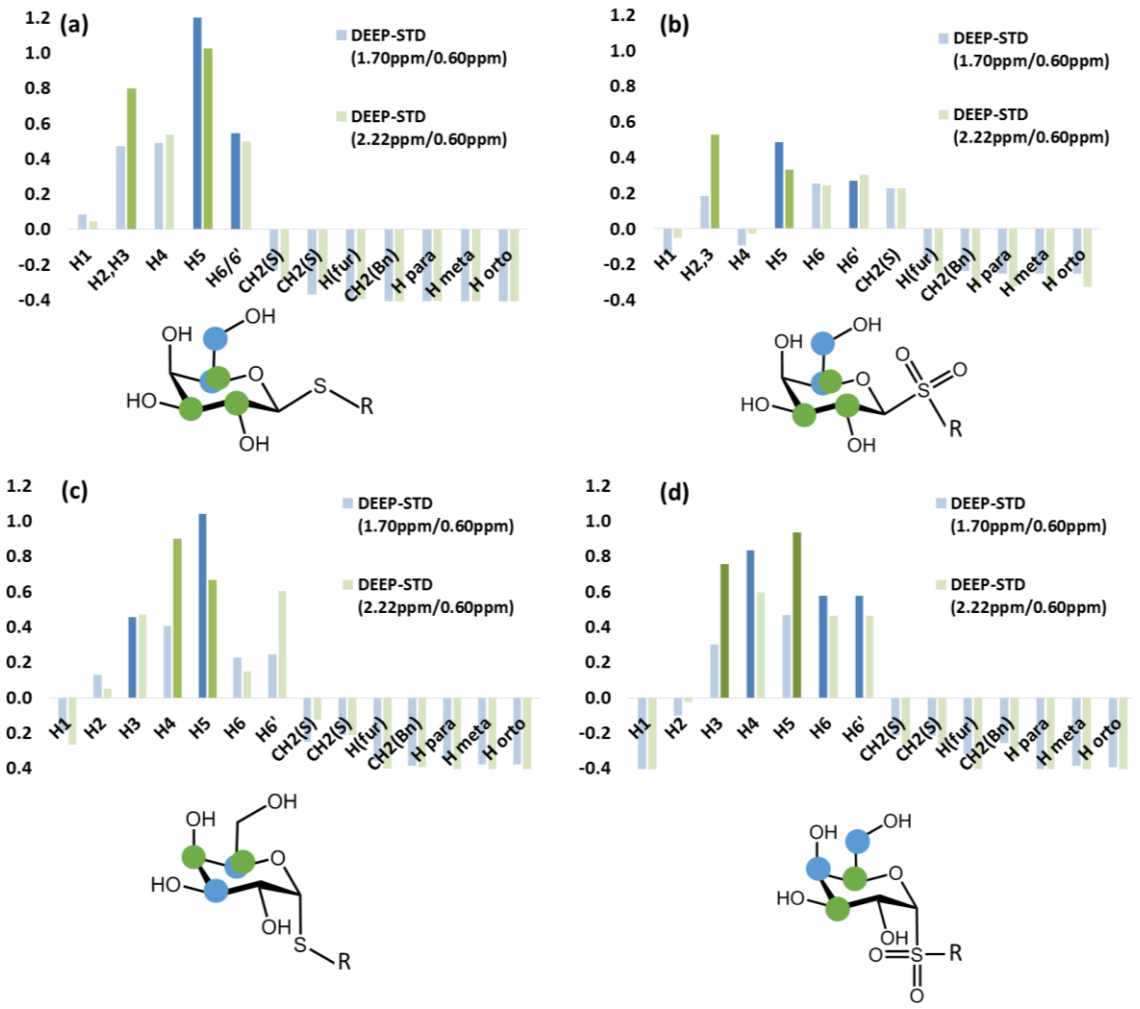

Figure 4.32. DEEP-STD histograms (2.22 ppm/0.6 ppm *vs.* 1.70ppm/0.60ppm) of **(a) Ligand 93**, **(b) Ligand 97**, **(c) Ligand 105** and **(d) Ligand 108** in complex with CTB. For each irradiation frequency (2.22 ppm/0.60 ppm in blue and 1.70 ppm/0.60 ppm in green), the two strongest ΔDEEP-STD are shown in darker tones. These are additionally plotted on the ligands structures, where blue dots show protons receiving more saturation at 1.70 ppm and therefore pointing towards Gln56, and green dots show protons receiving more saturation at 2.22 ppm and therefore pointing towards Glu51. R ≡ furaote-benzyl group.

In particular, **Ligand 93** and **97** orient their H2,3 and H5 protons closer to Glu51 (green), and H4 and both H6 protons towards Gln56 (blue), as shown in Figure 4.32a,b. On the contrary, **Ligand 105** orients H4 and H5 towards Glu51 (green) and H5 and H3 towards Gln56 (blue). **Ligand 108** points H4 and H6 towards Gln56 (blue) and H3 and H5 towards Glu51 (green).

It is important to highlight that the DEEP-STD NMR experiments at 2.25 ppm (fingerprint approach, Figure 4.29) already carried most of this information, *i.e.*, that protons H3 to H6s of the thio-galactose moieties were in close distance to protein residues resonating

at around 2.25 ppm (namely, Glu51, Gln56 and possibly Lys91). In fact, this is due to the broad CTB signals, so that irradiating at 2.25 ppm becomes close enough to excite them, thus giving rise to a very different STD factor compared to the irradiation at 0.60 ppm.

We consider that the on-resonance screening, albeit tedious, is a powerful expansion of the DEEP-STD NMR approach, which allows to identify the irradiation frequencies at which the STD values of each proton reach a maximum and giving finer details on the orientation of the ligands in the binding site.

4.3.5 Competition experiments: zooming in on the sialic acid/PHF subsites

The presence of a cluster of CTB residues resonating around 2.25 ppm in the galactose subsite, allowed us to apply DEEP-STD NMR to the thio-galactose moiety of the *SCJ* ligands. However, the same approach was not possible for the furoate-benzyl moiety, for which we could not identify frequencies where furoate and aromatic protons were responsive.

Nevertheless, as the *SCJ* ligands occupy the galactose subsite with the thio-galactose moiety, the furoate-benzyl moiety must either occupy the adjacent sialic acid subsite or point towards the novel subsite (given the "small" size of the aglycon moiety).

From our previous study (Section 4.2), we know that **Ligand 33** occupies the novel subsite (for which the 3D model of the complex is available, Figure 4.19b). Thus, we resorted to STD NMR competition experiments between the *SCJ* ligands and **Ligand 33**, to find out whether the novel binding subsite is involved in accommodating these small ligands. The competitions were performed by the addition of the *SCJ* ligands to four samples containing the **Ligand 33**/CTB complex. As **Ligand 33** has lower affinity than the small ligands (1.35 mM *vs.* 0.25-0.39 mM, respectively), the former would be displaced by any of the latter being able to occupy the novel binding subsite.

The results of the competitions experiments showed that all the reporter signals from **Ligand 33** (*i.e.* those that do not overlap with any signal coming from the *SCJ* ligands) retained their STD intensity upon addition of the *SCJ* ligands (Table 4.3). The lack of competition between them suggests that the *SCJ* ligands do not hinder the binding space of **Ligand 33** and they most likely are oriented towards the sialic acid subsite.

		Ligand 93		Ligand 97		Ligand 105		Ligand 108	
δ ^{1}H (ppm)	**Proton ID**	STD % before addition	STD % after addition	STD % before addition	STD % after addition	STD % before addition	STD % after addition	STD % before addition	STD % after addition
2.24	**CH_3**	9.16	-	10.29	12.15	9.34	-	8.72	9.8
3.48	**H2''**	3.01	3.84	2.92	-	2.5	-	2.78	-
3.79	**H3''**	3.01	4.36	4.28	5.2	4.49	-	3.35	-
4.23	**H4''b**	4.24	4.3	3.45	4.28	2.68	2.87	3.45	3.1
4.41	**H4''a**	2.92	-	-	3.88	2.1	2.38	2.12	3.62
5.03	**CH_2(Bn)**	3.96	-	4.54	-	3.77	6.2	3.66	-
6.36	**H3'''**	9.52	8.98	11.13	12.63	10.29	9.25	9.43	10.7
7.12	**Hpara**	9.16	-	10.21	-	9.24	-	9.01	-
7.16	**Hmeta**	9.8	-	10.98	-	9.81	-	9.23	-
7.21	**Horto**	9.43	-	11.31	-	9.56	-	9.38	-
7.68	**Htriazol**	4.81	5.57	6.7	8.98	4.9	6.32	4.54	7.46

Table 4.3. STD NMR competition experiments between **Ligand 33** and the *SCJ* ligands. The STD intensities (%) of the furoate-benzyl moiety of **Ligand 33** are given for each sample before and after the addition of the competitor (**Ligand 93** to **108**). No significant variation in the STD intensities of **Ligand 33** was observed when adding any of the *SCJ* ligands. This shows that there is no competition between them and that the binding of the *SCJ* ligands to CTB does not occur at the novel binding subsite.

4.3.6 Tr-NOESY/NOESY distance measurements: zooming in on the rotatable bonds

Determining intra-ligand distances in the bound state by tr-NOESY build-up curves using the isolated spin-pair approximation is a powerful way to obtain structural information (see Materials and Methods, Sub-section 4.5.2). Running similar NOESY experiments on the free ligands in the absence of the receptor allows to assess whether there are conformational rearrangements upon binding.

We focused on the protons at the glycosidic linkage, around the sulphur atom, determining for each ligand the CH_2(S)-H1 and CH_2(S)-Hfur'' distances (carbon nomenclature is given in Figure 4.33). The intra-sugar-ring H1-H5 or H1-H2 NOEs were chosen as reference of the β- and α-protons respectively, as they correspond to fixed inter-proton distances.

Remarkably, the C1-S, S-C(S) and C(S)-Hfur' bonds represent the centre of the molecules and the main source of flexibility as they constitute three rotatable bonds (red arrows in Figure 4.33). Thus, defining the distances between the protons surroundings these bonds is of specific interest to elucidate the bound conformation and to compare it with the free state. Excluding **Ligand 97**, in which the two CH(S) protons resonate at the same frequency, for the other three compounds it was possible to determine the distances of H1 and Hfur'' from both the CH_2(S) protons (Table 4.4).

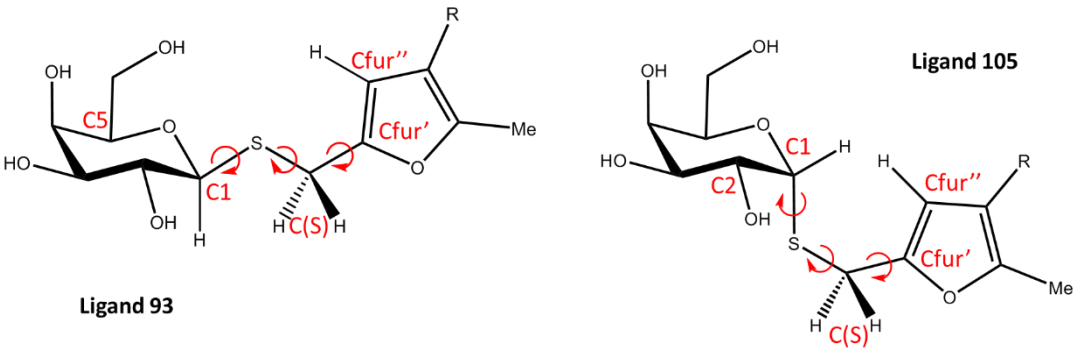

Figure 4.33. Carbon nomenclature for **Ligands 93** and **105**, which can be extended to **Ligand 97** and **105**. The central rotatable bonds are indicated by red curly arrows. R ≡ benzyl group.

The results shown on Table 4.4 can be summarised as follows:

1) both β-compounds (**Ligands 93** and **97**) have shorter H(S)-H1 distances than the α-compounds;
2) both thioethers (**Ligands 93** and **105**) have similar short H(S)-Hfur'' distances (and the same to both H(S)s);
3) **Ligand 93**'s H(S)-H1 distances are comparable to its H(S)-Hfur'' distances, while Ligand 105's H(S)-H1 distances are larger than its H(S)-Hfur'' distances;
4) for the sulphones (**Ligands 97** and **108**), the distance H(S)-Hfur'' is larger than for the thioethers (above 3.3 Å *vs.* 2.0-2.2 Å for the thioethers);
5) **Ligand 93**, **97** and **105** show some (small) degree of conformational rearrangement upon binding;
6) **Ligand 108** seems to undergo a significant conformational rearrangement, as its H(S)2 cross peak to H1 and Hfur'' disappear completely in the free state.

	Reference (Å)		Distances (Å)				Distances (Å)		
Ligand 93	**H1-H5**		**H1-H(S)1**		**H1-H(S)2**		**Hfur''-H(S)1**		**Hfur''-H(S)2**
Bound	2.4		2.3	=	2.3	=	2.2	=	2.2
Free	2.4		2.6	=	2.5	>	2.3	=	2.2
Ligand 97	**H1-H5**		**H1-H(S)s**				**Hfur''-H(S)s**		
Bound	2.4		1.9			<	3.6		
Free	2.4		1.9			<	3.1		
Ligand 105	**H1-H2**		**H1-H(S)1**		**H1-H(S)2**		**Hfur''-H(S)1**		**Hfur''-H(S)2**
Bound	2.5		2.8	=	2.7	>	2.1	=	2.0
Free	2.5		2.7	=	2.6	>	2.1	>	1.9
Ligand 108	**H1-H2**		**H1-H(S)1**		**H1-H(S)2**		**Hfur''-H(S)1**		**Hfur''-H(S)2**
Bound	2.5		2.9	=	3.0	<	3.3	=	3.2
Free	2.5		2.8	<	>4	<	3.7	<	>4

Table 4.4. Experimental conformation-defining distances of the *SCJ* ligands in the bound and free state as extracted from tr-NOESY and NOESY experiment, respectively. The reference distances used where measured on the galactose protons of 3NPG from the X-ray structure of the 3NPG/CTB complex (PDB ID: 1EEI[88]). Note that the proR and proS CH_2(S) protons are not assigned, therefore they are simply referred to as H(S)1 and H(S)2.

The conformational study by tr-NOESY and NOESY experiments allowed us to define distances around the sulphur, regarded as the central element of this class of ligands, and to investigate for conformational rearrangement upon binding. In conclusion, as already argued by STD epitope mapping and DEEP-STD data, the four ligands behave very differently, and it is hard to find a consistent pattern across the *SCJ set*, confirming that the anomeric configuration and the presence or absence of the oxygens on the sulphur have a strong influence, especially on the central (very flexible) part of the ligands.

4.3.7 Molecular docking: 3D molecular models for the complexes of CTB with SCJ ligands

Although we demonstrated by STD NMR that the novel binding subsite is not involved in the molecular recognition of the *SCJ* ligands, and that the thio-galactose moieties

present different orientations in the galactose binding subsite, a 3D molecular model of the binding of each ligand to CTB is lacking.

For the purposes of further three-finger ligands development, simply providing a qualitative model of the binding mode of the *SCJ* ligands validated by NMR experiments can be enough, as these ligands can be considered as simple intermediates. Therefore, we performed molecular docking of the four *SCJ* ligands on the "closed" CTB structure (the GM1/CTB XRD structure, PDB ID: 3CHB).

For each ligand, the first 5 to 10 lowest energy solutions converged to a binding mode in which the thio-galactose moiety sits in the galactose binding subsite and the furoate-benzyl group fit in the sialic acid binding subsite (Figure 4.34, docking tables are reported in Section A.6 of the Appendix). Remarkably, the galactose moiety of the four ligands did not converge to a single orientation, in good agreement with what we observed experimentally. On the contrary, the four benzyl groups precisely converged in the cleft accommodating the sialic acid of GM1 in the PDB ID: 3CHB (blue in Figure 4.34), where a π-stacking to Tyr12 can take place (Figure 4.36b). The "closed" conformation of the protein certainly introduces a bias, preventing the ligands to point towards the Ile58 and Lys34 residues, which are responsible of "opening" the novel subsite. However, as previously discussed, our experimental data showed that the novel subsite is not involved in the *SCJ* ligands binding mode, and the π-stacking to the Tyr12 appears to be a reasonable driving force to orient the benzyl group in the sialic acid subsite.

A closer look to the galactose subsite shows the different orientations of the thio-galactose moieties. The galactose moiety of lowest energy docking solutions of the α-ligands (**Ligands 105** and **108**) precisely overlap with the galactose moiety of GM1 and 3NPG, as observed in their XRD structure (Figure 4.35a). On the contrary, the galactose moiety of the lowest energy docking solutions of the β-ligands (**Ligands 93** and **97**) do not converge either between themselves, or with their α-analogues (Figure 4.35b).

Disappointingly, the orientation of the thio-galactose moieties obtained by rigid docking do not match with our experimental findings, *i.e.*, i) the CH_2(S)-H1 and CH_2(S)-$H_{fur''}$ distances found by tr-NOESY experiments do not match with those obtained from docking, and ii) the single residue-ligand contacts between H3-H6 and Glu51 and Gln56

(suggested by the DEEP-STD on-resonance scanning data) do not lie within the range of distances measured in the best docking poses.

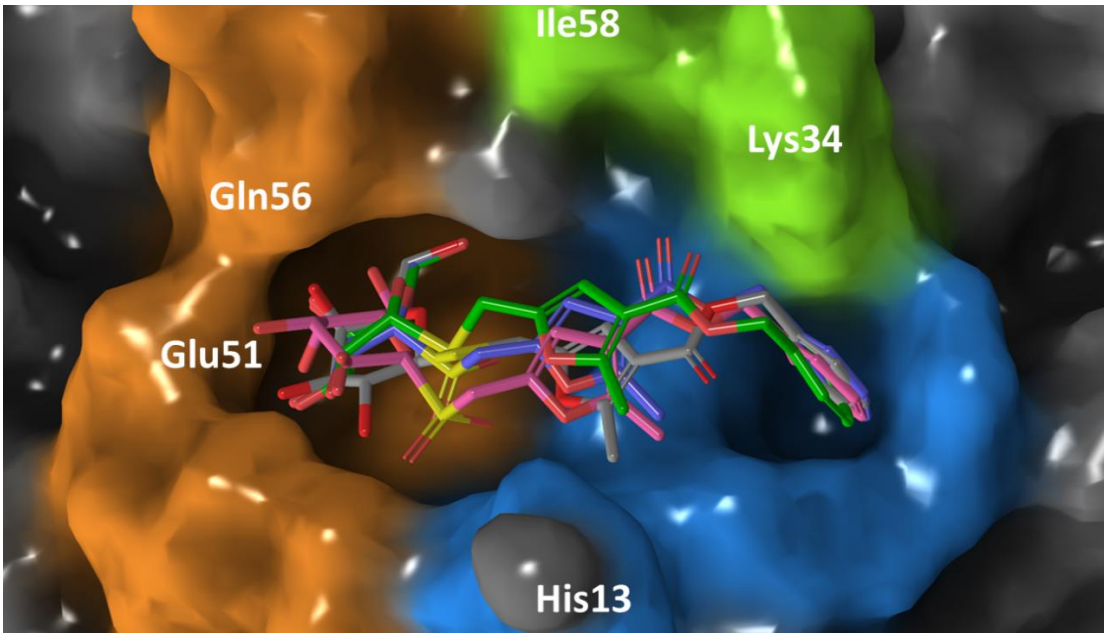

Figure 4.34. Lowest energy docking solutions for **Ligand 93** (grey), **Ligand 97** (pink), **Ligand 105** (green) and **Ligand 108** (violet) with CTB ("closed" conformation). The residues lining the galactose, sialic acid and novel binding subsites are enclosed in orange, blue and green surfaces, respectively. Relevant amino acids are labelled on top of the surfaces.

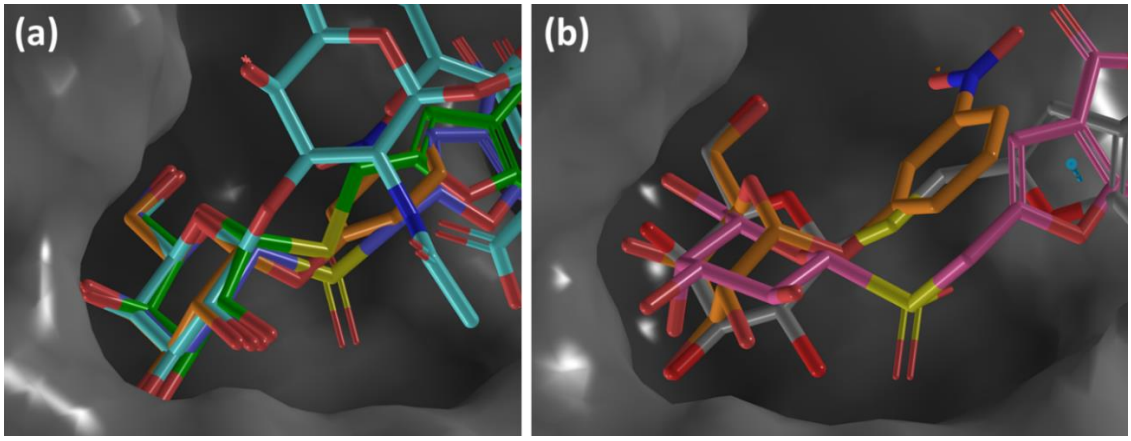

Figure 4.35. Zoom on the galactose binding subsite for the lowest energy docking solutions for **(a) Ligand 105** (green) and **Ligand 108** (violet) superimposed to GM1 (turquoise) and 3NPG (orange); **(b) Ligand 93** (grey), **Ligand 97** (pink) superimposed to 3NPG only (orange) with CTB ("closed" conformation).

On the other hand, it is reasonable to think that the orientation of the thio-galactose moiety might adjust by establishing favourable contacts to the amidic group of Asn14 and/or the NH group of Trp88 present at the protein surface (turquoise in Figure 4.36a).

The rotatable bonds S-C(S) and C(S)-Cfur' allow the methyl group on the furoate group to point towards the His13 (red sticks in Figure 4.36a). This can be justified by the presence of the Gln61 on the other side of the protein surface. The glutamine is rather hydrophilic and it hydrogen-bonds to a water molecule in the GM1 crystal structure (Figure 4.36a), therefore it could repel the methyl group, directing it towards the opposite histidine. Finally, the semi-rotatable bonds in between the furoate and the phenyl ring adjust to allow the best contact between the aromatic moiety and Tyr12 (Figure 4.36b).

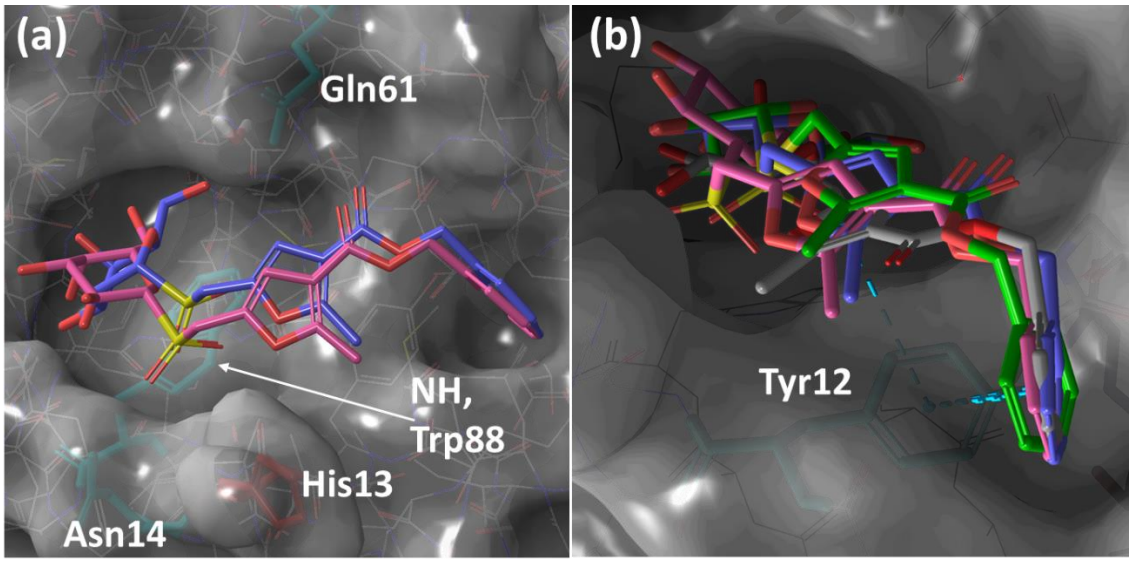

Figure 4.36. Lowest energy docking solutions for the *SCJ* ligands on a semi-transparent surface of CTB to show the interactions with key lipophilic/hydrophilic residues. **(a)** Only the sulphones (**Ligand 97**, pink and **Ligand 108**, violet) are included for simplicity, to show the sulphonic oxygens pointing towards Asn14 and Trp88 (in turquoise sticks), and the methyl on the furoate pointing parallel to His13 (red sticks), rather than towards Gln61 (turquoise sticks). **(b)** The four *SCJ* ligands are shown, zooming on the π-stacking interaction between the phenyl group of the ligands and Tyr12 (turquoise sticks).

4.3.8 Implications of the study and future directions

The 3D molecular models from docking solutions provided here for the binding of *SCJ* ligands to CTB should be taken as qualitative models of binding. Despite lacking quantitative agreement with the STD NMR and tr-NOESY data, they confirm the impact of the variations at the thio-glycosidic linkage (the α- or β- configuration and the oxidation state of the sulphur) on the orientation of the galactose in the galactose subsite. Interestingly, our models highlight the specificity of the phenyl group for the

sialic acid subsite. This structural element had been already reported by the research group of by Jimenez-Barbero for **Ligand 7** (Figure 4.3), in which the phenyl group fitted nicely in the sialic acid subsite, being held in place by a π-stacking interaction with the underlying Tyr12, as shown by molecular docking[128]. The same finding was reported by Fan and co-workers in studying the interaction of CTB with 3NPG and its aromatic derivatives. In particular, the crystal structure of BAPG (a 3NPG derivative, in which the an amidic group with a benzyl group in α takes the place of the nitro group) shows the same π-stacking between the terminal phenyl ring and Tyr12 (PDB ID: 1FD7)[88].

In fact, our findings are of paramount importance for the design of three-finger ligands.

The structures investigated exclude the possibility of decoration at the sulphoxide moiety, as the two oxygens are pointing towards the protein surface. However, our results clearly show that the furoate Cfur'' could be decorated with a PHF-like fragment ("**Ligand 33**-like") which would point towards the novel binding subsite. To this purpose, the α-ligands of the *SCJ set* could be the best choice, as they resemble 3NPG more closely, and most likely the α-thioether (**Ligand 105**) would do the job best, also considering that it has the lowest K_D. Even so, to explore the GM1 binding pocket even further it would be interesting to investigate three-finger analogues of the four *SCJ* ligands studied here.

Once and again, the combination of NMR and molecular modelling has proved to be an excellent tool to elucidate the binding mode of flexible ligands, magnifying binding differences of ligands with similar chemical structures and providing quick answer.

As a final remark, our CTB – *SCJ* ligands systems had a very good behaviour in terms of DEEP-STD NMR experiments, due to the presence of a cluster of residues in CTB resonating around 2.25 ppm (Glu51 and Gln56). The DEEP-STD NMR fingerprint approach was very useful to clarify the position and orientation of binding of the thio-galactose moieties, using 3NPG as a reference. Additionally, we explored the potentiality of DEEP-STD on-resonance scanning to detect single residue-ligand interactions by identifying the maxima in the STD % *vs.* irradiation frequency curves. It might be potentially useful to implement this latter approach to restrained docking calculations.

4.4 Conclusions

By a combination of STD NMR, a novel Inter-Ligand (IL) STD NMR approach, docking calculations and molecular dynamics simulations, we describe a novel binding sub-site on the GM1 binding pocket of the CTB receptor, adjacent to the galactose and the sialic acid binding sites, showing specificity of the non-carbohydrate moieties of our ligands and hitherto unknown. Our polyhydroxyalkylfuroate binders, **Ligands 30** and **Ligand 33**, do not bind in the proposed sialic acid subsite, as it was proposed in previous studies, but instead they show specificity for the novel groove through their non-carbohydrate polihydroxyfuroate moiety. **Ligand 33** fits deeper than Ligand 30 and its binding is compatible with the binding of 3NPG in the adjacent galactose subsite (as observed by ILOE and IL-STD NMR).

Despite the failed attempts of crystallization, the combination of STD NMR and tr-NOESY experiments with classic MD, HREMD and molecular docking resulted in the validation of a dynamic model based on 100 MD frames for each ligand, as confirmed by CORCEMA-ST.

The presence of the new non-carbohydrate specific binding groove, accessible also when the adjacent subsites are occupied, is of paramount importance to pave the way to design future classes of CTB inhibitors based on the concept of three-finger compounds, with the prospective of increased affinity relative to the ligands studied here. The ternary complex of **Ligand33**/CTB/3NPG was a good system for the application and validation of the novel IL-STD methodology, introduced in Chapter 3, Section 3.4. This novel STD NMR approach allowed us to reproduce the information obtained by ILOE, and together with it could be a promising and time-efficient technique for knowledge-based drug design and to explore the orientation of binding sub-sites and bound fragments in known and unknown binding pockets.

A thorough STD NMR analysis of "small **Ligand 30** analogues" lacking the triazole-polyhydroxy chain, the *SCJ set*, allowed to investigate the specificity of the thio-galactose and furoate-benzoyl fragments, leading towards the design of three-finger ligands. We also provided 3D models of the binding obtained by molecular docking, showing qualitative agreement with STD NMR binding epitope mapping.

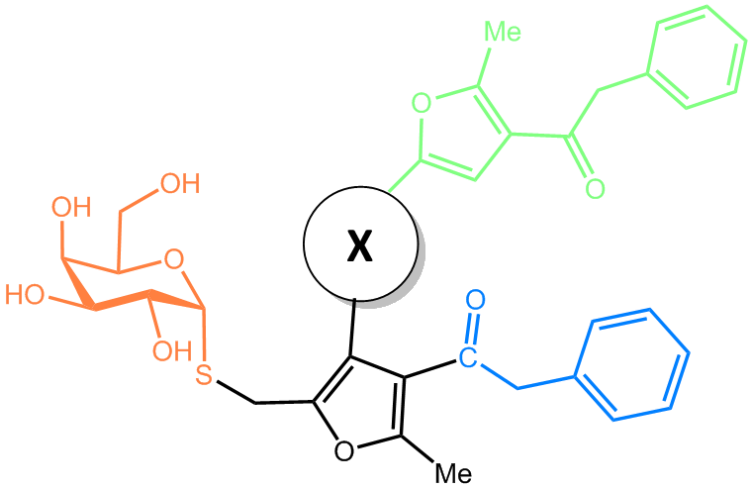

Figure 4.37. Schematic representation of a novel class of three-finger non-carbohydrate CTB inhibitors proposed by us, potentially able to occupy simultaneously the galactose, sialic acid and novel binding subsites (orange, blue and green fragments, respectively). The length and the flexibility of the X linker, should be investigated.

Based on our 3D molecular models (Figures 4.19 and 4.34), we propose that the decoration of the *SCJ* ligands with a "**Ligand 33**-like" fragment at position Cfur'' (the only CH of the furoate ring) could yield powerful three-finger ligands, with optimal affinity to the three binding subsites (Figure 4.37). The length and flexibility of the linker should be tuned to occupy the three subsites simultaneously, specifically and with the minimal loss of enthalpy. This could result into a novel approach to CTB inhibition, building on a promising class of non-carbohydrate, non-hydrolysable, innovative compounds.

4.5 Materials and methods

4.5.1 Chemicals, ligands and proteins

Ligand 3-*nitro*-phenyl α-D-galactopyranoside, Cholera Toxin Subunit B, deuterium oxide (99.9% ^{2}H), dimethyl sulphoxide-d6 (DMSO-d6), disodium hydrogen phosphate (Na_2HPO_4), potassium dihydrogen phosphate (KH_2PO_4), sodium chloride (NaCl) and potassium chloride (KCl) were purchased from Sigma Aldrich. The PHF ligands (**Ligands 20**, **30** and **33**) and the *SCJ* ligands (**Ligand 93**, **97**, **98**, **105**, **108** and **109**) were provided by our collaborators at University of Seville (research group of Prof. Inmaculada Robina). The synthesis of the PHF ligands has been described previously[92], while the synthesis of the *SCJ* ligands is still unpublished. All the synthetic ligands were prepared in stock solutions 3% DMSO-d6 in D_2O.

4.5.2 NMR measurements and processing

All the STD NMR experiments were performed on samples of 1 mM ligand and 5 µM CTB in 10 mM PBS and 137 mM NaCl buffer pH 7.4 in D_2O. As each pentamer contains 5 equivalent binding sites, the protein:ligand ([P]:[L]) ratio obtained was 1:40. For the STD NMR competitions of the PHF ligands (Sub-section 4.2.2), the STD NMR experiment of the ligand under investigation (**Ligand 30** and **Ligand 33**) in the presence of CTB was first performed (irradiation frequency 0.0 ppm, saturation time 2 s, delay between experiments of 5 s and 512 scans). Then the competitor ligand in equimolar concentration was freeze-dried and dissolved in the initial sample, and the same STD NMR experiment was subsequently carried out. For the IL-STD NMR experiments on the ternary complex 3NPG/CTB/**Ligand 33** (Sub-section 4.2.6), STD NMR experiments at 2 s saturation time with a [P]:[L] of 1:40 (1 mM of each ligand and 5 µM CTB) were performed by irradiating at 0.60 ppm (*off-ligand* experiment) and 7.27 ppm (*on-ligand* experiment). As a control, the same couple of experiments in the same conditions were acquired on the binary complexes, 3NPG/CTB and **Ligand 33**/CTB. For the ILOE on the ternary complex 3NPG/CTB/**Ligand 33** (Sub-section 4.2.3), transferred-NOESY (tr-NOESY) experiments were performed with a [P]:[L] of 1:10 (1 mM of each ligand and 20 µM CTB), with a mixing time of 1.2 s, a relaxation delay of 2 s and 160 scans. As a control, the same experiment was performed on 3NPG/CTB, **Ligand 33**/CTB and CTB alone under the same conditions. In addition, IL-STD NMR experiments were performed on the ternary complex with a [P] : [L] ratio of 1:40 under the same conditions as the STD NMR competition experiments.

For the *SCJ* ligands (**Ligands 93**, **97**, **98**, **105**, **108** and **109**), STD NMR build-up curves were acquired with irradiation frequency 0.60 ppm, at 0.5 s, 1 s, 2 s, 3 s, 4 s and 5s, with a delay between experiment of 5 s and 256 scans (Sub-section 4.3.2). For the DEEP-STD NMR fingerprint approach and on-resonance scanning of **Ligands 93**, **97**, **105** and **108** (Sub-section 4.3.3-.4), a series of STD NMR experiments at 2 s saturation time and 32 scans were performed at the following irradiation frequencies: 0.60 ppm, 1.464 ppm, 1.70 ppm, 1.80 ppm, 1.90 ppm, 2.0 ppm, 2.1 ppm, 2.20 ppm, 2.222 ppm, 2.25 ppm, 2.30 ppm, 2.40 ppm, 2.50 ppm, 2.60 ppm, 2.70 ppm and 2.94 ppm. For the STD NMR competitions of the *SCJ* ligands (Sub-section 4.3.5), STD NMR experiments of four sample containing 1 mM **Ligand 33** and 5 µM CTB were firstly performed (irradiation

frequency 0.60 ppm, saturation time 2 s, relaxation delay between experiments of 5 s and 32 scans), then the competitor ligand in equimolar concentration was added and the same STD NMR experiment was repeated. For the distance determination of **Ligands 93**, **97**, **105** and **108** in the bound state (Sub-section 4.3.6), NOE build-up curves were acquired on the STD NMR samples (1 mM ligands and 5 µM CTB) running tr-NOESY experiments at 0.3 s, 0.4 s, 0.6 s and 0.8 s mixing time, delay between experiments 1.5 s delay between experiments and 16 scans. To determine distances in the free state, the same NOESY experiments were run on samples containing 1 mM of each ligands, in the absence of the protein. All the NMR experiments were performed at 278 K.

All the STD NMR experiments were acquired using an STD NMR pulse sequence that included 2.5 ms and 5 ms trim pulses, a spoil gradient (3 ms), and water suppression by excitation sculpting with gradients (*stddiffesgp.3*). Saturation was achieved applying a train of 50 ms Gaussian pulses (0.40 mW) on the f2 channel, at the required on-resonance irradiation frequency (depending on the experiment) and 40 ppm (off-resonance experiments). The broad protein signals were removed using a 40 ms spinlock (T1ρ) filter (*stddiff.3*). Tr-NOESY and NOESY experiments were performed using a phase-sensitive pulse program with gradient pulses in the mixing time and a relaxation delay of 1.5 s or 2 s, with water suppression using the 3-9-19 scheme with gradients (noesyfpgpph19) when necessary. All the experiments were recorded at ^{1}H frequency of 800.23 MHz on a Bruker Avance III spectrometer equipped with a 5-mmD probe TXI 800 MHz H-C/N-D-05 Z BTO.

To obtain the binding epitope mapping of the ligands, we used Origin 10.4.12 to mathematically fit the STD NMR build-up curves to a mono-exponential equation, from which the initial slopes were obtained. The binding epitope was obtained by dividing all the initial slopes by the one of a convenient proton of the sialic acid residue (depending on the ligand), to which an arbitrary value of 100% was assigned. DEEP-STD NMR maps were calculated using the DEEP-STD NMR equation[87] (Equation 3.5):

$$\Delta\mathrm{DEEP}-\mathrm{STD}_i = \frac{STD_{\ exp1,i}}{STD_{\ exp2,i}} - \frac{1}{n}\sum_i^n \left(\frac{STD_{\ exp1,i}}{STD_{\ exp2,i}}\right).$$

Intermolecular distances were calculated according to the initial rate approximation, which establishes that the evolution of the initial build-up of NOE enhancements with

mixing time is approximately linear[56]. The longitudinal cross-relaxation rate can therefore be determined from the initial slope of the build-up curve (I_{IS} as function of τ_m), where the NOE enhancement (I_{IS}) is defined as the ratio between the intensity of the cross-peak at a certain mixing time and the intensity of the diagonal peak at zero mixing time [139]. In turn, the cross-relaxation rate is proportional to the inverse sixth power of the inter-nuclear distance (σ_{IS} is proportional to r_{IS}^{-6}). The observed NOE intensities were normalised relatively to a known spin pair distance as indicated in the text. For distances in the bound state, the slopes were calculated on the 0.3 s, 0.4 s and 0.6 s mixing time points (as at 0.8 s the line was not linear anymore); for the distances in the free state, the slopes were calculated on the 0.4 s, 0.6 s and 0.8 s mixing time points, due to the different correlation time of the ligands in the bound and free state.

4.5.3 Molecular Docking and CORCEMA-ST validation

All molecular modelling was performed with the module Glide within Schrödinger's Maestro software suite[140,76,141]. For the docking of the PHF ligands, coordinates for the receptor were obtained from the Protein Data Bank (PDB ID: 3CHB) for the closed conformation and from a representative frame of our HREMD trajectory of the GM1/CTB complex for the open conformation. Where necessary, coordinates for missing atoms were added according to known protein chemistry and sidechain protonation was optimised for neutral pH. A short minimisation was run using the OPLS3 force field, converging heavy atoms to a RMSD of 0.3 Å. The receptor grid was then calculated, using the centroid of GM1 as geometric centre of the rigid box and a length of 30 Å in order to encompass all subsites. Three dimensional structures of all ligands were generated using a conformational search, implementing Monte-Carlo torsional sampling, keeping only unique structures (RMSD > 0.5 Å) and eliminating all structure with an energy 21 kJ mol^{-1} greater than the lowest energy structure. All resulting structures were then minimised using conjugate gradient minimisation, converging on a threshold of 0.05 kJ mol^{-1} $Å^{-1}$. For each ligand, the 10 lowest energy conformations were used to initiate docking. The docking consisted of further conformer generation, docking and then minimisation. Conformers were generated with 4x enhanced sampling and during docking, the non-bonded term of the potential energy function was softened for nonpolar ligand atoms (charge < |0.15|) by applying a scaling factor of 0.8. Finally, minimisation was performed using implicit solvent with a distance dependant dielectric

constant of 4. For the molecular docking of the *SCJ set*, a 20 Å was built using the coordinates of the receptor in the closed conformation from the Protein Data Bank (PDB ID: 3CHB), otherwise the same settings as for the PHF ligands were used.

For CORCEMA-ST[66], based on the experimental conditions, the concentration of ligand was 1 mM and the [P] : [L] ratio was kept fixed at 25:1. The cut-off distance around the binding pocket was 13 Å. k_{on} was set to 1 × 10-8 M^{-1} and the irradiation frequency to the range -0.8 ppm to 0.8 ppm. The bound ligand correlation time (τ) was 65 ns for both ligands, whereas the free correlation time was 1.2 ns, 1 ns and 1 ns respectively for **Ligand 30**, **Ligand 20** and **Ligand 33** (these correlation time values are coming from a rough approximation based on the molecular weight of protein and ligands). The equilibrium constant used for Ligand 30 was 2000 M^{-1}, 1000 M^{-1} for Ligand 20 and 1500 M^{-1} for Ligand 33, with ρ leak of 0.25 s^{-1}. For **Ligand 30** and **33**, CORCEMA-ST calculations were run in parallel on the High Performance Computing Unit, and then averaged to provide a representative build up curve for each ligand. The NOE R-factor was calculated on the averaged simulated data obtained, with the following equation:

$$NOE\ R-Factor = \sqrt{\frac{\sum_k \left(STD_k^{exp} - STD_k^{cal}\right)^2}{\sum_k \left(STD_k^{exp}\right)^2}}.$$

Chapter 5

Investigation of the structural elements of sialic acid recognition by the intramolecular *trans*-sialidase *Rg*NanH from gut microbiota

Parts of the findings from this Chapter are published as:

Owen, D. C.*; Tailford, L. E.*; Monaco, S.; Šuligoj T.; Vaux, L.; Lallement, R.; Khedri, Z.; Yu, H.; Lecointe, K.; Walshaw, J.; Tribolo, S.; Horrex, M.; Bell, A.; Chen, X.; Taylor, G. L.; Varki, A.; Angulo, J.; Juge, N., Unravelling the specificity and mechanism of sialic acid recognition by the gut symbiont *Ruminococcus gnavus*. *Nature Communications* **2017**, 8, 2196.

5.1 Introduction

5.1.1 Gut microbiota: general features and human health

The human body provides a suitable environment for microorganisms to colonise in numbers comparable to the number of the human cells in the body[142]. Most of these microorganisms resides in the gastrointestinal tract (GIT) where they are referred to as the gut microbiota[143].

These resident species comprise non-pathogenic bacteria, eukaryotic microorganisms, *archea*, viruses and bacteriophages; and the number of overall species is estimated around several hundred[144]. Among the bacteria, the dominating *phyla* are: Bacteroidetes, Firmicutes and Actinobacteria[145]. Many are commensals, *i.e.* they benefit from the nutrient-rich environment of the intestine while at the same time, improving the digestion by the host by fermenting otherwise non-digestible nutrients (humans produce about 17 classes of host enzymes in the GIT mostly targeting starch degradation, while the microbiota produces thousands of complementary hydrolytic enzymes with diverse specificities[146]). Other resident bacteria include opportunistic pathogens, which can cause diseases when the host resistance is lowered. Overall, the

presence of the microorganisms keeps the system in a state of "physiological inflammation": this provides immune-defence and ensures a quick and adapted response to pathogens[147].

The microbiota ecosystem in the GIT evolves throughout the human life (Figure 5.1). The unborn child is sterile, or almost so, and receives the first microbiota at birth, with the delivery mode (caesarean section or vaginal delivery) being the most important factor affecting the initial microbiota composition[144]. The composition evolves along the first years of life both in diversity and richness: the feeding mode (breast- or formula-feeding) is a crucial factor, and the introduction of solid food has a very strong effect on the gut microbiota evolution in the baby. Once the balance is reached, past the toddler age, there are many factors which can perturb it: antibiotic treatments, malnutrition, obesity and general health condition[148].

The role of the human gut microbiota is critical for the well-being of the host, through its contribution to polysaccharide digestion, immune system development, vitamin production and pathogen defence[149]. A dysbiosis of the gut microbiota is associated with numerous intestinal diseases including cancers, infections, and inflammatory bowel diseases [150], stressing the importance of understanding the mechanisms of host-microbe interactions, so to devise microbiota-targeted strategies to restore health.

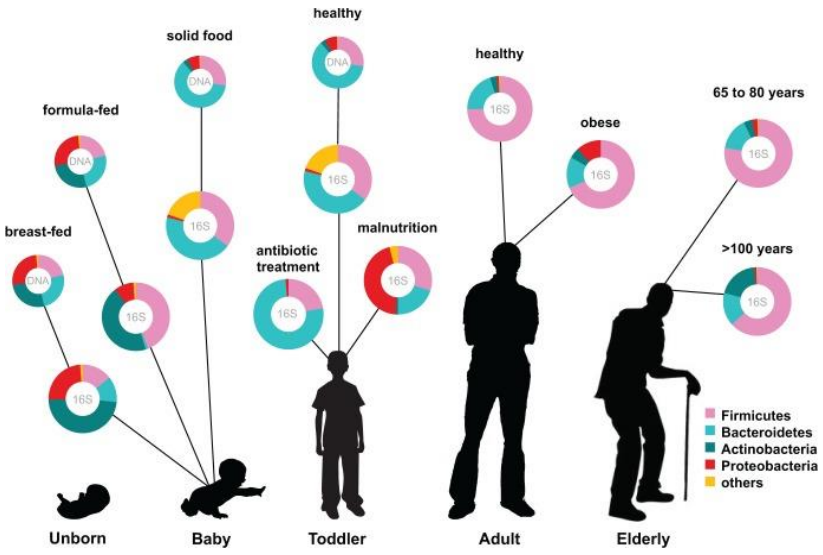

Figure 5.1. Dynamic composition of the gut microbiota and factors affecting it. The "cake representation"' provides a global overview of the relative abundance of key phyla of the human microbiota composition in various stages of life, measured by either 16S RNA or metagenomic approaches. Figure from 151.

5.1.2 Mucin glycans and gut microbiota: role of the sialic acid and sialidases

The microbiota composition varies longitudinally along the GIT but also transversally from the lumen to the mucosa[152], with most of the gut bacteria residing in the colon, where they compete for dietary and host glycans[153].

The gradient and availability of glycans within discrete physical niches is among the several factors influencing the biogeography of symbiotic bacteria within the gut[152]. The mucus layer covering the GIT is at the interface between the gut microbiota and the host[150]. In the colon, the mucus layer is divided into an outer layer providing a habitat to commensal bacteria and an inner layer firmly attached to the epithelium and providing a protection from bacterial invasion[150]. Large highly glycosylated gel-forming mucin secreted by goblet cells are the main components of the intestinal mucus layers[154]. Importantly, the mucin oligosaccharides provide binding sites and nutrients to the bacteria which have adapted to the mucosal environment[155]. Commensal and pathogenic microbes have evolved a range of adhesins allowing their interaction with the mucus in the gut, reflecting the structural diversity of mucin glycans and their location[156]. Variations in mucosal carbohydrate availability lead to variations in the composition of the resident microbiota[157].

The glycan structures present in mucins are complex, consisting of core *O*-glycans elongated by diverse carbohydrate sequences. These are terminated by fucose and sialic acid such as *N*-acetylneuraminic acid (Neu5Ac) sugar residues, *via* α1–2/3/4 and α2–3/6 linkages, respectively. Hence, these sugar residues are prominent targets for commensal and pathogenic bacteria[158]. The release of sialic acid by microbial sialidases allows the bacteria present in the mucosal environment to access free sialic acid for catabolism, decrypt host ligands for adherence, participate in biofilm formation, modulate immune function by metabolic incorporation, and expose the underlying glycans for further degradation[154,159,18].

Based on their substrate specificity and mechanism, bacterial exo-sialidases can be divided into three classes: hydrolytic, *trans*-sialidases and intramolecular *trans*-sialidases (IT-sialidases). Hydrolytic sialidases release free sialic acid (Figure 5.2a), *trans*-sialidases transfer the cleaved sialic acid to other glycoconjugates, while IT-sialidases release 2,7-anhydro-sialic acid (Figure 5.2b)[18]. The capacity to carry out one or more of

these enzymatic reactions, together with the ability or not of catabolising sialic acid and its derivatives, has a large impact on the adaptation of the microorganisms to the mucosal environment[160].

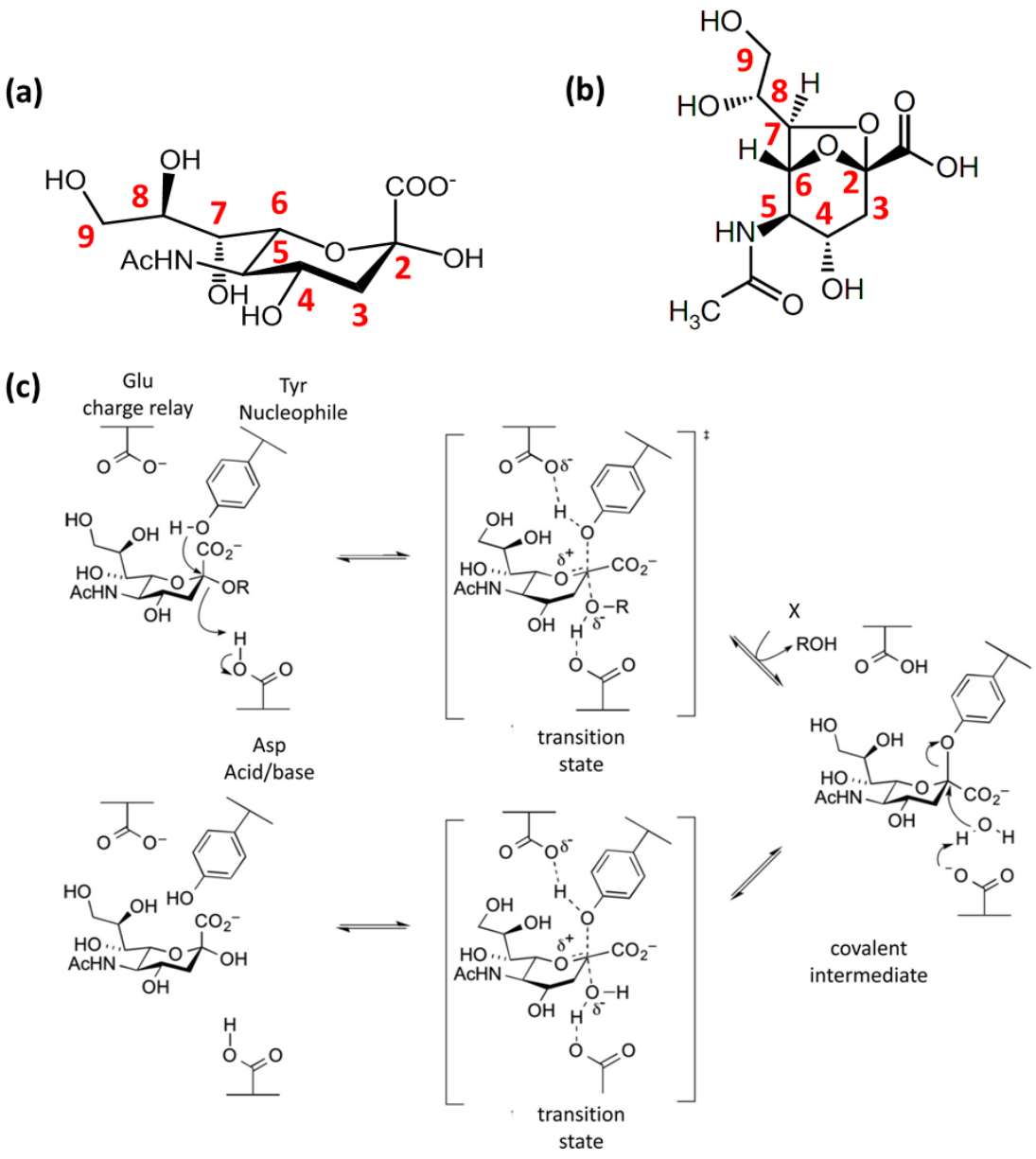

Figure 5.2. Structures of **(a)** sialic acid (NEu5Ac) and **(b)** 2,7-anhydro-sialic acid (2,7-anhydro-Neu5Ac), with their atom nomenclature. The nomenclature is given for the carbons of Neu5Ac and it is applied to all the ligands. **(c)** Mechanism of action hydrolytic/*trans*-/IT-sialidases via a two-step double-displacement. The glycosylation step is the same for all three classes of sialidase, but for the deglycosylation step the incoming molecule can be water, another sugar or the internal oxygen atom, as indicated. Adapted from http://www.cazy.org[27,18].

All sialidases follow a catalytic mechanism of double inversion (so that the α-configuration of Neu5Ac is retained) via the formation of a covalent intermediate, which

sees the Neu5Ac bound to the protein. A nucleophile tyrosine, a charge relay glutamate and an acid/base aspartate are the essential catalytic residues involved in the enzymatic reaction[170], as schematised in Figure 5.2c.

As often with carbohydrate-active enzymes, sialidases are modular proteins consisting of:

- A catalytic domain, classified as a glycoside hydrolase family 33 (GH33).
- Carbohydrate binding modules (CBMs) such as sialic acid specific CBM40[159,161] and broadly specific CBM32[162].

As defined by the Carbohydrate Active Enzymes Database, CAZy, http://www.cazy.org[27]. The CAZy classification is based on protein sequences and structure similarities, so to help reflecting enzymatic structural features (better than if it was based only on substrate specificity), providing a framework to understand mechanistic properties[163].

5.1.3 The intramolecular *trans*-sialidase from *Ruminococcus gnavus* (*Rg*NanH)

The biological system of interest in this section is *Rg*NanH, an IT-sialidase from *Ruminococcus gnavus*. *R. gnavus* is a prominent member of the gut microbiota of the healthy human gut[164] and one of the few commensal species of bacteria which have been studied for their ability to utilise intestinal mucin glycans[165,17]. Besides, *R. gnavus* has been found to be highly over-represented in individuals affected by Inflammatory Bowel Disease[68]. As an IT-sialidase, *Rg*NanH has high specificity for cleaving α2/3 sialic acid capping the mucin oligosaccharides[18]. The hydrolysis of the terminal sialic acid is followed by intramolecular transglycosylation at the C2 position of Neu5Ac, attacked by its O7-hydroxy group at the glycerol moiety. The product of the reaction is 2,7-anhydro-Neu5Ac[166].

*Rg*NanH is the third IT-sialidase to be characterised, after NanL from *Macrobella decora*[167] and NanB from *Streptococcus pneumoniae*[168], and the first from a gut commensal species [4]. Despite their low sequence similarity, all IT-sialidases share the canonical six-bladed β-propeller typical of sialidases (although the sequence homology across sialidases is also low)[168]. As an example, Figure 5.3 shows the structure of the IT-sialidase from *Streptococcus pneumoniae*.

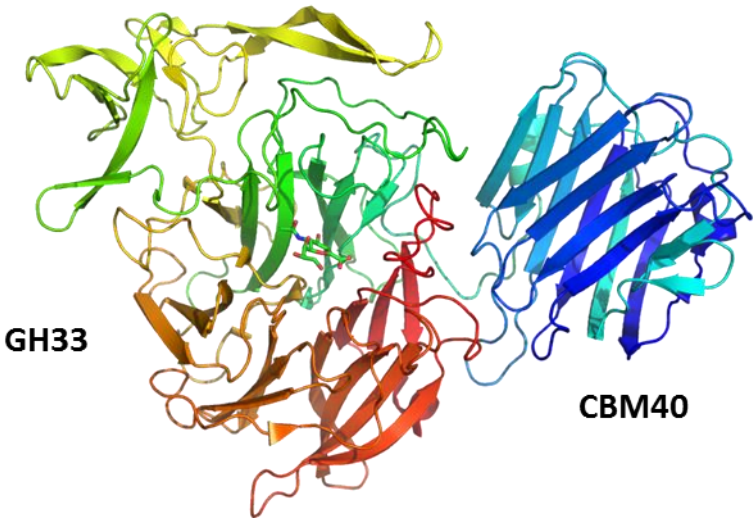

Figure 5.3. General structure of an IT-sialidase (here from *Streptococcus pneumoniae* PDB ID: 2VW1[169]). CBM40 in blue and GH33 in a six-bladed β-propeller shape in green to orange (complexing Neu5Ac, in sticks).

Interestingly, *R. gnavus* can grow on 2,7-anhydro-Neu5Ac as a sole carbon source[111]. This finding strengthens the proposed adaptation strategy of *R. gnavus* to the mucosal environment by releasing a molecule, 2,7-anhydro-Neu5Ac, which it can preferentially catabolise, while uncapping the mucin glycan chains. This allows *R. gnavus* to access this rich source of nutrient in the colonic mucus for its own benefit, providing a nutritional advantage over the competitors, which are usually Neu5Ac catabolisers[18]. At the same time, it opens the way for further degradation of the mucin glycan chain for the rest of the mucosal bacteria community.

In this study, led by the research group of Nathalie Juge at the Quadram Institute Bioscience (QIB), we analysed both domains of *Rg*NanH: the catalytic domain GH33, and the carbohydrate binding module CBM40. We used STD NMR spectroscopy to contribute to the characterization of the molecular recognition preferences of the two domains in solution against oligosaccharides terminating with sialic acid (or its derivatives). The purified recombinant domains, *Rg*GH33 and *Rg*CBM40 (wild-type and mutants produced in *E. coli*) were provided by Louise Tailford at QIB (Juge Group), who also provided the sialylated ligands including the enzymatically-synthesised 2,7-anhydro-Neu5Ac.

In the structural investigation of GH33, we used both traditional STD NMR (binding epitope mappings determination, STD NMR competition experiments and variable

temperature STD NMR experiments), and the new DEEP-STD NMR approach developed in the framework of this PhD thesis. This allowed us to elucidate on the relative affinity and specificity of 3'-sialyllactose (Neu5Acα3Lac or 3'SL) and 6'-sialyllactose (Neu5Acα6Lac or 6'SL) and, in combination with molecular docking, to provide for the first time a 3D model of the Michaelis complex of an IT-sialidase.

In the structural investigation of CBM40, we used traditional STD NMR binding epitope mappings to study a larger set of sialylated ligands, with the aim to understand the effect of ligands deletions and decorations relative to the native ligands 3'SL and 6'SL. Our NMR analysis fits in a much wider biological study, including X-Ray diffraction structures, glycan microarrays studies, microcalorimetry analyses, immunogold labelling and western blotting[97].

5.2 GH33 D282A structural studies

To study the binding of sialoglycans to the catalytic domain of *Rg*NanH, an inactive GH33 was engineered at QIB, by substituting the acid/base aspartate in position 282 (D282) with an alanine to produce GH33 D282A.

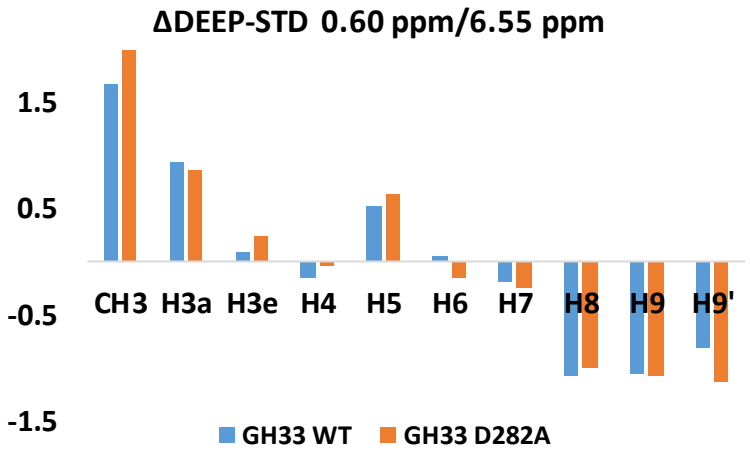

Figure 5.4. ΔDEEP-STD pattern for 2,7-anhydro-Neu5Ac in complex with GH33 WT[87], blue bars, and with GH33 D282A, orange bars. The graph demonstrate that the effect of the mutation is negligible from the STD NMR spectroscopic point of view.

Importantly, we repeated the differential irradiation DEEP-STD experiments on the 2,7-anhydro-Neu5Ac/GH33 complex reported in reference 87, on the 2,7-anhydro-Neu5Ac/GH33 D282A complex (Figure 5.4). The two ΔDEEP-STD patterns are very similar within the experimental error, showing that the effect of the mutation from aspartate

to alanine in position 282 is negligible from the STD NMR spectroscopic point of view. This also implies that the data reported in the present section are easily comparable with the data reported in Chapter 3, Section 3.3, and in reference 87.

Also, as discussed at the end of this section and shown by molecular docking in Section A.9 of the Appendix, the mutation did not affect the binding mode of the ligands so that both, GH33 or GH33 D282A, can be used for binding analyses by STD NMR.

5.2.1 STD NMR binding epitope mapping and variable temperature experiments: investigating relative affinities and kinetics of binding of 3'SL and 6'SL to GH33 D282A

The interactions of 3'SL and 6'SL with GH33 D282A were analysed by STD NMR on samples containing each sialoglycan in the presence of the enzyme. Binding was detected for both oligosaccharides. Figure 5.5 shows the resulting STD NMR spectra of the two complexes, carried out with different irradiation frequencies (0.60 ppm and 6.55 ppm).

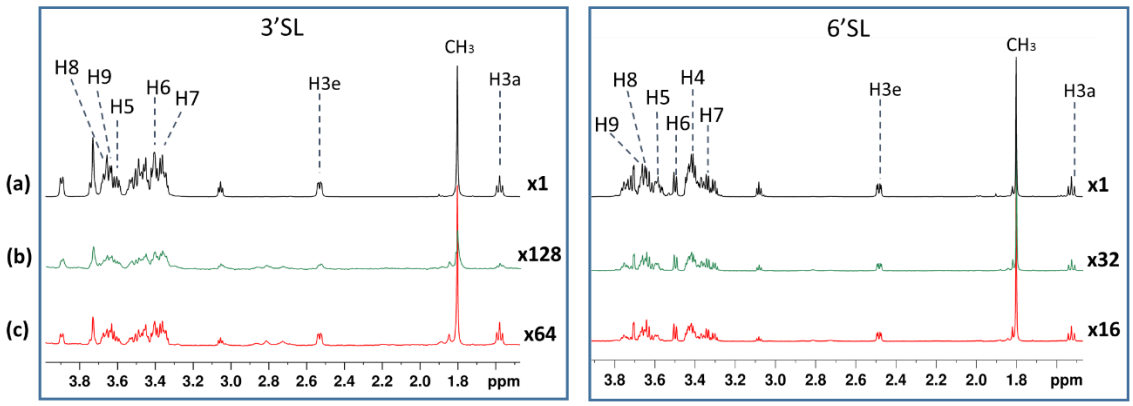

Figure 5.5. 50 μM GH33 D282A in the presence of 1 mM of 3'SL (left panel) and 1 mM 6'SL (right panel), 285 K. **(a)** Reference spectra; **(b)** STD NMR spectra with on-resonance irradiation at 6.55 ppm; **(c)** STD NMR spectra with on-resonance irradiation at 0.60 ppm. Spectra were acquired at 2 s saturation time. The magnification relative to the reference spectrum is given for each difference spectrum.

The observation of binding for 3'SL was in line with this ligand being the preferential substrate of GH33[17], as the sialoglycan microarray analyses conducted with GH33 D282A also showed (34). However, the binding of 6'SL was not expected (it was not detected by microarrays). Next, STD NMR build-up curves at two different frequencies were run for 3'SL and 6'SL and binding epitopes were obtained for the two ligands upon saturation at different frequencies (Figure 5.6).

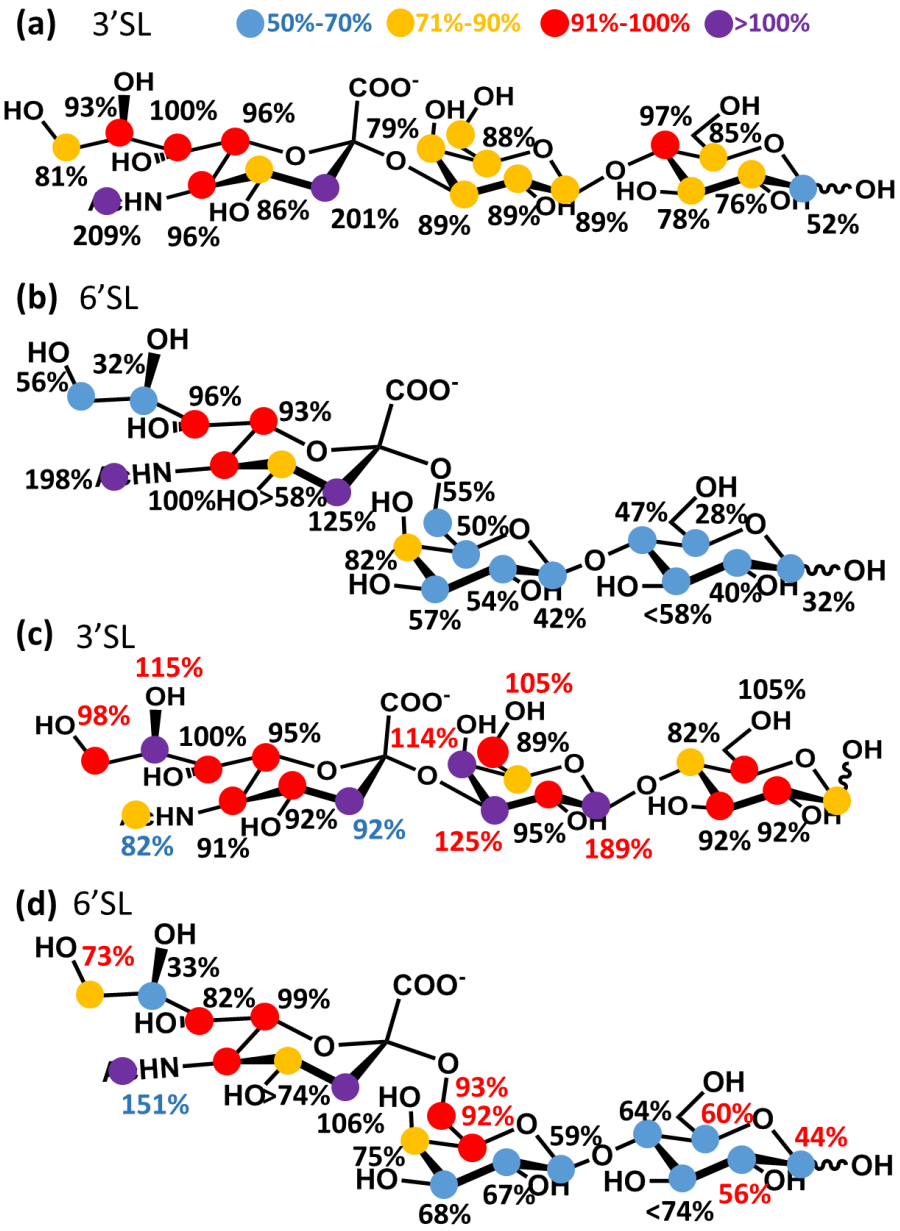

Figure 5.6. Binding epitope mappings of 3'SL and 6'SL as bound to GH33 D282A from STD NMR experiments with irradiation frequency at 0.60 ppm **(a,b)**, or 6.55 ppm **(c,d)**. Relative STD intensities are shown for each proton. Please note that, due to a slight "STD incremental effect" at 0.60 ppm irradiation frequency, the STD intensity of the methyl group is largely over-scale; therefore, we preferred to normalise to the second strongest STD signal (H7 for 3'SL and H5 for 6'SL). To graphically appreciate variations of the binding epitopes upon changing the saturation frequency, in **(c)** and **(d)** the numerical values of the relative STD intensities which decrease from 0.60 ppm to 6.55 ppm are reported in light blue and those which increase are reported in red (unrelatedly to the colour of the dots).

Traditional binding epitope mapping analysis is a very powerful tool in structural studies of ligands for which the structures of their complexes are unknown[70b]. In addition,

comparison of epitope maps acquired at different saturation frequencies can provide further insights into specific intermolecular contacts in ligand/protein complexes, as a preliminary step, before a DEEP-STD NMR analysis.

By comparing the STD NMR data for 3'SL and 6'SL at 0.60 ppm (Figure 5.6a,b), the following considerations could be made:

i) the sialic acid is the main recognition element and the STD pattern for this residue is similar for both oligosaccharides (except for the glycerol moiety which appeared to be more solvent exposed in 6'SL);

ii) the binding epitope of 3'SL is "flatter" than 6'SL (lowest values being 52% and 28%, respectively), with 6'SL showing less contacts with the enzyme than 3'SL;

iii) the absolute STD values are higher for 6'SL than for 3'SL (spectra in Figure 5.5).

In Figure 5.6c,d, the epitopes of the same ligands at 6.55 ppm irradiation frequency are reported. A comparison of the binding epitopes at 0.60 ppm (Figure 5.6a,b) and at 6.55 ppm (Figure 5.6c,d) revealed a further difference:

iv) the pattern of values increasing or decreasing upon changing irradiation frequency differs slightly for the two ligands: for both, the CH_3 and H3 protons receive more saturation at the irradiation frequency of 0.60 ppm and the glycerol protons H8-H9s receive more saturation at the irradiation frequency of 6.55 ppm.

This point will be expanded in Sub-section 5.2.2, where the DEEP-STD NMR fingerprint approach is applied to these ligands. For comparison, the binding epitope mappings for the complex 2,7-anhydro-Neu5Ac/GH33 at 0.60 ppm and 6.55 ppm are reported in Section A.7 of the Appendix.

From the NMR perspective, and as a first interpretation, we can consider that points i), ii) and iii) suggest differences in *relative affinities* and *kinetics*; while point iv) reports on interesting features about the *mode* and *location of binding.* To determine the relative affinity of 3'SL and 6'SL, variable temperature STD NMR experiments were performed, acquiring 2 s saturation time STD NMR experiments with protein saturation at 0.60 ppm and following the intensity of the best isolated STD signals. Variable temperature

experiments rely on the fact that the kinetics of exchange increases with temperature and that STD intensities depend heavily on kinetics[66]. Thus, we can associate the observation of increases in STD signals at increasing temperatures to strong interactions (in STD NMR standards), as they signal a slow kinetics (for which STDs are weak due to the poor accumulation of saturated ligand in the bulk free state) getting faster and hence becoming more "STD-visible" upon heating the sample. This is what we observed for 3'SL (Figure 5.7a).

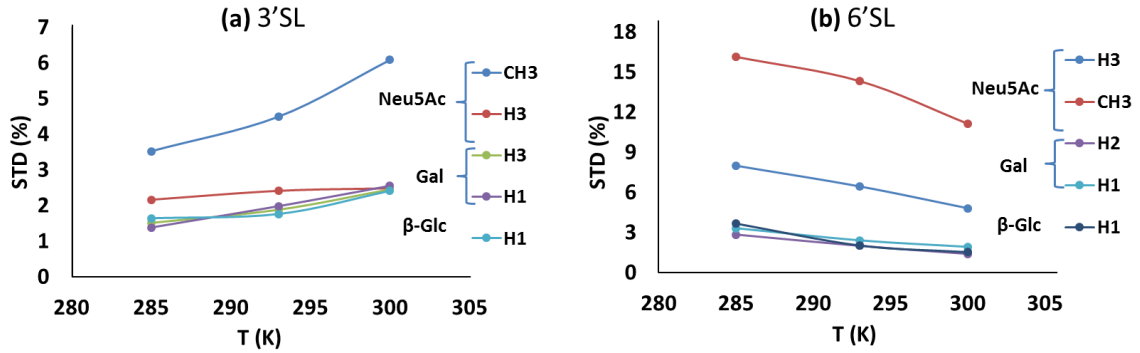

Figure 5.7. Variable temperature STD NMR experiments. STD intensities for 3'SL, **(a)**, and 6'SL, **(b)**, upon binding to GH33 D282A at increasing temperatures. Upon increases in temperature, the STD signals are increasing for 3'SL and decreasing for 6'SL. This strongly support that both ligands are located at different extremes of the kinetic window favourable for STD NMR observation (3'SL in the low kinetics range, and 6'SL in the high kinetics one).

On the other hand, decreasing STD signals at increasing temperature reflect weak interactions, that is, a fast kinetics (favourable for STD NMR observation) getting even faster and less "STD-visible", due to the very short residence time in the bound state precluding significant transfer of saturation by intermolecular NOE[66]. This is what we observed for 6'SL (Figure 5.7b).

Based on these data, we propose that the two ligands fall into two different kinetics windows: i) 3'SL binds to GH33 D282A with slow kinetics (higher affinity) and hence a long residence time in the bound state, compatible with the "flatter" binding epitope mapping observed, whereas ii) 6'SL binds with faster kinetics (shorter bound residence time), as the magnetisation spreads less homogeneously along the ligand.

It is worth noting that this ranking of affinity was properly elucidated by the variable temperature STD NMR experiments, as the wrong conclusion could have been drawn if

the discussion had been focused on their respective STD intensities at a given temperature. As mentioned before, STD intensities were stronger, under similar experimental conditions, for 6'SL (Figure 5.5), which could have been misinterpreted as a sign of a higher affinity of this ligand for GH33. This is an important characteristic of STD NMR experiments: the comparison of intensities between different ligands, under similar experimental conditions, cannot be directly used to rank them in terms of affinities. The absolute intensity of an STD NMR signal can be low either for strong ligands or for weak ones, provided their kinetics of binding are placed in the extreme regions of the appropriate kinetics window for STD NMR observation[66].

5.2.2 DEEP-STD NMR fingerprint: probing the ligands orientation in the binding pocket

While gathering experimental evidences to rank the ligands in terms of relative affinity was relatively straightforward, providing answers about their binding modes requires a more profound analysis. We resorted to our newly developed DEEP-STD NMR analysis and to STD NMR competition experiments. As outlined in Chapter 4, Sub-section 4.3.3, DEEP-STD NMR can be used to determine a fingerprint of the binding pocket, which can aid to discriminate whether two ligands bind in the same location and, if they do, whether their orientation in the bound state is comparable.

In this study, the DEEP-STD factors (with differential frequencies 0.60 ppm/6.55 ppm, and 0.5 s saturation time) were calculated for 3'SL and 6'SL with GH33 D282A and compared to the differential epitope map obtained for 2,7-anhydro-Neu5Ac and GH33 in the same conditions. Their differential epitope maps are reported in Figure 5.8.

For this analysis, the histogram representation is preferable to the differential epitope map onto the ligand structures, as it is the whole ΔDEEP-STD pattern (including values below a given threshold) that drives the data interpretation.

A similar ΔDEEP-STD pattern was obtained for the Neu5Ac moiety of the three ligands, 3'SL, 6'SL and 2,7-anhydro-Neu5Ac. Overall, this suggests that the three ligands occupy the same binding pocket, most likely with the same orientation of the Neu5Ac moiety, as they "respond" to the differential irradiations in an analogous way.

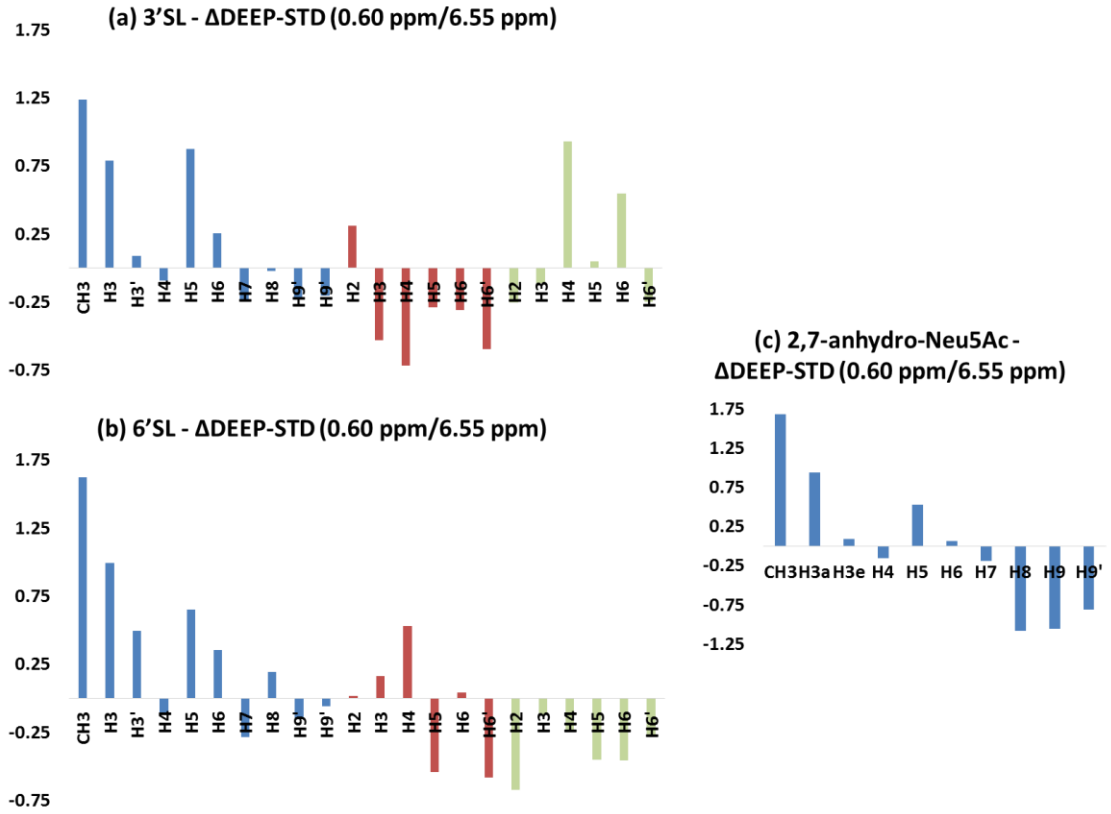

Figure 5.8. Differential Epitope Mapping (0.60 ppm/6.55 ppm) of 3'SL **(a)** and 6'SL **(b)**, in complex with GH33 D282A, and of 2,7-anhydro-Neu5Ac **(c)** in complex with GH33[87]. Protons from the Neu5Ac moiety are shown in light blue, protons from the galactose moiety are shown in red, and protons from the glucose moiety are shown in green.

On closer inspection, the main differences were found at the protons of the polyhydroxy chain (H7 to H9): this is compatible with the different geometry of both sialic acid rings at the non-reducing ends of 3'SL and 6'SL relative to 2,7-anhydro-Neu5Ac. The comparison between 3'SL and 2,7-anhydro-Neu5Ac is particularly interesting: we can look at them as the substrate and the product of the catalytic reaction, respectively. During hydrolysis, the hydroxyl group at C7 attacks and covalently binds to C2, while the lactose is leaving. Arguably, the smaller |ΔDEEP-STD| values of protons H8 and H9s of 3'SL, relative to those of H8 and H9s of 2,7-anhydro-Neu5Ac, may be reporting on the movement of the polyhydroxy chain further from the aromatic residues resonating at 6.55 ppm (Tyr667, and Trp698), as the reaction takes place.

The DEEP-STD factors for the H8-H9 of 6'SL are even more different from those of 2,7-anhydro-Neu5Ac. For 6'SL, the ΔDEEP-STD of H8 is positive and the |ΔDEEP-STD| of H9 are even smaller if compared to those of 3'SL and 2,7-anhydro-Neu5Ac. This suggests

that the polyhydroxy chain of 6'SL is oriented in a different way from the one of 3'SL (arguably pointing slightly further from the patch of aromatic residues).

More remarkably, the DEEP-STD factors of the galactose and glucose moieties differ considerably among the two ligands, reporting on a divergent orientation of their respective lactose parts in the binding pocket. Particularly, homogeneously negative ΔDEEP-STD were observed for the galactose moiety of 3'SL, suggesting vicinity of this galactose ring to one or more aromatic residues (for 6'SL, significantly negative ΔDEEP-STD are only observed for H5 and H6 of galactose and H2, H5 and H6 of glucose).

The analysis of the DEEP-STD NMR fingerprints of the two ligands 3'SL and 6'SL against 2,7-anyhydro-Neu5Ac gives us ground to propose that both ligands binds in the same binding pocket with different orientation of the lactose moiety. The diverse orientations were readily determined in the DEEP-STD maps in agreement with the different geometry of the glyosidic link. To further confirm that the binding takes place in the same binding site, we carried out STD NMR competition experiments.

5.2.3 STD NMR competition experiments: validation of the DEEP-STD NMR data on the location of binding (and comparison of the two approaches)

STD NMR competition is a traditional approach to probe if two or more ligands are binding in the same location on a given protein[81,82]. Here, we first assessed the competition between 3'SL and 6'SL by performing STD NMR experiments of GH33 D282A in the presence of **a)** 3'SL (1 mM), **b)** 6'SL (1 mM) and **c)** 3'SL (1 mM) + 6'SL (1 mM). In all cases, the STD intensities of 7 well-isolated reporter signals were followed at 288 K, as shown in Figure 5.9. The STD intensities of the reporter signals of 3'SL were not significantly affected when 6'SL was present (Figure 5.9, left). In contrast, the STD intensities of the protons of 6'SL were considerably reduced in the presence of 3'SL (Figure 5.9, right). These findings support binding of the two ligands in the same site, with 3'SL having stronger binding affinity (it "displaces" 6'SL almost completely).

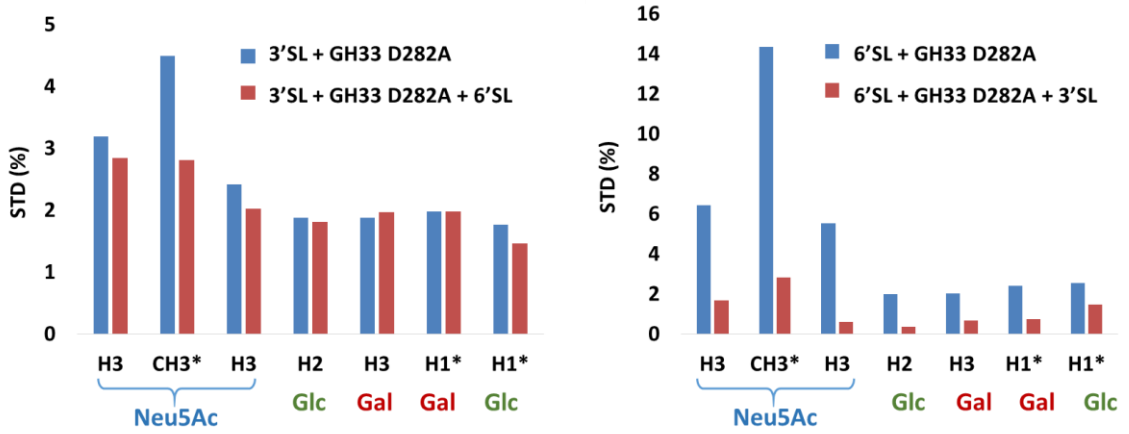

Figure 5.9. STD NMR competition experiments for 3'SL (left) and 6'SL (right). The samples containing the complexes 3'SL/GH33 D282A and 6'SL/GH33 D282A are compared with the sample containing both ligands 3'SL and 6'SL in the presence of GH33 D282A. The stars indicate overlapping of the signal of the two species in the sample containing both ligands. 3'SL displaces 6'SL, but its binding is not affected by the presence of 6'SL.

Another valuable application of STD NMR competition is to identify the position of binding of ligands with unknown binding site by using the "spy approach"[171]. This was developed to investigate strong interactions above the limit of STD NMR detection (slow chemical exchange, *i.e.* K_D below 10^{-7} M) and relies on the use of "STD-visible" weak-binders (ideally of known affinity and for which a 3D structure in complex with the protein of interest is available) to monitor their displacement from the binding site in the presence of other potential high-affinity ligands. Disappearance of the STD signals of the "spy" provides evidence that the tested ligands bind to the receptor more tightly and in the same binding pocket as the spy.

Alternatively, the approach can be used with a strong-binder of known 3D structure as an "inverted spy" to probe for interaction of the protein to weaker ligands. Fortuitously, while performing STD analysis of 3'SL/GH33 D282A on consecutive days, we found that 2,7-anhydro-Neu5Ac could be used for this purpose (Figure 5.10).

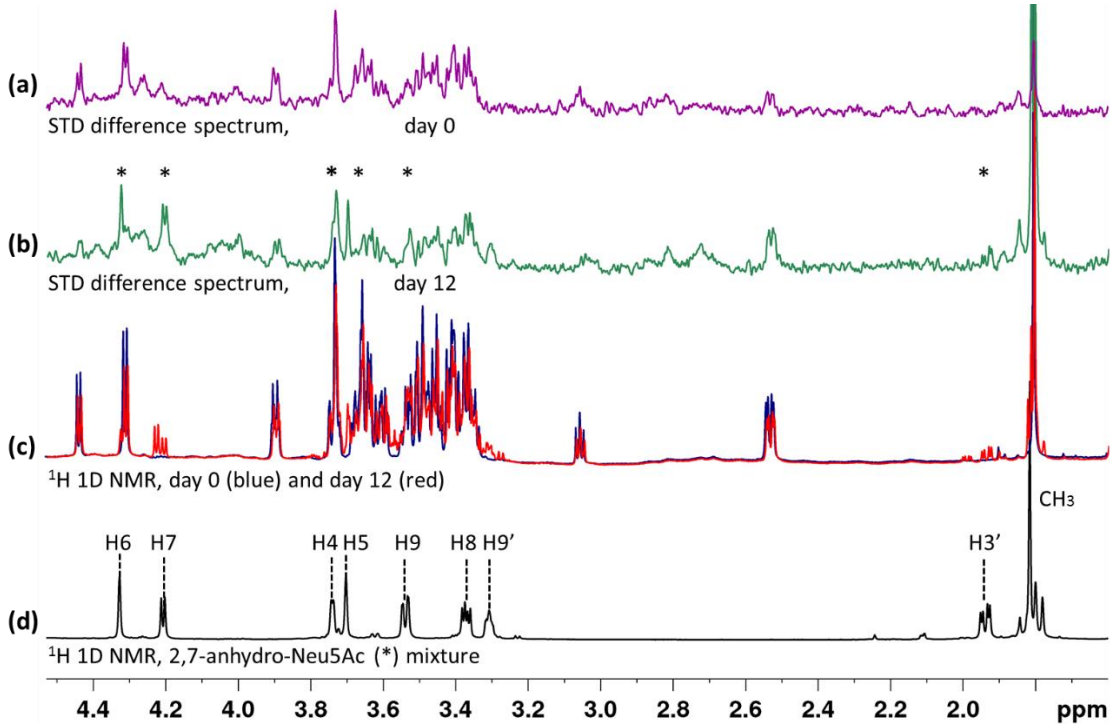

Figure 5.10. STD NMR spectra at day 0 **(a)** and day 12 **(b)** of a sample containing 3'SL in the presence of GH33 D282A (2 s saturation time, 0.60 ppm irradiation frequency). Hydrolysis of 3'SL to produce 2,7-anhydro-Neu5Ac was observed, and it led to a change in the pattern of STD signals. By overlapping the ^{1}H 1D spectra recorded at the two time-points **(c)**, the hydrolysis was estimated to be around 30%. The signals 2,7-anhydro-Neu5Ac are labelled in **(d)** and starred in **(b)**, where we see 2,7-anhydro-Neu5Ac competing with 3'SL. Due to non-equimolar concentration of the two ligands, no quantification can be done, but we can conclude that the product displaced the substrate.

Due to the residual activity of the catalytic mutant, new STD signals appeared in the STD difference spectrum, taking over the signals from 3'SL (Figure 5.10b). The new signals belonged to 2,7-anhydro-Neu5Ac, the product of the catalytic reaction. The enzymatic reaction was also evident in the ^{1}H 1D spectrum, but less appreciably, due to the resulting set of crowded signals. The STD data showed that 2,7-anhydro-Neu5Ac displaces the substrate 3'SL from the enzyme binding pocket. This confirms that both ligands bind at the same site, with the product 2,7-anhydro-Neu5Ac binding more strongly than the substrate 3'SL to the GH33 D282A mutant.

We then further studied the competition between 6'SL and 2,7-anhydro-Neu5Ac. As previously reported, GH33 does not hydrolyse α2/6 substrates; accordingly, GH33 D282A did not show any activity towards 6'SL, and this sample remained stable over time. However, as shown in Figure 5.11, upon addition of an equimolar concentration

of 2,7-anhydro-Neu5Ac to the 6'SL/GH33 D282A sample, the signals corresponding to 6'SL completely disappeared.

Together findings from these STD NMR competitions demonstrated that:

- 3'SL and 6'SL compete for the same binding site (3'SL displace 6'SL, Figure 5.9);
- 3'SL and 2,7-anhydro-Neu5Ac compete for the same binding site (2,7-anhydro-Neu5Ac displaces 3'SL, Figure 5.10);
- 6'SL and 2,7-anhydro-Neu5Ac compete for the same binding site (2,7-anhydro-Neu5Ac displaces 6'SL, Figure 5.11).

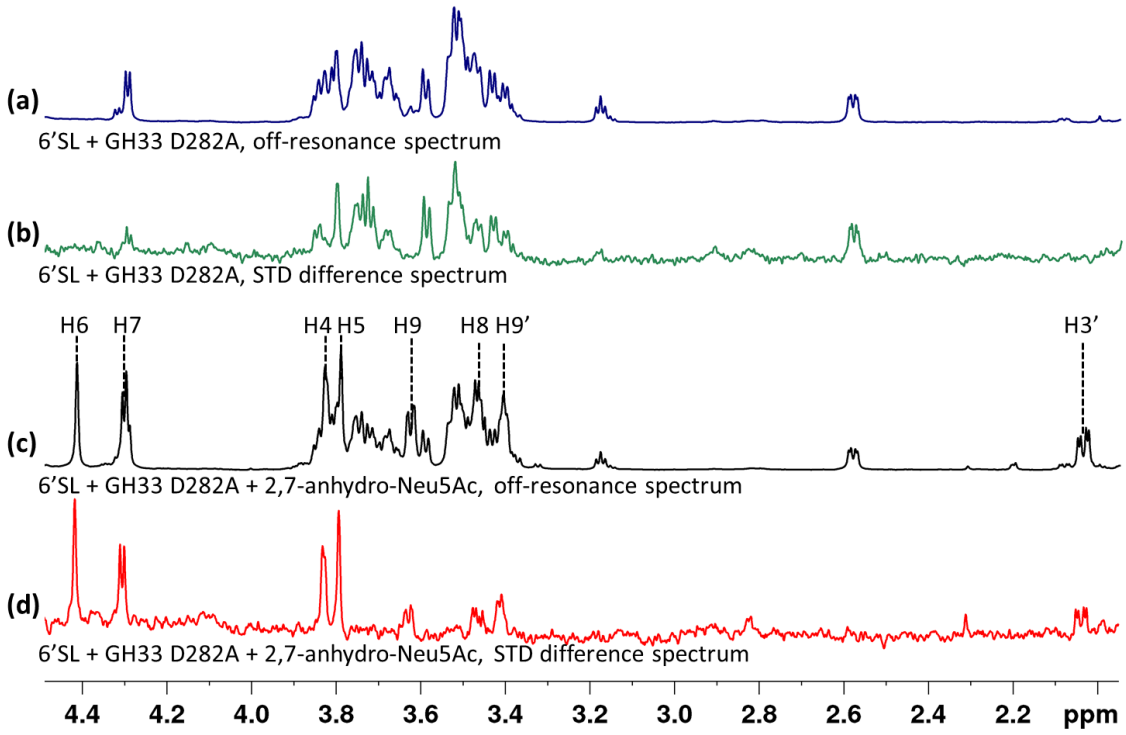

Figure 5.11. STD NMR competition experiment between 6'SL and 2,7-anhydro-Neu5Ac: off-resonance and STD NMR spectra prior addition, **(a)** and **(b)**, and after addition, **(c)** and **(d)**, of 2,7-anhydro-Neu5Ac in the presence of GH33 D282A. The STD signals of 6'SL disappear completely when 2,7-anhydro-Neu5Ac is present, proving that they bind in the same binding pocket. The signals of 2,7-anhydro-Neu5Ac are labelled in **(c)**.

The combination of these STD NMR competition experiments supports the existence of a single binding pocket for these sialoglycans in GH33, as also suggested by the DEEP-STD fingerprint approach (Sub-section 5.2.2). As both oligosaccharides are displaced by 2,7-anhydro-Neu5Ac, and 3'SL displaces 6'SL, a ranking of affinities could be experimentally established with 2,7-anhydro-Neu5Ac > 3'SL > 6'SL.

It is worth stressing that the information obtained by STD NMR competition experiments offers more definitive answers on the location of binding, when compared to the DEEP-STD NMR approach. This is so because similar DEEP-STD fingerprints may also be the result of ligands binding to two different binding sites that fortuitously showed similar geometrical distribution of residues contacting the ligands. The two approaches are therefore complementary: STD NMR competition experiments providing further evidence on the location of binding (with the starting point of a known 3D structure), while DEEP-STD NMR can provide information on the orientation of unknown ligands in known binding sites.

5.2.4 Docking of 3'SL and 6'SL onto GH33: towards the model of the 3D structure in solution

To provide a model of the structure in solution, we combined the STD NMR experimental information with molecular docking, using the crystal structure of 2,7-anhydro-Neu5Ac in complex with *Rg*NanH GH33 (PDB ID: 4X4A[17]) as a starting point. First, the docking conditions were optimised to reproduce the complex of 2,7-anhydro-Neu5Ac in the GH33 binding as observed in the crystal structure[17].

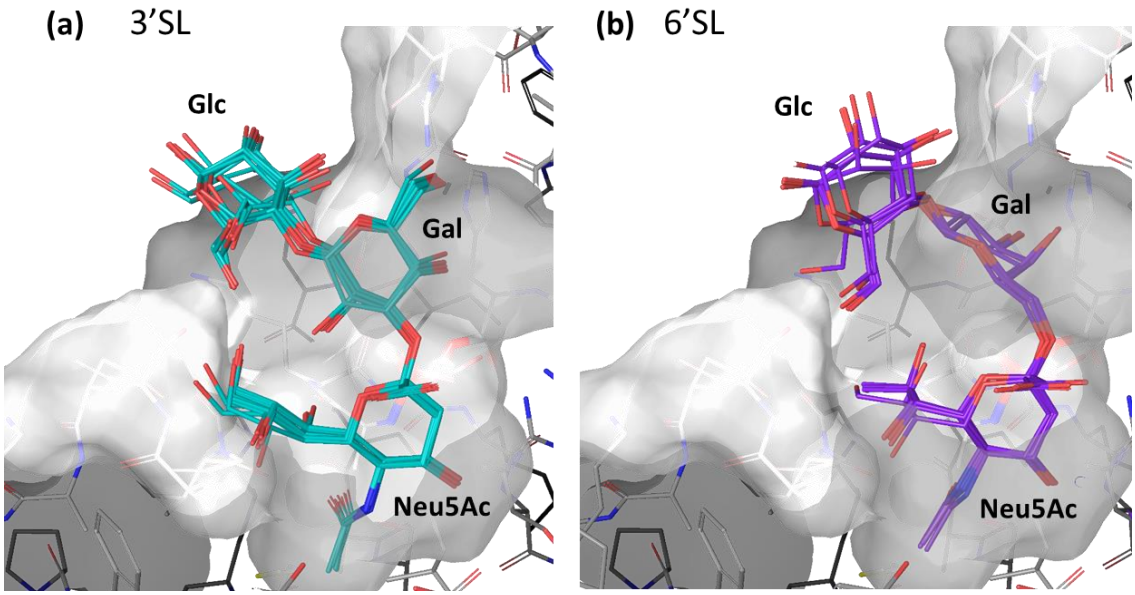

Figure 5.12. Superimposition of the lowest energy convergent docking solutions for GH33 in complex with 3'SL, **(a)** 10 poses, and 6'SL, **(b)** 6 poses.

The coordinate of GH33 from the X-Ray structure with 2,7-anhydro-Neu5Ac (PDB ID: 4X4A) were used to generate the receptor grid; then, 3'SL and 6'SL were docked. Both ligands converged to lowest energy solutions showing the Neu5Ac ring fitting deeply in the catalytic cavity (with *glide gscore* and *glide emodel* of *circa* -5.0 kcal/mol and -55 kcal/mol, respectively, comparable for both ligands). For 3'SL, convergence of the first 10 lowest poses was observed, whereas convergence of the first 6 poses was observed for 6'SL (Figure 5.12). The docking tables are reported in Section A.10 of the Appendix.

The solutions with higher energies presented the ligands flipped by 180°, fitting the glucose in the catalytic cavity, instead. These solutions could safely be excluded based on the experimental STD NMR evidences showing that GH33 does mainly recognise the sialic acid residue (Figure 5.6).

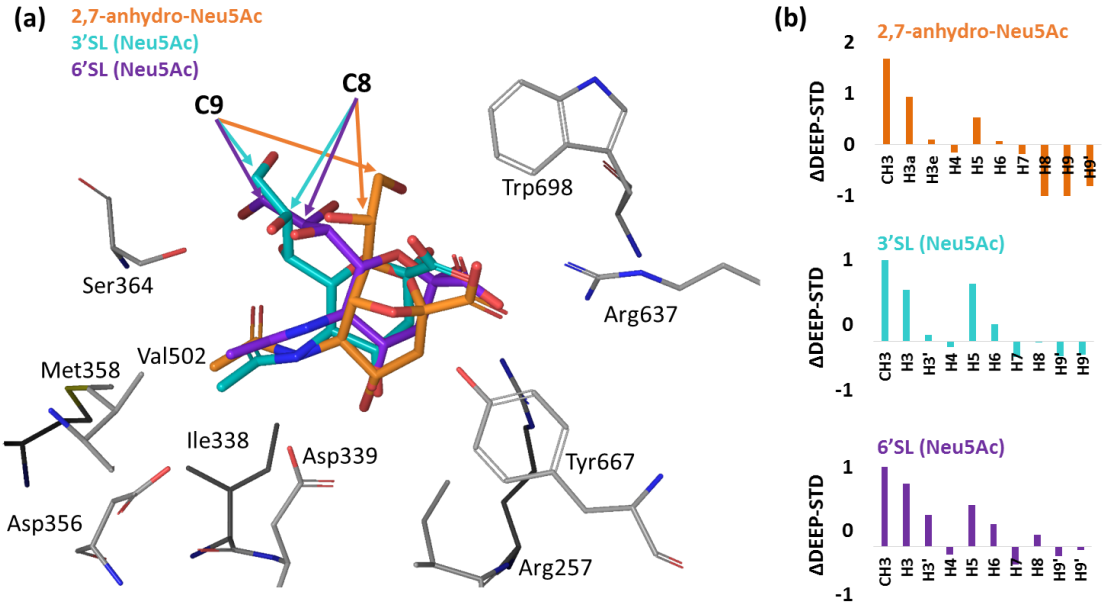

Figure 5.13. **(a)** 2,7-anhydro-Neu5Ac (in orange) from PDB ID: 4X4A, overlapped with the Neu5Ac rings of the best docking solutions for 3'SL (cyan) and 6'SL (purple). Only protein residues within 3 Å are shown and labelled. C8 and C9 are pointed by arrows. Protons are omitted for clarity (here and in the following molecular schemes). **(b)** ΔDEEP-STD for 2,7-anhydro-Neu5Ac[87] and the Neu5Ac rings of 3'SL and 6'SL, following the same colouring scheme as in **(a)** (adapted from Figure 5.8).

In both 3'SL/GH33 and 6'SL/GH33 models, the Neu5Ac ring was found in a comparable orientation to 2,7-anhydro-Neu5Ac in the crystal structure. Interestingly, the glycerol moieties of the tri-saccharides were further away from the Tyr667 and Trp698 residues

(resonating at 6.55 ppm) than the polyhydroxy chain of 2,7-anhydro-Neu5Ac (whose hydroxyl on C7 is covalently bound to C2 because of the intramolecular trans-glycosylation). In agreement with the DEEP-STD fingerprint data, H8 and H9s of 6'SL (purple in Figure 5.13) appear to point in the opposite direction relative to the same protons of 3'SL (cyan in Figure 5.13).

It is worth noticing that the main difference between the binding of 3'SL and 6'SL is in the orientation of the lactose moiety. The α2/3 glycosidic linkage directs the galactose of 3'SL to a very efficient π-stacking with Trp698. Figures 5.14a and 5.15a provide two views of 3'SL π-stacking to Trp698 and shows how the galactose region encompassing C3, C4 and C5 point towards the tryptophan side chain (in magenta, resonating around 6.55 ppm, as discussed in Chapter 3). This is in strong agreement with the DEEP-STD data reported with both figures (Figures 5.14a and 5.15a).

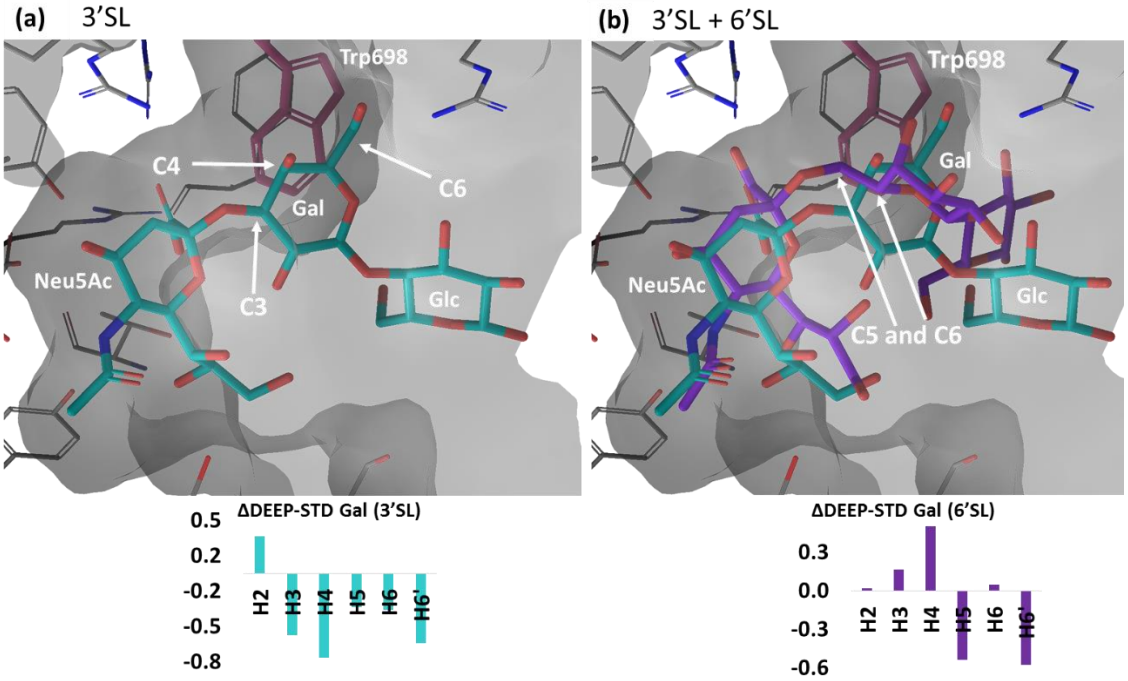

Figure 5.14. Lowest energy docking solutions for 3'SL **(a)** and 6'SL **(b)** bound to GH33. For the purposes of comparability, both 3'SL and 6'SL are shown in **(b)**. Top view of the galactose orientation relative to Trp698 (shown in magenta). The ΔDEEP-STD histograms (0.60 ppm/6.55 ppm) of the galactose protons of both ligands (adapted from Figure 5.8) are shown at the bottom. Positions of ligands protons with negative ΔDEEP-STDs are pointed by arrows and labelled to show their close contact to the aromatic residue Trp698.

The glucose moiety of 3'SL is pointing more towards the solvent, again in accordance with low ΔDEEP-STDs as shown in Figure 5.15a: if the ligand protons are solvent exposed, they will not be strongly affected by differential irradiation frequencies, although the strong positive ΔDEEP-STD observed for the glucose-H4 of 3'SL remains to be investigated.

For 6'SL, the larger flexibility of α2/6 glycosidic linkage makes the galactose ring to sit in an orientation rather perpendicular to the galactose of 3'SL and therefore to Trp698 (Figure 5.15b). This results in galactose H5 and H6s protons pointing towards the tryptophan, while the rest of the galactose protons point to the other side of the binding pocket. This is in agreement with the DEEP-STD data reported below both figures (Figures 5.14b and 5.15b).

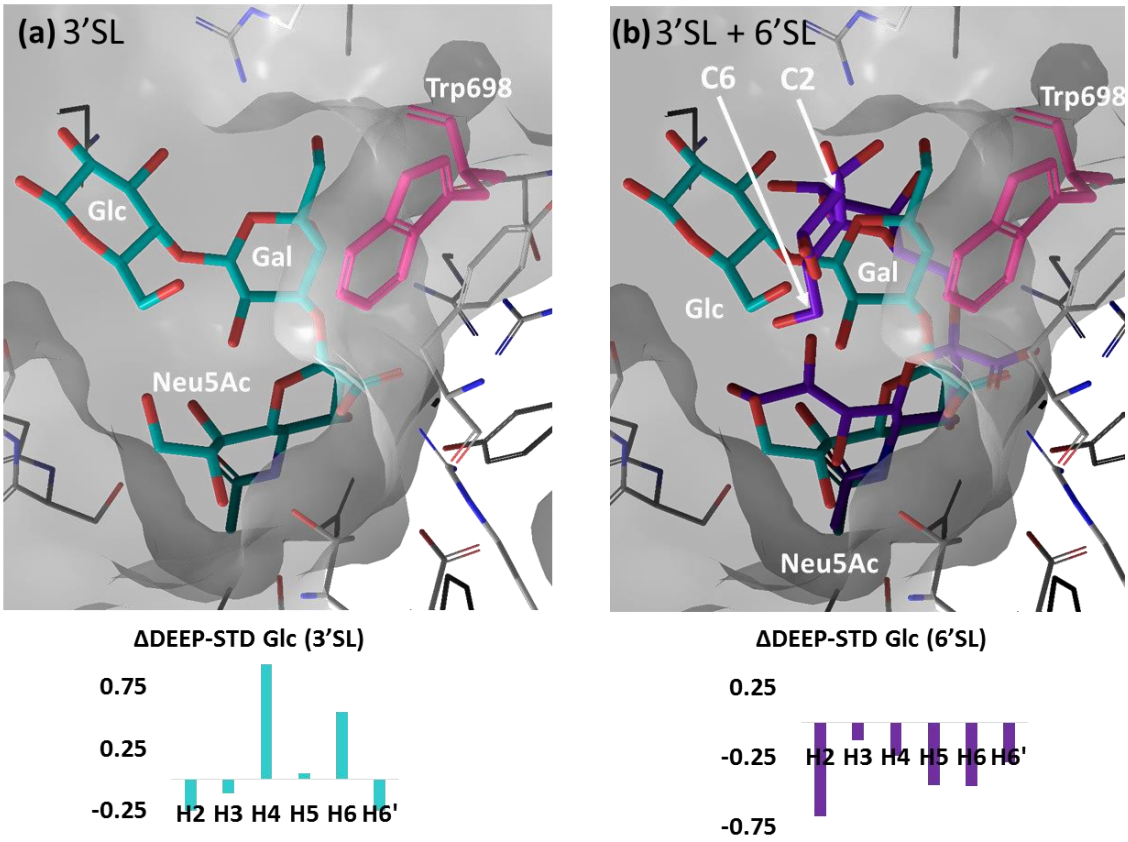

Figure 5.15. Lowest energy docking solutions for 3'SL **(a)** and 6'SL **(b)** bound to GH33. For the purposes of comparability, both 3'SL and 6'SL are shown in **(b)**. Side view of Trp698 (shown in magenta) relative to the galactose of 3'SL and the glucose of 6'SL. The ΔDEEP-STD histograms (0.60 ppm/6.55 ppm) of the glucose protons of both ligands (adapted from Figure 5.8) are shown below each figure. Ligands protons with negative ΔDEEP-STDs are pointed by arrows and labelled to show their close contact to the aromatic residue Trp698.

The distinct bending of 6'SL imposed by its flexible glycosidic linkage brings its glucose ring closer to Trp698 than the adjacent galactose ring: glucose C2 and C6 point towards the tryptophan, as one would expect from the negative DEEP-STD factors observed for H2 and one of the H6s (Figure 5.14b).

To further support that mutation of aspartate to alanine at position 282 did not affect the binding mode of the sialoglycans into the GH33 binding pocket, the mutant was reproduced computationally based on the crystal structure of GH33[17] by mutating the single aspartate 282 residue to an alanine on Maestro Schrodinger (refer to Materials and Methods, Sub-section 5.5.3 for details). From this structure, a new grid was generated, with size and coordinates identical to the WT grid. The docking of 3'SL and 6'SL was repeated in the newly obtained grid, and the resulting poses and clustering perfectly matched those obtained for the wild type GH33. This result suggests that, whereas the D282A mutation partially inactivates the enzymatic activity, it does not appear to affect the binding mode of 3'SL and 6'SL in the catalytic cleft. This is something important regarding the studies of ligand binding to enzymes: binding in many cases can be considered a separate event from the enzymatic reaction, meaning that the presence or absence of catalytic residues might not impact significantly the binding modes of ligands. This explain why ligands of similar nature can bind in the same catalytic site with very different enzymatic outcomes (reaction or no reaction at all). For example, this situation was also described by NMR regarding the binding of ligands to the human blood group B galactosyltransferase (GTB): UDP-Gal and UDP-Glc sugar nucleotides bind equally well to the enzyme, whereas the reaction only takes place on the first one (the natural substrate).

5.2.5 CORCEMA-ST simulations on 3'SL/GH33 and 6'SL/GH33: validating the 3D structures in solution

To validate the 3D structures of the 3'SL/GH33 or 6'SL/GH33 complex in solution, CORCEMA-ST simulations were run for the lowest energy solutions from docking calculations of 3'SL and 6'SL. The simulated build-up curves were compared against the experimental ones obtained at an irradiation frequency of 0.60 ppm (Figure 5.16 and 5.17). For both ligands, NOE R-factors lower than 0.30 were obtained, validating the docking solutions as good models for the complexes in solution.

It is worth noticing that, for 6'SL, the glucose moiety was excluded from the calculations, due to the high number of glucose overlapping proton signals in the NMR spectrum. Remarkably, the presence of overlapping protons does not have a major effect on the interpretation of DEEP-STD data, as the effect is cancelled by comparison of data sets in two conditions. Instead, proton overlapping brings to larger discrepancies when comparing the CORCEMA-ST simulated data with the experimental data: CORCEMA-ST predicts STD intensities for each single proton, while for overlapped protons we only appreciate the average of the STD intensities of the protons resonating at the same frequencies. In these situations, it is hard to fit the simulated data to the experimental one (for example, the NOE R-factor for 6'SL including the glucose was 0.35).

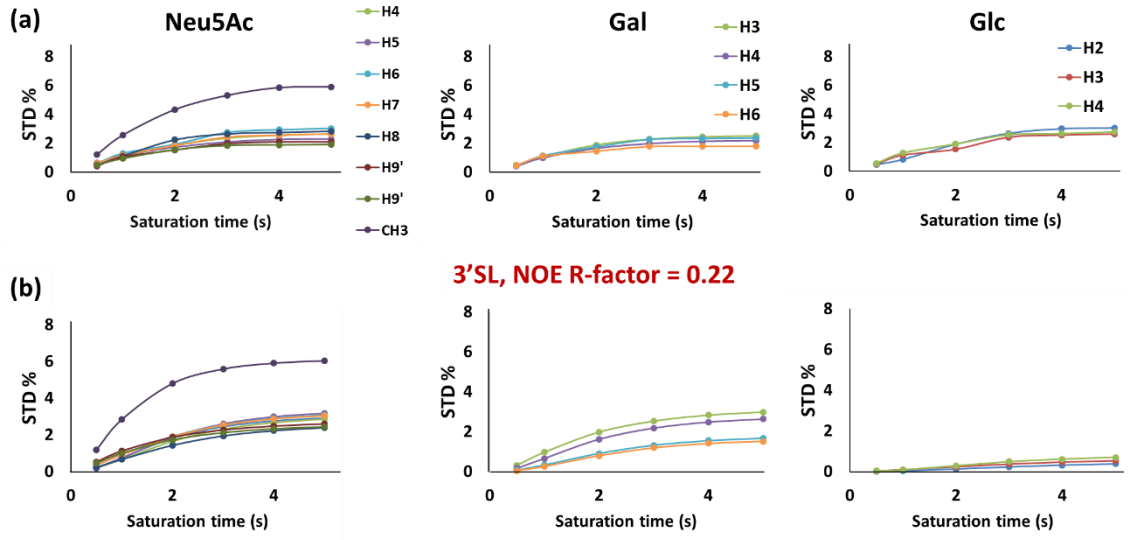

Figure 5.16. Experimental STD NMR build-up curves, **(a)**, and CORCEMA-ST simulated build-up curves, **(b)**, for the 3'SL/GH33 complex.

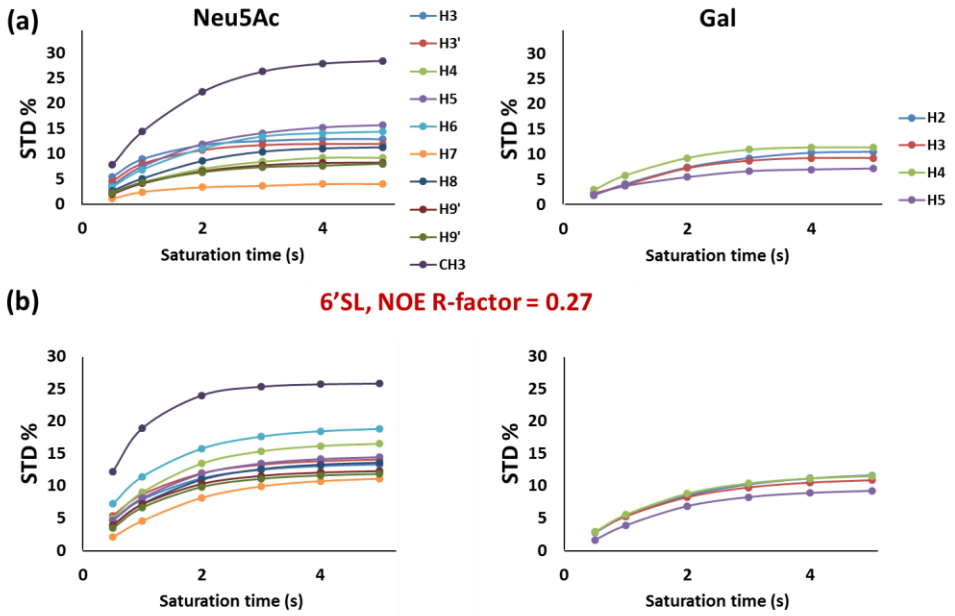

Figure 5.17. Experimental STD NMR build-up curves, **(a)**, and CORCEMA-ST simulated build-up curves, **(b)**, for the 6'SL/GH33 complex.

5.2.6 Some thermodynamics considerations

The intramolecular *trans*-glycosylation reaction of *Rg*NanH has been demonstrated to have a strict specificity for α2-3-linked sialic acid[17,97]. Therefore, the STD NMR findings supporting binding of both 3'SL and 6'SL to GH33 D282A was unexpected. In addition, STD NMR experiments (including variable temperature and competition experiments) demonstrated that the binding takes place with different affinities and with different kinetics: 3'SL inhibits the binding of 6'SL almost completely and binds with a considerably larger residence time in the bound state.

According to our docking models, the two ligands would bind with rather similar binding modes, particularly at the main recognition element which is the sialic acid ring at the non-reducing end, and differing mainly at the orientation of the lactose moiety. Whereas for 3'SL a π-stacking with Trp698 is observed, the α2/6 glyosidic linkage of 6'SL directs its galactose ring further from the tryptophan. Trp698 has been proposed to be one of the key residues involved in the *trans*-sialidase reaction, providing an hydrophobic environment that would repel water molecules from the catalytic cleft, and promote the closest nucleophilic hydroxyl (either from another sugar or from the substrate itself) to attack the covalent intermediate of the Neu5Ac bound to the nucleophilic tyrosine, therefore favouring an intramolecular reaction[18]. Trp698 has also

been associated to the strict specificity of *trans*-sialidases, and especially of the IT-sialidases, for α2/3 sialoglycans. Our 3D models showed that the galactose ring is in a favourable position for π-stacking with Trp698 only in the case of the 3'SL, in contrast to 6'SL. This could account for the lack of reaction with 6'SL, which is also recognised with much lower affinity.

From a thermodynamic point of view, the higher affinity of 3'SL for the catalytic domain of *Rg*NanH can be explained thermodynamically, as due to:

i) Favourable enthalpy contribution coming from the π-stacking with Trp698 (only possible for 3'SL);

ii) Unfavourable entropy contribution for 6'SL, due to the extra rotatable bond in the geometry of the α2/6 glycosidic bond.

Additionally, the STD NMR competition experiments show that the presence of 6'SL does not interfere with the interaction between 3'SL and GH33 D282A, even when they are in equimolar concentrations. Therefore, although both ligands are present in the mucus layer of the mammalian intestine, the potentiality of GH33 to recognise and weakly bind α2/6 sialoglycans is expected not to affect the efficiency of hydrolysis of α2/3 sugars in physiological conditions.

5.3 CBM40 structural studies

As opposed as to the structural study on *Rg*GH33 D282A, focused on solving the 3D model of the enzyme in complex with 3'SL and 6'SL, the study on CBM40 reported in the present section focuses on a much larger set of sialylated ligands. Here, the aim is to understand the effect of ligands deletions and decorations relative to the native ligands 3'SL and 6'SL in the molecular recognition of the carbohydrate binding module *Rg*CBM40.

5.3.1 Binding tests and binding epitope mapping: investigating the broad specificity of CBM40

In this study[97], Neu5Ac, *N*-glycolylneuraminic acid (Neu5Gc), 2,7-anhydro-Neu5Ac, Neu5Acα3Lac (3'SL), Neu5Acα6Lac (6'SL), Neu5Acα3Gal (3'SGal), Neu5Acα6Gal (6'SGal), Neu5Acα3LacNAc (3'SLN), Neu5Acα6LacNAc (6'SLN), Neu5Gcα3Lac (3'SLGc),

Neu5Gcα6Lac (6'SLGc), Neu5Acα6GalαOC$_3$H$_6$N$_3$ (Neu5Ac-STn), Neu5Gcα6GalαOC$_3$H$_6$N$_3$ (Neu5Gc-STn), and Neu5Acα3Galβ3GalNAcαOC$_3$H$_6$N$_3$ (STF) were tested as potential ligands for CBM40. The structures and atom nomenclature of Neu5Ac and 2,7-anhydro-Neu5Ac are shown in Figure 5.2a,b, and those of the di- and trisaccharides are reported in Figure 5.18 and Figure 5.20, together with their binding epitopes.

The set of sialylated ligands includes several elements of diversification, with the main motif being the α2/3 or α2/6 glycosidic linkage of sialic acid and its derivative to the rest of the molecule. Another essential element is the presence of the Neu5Gc as the non-reducing end. Neu5Gc is a Neu5Ac derivative, where the methyl of the acetyl group is substituted with a glycolyl group. Also, the deletion of the reducing glucose is another element under investigation, such as the presence of acetamide groups in position C2 of the reducing moieties.

STD NMR experiments at 2 s saturation time were performed on all the sugars. Binding to CBM40 was detected for all the di- and trisaccharides, while the three monosaccharides, Neu5Ac, Neu5Gc and 2,7-anhydro-Neu5Ac did not show binding. For the 11 binders, STD NMR build-up curves were run, and their binding epitope mappings were obtained as described in Materials and Methods (Sub-section 5.5.2). The sialic acid ring was found to be the main recognition element across all ligands tested and the binding mode was not affected by the nature of the glycosidic linkage (α2/3 or α2/6) of the sialoglycan.

Figure 5.18a shows the STD NMR spectra of 3'SL and 6'SL interacting with CBM40, and Figure 5.18b their corresponding binding epitope mappings. The binding epitopes of 3'SL and 6'SL from STD NMR were in good agreement with the crystal structures that were solved in parallel to this work of the complex between *Rg*CBM40 and 3'SL or 6'SL (Figure 5.19, PDB ID: 6ER3 and 6ER4[97]). In the crystal structures, the sialic acid shows the closest contacts to the protein surface while the lactose moiety is more solvent exposed, as also suggested from the low STD intensities observed for the galactose and glucose protons. The very strong STD intensity of the methyl group is in excellent agreement with the *N*-acetyl group sitting deep in the hydrophobic pocket surrounded by the protein protons from the side chains of Ile95, Tyr116, and Tyr210 (Figure 5.19), about which we discussed in Chapter 3, Sub-section 3.2.2. The high intensity on H7 compared to the

much lower one on the adjacent H8 agrees with H7 facing the hydrophobic side chains, while H8, in *trans*-conformation to it, is pointing towards the solvent.

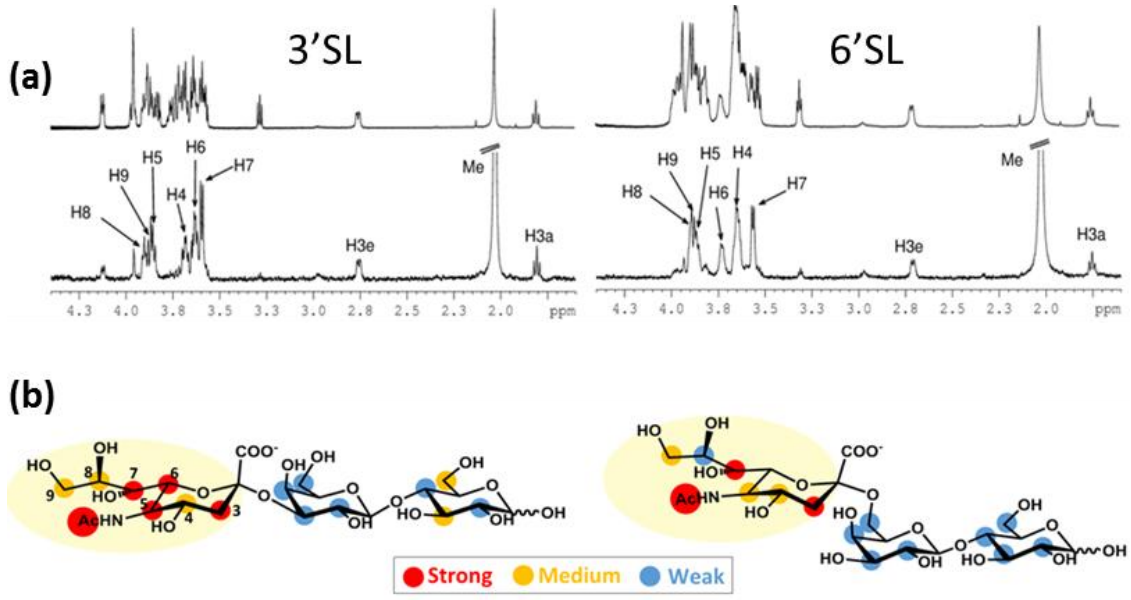

Figure 5.18. STD NMR analysis of CBM40 binding to sialoglycans, **(a)** Reference (top, x1) and difference (bottom, x32) spectra of 3'SL and 6'SL, 2 s saturation time. The strongest signals from Neu5Ac protons are labelled in the difference spectra. **(b)** Binding epitope mapping from STD NMR of 3'SL and 6'SL. Legend indicates relative STD intensities normalised at H7: blue, 0–24%; yellow, 25–50%, red 51–100%; larger red dots indicate values over 100%. Figure from 97.

The binding epitopes of the other oligosaccharides tested showed that there are no stark differences in the orientation of the sialic acid ring in the binding pocket of CBM40, within the STD NMR experimental error (Figure 5.20). The absence of the reducing glucose (for 3'SGal, 6'SGal, Neu5Ac-STn and Neu5Gc-STn) or the presence of decorations on the ligands (3'SLN and 6'SLN) did not significantly impact on the overall binding mode.

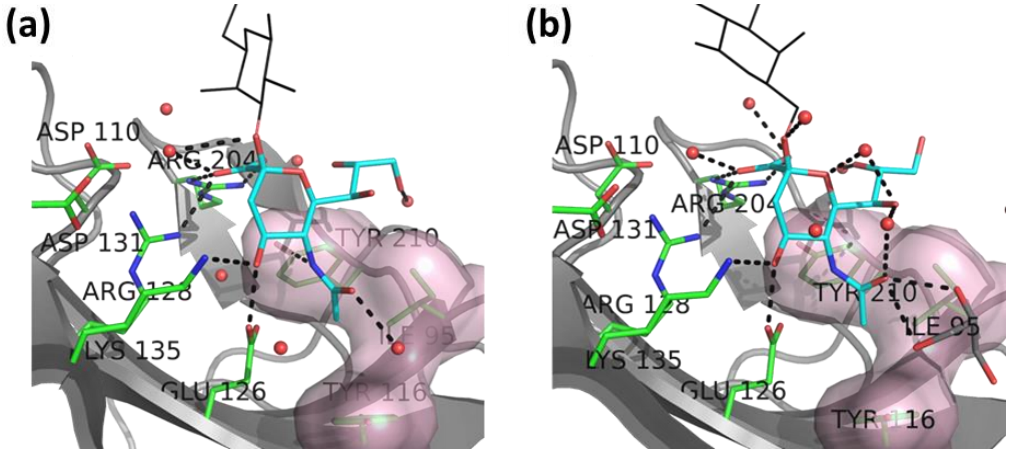

Figure 5.19. Close-up view of the crystal structures of CBM40 in complex with **(a)** 3'SL and **(b)** 6'SL, PDB ID: 6ER3 and 6ER4, respectively. The Neu5Ac residue is shown in cyan and the galactose residue as black lines, for clarity the glucose residue i s not shown. Interacting CBM40 residues are shown in green with black dashed lines indicating hydrogen bonding interactions. The semi-transparent surfaces indicate hydrophobic niche. Figure from 97.

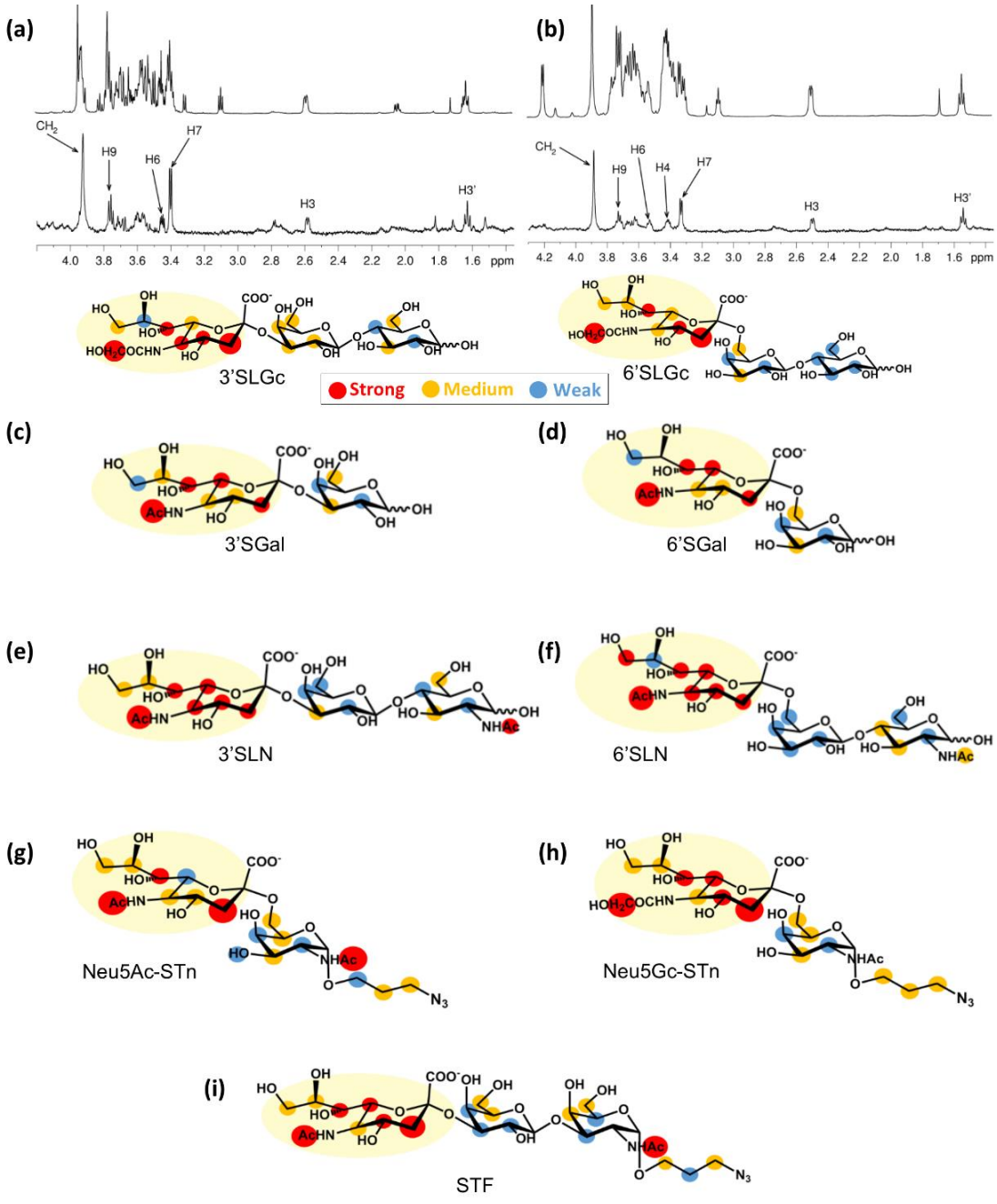

Figure 5.20. Binding epitope mapping of **(a)** Neu5Gcα3Lac (3'SLGc), **(b)** Neu5Gcα6Lac (6'SLGc), **(c)** Neu5Acα3Gal (3'SGal), **(d)** Neu5Acα6Gal (6'SGal), **(e)** Neu5Acα3LacNAc (3'SLN), **(f)** Neu5Acα6LacNAc (6'SLN), **(g)** Neu5Acα6GalαOC$_3$H$_6$N$_3$ (Neu5Ac-STn), **(h)** Neu5Gcα6GalαOC$_3$H$_6$N$_3$ (Neu5Gc-STn), and **(i)** Neu5Acα3Galβ3GalNAcαOC$_3$H$_6$N$_3$ (STF) as bound to CBM40 from STD NMR experiments. Legend as in Figure 5.18. STD NMR spectra (2 s saturation

time) of the binding of 3'SLGc and Neu5Gcα2-6Lac 6'SLGc to CBM40 are shown as a representative example (reference spectra on top x1, difference spectra on bottom x64). Figure from 97.

Interestingly, CBM40 also showed binding to Neu5Gc-ending oligosaccharides, albeit with a lower strength (overall lower STD signals in Figure 5.20a,b when compared to Figure5.18a).

Again, sialic acid was the main recognition element of these sialoglycans, but the binding epitope mappings were slightly different to those observed for 3'SL and 6'SL. For the Neu5Gc-ending ligands, stronger STD intensities on H3s and lower ones on H6 were observed, suggesting a small reorientation of the ring around C3, which would expose C6, to fit the bulkier hydroxyl group on the acetamide moiety.

5.3.2 Interpreting the different "STD incremental effect" of the methyl groups across the ligands

All the Neu5Ac-containing oligosaccharide ligands of CBM40 showed an over-the-scale STD value of the acetamide methyl group. In Chapter 3, Sub-section 3.2.2, we analysed the "STD incremental effect" for the methyl group of Neu5Ac in the 6'SL/CBM40 complex, and we used "on-resonance" scanning to explain this effect. We ascribed it to the combination of:

i) the small size of the molecule (short τ_c), making spin diffusion inefficient;

ii) the shallow binding pocket where only the methyl group of Neu5Ac is deeply buried in the hydrophobic niche;

iii) the direct irradiation at the aliphatic side chain of Ile95 within the hydrophobic niche;

iv) the fast kinetics of exchange, where the short bound residence time prevents the magnetisation to spread evenly from the methyl group, in very close contact with the protein protons, and the rest of the ligand protons, further from the protein surface (we will explain this point further at the end of this sub-section).

As mentioned above, the binding epitopes of all ligand tested are similar within the experimental error (Figure 5.18 and 5.20), and the methyl group is always largely over-

scale. For the Neu5Ac-containing ligands, the normalised STD initial slopes of the methyl group ranged from 450% to 1200% relative to H7 of the Neu5Ac. For the three Neu5Gc-containing ligands, where the glycolic group replaces the methyl group (Figure 5.20a,b,h), the normalised STD initial slope of the methylene group (-$\underline{CH_2}$-OH) was lower (167%, 106% and 108% for 3'SLGc, 6'SLGc and Neu5Gc-STn, respectively). These findings agree with the presence of the bulky hydroxyl of the glycolic group somehow hindering the interaction, bringing the CH_2 further from the protein surface, when compared to the Neu5Ac ligands.

One of our initial concerns was to exclude the possibility of direct irradiation on the methyl group. The analysis of the resulting binding epitopes provided further evidences that this is indeed the case. Firstly, the spectra of the 3'SLN and 6'SLN contain two methyl groups, one on the Neu5Ac and one on the lactosamine. Although their signals are only 0.01 ppm-0.05 ppm far from each other, the STD intensities of the lactosamine methyl group were in the same range as those of the rest of the molecule. Secondly, the H3 axial, which falls about 0.3 ppm high-field to the methyl group in analysis, does not show such a dramatic effect. Finally, because the methyl groups share the same chemical shift (between 1.70 and 1.80 ppm), normalised STD intensities ranging between 450% and 1200% indicate that the phenomenon is ligand-dependent and cannot simply be ascribed to a technical artefact.

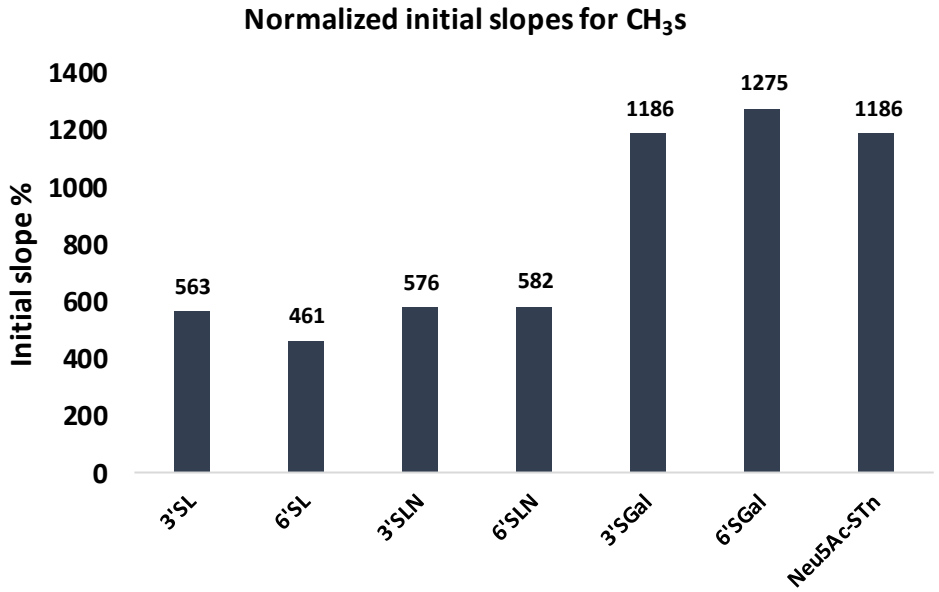

Figure 5.21. STD initial slopes of CH_3s normalised relative to H7 for the Neu5Ac-containing binders. The disaccharides show higher values, suggesting a shorter residence time in the binding pocket.

Since direct irradiation was excluded, we investigated whether the different STD intensities across the Neu5Ac-containing ligands could provide information on the kinetics of binding. The size of the protein, the binding mode and the irradiation frequency, points i) to iii), are constant for all the STD NMR experiments; instead, the residence time, point iv), can be considered the variable factor. All the ligands interact with different affinities and, most likely, different kinetics of binding. According to iv), the faster the kinetics of interaction, the larger the gap between the STD value of the methyl group and that of the H7 (H7 was used as a reporter signal as it generally showed the second strongest STD signal across the set of ligands).

To explain how the gap in STD intensities between the methyl groups and the H7 relates to the kinetics of the interaction, we can use an analogy to thermal diffusion. Let us imagine a bar of metal which is brought close to a flame or any other source of heat. The very extremity of the bar, the one closer to the flame, will be the first part to get hot, and the hottest one at any time point. Hence, the longer time the bar will spend close to the flame, the further the heat will diffuse and the hottest the rest of the bar will get (in a gradient-manner, but we will not get too formal about the mathematics of it). If we now imagine the magnetisation to be the heat, the protein surface to be the source of heath, and the ligands to be bar-like, we can translate the thermic diffusion analogy to explain the gap in STD intensities between the extremity of the ligands and the rest of the molecule. In fact, as we know from the X-Ray structure in Figure 5.19, the methyl groups of the sialylated ligands (extremity of the bar) fits inside a hydrophobic pocket at a very close distance to Ile95 (source of heat). Therefore, the amount of saturation (heat) which will reach the rest of the molecule (the rest of the bar) depends on the residence time spent by the ligand in the binding pocket.

As shown in Figure 5.21, the normalised values of the methyl protons group in two ranges: the trisaccharides with values around 500% and the disaccharides with values around 1200%. With the thermic diffusion analogy in mind, this pattern can be ascribed to a faster kinetics of binding for the disaccharides relative to the trisaccharides (albeit we do not have any thermodynamic data confirming this yet).

5.3.3 Discussion of the STD NMR data in comparison with the glycan microarray and ITC data

In summary, our STD NMR study reports CBM40 as a low affinity and broadly-specific binder of sialoglycans containing Neu5Ac at the non-reducing end. Albeit the sialic acid moiety is the main recognition element, sialic acid monomers were not recognised and at least a second moiety was necessary for recognition. The main role of the carbohydrate binding module is believed to be concentrating the IT-sialidase to potential substrates containing sialic acid. Neu5Ac, Neu5Gc and 2,7-anhydro-Neu5Ac are not hydrolysable and, consistently with its biological role, CBM40 does not recognise them. Accordingly, the minimal recognised sugars are the sialylgalactosides (3'SGal, 6'SGal, Neu5Ac-STn and Neu5Gc-STn).

Our STD NMR findings are in agreement with the sialoglycan microarray screening and isothermal titration calorimetry (ITC) analyses[97], showing that CBM40 bound to terminal Neu5Ac and Neu5Gc attached to aglycons by α2/3 and α2/6 linkages, with a preference for terminal Neu5Ac over Neu5Gc and for α2/3 over α2/6-linkages (34). However, although the microarray reported excellent binding also to 3'SLN but not to 6'SLN, both gave positive results in STD NMR experiments. The same was observed for 6'SLGc and Neu5Gc-STn (which gave STD signals but no binding on microarrays).

STD NMR is very sensitive towards binding events but is not the optimal technique to quantify binding affinities. In (34), ITC data were provided for Neu5Ac, Neu5Gc, 3'SL, 6'SL, 3'SLGc and 6'SLGc as bound to CBM40. While the ITC data were mostly in agreement with the STD NMR data, it is interesting to compare the outcome from the two techniques: whereas by STD NMR we could not detect significant differences in affinities between 3'SL and 6'SL, the two showed a K_D of 0.56 mM and 1.41 mM, respectively. On the other hand, we did notice a lower STD intensity for the SLGc sugars relative to the SL sugars, and, in fact, the former ones have K_Ds above 2 mM by ITC measurements. Noticeably, no STD signal could be observed at all for the monosaccharides, whose binding is estimated above 15 mM by ITC (i.e. beyond the detection limit in terms of kinetics for STD NMR).

From a biological perspective, the role of CBM40 has been proposed to target the enzyme (and hence the bacteria) towards sialic acid rich regions of the GIT (34). Using immunogold labelling and western blotting on mucins of different origin, our

collaborators at the Quadram Bioscience Institute provided further evidences for CBM40 from *R. gnavus* to be considered as a novel bacterial mucus adhesin[97]. Interactions between bacterial adhesins from gut commensals and mucin glycans are generally of low affinity[172,173]. This fits with the findings that CBM40 broadly and weakly recognises potential IT-sialidase substrates, with a slight preference for α2/3 over α2/6 sugars. This specificity differs from other characterised CBM40s showing a net preference for the α2/3 sialoglycans, such as those from *C. perfringens* or *M. decora*[174,167]. It is interesting to pinpoint the lack of recognition of monosaccharides Neu5Ac, Neu5Gc or 2,7-anhydro-Neu5Ac (the product of the enzymatic reaction) by CBM40.

5.4 Conclusions

The work presented in this chapter is a thorough structural investigation on the molecular recognition processes involving the two domains of the intramolecular *trans*-sialidase *Rg*NanH from the gut commensal symbiont *Ruminoccocus gnavus*, *i.e.* the GH33 and CBM40 domains, by STD NMR spectroscopy and computational tools. The specificity and the mechanisms of molecular recognition were investigated with a focus on the features underpinning the ligand specificity of the proteins and their ability to discriminate between α2/3 and α2/6 sialoglycans. Noticeably, the 3D structures in solution for 3'SL and 6'SL bound to GH33 were provided and validated against STD and DEEP-STD NMR data. Our NMR-validated 3'SL/GH33 3D model represents the first Michaelis complex of an IT-sialidase to date.

Arguably, the fact that some residual activity is retained in the inactive mutant makes it particularly difficult to crystallise the sialidases in the presence of the substrate, as in the long times of the crystallisation process the enzymatic reaction will take place to a significant extent, resulting in the complex of the protein with the enzymatic product. This was the case for the sialidases from *V. cholera* (PDB ID: 1W0P[161]) and for the IT-sialidase from *M. decora* (PDB ID: 3SLI[166]), where crystallisation attempts with the enzyme in the presence of the substrate were not successful and only complexes with Neu5Ac or 2,7-anhydro-Neu5Ac are available. The catalytic *trans*-sialidase mutant from *T. cruzi*, on the other hand, was successfully crystallised with both 3'SL (PDB ID: 1S0I[175]) and 2,3-dehoxy-Neu5Ac (DANA) and lactose (PDB ID: 1MS0[176]).

The structural analysis in this chapter informs on how the *Rg*NanH carbohydrate binding and catalytic domains bind to their ligands. For both domains, the Neu5Ac sialic acid moiety constitutes the main recognition element with comparable orientation in α2/3 and α2/6 ligands, although differing in binding affinities and kinetics. The carbohydrate binding module, CBM40, can recognize a wide range of di- and trisaccharides containing sialic acid, with a slight preference for α2/3 ligands. Although both, 3'SL and 6'SL bind to GH33, this enzyme binds to 3'SL with higher affinity than to 6'SL, in agreement with its substrate specificity. The product of the enzymatic reaction, 2,7-anhydro-Neu5Ac, which inhibits the catalytic domain GH33, does not bind to CBM40, nor do the other monosaccharides tested (including Neu5Ac and Neu5Gc).

The STD NMR competition experiments showed that the presence of 6'SL does not interfere with the interaction between 3'SL and GH33, even when they are in equimolar concentrations. Therefore, the potentiality of GH33 to recognise and weakly bind α2/6 sialoglycans is thought not to affect the efficiency of hydrolysis of α2/3 sugars in physiological conditions.

The main information which we can depict from the STD NMR studies on the mechanism of action of *Rg*NanH is:

- CBM40 broadly recognises α2/3- and α2/6-sialoglycans with a slight preference for the former, so it concentrates the enzyme in sialic-acid rich regions (addressing α2/3 rich regions preferentially), accordingly to its adhesin nature;
- GH33 has much higher affinity for 3'SL than for 6'SL, which still can bind transiently (he catalytic site is, in practice, inaccessible to 6'SL when 3'SL is present);
- The higher affinity of 3'SL can be explained by the presence of a π-stacking of the galactose of 3'SL with the tryptophan at position 698 (which is not possible with 6'SL), confirming the significant role of this residue for the *trans*-sialidase activity and the α2/3 specificity;
- 2,7-anhydro-Neu5Ac has higher affinity for GH33 than 3'SL, thus it can inhibit the hydrolysis when enough of the product has been generated.

From the NMR point of view, STD NMR has proven to be an optimal and versatile tool to unveil details of interactions of proteins with weak (*e.g.*, CBM40 with the sialic acid

containing di- and trisaccharides), or relatively strong ligands (*e.g.*, GH33 D282A with 3'SL or 2,7-anhydro-Neu5Ac).

On a parallel route to the biological study, the peculiar STD behaviour of the sialoglycan methyl group interacting with CBM40 inspired us to follow-up this analysis to rationalise this system and paved the way to the development of the DEEP-STD NMR approach. From there, the analysis of GH33 prompted the development of DEEP-STD NMR, given the presence of two opposite "patches" of aromatic and aliphatic residues (as described in Chapter 3, Section 3.3).

In particular, the DEEP-STD fingerprint approach was introduced and implemented to answer fundamental questions regarding the recognition of the non-hydrolysable 6'SL by GH33. DEEP-STD NMR provided information at the atomic detail and allowed to discriminate between the two binding modes, in excellent agreement with the lowest energy solutions obtained by molecular docking of the two ligands.

These results further supported the use of DEEP-STD NMR as a powerful technique to shed light on the specificity of protein-ligand interactions, to drive the generation of computational models and validate them, or precisely orienting unknown ligands in known binding sites.

5.5 Materials and methods

5.5.1 Chemicals and proteins

Neu5Ac, deuterium oxide (99.9% ^{2}H), sodium chloride and tris-(hydroxymethyl-d_3)-amino-d_2-methane (Tris-d_{11}, 98% ^{2}H) were purchased by Sigma Aldrich. 3'SLGc, 6'SLGc, Neu5Ac-STn, Neu5Gc-STn and STF were synthesised by our collaborators following a published methodology[177]. Neu5Gc, 3'SL, 6'SL, 3'SGal, 6'SGal, 3'SLacNAc, 6'SLacNAc, were from Carbosynth. Enzymatically synthesised 2,7-anhydro-Neu5Ac and recombinantly produced CBM40 and GH33 D282A were provided by our collaborators at QIB as described in Sub-section 3.5.1 (GH33 D282A was expressed similarly to GH33).

5.5.2 NMR measurements and processing

All the sugars were assigned on the bases of 1D ^{1}H, 2D ^{1}H,^{1}H-DQF-COSY, ^{1}H,^{1}H-TOCSY, ^{1}H,^{13}C-HSQC and ^{1}H,^{1}H-NOESY experiments ran on samples of the free ligands in

unbuffered D_2O, pH 7.0. For STD NMR experiments, all the samples consisted of 1 mM ligand and 50 µM GH33 D282A or CBM40 in a D_2O pH 7.8 buffer solution containing 10 mM Tris-d_{11} and 100 mM NaCl (ligand : protein ratio 20 : 1). To obtain the binding epitope mappings of the ligands (Sections 5.2 and 5.3), the STD NMR experiments were carried out at different saturation times (0.5, 1, 2, 3, 4 and 5 s) with 1024 scans for CBM40 and 512 scans for GH33. CBM40 experiments were carried out at 288 K, while GH33 D282A experiments were carried out variable temperature as specified in the text. A sequence that includes 2.5 ms trim pulses and 3 ms spoil gradient was used. Saturation was achieved applying a train of 50 ms Gaussian pulses (0.40 mW) on the f2 channel, at 0.60 ppm and/or 6.55 ppm (on-resonance experiments) and 40 ppm (off-resonance experiments). The protein signal was removed using a 40 ms spinlock (T1ρ) filter. All the experiments were recorded at ^1H frequency of 800.23 MHz on a Bruker Avance III spectrometer equipped with 5-mmD probe TXI 800 MHz H-C/N-D-05 Z BTO.

The STD NMR build-up curves were fitted mathematically to a mono-exponential equation, from which the initial slopes were obtained[70c]. The binding epitope was obtained by dividing all the initial slopes by the one of a convenient proton of the sialic acid residue (depending on the ligand), to which an arbitrary value of 100% was assigned.

DEEP-STD factors were calculated using the DEEP-STD equation[87] (Equation 3.5):

$$\Delta\mathrm{DEEP}-\mathrm{STD}_i = \frac{STD\ _{exp1,i}}{STD\ _{exp2,i}} - \frac{1}{n}\sum_i^n \left(\frac{STD\ _{exp1,i}}{STD\ _{exp2,i}}\right).$$

5.5.3 Docking calculations

All molecular modelling was performed with the module Glide within Schrödinger's Maestro software suite, version 11[140,141,76]. Coordinates for the receptor were obtained from the Protein Data Bank (PDB ID: 4X4A[17]). Where necessary, coordinates for missing atoms were added according to known protein chemistry and sidechain protonation was optimised for neutral pH. A short minimisation was run using the OPLS3 force field, converging heavy atoms to a RMSD of 0.3 Å. The receptor grid was then calculated, centring on the centroid of 2,7-anhydro-Neu5Ac and with a length of 30 Å (WT grid). Three dimensional structures of 3'SL and 6'SL were generated using a conformational search, implementing Monte-Carlo torsional sampling, keeping only unique structures

(RMSD > 0.5 A) and eliminating all structures with an energy 21 kJ mol^{-1} greater than the lowest energy structure. All resulting structures were then minimised using conjugate gradient minimisation. For each ligand, the lowest energy conformations were used to initiate a preliminary round of docking. The docking consisted of further conformer generation, docking and then minimisation. Ten conformers were generated with 4x enhanced sampling and during docking, the non-bonded term of the potential energy function was softened for non-polar ligand atoms (charge < |0.15|) by applying a scaling factor of 0.8. Finally, minimisation was performed using implicit solvent with a distance dependent dielectric constant of 4. In a second round of docking, the 5 most diverse poses of the first round of docking were exported and re-docked in the same conditions, generating about 50 poses. To determine if the mutation D282A does not affect the binding mode, the D282 of the minimised receptor was mutated to A282 and a new grid was generated (mutant grid) with the same feature as the first. The 3D builder toolbox was used to mutate the residue. A second round of docking was then repeated with the mutant grid and a similar set of poses was generated (reported in Section A.9 of the Appendix).

5.5.3 CORCEMA-ST predictions

For CORCEMA-ST[66], the cut-off distance around the protons of the ligand in the binding pocket was 10 Å. The concentration of ligand and protein used was respectively 2 mM and 50 µM. k_{on} was set to 1×10^{-8} M^{-1} s^{-1}. The bound ligand correlation time was 50 ns, whereas the free correlation time was 1 ns. The equilibrium constant used was 5000 M^{-1} and 2000 M^{-1} for 3'SL and 6'SL respectively. The non-specific relaxation leakage term (ρ leak) used was 0.4 Hz, to account for the effect of traces of dissolved paramagnetic oxygen, and the irradiation frequency was set to the range 0.1 - 1.1 ppm to simulate the conditions of irradiation at 0.60 ppm. The NOE R-factor was calculated on the averaged simulated data obtained, with the following equation:

$$NOE\ R-Factor = \sqrt{\frac{\sum(STD\ \%_{exp} - STD\ \%_{calc})^2}{\sum(STD\ \%_{exp})^2}}.$$

Chapter 6

General discussion, conclusions and future work

6.1 Beyond the limits of ligand-based NMR approaches: DEEP-STD and IL-STD NMR (Chapter 3)

The core result of this thesis is the implementation of two novel STD NMR approaches: DiffErential Epitope mapping STD NMR (DEEP-STD NMR) and Inter Ligand STD NMR (IL-STD NMR). In our view, these go in the direction of overcoming the limits of ligand-based NMR techniques, which traditionally, and by definition, only allow the ligand to be observed, leaving its surrounding (the protein or ideally other fragments) outside the permitted perspective.

In Chapter 3, we answered some of our initial fundamental questions, demonstrating that:

1. the inhomogeneity of spin diffusion can be exploited to track the different pathways for direct and indirect saturation transfer by differential irradiation STD NMR, getting precious information on the architecture of the binding pocket, through observation of the ligand;
2. the presence of protons from water affects the saturation transfer from polar amino acid side chains containing exchangeable protons, and this is a further source of information allowing to detect the presence of (slow) exchanging polar amino acids lining the binding site;
3. by direct irradiation on the proton frequencies of one of the ligands in a multi-ligand complex, we can observe both *intra*-molecular NOEs across the bound state and *inter*-ligand NOEs on STD NMR experiments, for systems containing two ligands bound to adjacent subsites.

Based on this, we provided two standardised protocols which we named DiffErential EPitope mapping STD NMR (DEEP-STD NMR) and *Inter*-Ligand NMR (IL-NMR) to be applied to a wide range of protein-ligand systems (provided they are suitable for STD NMR observation).

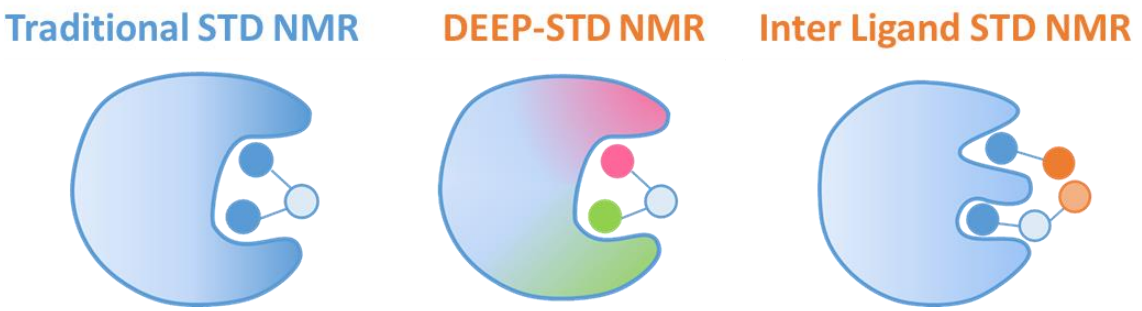

Figure 6.1. Cartoon schematising the information obtained on protein-ligand interactions by traditional STD NMR and by the two approaches proposed by us (DEEP-STD NMR and IL-STD NMR).

6.1.1 Conclusions on DEEP-STD NMR

DEEP-STD NMR gives access to a layer of structural information previously inaccessible: the nature of the amino acid residues surrounding the ligand in the binding site and an assessment of the *architecture of the binding site*. Among the strengths of the approach are the robustness and technical simplicity of its implementation, which relies on differential protein irradiation or differential deuteration of the polar solvent. Given the amphiphilic nature of many protein-ligand interactions, the approach is advantageous both in fragment based drug discovery and in fundamental biological research.

As a proof of concept, we have analysed high resolution crystal structures of published protein-ligand complexes (3NPG/CTB and 2,7-anhydro-Neu5Ac/GH33) to demonstrate that the DEEP-STD data have the potential to provide an extraordinary level of atomic detail, as we reported in Chapter 3, Section 3.3.

Hence, we implemented DEEP-STD NMR on unknown interactions involving the same proteins (CTB and GH33) in complex with, respectively:

i) a library of small thio-galactoside-furoate inhibitors (Chapter 4, Section 4.3);

ii) the natural sialoglycans 3'SL and 6'SL (Chapter 5, Section 5.2).

In both cases, we have shown that, when the 3D structure of the protein is known, the protocol allows, for the first time, information to be gained from STD NMR about the orientation of the ligand relative to the key residues of the receptor. We have also proved that we can use the ΔDEEP-STD patterns of the known complexes as finger-prints to locate and orient unknown ligands in known binding sites. In both studies i) and ii),

the DEEP-STD data was in very good agreement with the lowest energy docking solutions generated, aiding the validation of the quality of the computational models.

During the course of these studies, we have shown that the technique is versatile enough to be developed further and modified. In Chapter 4, Section 4.3, we expanded the scenario from two to many irradiation frequencies, allowing changes in the STD NMR intensities of each proton at variable saturation frequencies (arguably corresponding to single protein residues in a known receptor) to be mapped, in what we called the "on-resonance scanning" approach.

For receptors of unknown 3D structure, DEEP-STD NMR can help evaluating the architecture of the binding site, orienting, for instance, aromatic and aliphatic patches in the pocket relative to each other. This can be a valuable aid to identify unknown binding sites in receptors of known 3D structure, based on the response of known ligands to differential irradiation and solvent conditions.

6.1.2 Conclusions on IL-STD NMR

IL-STD NMR relies on a similar concept to DEEP-STD NMR: it exploits differential irradiation conditions to elucidate the "environment of binding", detecting secondary ligands surrounding the ligand of interest, and their orientation relative to it. In Chapter 4, Section 4.2 we have shown that, in a ternary complex composed of one receptor and two or more ligands, changing irradiation frequencies from "on-resonance" of one of the ligands to "off-resonance" (protein irradiation only) gives information on the proximity of the ligands and of their binding subsites. This also allows to define the "contact regions" between the fragments, that is, their relative orientation. It is important to remark that this is the same information which we can obtain by tr-NOESY following the ILOE approach[84], without the need of selecting appropriately the irradiation frequency, as we have to do in IL-STD. On the other hand, ideally, IL-STD can give the same information in a small fraction of the experimental time.

6.1.3 Perspectives and future work

Potentially, the combination of DEEP-STD and IL-STD NMR can be very fruitful in fragment based drug design contexts, allowing the ligands to be oriented in their own binding subsites and relative to each other.

We have demonstrated the potential of these approaches for relatively small ligands, which are more difficult to orient than larger ones. DEEP-STD NMR combined with the 3D structure of the protein is expected to define the orientation of larger ligands in the bound state even more easily, a feature which could turn out really useful in structural biological studies (the DEEP-STD NMR study on the 3'SL and 6'SL in complex with GH33, reported in Chapter 5, Section 5.2, is already a good example of successful analysis of larger ligands).

Currently, the binding of synthetic and natural fragments to an important ubiquitin-protein ligase receptor is being investigated by us. The 3D structure of the receptor has been solved by X-ray crystallography, but the binding site is not known as of yet, making this a very challenging problem to approach by the combination of DEEP-STD NMR and computational tools.

Further works undergoing in our group show very promising results in the application of DEEP-STD NMR to larger protein receptors (up until ~140 kDa). The magnitude of the ΔDEEP-STD (or ΔIL-STD) is smaller, but the information is still very clear to interpret.

In fact, the potential of this technique is not limited by the size nor the nature of the receptors. Projects currently under investigation in our laboratory show that DEEP-STD NMR can successfully be applied to lipopolysaccharides and even to block co-polymer micelles, moving out of the biological realm. The compartmented nature of these macromolecules (a large aliphatic environment adjacent to a large carbohydrate environment in the case of lipopolysaccharide; and the large blocks of different monomers assembled in an ordered way in block co-polymer micelles) makes them a perfect target for this technique.

The chance to apply DEEP-STD NMR to a wide number of diverse systems (and with promising results so far) is very important to strengthen the potential of the technique and our future perspectives include further methodological investigation on the effect of the modification of different parameters (including temperature, irradiation power, ligand concentration, *etc.*) on the DEEP-STD outcome.

So far, the significance and the crucial importance of this methodological work is that by exploiting intrinsic inhomogeneity in spin diffusion, we can differentiate between

different environments surrounding the ligands under observation. This is done by simple means of one-dimensional NMR spectra, and without requiring isotopic labelling (in contrast to, for example, SOS-NMR[178]) nor protein assignment (in contrast to receptor-based techniques). On these premises, we envision that DEEP-STD NMR, and in a smaller measure IL-STD NMR, will become popular approaches to characterise a wide range of ligand-receptor interactions, in pharmaceutical as well as biological research labs.

6.2 New avenues for cholera toxin inhibition (Chapter 4)

We started to investigate the field of CTB inhibitor design with the aim to confirm (or not) the binding mode of a polyhydroxyalkylfuroate thio-galactose ligand (**Ligand 30**) received from our collaborators in the research group of Inmaculada Robina, from the University of Seville. **Ligand 30** was qualitatively proposed to bind to the two well-known galactose and sialic acid subsites in the GM1 binding pocket of cholera toxin subunit[92], but the outcome of the study has been much more resounding than that.

In Section 4.2, we demonstrated the existence of a hitherto-unknown binding subsite in the GM1 binding pocket of CTB, by a combination of several NMR approaches, molecular docking and traditional and non-classical molecular dynamics (Figure 6.2). By means of CORCEMA-ST, we could successfully validate the quality of our 3D models of **Ligand 20**, **Ligand 30** and **Ligand 33** in solution and show the specificity of the novel binding subsite for the polyhydroxyalkylfurate-aromatic moiety of **Ligand 30** and **Ligand 33**.

The potential of the discovery is conspicuous, for two main reasons:

- the polyhydroxyalkylfuroate ligands are very drug-like, with a reduced carbohydrate nature (limited to the presence of non-hydrolisable thio-galactose moiety to anchor the ligands to the galactose binding subsite);
- the presence of an additional non-carbohydrate specific binding subsite paves the way for the design of three-finger ligands, which could occupy the three subsites at the same time, arguably increasing the affinity.

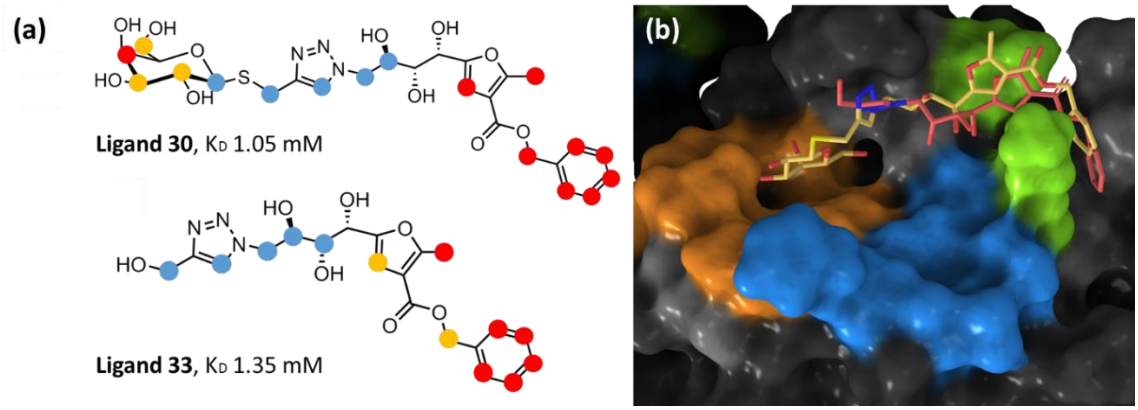

Figure 6.2. (a) The STD binding epitope mappings of **Ligand 30** and **33** in complex with CTB[92]. The STD values from the initial slopes of each proton relative to the most intense one are ranked according to their intensity: blue dots for weak contacts (45%-64%), yellow dots for medium contacts (65%-84%) and red dots for strong contacts (85%-100%). The K_D values for the interactions are reported below the structures. **(b)** Representative frames of the most populated clusters of **Ligand 30** (pale yellow) and **Ligand 33** (light red) with CTB in the "open" conformation. The residues forming the galactose binding subsite are shown as orange surface. The residues constituting the sialic acid binding subsite are shown as blue surface. Ile58 and Lys34, which constitute the novel binding subsite, are shown as green surface and labelled.

To further investigate the CTB binding pockets, in Section 4.3, we have investigated a small library of short analogues of **Ligand 30**, in which the thio-galactose is directly linked to the furoate-aromatic moiety, by means of different thio-glycosidic linkages (varying in anomeric configuration and oxidation state of the sulphur, in Figure 6.3a). This part of the study helped to discern and rank the specificities of the two moieties of the ligands (thio-galactose and furoate-aromatic), as the short size does not allow both the galactose and the novel subsite to be occupied, as for **Ligand 30**.

The main result was that the specificity of the thio-galactose residue for the galactose subsite drives the interaction, anchoring the ligands in the galactose subsite. Being too small to reach the novel binding subsite, the furoate-aromatic moiety is directed towards the sialic acid subsite, where it forms a π-stacking with Tyr12 of CTB (Figure 6.3b). The presence of the π-stacking for all of the four ligands, despite the nature of the thio-glycosidic bond, suggests that the interaction is particularly favourable and the sialic acid subsite has good selectivity for aromatic moieties, as previously proposed[128].

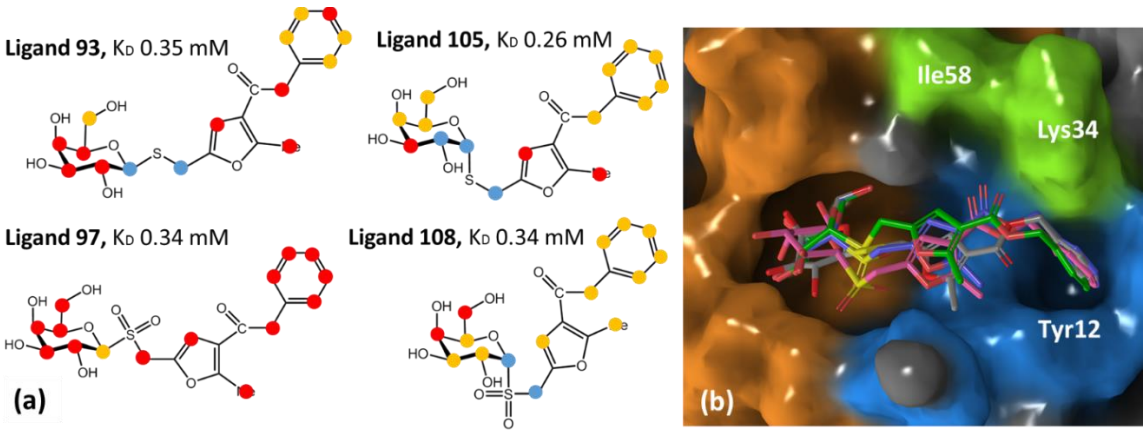

Figure 6.3. (a) The STD binding epitope mappings of the small analogues of **Ligand 30** in complex with CTB. The STD values from the initial slopes of each protons relative to the most intense one are ranked according to their intensity, following the same legend as in Figure 6.2. The K_D for the interactions is reported above the structures. **(b)** Lowest energy docking solutions for **Ligand 93** (grey), **Ligand 97** (pink), **Ligand 105** (green) and **Ligand 108** (violet) with CTB in the "closed" conformation. The surface colouring is the same as in Figure 6.2.

The two main pieces of structural information that we obtained investigating **Ligand 30**, **Ligand 33** and the small analogues of **Ligand 30** are:

- the proof that the polyhydroxyalkylfuroate moiety fits tightly in the novel binding subsite (provided a long enough linker);
- the hint that the sialic acid subsite can be specifically occupied by an aromatic moiety π-stacking with the underlying Tyr12.

This gave us ground to propose a novel three-finger scaffold for CTB inhibition, shown in Figure 6.4. Albeit the sulphur has been shown to be unideal for decoration, based on the docking poses, the furoate can be an optimal candidate for adding an additional fragment.

IL-STD NMR could be a valuable aid to assess the relative orientation of the fragments, directing the drug design process in a knowledge-based fashion. Once a novel library of three-finger ligands is synthetised by our collaborators at the University of Seville, combination of NMR and computational tools would again help elucidate the 3D structure, hopefully directing towards a high affinity and highly selective lead to progress towards an efficient CTB inhibitor.

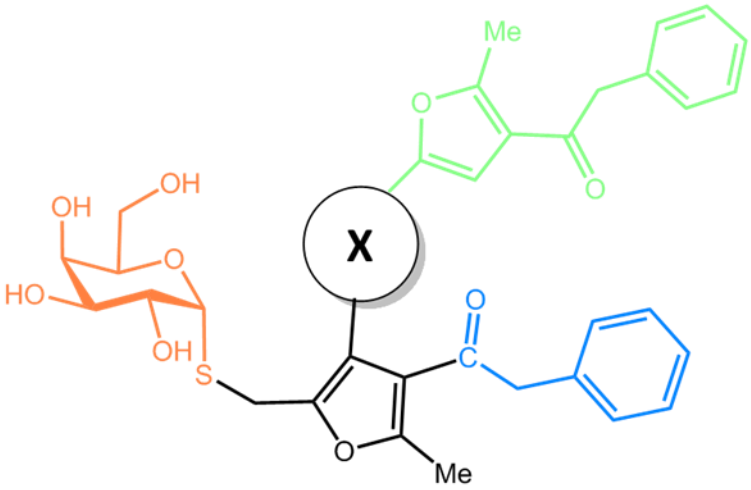

Figure 6.4. Future work: schematic representation of a novel class of three-finger non-carbohydrate CTB inhibitors proposed by us, potentially able to occupy the galactose, sialic acid and novel binding subsites simultaneously, through its orange, blue and green fragments, respectively. The length and the flexibility of the X linker, should be investigated.

6.3 Towards the structural understanding of intramolecular *trans*-sialidases from gut microbiota (Chapter 5)

In collaboration with the research group of Nathalie Juge from the Quadram Institute Bioscience, we undertook a thorough study of the two subdomains of the intramolecular *trans*-sialidase *Rg*NanH, CBM40 and GH33 D282A, combining STD NMR and computational tools with the biological assays, glycan microarray and ITC data performed by our collaborators. We investigated the specificity and the mechanism of molecular recognition, with a stress on the ability of discriminating between α2/3 and α2/6 sialoglycans.

By molecular docking, we provided the 3D molecular models of 3'SL and 6'SL as bound to GH33, which we validated against experimental STD and DEEP-STD data. This makes our 3D model of the 3'SL/GH33 the first Michaelis complex of an IT-sialidase to date (as the residual activity of the inactive mutants of sialidases makes crystallising the receptor-substrate complex challenging).

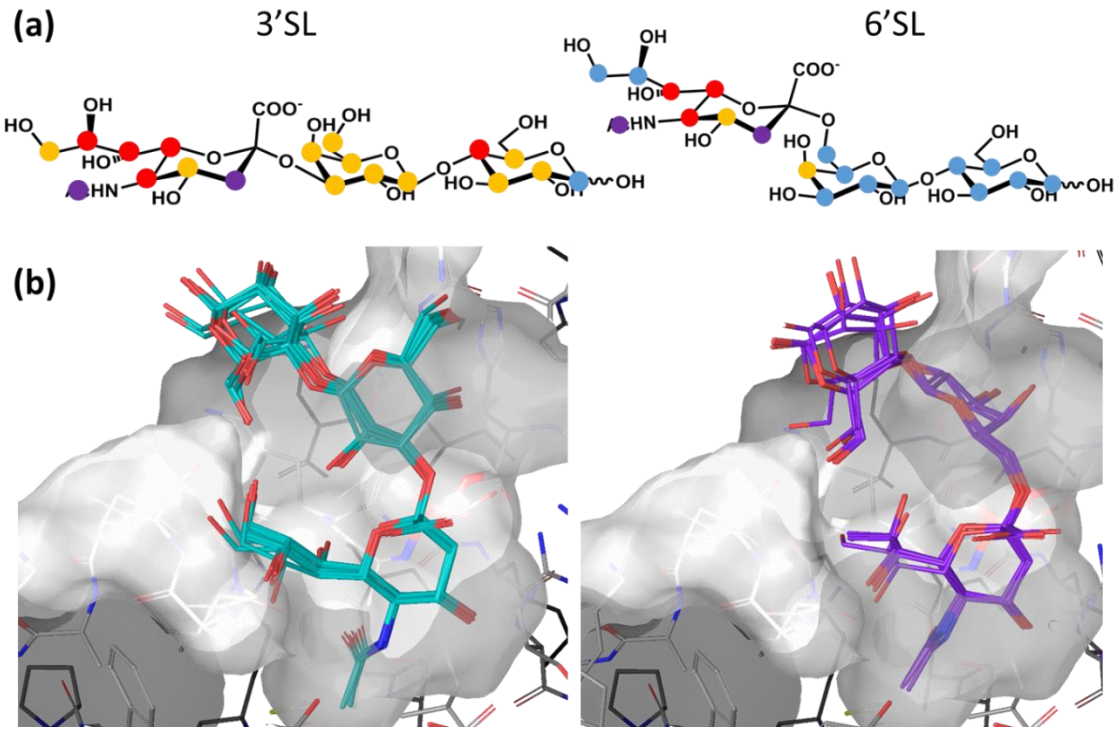

Figure 6.5. (a) The STD binding epitope mappings of 3'SL and 6'SL in complex with GH33, at 0.60 ppm. The STD values from the initial slopes of each protons relative to the most intense one are ranked according to their intensity: blue dots for weak contacts (50%-70%), yellow dots for medium contacts (71%-90%), red dots for strong contacts (91%-100%) and purple dots for very strong contacts (>100%). **(b)** The lowest energy converging docking solutions for GH33 in complex with 3'SL, 10 poses, and 6'SL, 6 poses.

For both subdomains, the sialic acid moiety is the main recognition element for both α2/3 and α2/6 sialoglycans. This seems to be accommodated in both binding sites with comparable orientation, despite the fact that the difference in the configuration of the glycosylic linkages results in different affinities and kinetics of binding for the two ligands. Expectedly, the catalytic domain GH33 shows higher affinity to its natural substrate 3'SL, than to 6'SL, to whom it still transiently binds. The "recruiting" carbohydrate binding module, CBM40, can recognize a wide range of di- and trisaccharides containing sialic acid (for all of which the binding epitope has been determined), although a slight preference for the α2/3 ligands is observed. 2,7-anhydro-Neu5Ac, which inhibits the catalytic domain, does not bind to CBM40, nor do the other monosaccharides (Neu5Ac and Neu5Gc).

This revealed the following structural aspects that can be correlated to the mechanism of action of *Rg*NanH:

- CBM40 broadly recognises 3'- and 6'-sialoglycans with a slight preference for the former. This allow to concentrate the enzyme in sialic acid-rich regions of the mucin niches (addressing α2/3 rich regions preferentially), accordingly to its adhesin nature;
- GH33 has much higher affinity for 3'SL than for 6'SL, which can still bind transiently;
- the higher affinity of 3'SL can be explained by the presence of a π-stacking of the galactose of 3'SL with Trp698 (which is not possible in the case of 6'SL), confirming the significant role of this residue for the *trans*-sialidase activity and the α2/3 selectivity;
- the catalytic site is, in practice, inaccessible to 6'SL when 3'SL is present;
- 2,7-anhydro-Neu5Ac has higher affinity for GH33 than 3'SL, thus it can inhibit the hydrolysis when enough of the product has been generated.

With the 3D model of the substrate as a starting point, the future work aims at elucidating the mechanism through which GH33 carries out the enzymatic IT-sialidase reaction from a molecular point of view. GH33 has been proved to be an exceptional candidate for the application of DEEP-STD NMR, given its amphiphilic nature combining aliphatic and aromatic patches with a highly polar cluster of arginine. We are confident that performing mutagenesis on key residues thought to be responsible for the enzymatic behaviour of GH33, and analysing the GH33 mutants by combination of STD and DEEP-STD NMR, as well as computational tools, will give further insights in the understanding of a fascinating, as much as an unusual, reaction.

From our point of view, providing the 3D structure of the 3'SL/GH33 complex in solution is a significant achievement. Hence, unravelling the actual mechanism of reaction on a solid experimental basis could be a milestone in the investigation of mucin degrader micro-organisms and a substantial step forward in the design of sialidase inhibitors, a class of drug with a huge potential.

Appendix

A.1 Full spectra of the STD NMR competition experiments a) and b)

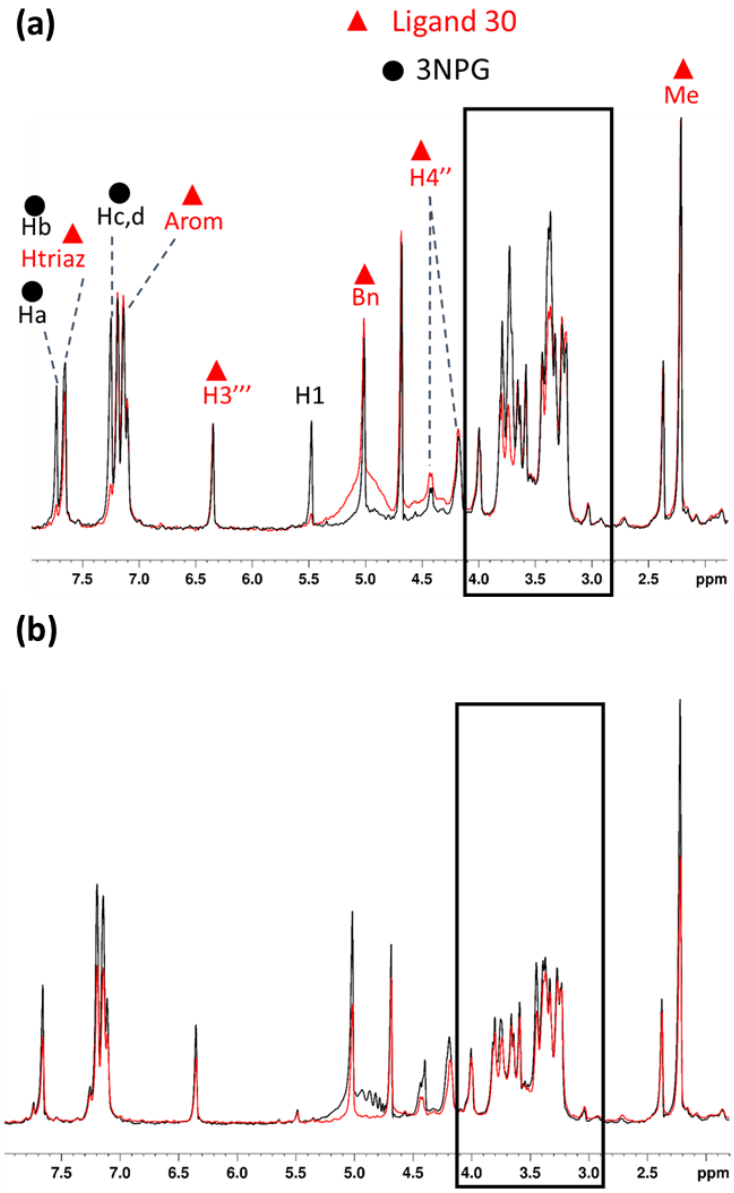

Figure A.1. Appendix to Figure 4.11a,b in Sub-section 4.2.2 Full width reference spectra of the STD competition experiments of **Ligand 30**/CTB/3NPG **(a)** and **Ligand 30**/CTB/**Ligand 33 (b)**. In red the spectra of the first binder alone and in black the spectra upon addition of the competitor. The assignment of 3NPG and Ligand 30 is given in **(a)** (except for the 3.0 ppm to 4.0 ppm region squared in black, assigned in the Figure 4.11a,b). The peaks of **Ligand 33** perfectly overlap to those of **Ligand 30**, therefore in **(b)** the assignment is not repeated.

A.2 STD NMR experiment on the complex 3'SL/CTB

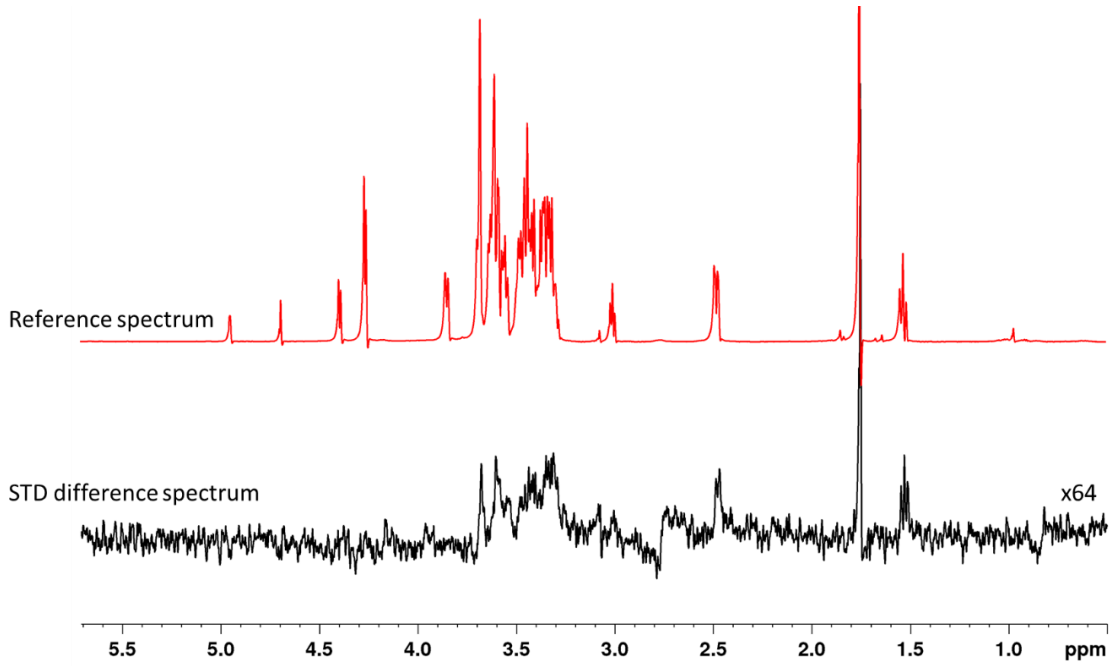

Figure A.2. Appendix to Figure 4.11d in Sub-section 4.2.2. 25 μM CTB in the presence of 1 mM of 3'SL, 278 K. The reference and the difference spectra are shown and labelled, the magnification of the difference spectra relative to the reference spectra is reported. Spectra are acquired at 2 s saturation time and 0.0 ppm saturation time. For the full assignment of 3'SL, see Table A.16.

A.3 Tr-NOESY control spectra (ILOE approach)

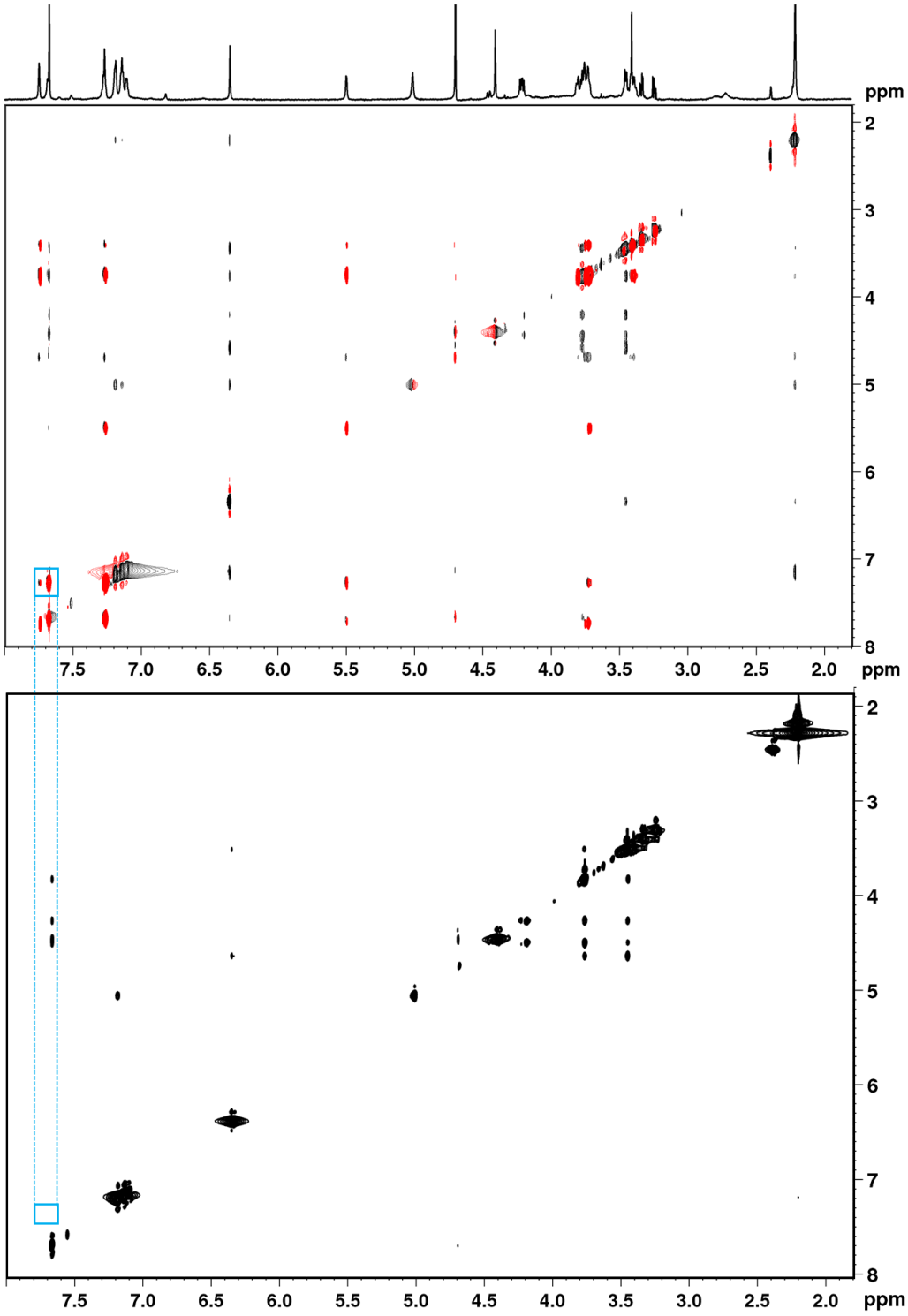

Figure A.3. Appendix to Figure 4.12 in Sub-section 4.2.3. Full width tr-NOESY experiments (Mixing time 1.2 s). Top, **Ligand 33**/CTB/3NPG (ternary complex, in black) and 3NPG/CTB (binary complex, in red), as in Figure 4.12. Bottom, **Ligand 33**/CTB (binary complex): the spectral area containing the inter-ligand signal is empty when the protein is alone, confirming that this is due to the ligands.

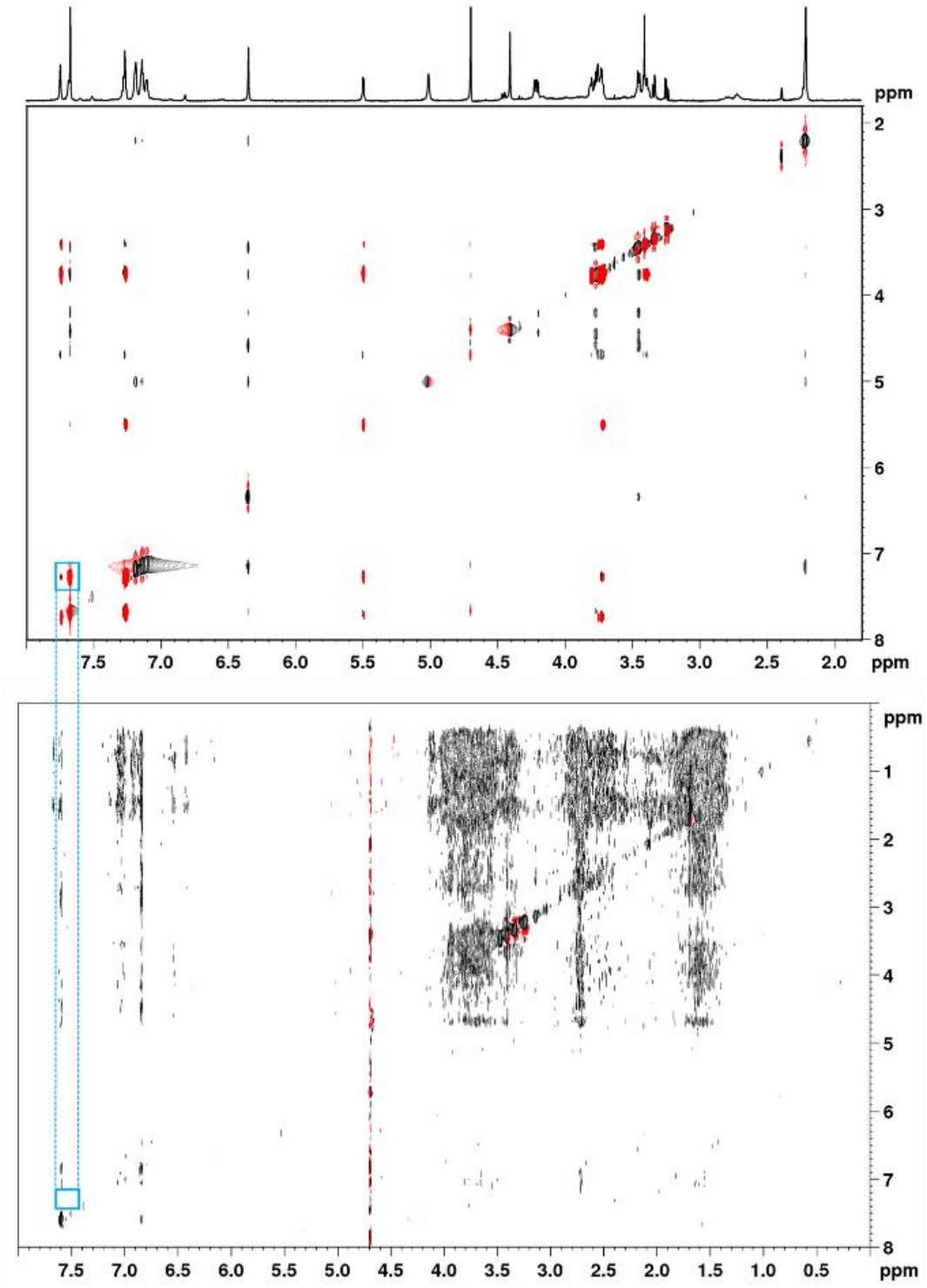

Figure A.4. Appendix to Figure 4.12 in Sub-section 4.2.3. Full width tr-NOESY experiments (Mixing time 1.2 s). Top, **Ligand 33**/CTB/3NPG (ternary complex, in black) and 3NPG-CTB (binary complex, in red), as in Figure 4.12. Bottom, CTB alone: the spectral area containing the inter-ligand signal is empty when the protein is alone, confirming that this is due to the ligands.

A.4 Assignment and STD NMR binding epitope mapping raw data for the *SCJ set* of ligands in complex with CTB

			STD (%) Ligand 93							
Proton ID	**^{1}H δ (ppm)**	**^{13}C δ (ppm)**	**0.25 s**	**0.5 s**	**0.75 s**	**1 s**	**2 s**	**3 s**	**4 s**	**5 s**
CH_3(fur)	2.22	13.3	2.41	5.51	8.39	10.5	16.6	19.6	21.19	21.82
H2	3.248	68.68	1.92	3.92	7.32	8.98	14.63	18.13	19.22	19.6
H3	3.245	73.45	1.92	3.92	7.32	8.98	14.63	18.13	19.22	19.6
H4	3.277	78.36	1.69	4.54	7.46	8.39	13.4	15.36	16.13	16.77
H6/6'	3.319	60.27	1.71	4.49	6.9	8.39	11.57	13.01	14.07	14.07
CH_2(S)	3.554	25.7	1.83	3.92	8.07	6.08	8.64	10.5	10.91	11.92
H5	3.617	68.3	1.41	4.67	8.07	9.52	15.06	17.44	17.78	19.79
CH_2(S)	3.693	25.7	1.92	2.95	6.64	6.57	8.47	10.5	10.91	11.92
H1	4.039	84.53	1.61	2.41	4.9	6.14	10.81	13.14	14.2	14.91
CH_2(Bn)	5.01	66.01	2.29	4.49	6.44	8.81	12.39	13.27	14.48	13.93
H(fur)	6.279	108.1	1.71	4.54	7.32	9.34	17.1	21.19	22.91	23.83
H para	7.1	127.9	2.14	4.49	7.83	9.8	16.44	20.99	22.47	23.83
H meta	7.14	128.3	1.83	4.36	7.17	9.34	15.81	20.38	22.04	23.83
H orto	7.187	127.9	2.1	4.11	6.51	8.89	15.36	19.22	21.4	22.04

Table A.1. Appendix to Figure 4.28 in Sub-section 4.3.2. Ligand 93. ^{1}H and ^{13}C assignment and the raw STD (%) at increasing saturation times. Irradiation frequency 0.0 ppm, 278 K.

			STD (%) Ligand 97							
Proton ID	**^{1}H δ (ppm)**	**^{13}C δ (ppm)**	**0.25 s**	**0.5 s**	**0.75 s**	**1 s**	**2 s**	**3 s**	**4 s**	**5 s**
CH_3(fur)	2.23	13	6.44	14.48	20.58	25.76	36.6	40.75	42.37	42.79
H2	3.408	72.2	5.96	12.51	18.31	23.36	31.31	35.2	36.25	36.25
H3	3.408	72.2	5.96	12.51	18.31	23.36	31.31	35.2	36.25	36.25
H6	3.446	60.8	6.14	12.63	18.67	22.25	28.12	31.71	31.93	31.93
H6'	3.531	80.1	6.08	12.51	19.22	21.61	28.12	29.53	30.71	30.11
CH_2(S)	3.531	60.8	6.38	13.15	18.13	21.81	27.58	29.81	30.41	30.41
H5	3.692	68.1	6.26	12.51	19.79	23.59	32.88	36.25	34.18	36.6
H4	3.82	64.9	5.04	13.14	18.85	23.59	38.43	46.26	44.06	46.71
H1	4.236	88.3	3.84	9.9	14.91	19.6	29.82	33.85	35.5	35.5
CH_2(Bn)	4.984	66.3	6.2	13.14	17.95	22.35	29.53	31.62	32.24	33.52
H(fur)	6.607	112.9	5.79	12.63	19.6	25.51	41.15	48.1	51	52.2
H para	7.085	128.15	5.9	13.14	20.58	25.76	39.57	45.37	47.17	48.57
H meta	7.119	128.4	5.73	13.66	20.58	25.76	39.57	45.37	47.17	48.57
H orto	7.168	127.8	5.62	12.88	19.79	24.74	38.43	44.06	46.26	47.63

Table A.2. Appendix to Figure 4.28 in Sub-section 4.3.2. Ligand 97. ^{1}H and ^{13}C assignment and the raw STD (%) at increasing saturation times. Irradiation frequency 0.0 ppm, 278 K.

			STD (%) Ligand 98							
Proton ID	**^{1}H δ (ppm)**	**^{13}C δ (ppm)**	**0.25 s**	**0.5 s**	**0.75 s**	**1 s**	**2 s**	**3 s**	**4 s**	**5 s**
CH_3(fur)	2.24	12.9	4.4	9.52	13.53	17.27	25.76	29.24	30.41	31.31
H6	3.44	60.83	3.04	7.53	11.35	14.07	20.18	21.61	22.47	22.25
H6'	3.541	60.83	3.32	6.57	9.61	13.14	16.13	16.77	17.61	17.95
H5	3.545	80.2	3.32	6.57	9.61	13.14	16.13	16.77	17.61	17.95
H2	3.682	66.27	3.01	6.57	10	13.66	20.99	23.59	26.27	26.27
H3	3.682	64.4	3.01	6.57	10	13.66	20.99	23.59	26.27	26.27
H4	3.829	87.9	4.11	7.91	14.77	17.27	25.26	28.12	29.1	29.82
CH_2(S)	4.14	44.09	3.13	6.2	6.9	8.98	13.93	14.34	15.36	15.36
H1	4.19	90.7	2.58	5.79	8.14	10	15.21	17.61	17.78	17.78
CH_2(S)	4.21	44.09	2.58	5.79	8.14	10	15.21	17.61	17.78	17.78
CH_2(Bn)	5.105	66.3	3.8	7.32	10.7	13.53	18.67	19.41	20.99	20.58
H(fur)	6.51	112.2	3.96	8.31	13.1	16.28	27.05	31.93	34.18	34.86
H para	7.103	128.13	3.99	7.83	11.69	15.06	25.76	31.01	32.88	34.18
H meta	7.14	128.39	3.38	7.53	11.69	15.51	25.76	31.01	32.88	34.18
H orto	7.189	127.83	3.32	7.32	11.24	14.48	24.29	29.53	31.31	32.24

Table A.3. Appendix to Figure 4.28 in Sub-section 4.3.2. Ligand 98. ^{1}H and ^{13}C assignment and the raw STD (%) at increasing saturation times. Irradiation frequency 0.0 ppm, 278 K.

			STD (%) Ligand 105							
Proton ID	**^{1}H δ (ppm)**	**^{13}C δ (ppm)**	**0.25 s**	**0.5 s**	**0.75 s**	**1 s**	**2 s**	**3 s**	**4 s**	**5 s**
CH_3(fur)	2.217	13	5.3	11.02	15.97	20.99	31.01	34.52	35.89	35.2
H6	3.379	60.5	5.05	8.39	14.63	16.28	19.79	20.69	21.4	21.4
H3	3.394	69.9	5.05	8.39	14.63	16.28	19.79	20.69	21.4	21.4
H6	3.417	60.5	3.77	7.68	12.63	15.06	21.19	21.19	21.4	21.82
CH_2(S)	3.433	24.3	3.77	7.68	12.63	15.06	21.19	21.19	21.4	21.82
CH_2(S)	3.526	24.3	4.19	5.41	9.25	11.24	15.97	16.77	17.2	17.95
H5	3.654	68.6	4.54	8.07	13.53	17.27	24.77	27.58	28.96	28.96
H2	3.767	67.3	4.32	8.07	12.15	13.66	25.26	31.31	30.41	32.28
H4	3.925	71.2	5.84	8.31	13.14	18.13	26.01	28.12	29.4	30.11
CH_2(Bn)	5.006	66.2	3.45	8.31	10.7	13.66	22.04	23.14	24.29	23.59
H1	5.09	83.9	3.66	7.56	9.9	13.66	22.47	26.52	26.52	27.85
H(fur)	6.27	107.2	3.99	9.25	15.06	18.85	29.82	35.55	37.32	37.69
H para	7.107	128.14	2.87	7.39	11.57	14.48	23.83	30.41	30.71	33.2
H meta	7.143	128.32	2.76	7.24	12.03	14.91	24.53	30.41	32.88	33.85
H orto	7.182	128.32	3.07	7.39	11.69	14.2	24.77	30.41	32.56	33.2

Table A.4. Appendix to Figure 4.28 in Sub-section 4.3.2. Ligand 105. ^{1}H and ^{13}C assignment and the raw STD (%) at increasing saturation times. Irradiation frequency 0.0 ppm, 278 K.

			STD (%) Ligand 108							
Proton ID	**^{1}H δ (ppm)**	**^{13}C δ (ppm)**	**0.25 s**	**0.5 s**	**0.75 s**	**1 s**	**2 s**	**3 s**	**4 s**	**5 s**
CH_3(fur)	2.23	13.2	2.81	6.2	9.52	12.3	19.41	22.69	24.53	25.02
H6	3.408	60.6	4.15	7.83	10.5	14.2	18.67	17.27	18.85	18.85
H6'	3.492	60.6	4.49	8.89	12.27	15.06	19.22	18.13	19.99	19.99
H4	3.574	77.9	3.59	7.53	11.13	13.27	18.85	21.19	21.19	21.19
H5	3.764	67.6	2.48	6.44	9.71	13.27	20.58	23.83	24.77	26.01
H3	3.764	70.3	2.48	6.44	9.71	13.27	20.58	23.82	24.77	26.01
CH_2(S)	4.12	44.6	2.68	4.36	7.1	8.14	12.15	13.86	13.4	13.52
H2	4.064	67.3	3.07	5.31	10.5	11.27	19.99	23.83	24.77	25.01
CH_2(S)	4.126	44.6	2.68	4.4	7.1	8.39	11.92	12.35	13.01	11.35
H1	4.468	92.3	3.35	2.89	6.77	9.71	15.21	17.61	19.22	19.41
CH_2(Bn)	5.006	66.3	2.12	5.68	7.99	10.09	14.77	16.44	16.6	17.1
H(fur)	6.528	112.67	2.45	6.64	9.8	13.14	22.91	28.4	30.71	31.93
H para	7.104	128.1	2.68	6.2	9.43	12.39	21.4	26.27	28.68	30.41
H meta	7.14	128.37	2.68	5.9	9.16	12.03	20.78	26.27	28.68	30.41
H orto	7.195	127.7	2.38	5.57	8.81	11.46	19.6	24.53	27.05	28.4

Table A.5. Appendix to Figure 4.28 in Sub-section 4.3.2. Ligand 108. ^{1}H and ^{13}C assignment and the raw STD (%) at increasing saturation times. Irradiation frequency 0.0 ppm, 278 K.

			STD (%) Ligand 109							
Proton ID	**^{1}H δ (ppm)**	**^{13}C δ (ppm)**	**0.25 s**	**0.5 s**	**0.75 s**	**1 s**	**2 s**	**3 s**	**4 s**	**5 s**
CH_3(fur)	2.2	12.8	3.99	6.08	7.68	12.39	14.77	16.44	17.1	17.27
H4	3.37	74.9	1.96	2.65	3.62	5.3	5.79	7.04	7.24	6.83
H6/6'	3.422	60.8	3.59	4.32	5.57	7.99	8.64	9.25	9.34	8.89
H5	3.735	67.9	3.73	5.73	7.1	12.51	13.93	13.53	15.06	13.93
H3	3.945	76.7	3.29	4.49	4.76	10.81	10.7	11.8	13.01	11.8
H2	4.05	66.6	3.79	4.28	4.76	10.19	10.5	14.2	14.07	14.34
CH_2(S)	4.309	51.5	0.75	1.45	1.85	2.6	3.22	4.49	4.11	4.95
CH_2(S)	4.43	51.5	-	-	-	-	-	-	-	-
CH_2(Bn)	4.97	66.4	3.29	4.49	6.2	9.43	10.91	11.46	11.92	11.24
H1	4.97	89.2	3.29	4.49	6.2	9.43	10.91	11.46	11.92	11.24
H(fur)	6.53	113.27	4.28	6.2	8.31	15.66	20.39	23.14	23.83	23.59
H para	7.066	128.11	3.32	4.76	6.83	11.46	15.06	17.1	17.95	17.44
H meta	7.098	128.55	3.45	5.41	7.39	12.51	15.66	17.95	18.85	19.41
H orto	7.14	127.7	3.1	5.1	6.7	12.03	15.66	17.1	18.13	18.31

Table A.6. Appendix to Figure 4.28 in Sub-section 4.3.2. Ligand 109. ^{1}H and ^{13}C assignment and the raw STD (%) at increasing saturation times. Irradiation frequency 0.0 ppm, 278 K.

	Normalised initial slopes %					
	Ligand 93	**Ligand 97**	**Ligand 98**	**Ligand 105**	**Ligand 108**	**Ligand 109**
H1	57	71	60	62	54	76
H2	85	90	72	67	73	67
H3	85	90	72	91	79	73
H4	85	90	100	84	90	41
H5	90	96	77	80	79	94
H6	83	94	83	91	100	76
H6'		96	77	78		
CH_2(S)	59	96	59	78	56	17
CH_2(S)	60		60	56	62	
H(fur)	90	91	91	86	77	100
CH_3 (fur)	100	100	99	100	73	89
CH_2(Bn)	87	93	82	70	65	76
H orto	81	91	79	66	65	78
H meta	84	95	84	66	71	81
H para	90	95	84	65	73	76

Table A.7. Appendix to Figure 4.28 in Sub-section 4.3.2. STD NMR binding epitope mapping of the *SCJ set* of ligands. Normalised initial slope values for the six *SCJ* ligands, obtained from the fitting of the raw data in the Tables A.1 to A.6. For each ligand, the epitope is normalised against the proton with the strongest initial slope, arbitrarily.

A.5 DEEP-STD finger-print raw data for the *SCJ set* of ligands in complex with CTB

Proton	STD % 0.60 ppm	STD % 2.25 ppm	Ratio STD 2.25 ppm/0.60 ppm	ΔDEEP-STD
H1	13.79	19.99	1.45	0.05
H2,H3	17.1	37.69	2.20	0.80
H4	16.44	31.93	1.94	0.54
H5	17.1	41.55	2.43	1.03
H6/6'	15.36	29.24	1.90	0.50
CH_2(S)	14.07	15.66	1.11	-0.29
CH2(S)	14.07	15.66	1.11	-0.29
H(fur)	22.25	22.47	1.01	-0.39
CH_2(Bn)	16.6	16.44	0.99	-0.41
H para	23.36	19.99	0.86	-0.55
H meta	22.25	19.99	0.90	-0.51
H orto	21.19	19.99	0.94	-0.46
			STD average	
			1.40	

Table A.8. Appendix to Figure 4.29 in Sub-section 4.3.2. Ligand 93. DEEP-STD NMR using Differential Irradiation (2.25 ppm/0.60 ppm). Protons assignment is given in Table A.1.

Proton	STD % 0.60 ppm	STD % 2.25 ppm	Ratio STD 2.25 ppm/0.60 ppm	ΔDEEP-STD
H1	29.53	35.89	1.22	-0.05
H2,3	30.41	54.61	1.80	0.53
H4	36.96	45.81	1.24	-0.03
H5	33.52	53.55	1.60	0.33
H6	28.68	43.21	1.51	0.24
H6'	26.27	41.21	1.57	0.30
CH_2(S)	28.12	41.96	1.49	0.23
H(fur)	41.55	42.37	1.02	-0.25
CH_2(Bn)	30.71	28.68	0.93	-0.33
H para	40.53	38.06	0.94	-0.33
H meta	40.53	38.06	0.94	-0.33
H orto	40.53	38.06	0.94	-0.33
			STD average	
			1.27	

Table A.9. Appendix to Figure 4.29 in Sub-section 4.3.2. Ligand 97. DEEP-STD NMR using Differential Irradiation (2.25 ppm/0.60 ppm). Protons assignment is given in Table A.2.

Proton	STD % 0.60 ppm	STD % 2.25 ppm	Ratio STD 2.25 ppm/0.60 ppm	ΔDEEP-STD
H1	26.52	30.41	1.15	-0.27
H2	26.52	38.81	1.46	0.05
H3	24.53	46.26	1.89	0.47
H4	27.58	63.83	2.31	0.90
H5	26.52	55.14	2.08	0.67
H6	21.26	33.2	1.56	0.15
H6	20.99	42.37	2.02	0.61
CH_2(S)	15.21	19.6	1.29	-0.12
CH_2(S)	14.63	17.6	1.20	-0.21
H(fur)	34.18	34.52	1.01	-0.40
CH_2(Bn)	23.59	24.06	1.02	-0.39
H para	21.4	19.99	0.93	-0.48
H meta	27.05	25.02	0.92	-0.49
H orto	27.05	25.02	0.92	-0.49
			STD average	
			1.41	

Table A.10. Appendix to Figure 4.29 in Sub-section 4.3.2. Ligand 105. DEEP-STD NMR using Differential Irradiation (2.25 ppm/0.60 ppm). Protons assignment is given in Table A.4.

Proton	STD % 0.60 ppm	STD % 2.25 ppm	Ratio STD 2.25 ppm/0.60 ppm	ΔDEEP-STD
H1	12.21	12.03	0.99	-0.42
H2	19.22	26.52	1.38	-0.02
H3	22.47	48.51	2.16	0.76
H4	21.19	42.37	2.00	0.60
H5	21.82	51	2.34	0.94
H6	19.04	35.55	1.87	0.46
H6'	19.04	35.55	1.87	0.46
CH_2(S)	10.09	11.46	1.14	-0.27
CH_2(S)	10.09	11.46	1.14	-0.27
H(fur)	22.91	22.69	0.99	-0.41
CH_2(Bn)	11.57	12.27	1.06	-0.34
H para	18.67	16.13	0.86	-0.54
H meta	18.67	17.27	0.93	-0.48
H orto	17.44	16.13	0.92	-0.48
			STD average	
			1.40	

Table A.11. Appendix to Figure 4.29 in Sub-section 4.3.2. Ligand 108. DEEP-STD NMR using Differential Irradiation (2.25 ppm/0.60 ppm). Protons assignment is given in Table A.5.

A.6 Glide docking tables for the *SCJ set* of ligands in complex with CTB

Title	Potential Energy-OPLS-2005	Relative Potential Energy-OPLS-2005	docking score	glide emodel
SCJ-93	-10.465	2.398	-4.05	-47.621
SCJ-93	-10.694	2.169	-4.05	-47.621
SCJ-93	-12.753	0.11	-4.05	-47.621
SCJ-93	-12.863	0	-4.05	-47.621
SCJ-93	-10.465	2.398	-3.997	-48.631
SCJ-93	-10.694	2.169	-3.997	-48.631
SCJ-93	-12.753	0.11	-3.997	-48.631
SCJ-93	-12.863	0	-3.997	-48.631
SCJ-93	-10.465	2.398	-3.839	-47.479
SCJ-93	-10.694	2.169	-3.839	-47.479
SCJ-93	-12.753	0.11	-3.839	-47.479
SCJ-93	-12.863	0	-3.839	-47.479
SCJ-93	-10.465	2.398	-3.703	-46.58
SCJ-93	-10.694	2.169	-3.703	-46.58
SCJ-93	-12.753	0.11	-3.703	-46.58
SCJ-93	-12.863	0	-3.703	-46.58
SCJ-93	-10.465	2.398	-3.304	-44.163
SCJ-93	-10.694	2.169	-3.304	-44.163
SCJ-93	-12.863	0	-3.304	-44.163

Table A.12. Appendix to Figure 4.34 in Sub-section 4.3.7. Glide docking tables for the poses obtained by molecular docking of **Ligand 93** on CTB (protein structure from PDB ID: 3CHB). The docking poses are sorted according to their docking score.

Title	Potential Energy-OPLS-2005	Relative Potential Energy-OPLS-2005	docking score	glide emodel
SCJ-97	-19.452	0.46	-4.05	-48.197
SCJ-97	-19.592	0.32	-4.05	-48.197
SCJ-97	-19.721	0.191	-4.05	-48.197
SCJ-97	-19.912	0	-4.05	-48.197
SCJ-97	-19.452	0.46	-4.037	-48.264
SCJ-97	-19.592	0.32	-4.037	-48.264
SCJ-97	-19.721	0.191	-4.037	-48.264
SCJ-97	-19.912	0	-4.037	-48.264
SCJ-97	-19.452	0.46	-4.007	-49.616
SCJ-97	-19.592	0.32	-4.007	-49.616
SCJ-97	-19.721	0.191	-4.007	-49.616
SCJ-97	-19.912	0	-4.007	-49.616
SCJ-97	-19.452	0.46	-3.972	-50.71
SCJ-97	-19.592	0.32	-3.972	-50.71
SCJ-97	-19.721	0.191	-3.972	-50.71
SCJ-97	-19.912	0	-3.972	-50.71
SCJ-97	-19.452	0.46	-3.513	-47.174
SCJ-97	-19.592	0.32	-3.513	-47.174
SCJ-97	-19.721	0.191	-3.513	-47.174
SCJ-97	-19.912	0	-3.513	-47.174

Table A.13. Appendix to Figure 4.34 in Sub-section 4.3.7. Glide docking tables for the poses obtained by molecular docking of **Ligand 97** on CTB (protein structure from PDB ID: 3CHB). The docking poses are sorted according to their docking score.

Title	Potential Energy-OPLS-2005	Relative Potential Energy-OPLS-2005	docking score	glide emodel
SCJ-105	-13.78	0.207	-4.114	-49.046
SCJ-105	-13.787	0.2	-4.114	-49.046
SCJ-105	-13.892	0.094	-4.114	-49.046
SCJ-105	-13.986	0	-4.114	-49.046
SCJ-105	-13.78	0.207	-4.095	-49.801
SCJ-105	-13.787	0.2	-4.095	-49.801
SCJ-105	-13.892	0.094	-4.095	-49.801
SCJ-105	-13.986	0	-4.095	-49.801
SCJ-105	-13.78	0.207	-4.075	-49.555
SCJ-105	-13.787	0.2	-4.075	-49.555
SCJ-105	-13.892	0.094	-4.075	-49.555
SCJ-105	-13.986	0	-4.075	-49.555
SCJ-105	-13.78	0.207	-3.975	-47.795
SCJ-105	-13.787	0.2	-3.975	-47.795
SCJ-105	-13.892	0.094	-3.975	-47.795
SCJ-105	-13.986	0	-3.975	-47.795
SCJ-105	-13.78	0.207	-3.682	-46.168
SCJ-105	-13.787	0.2	-3.682	-46.168
SCJ-105	-13.892	0.094	-3.682	-46.168
SCJ-105	-13.986	0	-3.682	-46.168

Table A.14. Appendix to Figure 4.34 in Sub-section 4.3.7. Glide docking tables for the poses obtained by molecular docking of **Ligand 105** on CTB (protein structure from PDB ID: 3CHB). The docking poses are sorted according to their docking score.

Title	Potential Energy-OPLS-2005	Relative Potential Energy-OPLS-2005	docking score	glide emodel
SCJ-108	-23.929	0.143	-4.664	-55.287
SCJ-108	-24.019	0.053	-4.664	-55.287
SCJ-108	-24.071	0.001	-4.664	-55.287
SCJ-108	-24.072	0	-4.664	-55.287
SCJ-108	-23.929	0.143	-4.413	-53.432
SCJ-108	-24.019	0.053	-4.413	-53.432
SCJ-108	-24.071	0.001	-4.413	-53.432
SCJ-108	-24.072	0	-4.413	-53.432
SCJ-108	-23.929	0.143	-4.371	-52.992
SCJ-108	-24.019	0.053	-4.371	-52.992
SCJ-108	-24.071	0.001	-4.371	-52.992
SCJ-108	-24.072	0	-4.371	-52.992
SCJ-108	-23.929	0.143	-4.309	-52.898
SCJ-108	-24.019	0.053	-4.309	-52.898
SCJ-108	-24.071	0.001	-4.309	-52.898
SCJ-108	-24.072	0	-4.309	-52.898
SCJ-108	-23.929	0.143	-4.094	-53.746
SCJ-108	-24.019	0.053	-4.094	-53.746
SCJ-108	-24.071	0.001	-4.094	-53.746
SCJ-108	-24.072	0	-4.094	-53.746

Table A.15. Appendix to Figure 4.34 in Sub-section 4.3.7. Glide docking tables for the poses obtained by molecular docking of **Ligand 108** on CTB (protein structure from PDB ID: 3CHB). The docking poses are sorted according to their docking score.

A.7 Binding epitope mapping of 2,7-anhydro-Neu5Ac in complex with GH33

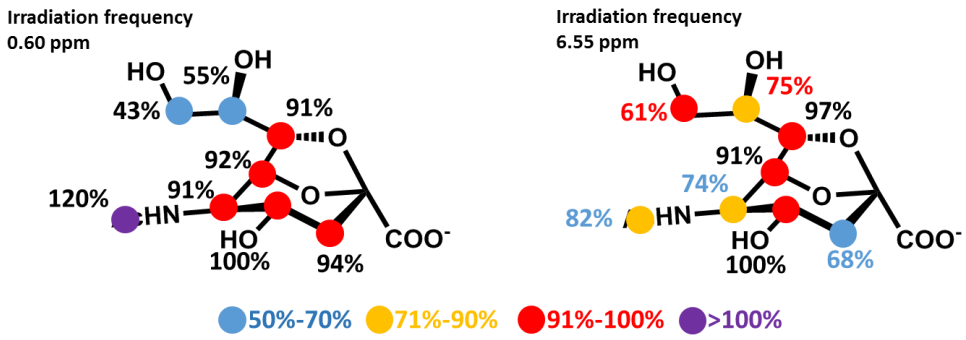

Figure A.5. Appendix to Sub-section 5.2.1 Binding epitope mapping of 2,7-anhydro-Neu5Ac as bound to GH33 from STD NMR experiments at the irradiation frequency of 0.60 ppm (left) and 6.55 ppm (right). Relative STD intensities normalised at H4 are shown for each atom (please note that for Neu5Ac and its derivative the pseudo-anomeric carbon is numbered C2).

A.8 Assignment and STD NMR binding epitope mapping raw data for 3'SL and 6'SL in complex with GH33 D282A at two different irradiation frequencies

				STD (%) 3'SL, 0.60 ppm					
Proton ID	**Spin system**	**^{1}H δ (ppm)**	**^{13}C δ (ppm)**	**0.5 s**	**1 s**	**2 s**	**3 s**	**4 s**	**5 s**
H3a	Neu5Ac	1.669	39.37	1.79	2.98	3.84	4.03	4.07	4.07
CH_3	Neu5Ac	1.898	21.68	1.21	2.55	4.32	5.3	5.84	5.89
H3e	Neu5Ac	2.627	39.37	1.15	2.04	2.98	3.16	3.19	3.19
H2	βGlc	3.150	73.79	0.32	0.81	1.87	2.63	2.95	3.01
H2	Gal	3.443	69.22	0.45	1.02	2.25	2.92	3.35	3.48
H2	αGlc	3.446	70.86	0.45	1.02	2.25	2.92	3.35	3.48
H5	βGlc	3.462	67.88	0.44	1.03	2.22	2.84	3.21	3.29
H8	Neu5Ac	3.472	74.58	0.47	1.1	2.04	2.63	2.73	2.81
H6	Neu5Ac	3.498	72.62	0.57	1.29	2.22	2.73	2.92	3.01
H3	βGlc	3.507	74.09	0.49	1.11	1.91	2.36	2.51	2.58
H9	Neu5Ac	3.512	62.39	0.41	1.04	1.52	1.96	2.08	2.08
H4	βGlc	3.536	78.01	0.54	1.26	1.89	2.53	2.61	2.73
H4	αGlc	3.538	77.97	0.54	1.26	1.89	2.53	2.61	2.73
H4	Neu5Ac	3.551	68.23	0.45	1.09	1.83	2.34	2.53	2.63
H5	Gal	3.579	74.95	0.42	1.12	1.73	2.25	2.32	**2.35**
H6	Gal	3.590	60.8	0.44	1.08	1.43	1.76	1.76	1.78
H6'	Gal	3.630	"	0.42	1.18	1.49	1.76	1.76	1.78
H6	αGlc	3.690	59.8	0.4	0.78	1.52	1.87	2.04	2.04
H6	βGlc	3.700	59.66	0.57	1.28	2.43	3.13	3.38	3.45
H3	αGlc	3.705	71.5	0.57	1.28	2.43	3.13	3.42	3.48
H5	Neu5Ac	3.723	51.49	0.63	1.16	1.71	2.08	2.25	2.27
H9'	Neu5Ac	3.744	62.39	0.49	0.93	1.54	1.83	1.87	1.91
H6'	βGlc	3.760	59.66	0.53	1.21	1.88	2.430	2.480	2.55
H7	Neu5Ac	3.776	71.58	0.59	1.15	1.83	2.41	2.55	2.63
H5	αGlc	3.810	69.87	-	-	-	-	-	-
H4	Gal	3.826	66.95	0.38	0.98	1.63	1.96	2.12	2.16
H6'	αGlc	3.830	59.8	0.33	0.84	1.12	1.48	1.64	1.65
H3	Gal	3.985	75.13	0.48	1.06	1.87	2.27	2.43	2.49

Table A.16. Appendix to Figure 5.5a in Sub-section 5.2.1 3'SL, 0.60 ppm. ^{1}H and ^{13}C assignment and the raw STD (%) at increasing saturation times. Experiments acquired at 285 K.

				STD (%) 6'SL, 0.60 ppm					
Proton ID	**Spin system**	**^{1}H δ (ppm)**	**^{13}C δ (ppm)**	**0.5 s**	**1 s**	**2 s**	**3 s**	**4 s**	**5 s**
H3a	Neu5Ac	1.664	40.00	5.41	8.89	11.69	12.51	12.88	12.88
CH_3	Neu5Ac	1.951	22.2	7.76	14.34	22.25	26.27	27.85	28.4
H3e	Neu5Ac	2.636	40.00	4.54	7.99	10.7	11.69	11.92	11.92
H2	βGlc	3.231	73.67	0.97	2.53	4.67	6.26	6.83	7.04
H2	Gal	3.454	70.66	1.87	4.07	7.39	9.25	10.29	10.49
H8	Neu5Ac	3.487	68.24	2.59	5.05	8.55	10.39	11.02	11.24
H2	αGlc	3.520	71.09	1.78	3.99	5.84	6.9	6.97	7.17
H6	Gal	3.523	63.53	1.21	2.92	4.54	5.84	6.39	6.43
H5	βGlc	3.536	74.58	1.35	3.01	4.72	5.73	6.02	6.04
H4	βGlc	3.542	79.52	1.68	3.48	5.51	6.51	6.89	6.97
H4	αGlc	3.550	79.53	1.68	3.48	5.51	6.51	6.89	6.97
H9	Neu5Ac	3.560	62.54	1.99	4.07	6.51	7.68	8.14	8.22
H3	βGlc	3.575	74.5	2.16	4.28	6.97	8.39	9.16	9.16
H4	Neu5Ac	3.578	68.25	2.16	4.28	6.97	8.39	9.16	9.16
H3	Gal	3.586	72.1	2.22	3.88	7.24	8.72	9.25	**9.25**
H6	Neu5Ac	3.644	72.33	3.42	6.83	11.02	13.41	14.07	14.34
H6	βGlc	3.717	60.2	0.78	1.81	2.55	3.22	3.29	3.29
H5	Gal	3.733	73.83	2.01	3.69	5.46	6.64	6.97	7.17
H3	αGlc	3.764	71.47	2.36	4.95	7.99	9.34	10.58	11.05
H6	αGlc	3.764	59.9	2.36	4.95	7.99	9.34	10.58	11.05
H5	Neu5Ac	3.787	51.6	3.78	7.46	11.92	14.07	15.21	15.66
H9'	Neu5Ac	3.800	62.54	2.22	4.28	6.32	7.32	7.61	7.99
H5	αGlc	3.81	71.75	3.01	5.05	7.76	9.430	9.710	9.71
H4	Gal	3.856	68.46	2.92	5.79	9.25	10.91	11.35	11.35
H6'	βGlc	3.866	60.2	1.03	2.38	3.32	3.59	3.96	3.96
H7	Neu5Ac	3.876	69.84	1.03	2.38	3.32	3.59	3.96	3.96
H6'	αGlc	3.878	59.9	1.03	2.38	3.32	3.59	3.96	3.96
H6'	Gal	3.900	63.53	2.08	4.19	6.51	6.38	7.24	7.24
H1	Gal	4.345	103.12	1.52	3.16	4.95	6.02	6.32	6.57
H1	βGlc	4.585	95.62	0.85	2.02	3.04	3.96	4.39	4.58
H1	αGlc	5.145	91.65	0.86	2.45	4.45	5.84	6.69	7.1

Table A.17. Appendix to Figure 5.5b in Sub-section 5.2.1 6'SL, 0.60 ppm. ^{1}H and ^{13}C assignment and the raw STD (%) at increasing saturation times. Experiments acquired at 298 K.

		STD (%) 3'SL, 6.55 ppm					
Proton ID	**Spin system**	**0.5 s**	**1 s**	**2 s**	**3 s**	**4 s**	**5 s**
H3a	Neu5Ac	0.28	0.62	1.01	1.15	1.19	1.19
CH_3	Neu5Ac	-	0.49	1.29	1.92	2.18	2.18
H3e	Neu5Ac	0.27	0.7	1.07	1.16	1.21	1.21
H2	βGlc	0.23	0.59	1.29	1.89	2.01	2.04
H2	Gal	0.23	0.72	1.46	1.94	2.41	2.45
H2	αGlc	0.23	0.72	1.46	1.94	2.41	2.45
H5	βGlc	0.26	0.78	1.52	1.87	2.25	2.25
H8	Neu5Ac	0.29	0.73	1.37	1.73	1.92	2.02
H6	Neu5Ac	0.3	0.71	1.25	1.61	1.78	1.87
H3	βGlc	0.32	0.65	1.19	1.52	1.64	1.74
H9	Neu5Ac	0.29	0.66	1.13	1.45	1.46	1.49
H4	βGlc	0.21	0.65	1.18	1.35	1.67	1.74
H4	αGlc	0.21	0.65	1.18	1.35	1.67	1.74
H4	Neu5Ac	0.29	0.69	1.23	1.52	1.73	1.81
H5	Gal	0.31	0.64	1.1	1.29	1.42	**1.51**
H6	Gal	0.33	0.67	1.03	1.18	1.23	1.26
H6'	Gal	0.4	0.75	1.19	1.38	1.45	1.48
H6	αGlc	0.35	0.78	1.24	1.33	1.49	1.55
H6	βGlc	0.26	0.74	1.45	1.55	1.68	1.85
H3	αGlc	0.26	0.74	1.45	1.55	1.68	1.85
H5	Neu5Ac	0.25	0.72	1.29	1.54	1.79	1.94
H9'	Neu5Ac	0.34	0.66	1.16	1.23	1.38	1.39
H6'	βGlc	0.38	0.72	1.39	1.610	1.740	1.88
H7	Neu5Ac	0.42	0.75	1.59	1.66	1.71	1.79
H5	αGlc	-	-	-	-	-	-
H4	Gal	0.38	0.78	1.38	1.71	1.83	1.86
H6'	αGlc	0.41	0.73	1.12	1.48	1.52	1.54
H3	Gal	0.43	0.85	1.58	1.83	2.08	2.08

Table A.18. Appendix to Figure 5.5c in Sub-section 5.2.1 3'SL, 6.55 ppm. Raw STD (%) at increasing saturation times. Experiments acquired at 285 K.

		STD (%) 6'SL, 6.55 ppm					
Proton ID	**Spin system**	**0.5 s**	**1 s**	**2 s**	**3 s**	**4 s**	**5 s**
H3a	Neu5Ac	2.12	4.15	6.51	7.04	7.46	7.46
CH_3	Neu5Ac	2.34	6.26	12.51	15.66	17.1	17.44
H3e	Neu5Ac	2.21	4.45	6.77	7.46	7.68	7.75
H2	βGlc	1.1	2.41	4.58	5.9	6.57	6.89
H2	Gal	1.19	2.81	5.73	7.24	8.22	8.47
H8	Neu5Ac	1.48	3.38	6.38	7.83	8.47	8.64
H2	αGlc	2.08	3.88	5.84	6.57	6.83	6.97
H6	Gal	1.24	2.48	4.28	5.15	5.57	5.79
H5	βGlc	1.22	2.53	4.15	5	5.35	5.35
H4	βGlc	1.28	2.68	4.54	5.51	5.84	5.89
H4	αGlc	1.28	2.68	4.54	5.51	5.84	5.89
H9	Neu5Ac	1.41	2.98	5.15	6.21	6.7	6.81
H3	βGlc	1.51	3.16	5.46	6.57	7.11	7.39
H4	Neu5Ac	1.51	3.16	5.46	6.57	7.11	7.39
H3	Gal	1.29	2.84	5.35	6.64	7.39	**7.39**
H6	Neu5Ac	1.79	4.15	7.68	9.43	10.29	10.39
H6	βGlc	0.71	1.48	2.18	2.61	2.81	2.87
H5	Gal	1.98	3.73	6.52	6.26	6.64	6.77
H3	αGlc	1.49	3.32	6.26	7.68	8.31	8.55
H6	αGlc	1.49	3.32	6.26	7.68	8.31	8.55
H5	Neu5Ac	1.71	4.15	7.91	9.89	10.7	10.81
H9'	Neu5Ac	1.48	3.16	5.15	6.02	6.32	6.44
H5	αGlc	1.68	3.73	6.19	7.610	7.990	8.07
H4	Gal	1.33	3.29	6.21	7.68	8.39	8.47
H6'	βGlc	0.81	1.48	2.53	3.04	3.19	3.84
H7	Neu5Ac	0.81	1.48	2.53	3.04	3.19	3.84
H6'	αGlc	0.81	1.48	2.53	3.04	3.19	3.84
H6'	Gal	2.27	4.24	6.14	6.97	7.53	7.53
H1	Gal	1.3	2.6	4.4	5.29	5.68	5.68
H1	βGlc	0.79	1.98	3.42	3.99	4.58	4.62
H1	αGlc	1.15	2.11	4.67	6.08	6.57	6.89

Table A.19. Appendix to Figure 5.5d in Sub-section 5.2.1 6'SL, 6.55 ppm. Raw STD (%) at increasing saturation times. Experiments acquired at 298 K.

A.9 Docking of 3'SL and 6'SL on GH33 D282A as compared to GH33 WT

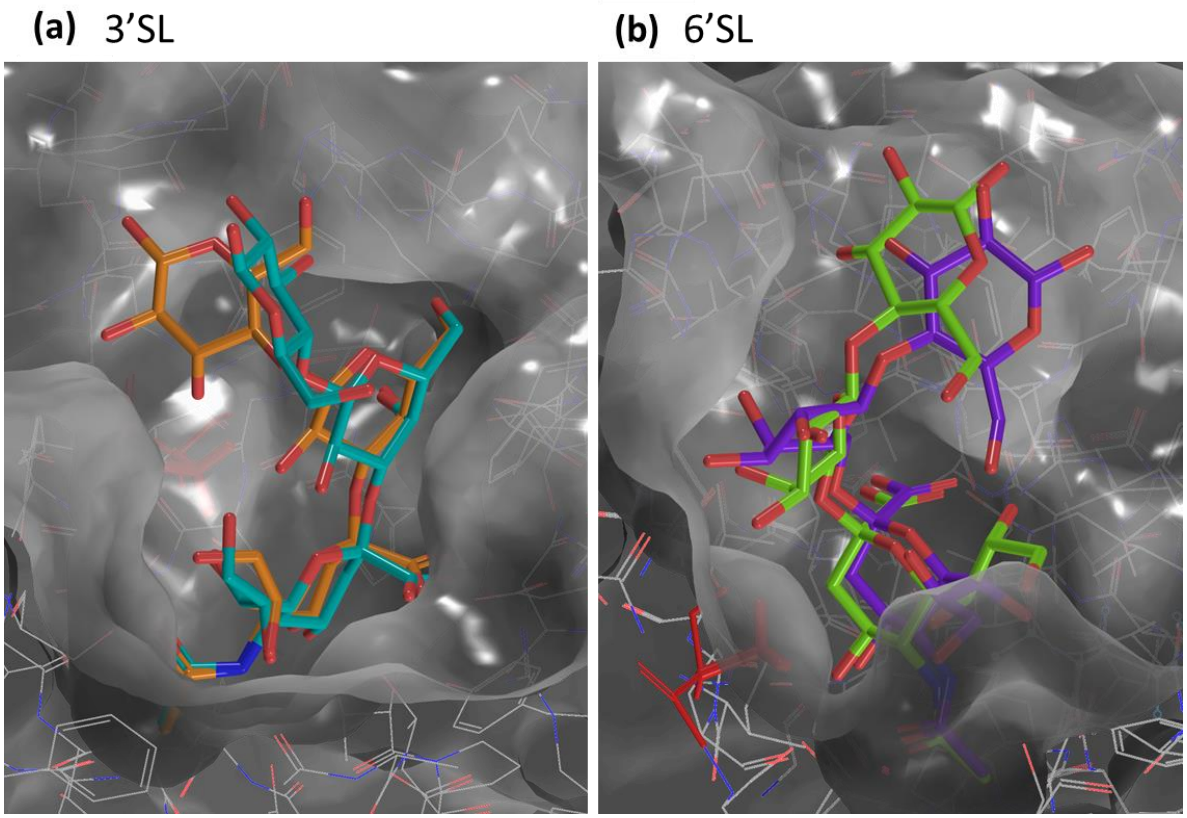

Figure A.6. Appendix to Sub-section 5.2.4. The lowest energy converging docking solutions for **(a)** 3'SL as bound to GH33 WT (cyan) and to GH33 D282A (orange); and **(b)** 6'SL as bound to GH33 WT (purple) and to GH33 D282A (green). The grey surface encloses the mutant grid only, and the Ala282 is shown in red.

A.10 Glide docking tables for 3'SL and 6'SL in complex with GH33 WT and GH33 D282A

	GH33 WT		GH33 D282A	
Title	glide gscore	glide emodel	glide gscore	glide emodel
3'SL_lowest	-4.851	-55.108	-5.714	-69.097
3'SL_lowest	-4.599	-52.153	-5.678	-62.32
3'SL_lowest	-4.378	-51.585	-5.555	-67.041
3'SL_lowest	-4.358	-49.368	-5.523	-65.15
3'SL_lowest	-4.336	-48.47	-5.471	-63.716
3'SL_lowest	-4.212	-46.64	-5.461	-63.415
3'SL_lowest	-4.165	-51.572	-5.456	-67.82
3'SL_lowest	-4.129	-50.611	-5.43	-59.581
3'SL_lowest	-4.091	-48.006	-5.404	-67.586
3'SL_lowest	-4.048	-48.877	-5.342	-61.495
3'SL_lowest	-4.027	-50.191	-5.297	-66.962
3'SL_lowest	-3.916	-49.995	-5.159	-62.441
3'SL_lowest	-3.622	-52.849	-5.149	-66.088
3'SL_lowest	-3.597	-51.321	-5.139	-58.56
3'SL_lowest	-3.579	-47.157	-5.113	-63.42
3'SL_lowest	-3.52	-47.799	-5.11	-60.602
3'SL_lowest	-3.435	-49.426	-5.107	-61.548
3'SL_lowest	-3.432	-47.374	-5.099	-58.338
3'SL_lowest	-3.421	-49.284	-5.073	-65.733
3'SL_lowest	-3.409	-47.612	-5.06	-63.575
3'SL_lowest	-3.387	-51.591	-5.012	-60.545
3'SL_lowest	-3.357	-48.439	-4.987	-61.824
3'SL_lowest	-3.351	-48.784	-4.904	-61.937
3'SL_lowest	-3.349	-49.877	-4.883	-64.155
3'SL_lowest	-3.341	-43.892	-4.871	-65.052

Table A.20. Appendix to Sub-section 5.2.4. *Glide gscore* and *glide emodel* tables for the the first 25 poses obtained by molecular docking of 3'SL on GH33 WT and GH33 D282A (protein structure from PDB ID: 4X4A). The docking poses are sorted according to their *gscore*.

	GH33 WT		GH33 D282A	
Title	glide gscore	glide emodel	glide gscore	glide emodel
6'SL_lowest	-5.473	-66.403	-5.444	-67.843
6'SL_lowest	-5.305	-62.096	-5.325	-70.796
6'SL_lowest	-5.06	-59.18	-5.286	-68.337
6'SL_lowest	-4.862	-59.249	-5.254	-68.351
6'SL_lowest	-4.815	-58.049	-5.226	-69.916
6'SL_lowest	-4.792	-56.344	-5.21	-69.48
6'SL_lowest	-4.064	-60.326	-5.207	-68.591
6'SL_lowest	-4.038	-57.426	-5.197	-67.927
6'SL_lowest	-3.874	-55.414	-5.071	-64.32
6'SL_lowest	-3.847	-51.886	-5.027	-66.51
6'SL_lowest	-3.648	-51.936	-4.871	-59.65
6'SL_lowest	-3.636	-53.673	-4.863	-64.773
6'SL_lowest	-3.426	-50.999	-4.406	-59.468
6'SL_lowest	-3.414	-51.824	-4.375	-58.224
6'SL_lowest	-3.386	-52.367	-4.316	-54.875
6'SL_lowest	-3.384	-49.545	-4.3	-60.301
6'SL_lowest	-3.353	-46.825	-4.003	-48.866
6'SL_lowest	-3.322	-46.886	-3.925	-53.695
6'SL_lowest	-3.223	-49.529	-3.817	-56.252
6'SL_lowest	-3.219	-49.444	-3.78	-53.083
6'SL_lowest	-3.119	-45.753	-3.751	-53.553
6'SL_lowest	-3.113	-43.44	-3.732	-52.327
6'SL_lowest	-3.081	-48.839	-3.691	-52.815
6'SL_lowest	-3.046	-47.699	-3.689	-53.024
6'SL_lowest	-2.938	-46.854	-3.638	-51.038

Table A.21. Appendix to Sub-section 5.2.4. *Glide gscore* and *glide emodel* tables for the first 25 poses obtained by molecular docking of 6'SL on GH33 WT and GH33 D282A (protein structure from PDB ID: 4X4A). The docking poses are sorted according to their *gscore*.

A.11 Assignment and STD NMR binding epitope mapping raw data for the sialoglycans in complex with CBM40

		STD (%) 3'SL					
Proton ID	**Spin system**	**0.5 s**	**1 s**	**2 s**	**3 s**	**4 s**	**5 s**
H3a	Neu5Ac	4.72	7.17	8.98	9.82	10.6	10.91
CH_3	Neu5Ac	29.24	44.93	57.9	60.85	62.6	62.6
H3e	Neu5Ac	4.03	7.17	9.52	10.21	10.81	10.91
H2	βGlc	/	/	1.03	1.11	1.81	1.81
H2	Gal	1.41	1.71	2.71	2.92	3.16	3.59
H2	αGlc	"	"	"	"	"	"
H5	βGlc	4.32	7.53	10.7	12.15	12.88	12.88
H8	Neu5Ac	5.51	9.52	14.34	15.97	16.93	16.93
H6	Neu5Ac	4.03	6.51	9.61	11.46	12.51	12.88
H3	βGlc	2.73	4.32	7.04	8.31	8.64	8.64
H9	Neu5Ac	2.9	4.24	6.92	7.53	8.07	8.98
H4	βGlc	1.33	2.55	4.39	5.9	6.26	6.32
H4	αGlc	"	"	"	"	"	"
H4	Neu5Ac	1.33	3.13	5.05	6.08	6.44	6.57
H5	Gal	0.73	0.9	1.54	1.57	1.98	1.98
H6	Gal	0.73	0.9	1.48	2.08	2.14	2.14
H6'	Gal	0.73	0.9	1.71	2.61	2.76	2.76
H6	αGlc	1.01	1.21	1.71	2.04	2.06	2.06
H6	βGlc	2.16	4.03	7.68	9.25	10.09	10.09
H3	αGlc	"	"	"	"	"	"
H5	Neu5Ac	3.25	5.57	9.99	12.88	13.42	13.93
H9'	Neu5Ac	2.79	5.35	8.14	9.71	10.09	10.6
H6'	βGlc	1.96	3.01	4.07	4.89	4.95	5.25
H7	Neu5Ac	2.38	4.99	6.97	9.89	10.81	11.35
H5	αGlc	1.28	1.63	2.48	2.61	2.61	2.61
H4	Gal	0.87	1.26	1.76	2.18	2.60	2.65
H6'	αGlc	0.92	1.26	1.76	1.98	1.98	1.98
H3	Gal	1.23	1.88	3.04	3.73	3.99	3.99

Table A.22. Appendix to Figure 5.20 in Sub-section 5.3.1 3'SL. Raw STD (%) at increasing saturation times. See Table A.16 for assignment. Experiments acquired at 293 K.

		STD (%) 6'SL					
Proton ID	**Spin system**	**0.5 s**	**1 s**	**2 s**	**3 s**	**4 s**	**5 s**
H3a	Neu5Ac	8.64	13.01	17.61	18.85	20.78	20.99
CH_3	Neu5Ac	53.55	71.07	77.59	77.59	77.59	77.59
H3e	Neu5Ac	8.64	13.43	18.49	19.99	20.78	20.99
H2	βGlc	1.01	2.02	3.92	4.76	5.1	5.41
H2	Gal	1.01	2.02	2.68	3.38	4.81	4.81
H8	Neu5Ac	14.34	23.83	30.11	31.01	31.92	32.56
H2	αGlc	1.53	2.98	3.32	4.03	4.58	5.2
H6	Gal	1.53	2.98	3.32	4.03	4.58	5.2
H5	βGlc	1.34	2.36	2.48	3.52	3.92	4.2
H4	βGlc	3.66	5.41	9.71	10.19	10.49	11.13
H4	αGlc	"	"	"	"	"	"
H9	Neu5Ac	4.24	7.32	11.57	12.63	13.39	13.66
H3	βGlc	2.87	5.01	7.99	8.89	9.24	9.34
H4	Neu5Ac	"	"	"	"	"	"
H3	Gal	1.46	2.98	4.95	6.02	6.9	7.17
H6	Neu5Ac	7.61	14.77	23.59	27.31	28.11	28.68
H6	βGlc	1.34	1.98	2.41	3.01	3.52	3.62
H5	Gal	1.28	2.73	3.55	3.66	4.4	4.72
H3	αGlc	3.73	6.9	10.01	12.27	12.63	13.4
H6	αGlc	/	/	/	/	/	/
H5	Neu5Ac	4.4	8.64	13.27	14.77	15.81	15.97
H9'	Neu5Ac	5.46	9.9	15.66	17.95	18.84	18.84
H7	Neu5Ac	3.28	6.09	7.97	10.01	11.81	12.36
H4	Gal	1.01	1.81	3.01	3.96	4.11	4.36
H6'	βGlc	1.43	2.02	3.25	3.45	3.59	3.77
H5	αGlc	1.43	2.02	3.25	3.45	3.59	3.77
H6'	αGlc	1.43	2.87	4.07	5.68	5.46	5.46
H6'	Gal	1.43	2.87	4.07	5.68	5.46	5.46
H1	Gal	0.83	1.45	2.1	2.81	3.45	3.69

Table A.23. Appendix to Figure 5.20 in Sub-section 5.3.1 6'SL. Raw STD (%) at increasing saturation times. See Table A.17 for assignment. Experiments acquired at 293 K.

				STD (%) 3'SLGc					
Proton ID	**Spin system**	**^{1}H δ (ppm)**	**^{13}C δ (ppm)**	**0.5 s**	**1 s**	**2 s**	**3 s**	**4 s**	**5 s**
H3a	Neu5Gc	1.672	39.39	2.29	3.29	3.88	4.62	4.89	4.89
H3e	Neu5Gc	2.583	39.39	1.78	2.7	3.55	3.99	4.24	4.07
H2	βGlc	3.090	73.29	-	0.52	0.72	0.95	1.05	1.15
H2	Gal	3.387	69.11	-	0.59	0.81	1.73	1.79	1.85
H2	αGlc	3.387	70.83	-	0.59	0.81	1.73	1.79	1.85
H8	Neu5Gc	3.397	67.63	1.35	2.7	4.32	5.20	5.57	5.68
H5	βGlc	3.402	74.51	0.41	1.57	2.95	3.25	3.59	3.84
H3	βGlc	3.441	74.05	0.58	1.3	1.69	2.34	2.58	2.58
H9	Neu5Gc	3.449	62.18	0.58	1.3	1.69	2.34	2.58	2.58
H4	αGlc	3.485	77.88	0.45	0.61	0.71	1.12	1.18	1.26
H4	βGlc	3.494	77.88	0.45	0.61	0.71	1.12	1.18	1.26
H7	Neu5Gc	3.520	75.07	0.38	0.67	0.82	1.29	1.34	1.39
H6	Gal	3.541	60.68	0.67	1.03	1.29	1.93	2.09	2.13
H6	Neu5Gc	3.561	72.23	0.47	0.96	1.63	2.43	2.53	2.53
H4	Neu5Gc	3.590	67.72	0.96	2.25	3.66	4.36	4.86	5.15
H6	βGlc	3.630	59.66	0.72	1.43	1.58	1.71	2.04	2.14
H5	αGlc	3.635	71.01	0.41	0.72	0.77	1.18	1.24	1.27
H9'	Neu5Gc	3.677	62.18	0.72	1.51	1.89	2.18	2.35	2.41
H6	αGlc	3.694	59.45	0.44	0.98	1.35	1.54	1.54	1.54
H5	Gal	3.697	71.55	0.63	1.53	2.16	2.68	2.92	2.95
H5	Neu5Gc	3.754	51.09	1.04	1.73	2.87	4.32	4.40	4.54
H4	Gal	3.756	67.12	0.51	1.16	1.71	1.78	1.94	2.14
H3	αGlc	3.757	69.87	0.51	1.16	1.71	1.78	1.94	2.14
H6'	βGlc	3.772	59.66	0.51	1.16	1.71	1.78	1.94	2.14
CH_2	Neu5Gc	3.926	60.68	1.92	3.77	6.26	7.53	7.87	8.07
H6'	Gal	3.926	60.68	1.92	3.77	6.26	7.53	7.87	8.07
H3	Gal	3.936	75.11	0.33	0.64	0.78	1.03	1.06	1.06

Table A.24. Appendix to Figure 5.20 in Sub-section 5.3.1 3'SLGc. ^{1}H and ^{13}C assignment and the raw STD (%) at increasing saturation times. Experiments acquired at 293 K.

				STD (%) 6'SLGc					
Proton ID	**Spin system**	**^{1}H δ (ppm)**	**^{13}C δ (ppm)**	**0.5 s**	**1 s**	**2 s**	**3 s**	**4 s**	**5 s**
H3a	Neu5Gc	1.571	39.90	2.1	3.1	3.73	4.19	4.19	4.19
H3e	Neu5Gc	2.530	39.90	1.31	2.2	2.81	2.78	3.19	3.35
H2	βGlc	3.122	73.40	-	0.36	0.66	0.83	1.03	1.04
H2	Gal	3.341	70.48	0.3	0.47	0.82	1.02	1.12	1.37
H8	Neu5Gc	3.369	67.94	1.31	2.2	4.28	5.11	5.41	5.62
H6	Gal	3.403	63.44	0.26	0.43	0.60	0.72	0.78	0.83
H2	αGlc	3.421	70.70	0.26	0.43	0.6	0.72	0.78	0.83
H5	βGlc	3.435	72.14	-	0.43	0.61	0.71	0.71	0.77
H9	Neu5Gc	3.443	62.28	0.44	0.61	1.01	1.13	1.19	1.24
H4	αGlc	3.446	79.33	0.34	0.51	0.72	0.92	1.08	1.11
H4	βGlc	3.446	79.33	0.34	0.51	0.72	0.92	1.08	1.11
H3	βGlc	3.460	72.14	0.33	0.43	0.76	0.86	0.98	0.98
H3	Gal	3.467	72.12	0.51	0.6	0.65	0.76	0.85	0.85
H4	Neu5Gc	3.563	67.79	0.63	1.49	2.27	2.87	3.16	3.45
H6	βGlc	3.601	59.88	/	/	/	/	/	/
H5	Gal	3.628	73.27	0.41	0.59	0.82	0.93	0.98	1.12
H3	αGlc	3.650	71.81	0.48	0.73	1.37	1.79	2.08	2.08
H6	Neu5Gc	3.651	71.70	0.48	0.73	1.37	1.79	2.08	2.08
H6	αGlc	3.669	59.82	0.32	0.64	1.03	1.15	1.48	1.48
H6'	αGlc	3.669	59.82	0.32	0.64	1.03	1.15	1.480	1.48
H9'	Neu5Gc	3.683	62.28	0.45	0.69	1.11	1.38	1.580	1.68
H7	Neu5Gc	3.710	71.70	0.52	0.91	1.55	1.76	1.980	2.14
H4	Gal	3.741	68.28	0.31	0.41	0.79	0.88	0.91	1.01
H6'	βGlc	3.758	59.88	0.51	0.73	1.480	1.73	1.83	1.9
H5	Neu5Gc	3.761	51.18	0.51	0.73	1.480	1.73	1.83	1.90
H5	αGlc	3.764	69.90	0.51	0.73	1.480	1.73	1.830	1.90
H6'	Gal	3.801	63.44	0.6	0.54	0.880	0.95	1.040	1.04
CH_2	Neu5Gc	3.926	60.57	1.39	2.22	4.32	5.05	5.200	5.30

Table A.25. Appendix to Figure 5.20 in Sub-section 5.3.1 6'SLGc. ^{1}H and ^{13}C assignment and the raw STD (%) at increasing saturation times. Experiments acquired at 293 K.

Proton ID	Spin system	^{1}H δ (ppm)	^{13}C δ (ppm)	STD (%) 3'SGal					
				0.5 s	1 s	2 s	3 s	4 s	5 s
H3	Neu5Ac	1.571	39.90	2.45	3.66	4.32	4.36	4.67	4.72
CH_3	Neu5Ac	1.789	21.75	26.52	40.75	51	54.08	55.14	55.14
H3	Neu5Ac	2.519	39.90	1.52	2.95	3.48	3.92	4.03	4.12
H2	βGal	3.288	69.83	0	0	0	0.00	1.09	1.23
H7	Neu5Ac	3.359	67.62	1.92	4.54	7.39	8.81	9.16	9.16
H6	Neu5Ac	3.381	72.45	1.45	2.9	5.46	6.32	6.83	7.24
H6	βGal	3.396	62.10	1.01	2.12	3.66	3.77	4.24	4.76
H4	Neu5Ac	3.443	67.98	1.25	1.71	3.10	3.84	4.19	4.19
H5	βGal	3.452	74.59	0.81	1.43	2.02	3.16	3.25	3.57
H9	Neu5Ac	3.479	60.66	0.56	0.75	1.58	1.87	1.87	2.18
H6	αGal	3.479	60.66	0	0.41	0.81	0.85	0.92	1.02
H2	αGal	3.607	66.56	0.78	1.12	2.18	2.95	3.22	3.22
H5	Neu5Ac	3.616	51.10	0.78	1.31	2.55	3.22	3.59	3.69
H6'	βGal	3.633	62.10	0.78	1.39	2.31	2.84	2.98	3.19
H8	Neu5Ac	3.649	71.38	0.78	1.39	2.12	2.36	2.53	2.58
H4	βGal	3.707	67.26	0.6	0.77	1.08	1.12	1.41	1.43
H4	αGal	3.778	67.34	0.5	0.77	1.08	1.11	1.15	1.17
H5	αGal	3.835	70.13	0.7	0.92	1.08	1.15	1.21	1.21
H3	βGal	3.849	75.61	0.9	1.04	1.45	1.58	1.61	1.88

Table A.26. Appendix to Figure 5.20 in Sub-section 5.3.1 3'SGal. ^{1}H and ^{13}C assignment and the raw STD (%) at increasing saturation times. Experiments acquired at 293 K.

Proton ID	Spin system	^{1}H δ (ppm)	^{13}C δ (ppm)	STD (%) 6'SGal					
				0.5 s	1 s	2 s	3 s	4 s	5 s
H3	Neu5Ac	1.471	39.94	3.19	5.05	5.84	6.38	6.77	6.77
CH_3	Neu5Ac	1.795	21.59	29.24	44.49	55.14	57.34	57.9	57.9
H3	Neu5Ac	2.488	39.94	2.04	3.01	4.19	4.36	4.49	4.54
H2	βGal	3.233	71.5	/	/	/	/	/	/
H5	βGal	3.336	67.90	2.1	4.67	7.83	8.81	9.34	9.34
H9'	Neu5Ac	3.346	63.43	1.9	3.55	6.19	7.09	7.53	7.53
H3	βGal	3.394	72.21	0.98	1.41	1.94	2.45	2.55	2.55
H6	βGal	3.395	62.12	1.52	1.88	3.01	3.25	3.45	3.76
H4	Neu5Ac	3.417	67.98	1.29	2.27	4.28	4.03	5.05	5.51
H6	Neu5Ac	3.495	72.33	1.42	2.55	4.19	5.25	5.96	6.02
H2	αGal	3.538	68.00	0.3	0.66	0.98	1.12	1.12	1.12
H5	αGal	3.543	72.52	0.3	0.66	0.98	1.12	1.12	1.12
H5	Neu5Ac	3.598	51.59	0.94	2.12	3.92	4.89	5.41	5.46
H3	αGal	3.617	68.66	0.8	1.24	2.27	2.48	2.69	2.93
H6'	βGal	3.641	62.12	0.91	1.54	2.41	2.55	2.62	2.62
H9'	Neu5Ac	3.651	63.43	0.5	1.16	2.20	2.27	2.68	2.78
H8	Neu5Ac	3.651	71.37	0.5	1.16	2.20	2.27	2.68	2.78
H6/6'	αGal	3.687	63.36	0.4	0.67	0.90	0.90	0.90	0.90
H4	βGal	3.687	68.44	0.4	0.67	0.90	0.90	0.90	0.90

Table A.27. Appendix to Figure 5.20 in Sub-section 5.3.1 6'SGal. ^{1}H and ^{13}C assignment and the raw STD (%) at increasing saturation times. Experiments acquired at 293 K.

				STD (%) 3'SLN					
Proton ID	**Spin system**	**^{1}H δ (ppm)**	**^{13}C δ (ppm)**	**0.5 s**	**1 s**	**2 s**	**3 s**	**4 s**	**5 s**
H3	Neu5Ac	1.573	39.20	7.1	10.81	16.13	16.77	18.31	19.04
CH_3	Neu5Ac	1.796	21.6	40.35	56.78	65.73	65.73	65.73	65.73
CH_3	βGlcNAc	1.806	21.6	7.91	9.25	11.57	13.53	14.34	14.48
CH_3	αGlcNAc	1.813	21.6	/	/	/	/	/	/
H3	Neu5Ac	2.524	39.20	8.31	12.88	16.13	18.31	20.18	21.61
H2	Gal	3.341	69.17	1.78	3.22	4.49	6.51	7.46	8.35
H8	Neu5Ac	3.359	67.69	11.24	12.88	18.85	22.04	22.91	23.59
H3	βGlc	3.362	74.52	7.61	10.19	13.83	17.1	18.31	19.54
H6	Neu5Ac	3.392	72.62	6.14	9.07	11.46	14.63	16.93	18.01
H9	Neu5Ac	3.406	62.20	7.32	12.03	16.13	19.6	19.99	20.78
H4	Neu5Ac	3.446	68.10	4.54	7.68	13.53	14.77	15.81	15.81
H5	βGlcNAc	3.450	72.12	2.48	4.32	7.76	9.16	9.52	9.91
H2	βGlcNAc	3.495	54.66	0.92	1.41	2.16	2.78	2.87	3.25
H6	Gal	3.496	60.82	"	"	"	"	"	"
H6'	Gal	3.496	60.82	"	"	"	"	"	"
H4	βGlcNAc	3.516	77.86	1.11	1.74	2.48	2.84	3.22	3.22
H4	αGlcNAc	3.520	78.04	"	"	"	"	"	"
H3	αGlcNAc	3.527	69.88	1.68	2.29	2.41	2.59	2.95	2.95
H6	βGlcNAc	3.615	59.58	3.92	7.11	12.27	13.93	14.48	15.36
H5	Neu5Ac	3.624	51.19	4.45	7.61	13.27	15.06	15.66	15.97
H9'	Neu5Ac	3.638	62.20	5.41	9.61	16.44	17.44	17.81	17.95
H6/6'	αGlcNAc	3.664	59.38	1.83	4.15	7.76	10.01	10.81	11.13
H2	αGlcNAc	3.668	53.55	"	"	"	"	"	"
H7	Neu5Ac	3.668	68.82	2.45	4.81	7.83	8.14	10.01	1.29
H4	Gal	3.722	67.05	1.94	3.55	5.01	5.84	6.14	6.83
H6'	βGlcNAc	3.742	59.58	1.94	2.71	2.95	3.11	3.38	3.38
H5	αGlcNAc	3.742	69.94	"	"	"	"	"	"
H3	Gal	3.894	75.11	2.22	4.98	5.79	7.39	7.76	8.55

Table A.28. Appendix to Figure 5.20 in Sub-section 5.3.1 3'SLN. ^{1}H and ^{13}C assignment and the raw STD (%) at increasing saturation times. Experiments acquired at 293 K.

				STD (%) 6'SLN					
Proton ID	**Spin system**	**^{1}H δ (ppm)**	**^{13}C δ (ppm)**	**0.5 s**	**1 s**	**2 s**	**3 s**	**4 s**	**5 s**
H3e	Neu5Ac	1.511	39.62	5.15	8.22	11.13	11.35	12.15	12.27
CH_3	Neu5Ac	1.806	21.65	32.24	49.05	60.81	62.6	62.6	62.6
CH_3	αGlcNAc	1.824	21.65	/	/	/	/	/	/
CH_3	βGlcNAc	1.844	21.73	1.63	2.84	4.45	5.21	5.57	5.89
H3a	Neu5Ac	2.443	39.62	5.57	7.68	11.54	11.57	11.69	12.03
H6	Gal	3.309	63.05	0.57	1.48	2.63	2.98	2.98	2.98
H2	Gal	3.324	70.30	0.98	1.64	3.16	3.73	3.89	3.99
H8	Neu5Ac	3.340	68.12	6.64	11.46	16.13	17.78	19.79	20.38
H5	βGlcNAc	3.404	74.09	2.87	5.25	7.99	8.81	8.89	9.43
H4	βGlcNAc	3.415	80.46	3.32	6.38	9.52	10.91	11.57	12.39
H9	Neu5Ac	3.416	62.28	"	"	"	"	"	"
H4	Neu5Ac	3.428	67.89	3.32	5.79	8.89	10.09	10.99	11.46
H3	βGlcNAc	3.428	67.89	"	"	"	"	"	"
H4	αGlcNAc	3.447	80.58	0.69	1.18	2.04	2.11	2.14	2.58
H3	Gal	3.458	72.12	0.69	1.29	2.36	2.41	2.52	2.87
H6	Neu5Ac	3.486	72.30	3.73	6.83	11.81	13.79	14.48	15.97
H2	βGlcNAc	3.516	55.50	0.99	1.09	2.22	2.92	2.98	3.32
H5	Neu5Ac	3.601	51.54	4.49	6.32	12.15	15.06	15.36	16.93
H6	βGlcNAc	3.606	60.00	1.45	3.13	4.86	6.14	6.77	6.97
H3	αGlcNAc	3.611	73.34	"	"	"	"	"	"
H9'	Neu5Ac	3.656	62.28	2.58	4.76	7.24	8.64	8.67	9.16
H6/H6'	αGlcNAc	3.656	59.79	1.39	2.53	3.92	4.67	4.81	4.86
H5	Gal	3.679	71.42	1.98	4.45	6.19	7.53	8.07	8.64
H4	Gal	3.705	68.60	0.37	0.72	1.31	1.63	1.73	1.88
H2	αGlcNAc	3.712	53.12	0.76	0.92	1.63	2.18	2.41	2.48
H7	Neu5Ac	3.715	69.17	"	"	"	"	"	"
H6'	βGlcNAc	3.745	60.00	1.08	1.28	2.34	2.45	2.71	3.22
H5	αGlcNAc	3.757	69.66	1.08	1.08	1.45	1.55	1.98	2.10
H6'	Gal	3.782	63.05	0.98	1.66	2.22	2.55	2.84	3.22
H1	Gal	4.228	103.30	0.77	1.61	1.74	1.79	1.88	1.96

Table A.29. Appendix to Figure 5.20 in Sub-section 5.3.1 6'SLN. ^{1}H and ^{13}C assignment and the raw STD (%) at increasing saturation times. Experiments acquired at 293 K.

				STD (%) Neu5Ac-STn					
Proton ID	**Spin system**	**^{1}H δ (ppm)**	**^{13}C δ (ppm)**	**0.5 s**	**1 s**	**2 s**	**3 s**	**4 s**	**5 s**
H3e	Neu5Ac	1.393	39.88	2.95	4.28	5.57	5.68	5.96	5.960
CH_2	Chain	1.661	27.64	1.18	1.74	2.2	2.24	2.45	2.290
CH_3	Neu5Ac	1.736	21.51	19.99	31.01	41.15	44.49	44.93	44.930
CH_3	GalNAc	1.744	21.51	3.01	4.45	6.02	7.17	6.38	6.64
H3a	Neu5Ac	2.434	39.88	1.79	3.07	3.92	4.11	4.24	4.24
CH_2(N3)	Chain	3.136	47.66	0.57	0.66	0.71	0.74	0.74	0.740
CH_2(N3)	Chain	3.204	47.66	0.46	0.6	0.76	0.89	0.99	1.010
CH_2(O)	Chain	3.214	64.66	0.46	0.6	0.76	0.89	0.99	1.010
H8	Neu5Ac	3.282	67.80	1.98	4.32	7.17	9.16	9.61	9.610
H6	GalNAc	3.301	63.48	0.56	0.87	1.02	1.11	1.35	1.35
H9	Neu5Ac	3.336	62.12	0.94	1.6	2.68	3.35	3.55	3.66
H4	Neu5Ac	3.350	67.88	1.260	1.9	3.01	4.36	4.45	4.55
H6	Neu5Ac	3.717	72.26	0.62	0.9	1.3	1.54	1.63	1.68
CH_2(O)	Chain	3.515	64.66	0.72	1.04	1.11	1.23	1.35	1.41
H5	Neu5Ac	3.539	51.44	1.09	1.68	2.9	3.92	4.11	4.44
H9'	Neu5Ac	3.584	62.12	0.66	0.98	1.63	1.98	2.12	2.18
H7	Neu5Ac	3.593	71.55	0.85	1.85	2.14	2.27	2.41	2.48
H6'	GalNAc	3.627	63.48	0.66	0.78	0.86	0.91	0.94	0.94
H3	GalNAc	3.634	66.98	0.66	0.78	0.86	0.91	0.94	0.94
H4	GalNAc	3.696	68.11	0.38	0.54	0.72	0.77	0.77	0.770
H5	GalNAc	3.757	69.14	0.49	0.98	1.07	1.2	1.24	1.24
H2	GalNAc	3.840	49.30	0.49	0.68	0.79	0.95	0.99	0.99

Table A.30. Appendix to Figure 5.20 in Sub-section 5.3.1 Neu5Ac-STn. ^{1}H and ^{13}C assignment and the raw STD (%) at increasing saturation times. Experiments acquired at 293 K.

				STD (%) Neu5Gc-STn					
Proton ID	**Spin system**	**^{1}H δ (ppm)**	**^{13}C δ (ppm)**	**0.5 s**	**1 s**	**2 s**	**3 s**	**4 s**	**5 s**
H3a	Neu5Gc	1.41	39.88	2.48	4.58	5.25	5.3	5.62	5.620
CH_2	Chain	1.617	27.64	1.48	2.41	2.63	2.78	2.95	2.950
CH_3	GalNAc	1.742	21.51	0.43	0.97	1.66	2.02	2.12	2.220
H3e	Neu5Gc	2.451	39.88	1.71	2.48	2.92	3.29	3.31	3.32
CH_2(N3)	Chain	3.143	47.66	0.62	0.78	0.86	0.910	0.99	0.99
CH_2(N3)	Chain	3.193	47.66	0.55	0.63	0.71	0.880	0.88	0.88
CH_2(O)	Chain	3.214	64.66	0.55	0.63	0.71	0.880	0.88	0.88
H8	Neu5Gc	3.270	67.56	1.29	2.92	5.25	6.260	6.83	6.83
H6	GalNAc	3.310	63.63	0.52	0.84	0.99	1.15	1.3	1.330
H9	Neu5Gc	3.337	62.12	0.81	1.16	1.69	1.830	1.87	2.1
H4	Neu5Gc	3.466	67.57	1.52	2.22	2.84	4.240	4.58	4.76
CH_2(O)	Chain	3.513	64.66	-	-	-	-	-	-
H6	Neu5Gc	3.534	71.82	1.05	1.57	2.29	2.650	2.76	2.87
H9'	Neu5Gc	3.577	62.12	0.78	0.94	1.19	1.420	1.78	1.78
H7	Neu5Gc	3.599	71.47	0.5	1.21	1.52	1.79	2.14	2.28
H3	GalNAc	3.627	67.05	0.52	0.85	1.68	2.36	2.38	2.38
H5	Neu5Gc	3.630	51.10	0.52	0.85	1.68	2.36	2.38	2.38
H6'	GalNAc	3.638	63.63	-	-	-	-	-	-
H4	GalNAc	3.696	68.11	0.24	0.54	0.72	0.84	0.86	0.86
H5	GalNAc	3.757	69.14	0.49	0.81	1.33	1.42	1.52	1.52
CH_2	Neu5Gc	3.815	60.57	1.64	3.07	4.81	5.57	5.73	6.020
H2	GalNAc	3.836	53.29	0.6	0.72	1.28	1.41	1.57	2.57

Table A.31. Appendix to Figure 5.20 in Sub-section 5.3.1 Neu5Gc-STn. ^{1}H and ^{13}C assignment and the raw STD (%) at increasing saturation times. Experiments acquired at 293 K.

				STD (%) STF					
Proton ID	**Spin system**	**^{1}H δ (ppm)**	**^{13}C δ (ppm)**	**0.5 s**	**1 s**	**2 s**	**3 s**	**4 s**	**5 s**
H3a	Neu5Ac	1.498	39.18	2.84	3.99	5.55	5.73	5.79	6.380
CH2	Chain	1.606	27.64	0.51	0.85	0.95	1.04	1.09	1.12
CH3	αGlcNAc	1.727	21.51	8.98	14.34	19.22	20.99	21.19	21.19
CH3	Neu5Ac	1.727	21.51	8.98	14.34	19.22	20.99	21.19	21.19
H3e	Neu5Ac	2.434	39.88	2.34	3.59	4.86	5.620	5.68	5.790
CH2(N3)	Chain	3.153	47.66	0.44	0.48	0.52	0.560	0.58	0.58
CH2(N3)	Chain	3.198	47.66	0.38	0.61	0.65	0.670	0.71	0.71
CH2(O)	Chain	3.235	64.39	0.27	0.61	0.92	1.05	1.15	1.210
H2	Gal	3.246	68.72	0.27	0.61	0.92	1.05	1.15	1.210
H8	Neu5Ac	3.294	67.60	1.83	3.96	6.26	7.240	8.22	8.22
H6	Neu5Ac	3.315	72.30	1.83	3.96	6.26	7.240	8.22	8.22
H9	Neu5Ac	3.338	62.02	0.75	1.6	2.41	2.680	2.92	3.01
H5	Gal	3.345	74.51	0.75	1.6	2.41	2.680	2.92	3.01
H4	Neu5Ac	3.376	68.07	0.56	2.55	3.8	4.32	5.2	5.41
H6/6'	Gal	3.410	60.69	0.43	0.69	0.75	0.790	0.790	0.790
H6/6'	αGlcNAc	3.450	60.68	0.43	0.69	0.75	0.790	0.790	0.790
CH2(O)	Chain	3.518	64.39	0.67	1.35	1.35	1.44	1.52	1.73
H5	Neu5Ac	3.552	51.20	0.78	1.74	2.91	3.42	3.96	4.52
H9'	Neu5Ac	3.556	62.02	0.78	1.74	2.91	3.42	3.96	4.520
H7	Neu5Ac	3.590	71.52	0.57	1.49	2.2	2.65	3.22	3.680
H4	Gal	3.636	67.02	0.38	0.7	1.21	1.26	1.46	1.740
H5	αGlcNAc	3.705	70.19	0.29	0.59	0.7	0.78	0.81	1.06
H3	αGlcNAc	3.752	76.96	0.44	0.7	0.75	0.780	0.92	1.07
H3	Gal	3.781	75.33	0.44	0.9	1.51	1.940	2.14	2.49
H4	αGlcNAc	3.959	68.37	0.55	0.79	0.97	1.160	1.16	1.210

Table A.32. Appendix to Figure 5.20 in Sub-section 5.3.1 STF. ^{1}H and ^{13}C assignment and the raw STD (%) at increasing saturation times. Experiments acquired at 293 K.

Bibliography

1. Venter, J. C.; Adams, M. D.; Myers, E. W.; Li, P. W.; Mural, R. J.; Sutton, G. G.; Smith, H. O.; Yandell, M.; Evans, C. A.; Holt, R. A., The sequence of the human genome. *Science* **2001,** *291* (5507), 1304-1351.
2. Goffeau, A.; Barrell, B. G.; Bussey, H.; Davis, R.; Dujon, B.; Feldmann, H.; Galibert, F.; Hoheisel, J.; Jacq, C.; Johnston, M., Life with 6000 genes. *Science* **1996,** *274* (5287), 546-567.
3. Clary, D. O.; Wolstenholme, D. R., The mitochondrial DNA molecule ofDrosophila yakuba: nucleotide sequence, gene organization, and genetic code. *Journal of Molecular Evolution* **1985,** *22* (3), 252-271.
4. Krishna, R. G.; Wold, F., Post-translational modifications of proteins. *Methods in protein sequence analysis*, Springer: 1993; pp 167-172.
5. Seitz, O., Glycopeptide synthesis and the effects of glycosylation on protein structure and activity. *ChemBioChem* **2000,** *1* (4), 214-246.
6. Erdman Jr, J. W.; MacDonald, I. A.; Zeisel, S. H., *Present knowledge in nutrition*. John Wiley & Sons: 2012.
7. Lindhorst, T. K., *Essentials of carbohydrate chemistry and biochemistry*. Wiley-Vch Weinheim: 2007.
8. Varki A, C. R., Esko JD, *et al.*, *Essentials of Glycobiology [Internet]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK310274/*. 3rd edition. ed.; 2015-2017.
9. Murphy, P. V.; André, S.; Gabius, H.-J., The third dimension of reading the sugar code by lectins: design of glycoclusters with cyclic scaffolds as tools with the aim to define correlations between spatial presentation and activity. *Molecules* **2013,** *18* (4), 4026-4053.
10. Taylor, M. E.; Drickamer, K., *Introduction to glycobiology*. Oxford university press: 2011.
11. Lichtenthaler, F. W., 100 Years "Schlüssel-Schloss-Prinzip": What Made Emil Fischer Use this Analogy? *Angewandte Chemie International Edition* **1995,** *33* (23-24), 2364-2374.
12. Gabius, H.-J.; André, S.; Jiménez-Barbero, J.; Romero, A.; Solís, D., From lectin structure to functional glycomics: principles of the sugar code. *Trends in Biochemical Sciences* **2011,** *36* (6), 298-313.
13. Crabtree, G. R.; Olson, E. N., NFAT signaling: choreographing the social lives of cells. *Cell* **2002,** *109* (2), S67-S79.
14. Zeng, X.; Andrade, C. A.; Oliveira, M. D.; Sun, X.-L., Carbohydrate–protein interactions and their biosensing applications. *Analytical and Bioanalytical Chemistry* **2012,** *402* (10), 3161-3176.
15. Mertz, J. A.; McCann, J. A.; Picking, W. D., Fluorescence analysis of galactose, lactose, and fucose interaction with the cholera toxin B subunit. *Biochemical and Biophysical Research communications* **1996,** *226* (1), 140-144.
16. Gill, D. M., Mechanism of action of cholera toxin. *Advances in cyclic nucleotide research* **1977,** *8*, 85.
17. Tailford, L. E.; Owen, C. D.; Walshaw, J.; Crost, E. H.; Hardy-Goddard, J.; Le Gall, G.; De Vos, W. M.; Taylor, G. L.; Juge, N., Discovery of intramolecular trans-sialidases in human gut microbiota suggests novel mechanisms of mucosal adaptation. *Nature Communications* **2015,** *6*, 7624.

18. Juge, N.; Tailford, L.; Owen, C. D., Sialidases from gut bacteria: a mini-review. *Biochemical Society Transactions* **2016,** *44* (1), 166-175.
19. Flynn, J. M.; Niccum, D.; Dunitz, J. M.; Hunter, R. C., Evidence and role for bacterial mucin degradation in cystic fibrosis airway disease. *PLoS Pathogens* **2016,** *12* (8), e1005846.
20. Colman, P. M.; Varghese, J.; Laver, W., Structure of the catalytic and antigenic sites in influenza virus neuraminidase. *Nature* **1983,** *303* (5912), 41-44.
21. Butcher, E. C., Leukocyte-endothelial cell recognition: three (or more) steps to specificity and diversity. *Cell* **1991,** *67* (6), 1033-1036.
22. Stanley, P.; Cummings, R. D., Structures common to different glycans. Cold Spring Harbour: **2017**.
23. Giangrande, P. L., The history of blood transfusion. *British Journal of Haematology* **2000,** *110* (4), 758-767.
24. Adamo, R.; Nilo, A.; Castagner, B.; Boutureira, O.; Berti, F.; Bernardes, G. J., Synthetically defined glycoprotein vaccines: current status and future directions. *Chemical Science* **2013,** *4* (8), 2995-3008.
25. Astronomo, R. D.; Burton, D. R., Carbohydrate vaccines: developing sweet solutions to sticky situations? *Nature reviews Drug discovery* **2010,** *9* (4), 308-324.
26. von Itzstein, M.; Colman, P., Design and synthesis of carbohydrate-based inhibitors of protein—carbohydrate interactions. *Current Opinion in Structural Biology* **1996,** *6* (5), 703-709.
27. Lombard, V.; Golaconda Ramulu, H.; Drula, E.; Coutinho, P. M.; Henrissat, B., The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Research* **2013,** *42* (D1), D490-D495.
28. Lee, Y. C.; Lee, R. T., Carbohydrate-protein interactions: basis of glycobiology. *Accounts of Chemical Research* **1995,** *28* (8), 321-327.
29. Hudson, K. L.; Bartlett, G. J.; Diehl, R. C.; Agirre, J.; Gallagher, T.; Kiessling, L. L.; Woolfson, D. N., Carbohydrate–aromatic interactions in proteins. *Journal of the American Chemical Society* **2015,** *137* (48), 15152-15160.
30. Peysselon, F.; Ricard-Blum, S., Heparin–protein interactions: from affinity and kinetics to biological roles. Application to an interaction network regulating angiogenesis. *Matrix Biology* **2014,** *35*, 73-81.
31. Siebert, H.-C.; André, S.; Lu, S.-Y.; Frank, M.; Kaltner, H.; Van Kuik, J. A.; Korchagina, E. Y.; Bovin, N.; Tajkhorshid, E.; Kaptein, R., Unique conformer selection of human growth-regulatory lectin galectin-1 for ganglioside GM1 versus bacterial toxins. *Biochemistry* **2003,** *42* (50), 14762-14773.
32. Merritt, E. A.; Sarfaty, S.; Akker, F. V. D.; L'Hoir, C.; Martial, J. A.; Hol, W. G., Crystal structure of cholera toxin B-pentamer bound to receptor GM1 pentasaccharide. *Protein Science* **1994,** *3* (2), 166-175.
33. Kopitz, J.; von Reitzenstein, C.; Burchert, M.; Cantz, M.; Gabius, H.-J., Galectin-1 is a major receptor for ganglioside GM1, a product of the growth-controlling activity of a cell surface ganglioside sialidase, on human neuroblastoma cells in culture. *Journal of Biological Chemistry* **1998,** *273* (18), 11205-11211.
34. Dam, T. K.; Brewer, C. F., Multivalent lectin—carbohydrate interactions: energetics and mechanisms of binding. *Advances in Carbohydrate Chemistry and Biochemistry* **2010,** *63*, 139-164.
35. Dam, T. K.; Gerken, T. A.; Brewer, C. F., Thermodynamics of multivalent carbohydrate–lectin cross-linking interactions: Importance of entropy in the bind and jump mechanism. *Biochemistry* **2009,** *48* (18), 3822-3827.
36. Dashnau, J. L.; Sharp, K. A.; Vanderkooi, J. M., Carbohydrate intramolecular hydrogen bonding cooperativity and its effect on water structure. *The Journal of Physical Chemistry B* **2005,** *109* (50), 24152-24159.
37. Breiten, B.; Lockett, M. R.; Sherman, W.; Fujita, S.; Al-Sayah, M.; Lange, H.; Bowers, C. M.; Heroux, A.; Krilov, G.; Whitesides, G. M., Water networks contribute to enthalpy/entropy

compensation in protein–ligand binding. *Journal of the American Chemical Society* **2013,** *135* (41), 15579-15584.
38. Lemieux, R. U., How water provides the impetus for molecular recognition in aqueous solution. *Accounts of Chemical Research* **1996,** *29* (8), 373-380.
39. Fersht, A., *Structure and mechanism in protein science: A Guide to enzyme catalysis and protein folding*. World Scientific: 2017; Vol. 9.
40. Dunitz, J. D., Win some, lose some: enthalpy-entropy compensation in weak intermolecular interactions. *Chemistry & Biology* **1995,** *2* (11), 709-712.
41. Qian, H., Entropy-enthalpy compensation: Conformational fluctuation and induced-fit. *The Journal of Chemical Physics* **1998,** *109* (22), 10015-10017.
42. Grunwald, E.; Steel, C., Solvent reorganization and thermodynamic enthalpy-entropy compensation. *Journal of the American Chemical Society* **1995,** *117* (21), 5687-5692.
43. J.W. Emsley, J. F., Milestones in the first fifty years of NMR. *Progress in Nuclear Magnetic Resonance Spectroscopy* **1995,** *28* (1), 1-9.
44. Purcell, E. M.; Torrey, H. C.; Pound, R. V., Resonance absorption by nuclear magnetic moments in a solid. *Physical Review* **1946,** *69* (1-2), 37.
45. Bloch, F., Nuclear induction. *Physical Review* **1946,** *70* (7-8), 460.
46. Proctor, W.; Yu, F., The dependence of a nuclear magnetic resonance frequency upon chemical compound. *Physical Review* **1950,** *77* (5), 717.
47. Becker, E. D., A brief history of nuclear magnetic resonance. *Analytical Chemistry* **1993,** *65* (6), 295A-302A.
48. Overhauser, A. W., Polarization of nuclei in metals. *Physical Review* **1953,** *92* (2), 411.
49. Gutowsky, H.; McCall, D.; Slichter, C. P., Nuclear magnetic resonance multiplets in liquids. *The Journal of Chemical Physics* **1953,** *21* (2), 279-292.
50. Anet, F. A.; Bourn, A., Nuclear Magnetic Resonance Spectral Assignments from Nuclear Overhauser Effects1. *Journal of the American Chemical Society* **1965,** *87* (22), 5250-5251.
51. Wüthrich, K.; Wider, G.; Wagner, G.; Braun, W., Sequential resonance assignments as a basis for determination of spatial protein structures by high resolution proton nuclear magnetic resonance. *Journal of Molecular Biology* **1982,** *155* (3), 311-319.
52. Felli, I. C.; Brutscher, B., Recent advances in solution NMR: fast methods and heteronuclear direct detection. *ChemPhysChem* **2009,** *10* (9-10), 1356-1368.
53. Levitt, M. H., *Spin dynamics: basics of nuclear magnetic resonance*. John Wiley & Sons: 2001.
54. Claridge, T. D., *High-resolution NMR techniques in organic chemistry*. Elsevier: 2016.
55. Keeler, J., *Understanding NMR spectroscopy*. John Wiley & Sons: 2011.
56. Neuhaus, D., Nuclear overhauser effect. *eMagRes* **2000**.
57. Clarkson, J.; Campbell, I., Studies of protein–ligand interactions by NMR. Portland Press Limited: 2003.
58. Hajduk, P. J.; Olejniczak, E. T.; Fesik, S. W., One-dimensional relaxation-and diffusion-edited NMR methods for screening compounds that bind to macromolecules. *Journal of the American Chemical Society* **1997,** *119* (50), 12257-12261.
59. Jahnke, W., Spin labels as a tool to identify and characterize protein–ligand interactions by NMR spectroscopy. *ChemBioChem* **2002,** *3* (2-3), 167-173.
60. Singh, M.; Tam, B.; Akabayov, B., NMR-Fragment Based Virtual Screening: A Brief Overview. *Molecules* **2018,** *23* (2), 233.
61. Dalvit, C.; Fogliatto, G.; Stewart, A.; Veronesi, M.; Stockman, B., WaterLOGSY as a method for primary NMR screening: practical aspects and range of applicability. *Journal of Biomolecular NMR* **2001,** *21* (4), 349-359.
62. Markwick, P. R.; Malliavin, T.; Nilges, M., Structural biology by NMR: structure, dynamics, and interactions. *PLoS computational biology* **2008,** *4* (9), e1000168.
63. Williamson, M. P., Using chemical shift perturbation to characterise ligand binding. *Progress in Nuclear Magnetic Resonance Spectroscopy* **2013,** *73*, 1-16.

64. Duus, J. Ø.; Gotfredsen, C. H.; Bock, K., Carbohydrate structural determination by NMR spectroscopy: modern methods and limitations. *Chemical Reviews* **2000,** *100* (12), 4589-4614.
65. Nurisso, A.; Kozmon, S.; Imberty, A., Comparison of docking methods for carbohydrate binding in calcium-dependent lectins and prediction of the carbohydrate binding mode to sea cucumber lectin CEL-III. *Molecular Simulation* **2008,** *34* (4), 469-479.
66. Jayalakshmi, V.; Krishna, N. R., Complete relaxation and conformational exchange matrix (CORCEMA) analysis of intermolecular saturation transfer effects in reversibly forming ligand–receptor complexes. *Journal of Magnetic Resonance* **2002,** *155* (1), 106-118.
67. Ramos-Soriano, J.; Niss, U.; Angulo, J.; Angulo, M.; Moreno-Vargas, A. J.; Carmona, A. T.; Ohlson, S.; Robina, I., Synthesis, Biological Evaluation, WAC and NMR Studies of S-Galactosides and Non-Carbohydrate Ligands of Cholera Toxin Based on Polyhydroxyalkylfuroate Moieties. *Chemistry – A European Journal* **2013,** *19* (52), 17989-18003.
68. Manichanh, C.; Borruel, N.; Casellas, F.; Guarner, F., The gut microbiota in IBD. *Nature Reviews Gastroenterology and Hepatology* **2012,** *9* (10), 599-608.
69. Bryant, R. G., The NMR time scale. *Journal of Chemical Education* **1983,** *60* (11), 933.
70. (a) Mayer, M.; Meyer, B., Characterization of ligand binding by saturation transfer difference NMR spectroscopy. *Angewandte Chemie International Edition* **1999,** *38* (12), 1784-1788; (b) 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* **2001,** *123* (25), 6108-6117; (c) Angulo, J.; Nieto, P. M., STD-NMR: application to transient interactions between biomolecules — a quantitative approach. *European Biophysics Journal* **2011,** *40* (12), 1357-1369.
71. Brand, T.; Cabrita, E. J.; Berger, S., Intermolecular interaction as investigated by NOE and diffusion studies. *Progress in Nuclear Magnetic Resonance Spectroscopy* **2005,** *46* (4), 159-196.
72. Mayer, M.; James, T. L., NMR-based characterization of phenothiazines as a RNA binding scaffold. *Journal of the American Chemical Society* **2004,** *126* (13), 4453-4460.
73. Ni, F., Recent developments in transferred NOE methods. *Progress in Nuclear Magnetic Resonance Spectroscopy* **1994,** *26*, 517-606.
74. Post, C. B., Exchange-transferred NOE spectroscopy and bound ligand structure determination. *Current Opinion in Structural Biology* **2003,** *13* (5), 581-588.
75. Andrew, R. L., Molecular modeling principles and applications. *2nd edition: Pearson Education Limited* **2001**.
76. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K., Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of Medicinal Chemistry* **2004,** *47* (7), 1739-1749.
77. Morris, G. M.; Lim-Wilby, M., Molecular docking. *Molecular modeling of proteins*, Springer: 2008; pp 365-382.
78. Enríquez-Navas, P. M. G., C.; Muñoz-Garcia, J.C.; Nieto, P.M.; Angulo J. , Glycan Structures Bound to Receptors from Saturation Transfer Difference NMR Spectroscopy: Quantitative Analysis by using CORCEMA-ST. Springer (Human Press): 2015; pp 475-487.
79. Jayalakshmi, V.; Biet, T.; Peters, T.; Krishna, N. R., Refinement of the conformation of UDP− galactose bound to galactosyltransferase using the STD NMR intensity-restrained CORCEMA optimization. *Journal of the American Chemical Society* **2004,** *126* (28), 8610-8611.
80. Kemper, S.; Patel, M. K.; Errey, J. C.; Davis, B. G.; Jones, J. A.; Claridge, T. D., Group epitope mapping considering relaxation of the ligand (GEM-CRL): including longitudinal relaxation rates in the analysis of saturation transfer difference (STD) experiments. *Journal of Magnetic Resonance* **2010,** *203* (1), 1-10.
81. Leung, I. K.; Demetriades, M.; Hardy, A. P.; Lejeune, C.; Smart, T. J.; Szöllössi, A.; Kawamura, A.; Schofield, C. J.; Claridge, T. D., Reporter ligand NMR screening method for 2-oxoglutarate oxygenase inhibitors. *Journal of Medicinal Chemistry* **2013,** *56* (2), 547-555.
82. Wang, Y. S.; Liu, D.; Wyss, D. F., Competition STD NMR for the detection of high-affinity ligands and NMR-based screening. *Magnetic Resonance in Chemistry* **2004,** *42* (6), 485-489.

83. Becattini, B.; Culmsee, C.; Leone, M.; Zhai, D.; Zhang, X.; Crowell, K. J.; Rega, M. F.; Landshamer, S.; Reed, J. C.; Plesnila, N., Structure–activity relationships by interligand NOE-based design and synthesis of antiapoptotic compounds targeting Bid. *Proceedings of the National Academy of Sciences* **2006,** *103* (33), 12602-12606.
84. Becattini, B.; Pellecchia, M., SAR by ILOEs: An NMR-Based Approach to Reverse Chemical Genetics. *Chemistry - A European Journal* **2006,** *12* (10), 2658-2662.
85. Yang, S.-Y., Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug Discovery Today* **2010,** *15* (11-12), 444-450.
86. Dror, O.; Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J., Predicting molecular interactions in silico: I. an updated guide to pharmacophore identification and its applications to drug design. *Frontiers in Medicinal Chemistry*, Bentham Science Publishers: 2006; Vol. 551, pp 551-584.
87. Monaco, S.; Tailford, L. E.; Juge, N.; Angulo, J., Differential Epitope Mapping by STD NMR Spectroscopy To Reveal the Nature of Protein–Ligand Contacts. *Angewandte Chemie International Edition* **2017,** *56* (48), 15289-15293.
88. Fan, E.; Merritt, E. A.; Zhang, Z.; Pickens, J. C.; Roach, C.; Ahn, M.; Hol, W. G., Exploration of the GM1 receptor-binding site of heat-labile enterotoxin and cholera toxin by phenyl-ring-containing galactose derivatives. *Acta Crystallographica Section D: Biological Crystallography* **2001,** *57* (2), 201-212.
89. Merritt, E. A.; Zhang, Z.; Pickens, J. C.; Ahn, M.; Hol, W. G.; Fan, E., Characterization and Crystal Structure of a High-Affinity Pentavalent Receptor-Binding Inhibitor for Cholera Toxin and E. c oli Heat-Labile Enterotoxin. *Journal of the American Chemical Society* **2002,** *124* (30), 8818-8824.
90. Cheshev, P.; Morelli, L.; Marchesi, M.; Podlipnik, Č.; Bergström, M.; Bernardi, A., Synthesis and affinity evaluation of a small library of bidentate cholera toxin ligands: towards nonhydrolyzable ganglioside mimics. *Chemistry - a European Journal* **2010,** *16* (6), 1951-1967.
91. Bergström, M.; Liu, S.; Kiick, K. L.; Ohlson, S., Cholera Toxin Inhibitors Studied with High-Performance Liquid Affinity Chromatography: A Robust Method to Evaluate Receptor–Ligand Interactions. *Chemical Biology & Drug Design* **2009,** *73* (1), 132-141.
92. Ramos-Soriano, J.; Niss, U.; Angulo, J.; Angulo, M.; Moreno-Vargas, A. J.; Carmona, A. T.; Ohlson, S.; Robina, I., Synthesis, Biological Evaluation, WAC and NMR Studies of S-Galactosides and Non-Carbohydrate Ligands of Cholera Toxin Based on Polyhydroxyalkylfuroate Moieties. *Chemistry - A European Journal* **2013,** *19* (52), 17989-18003.
93. Pîrnău, A.; Bogdan, M., Investigation of the interaction between naproxen and human serum albumin. *Romanian Journal of Biophysics* **2008,** *18* (1), 49-55.
94. Fielding, L.; Rutherford, S.; Fletcher, D., Determination of protein–ligand binding affinity by NMR: observations from serum albumin model systems. *Magnetic Resonance in Chemistry* **2005,** *43* (6), 463-470.
95. Angulo, J.; Enríquez-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 - A European Journal* **2010,** *16* (26), 7803-7812.
96. Bujacz, A.; Zielinski, K.; Sekula, B., Structural studies of bovine, equine, and leporine serum albumin complexes with naproxen. *Proteins: Structure, Function, and Bioinformatics* **2014,** *82* (9), 2199-2208.
97. Owen, C. D.; Tailford, L. E.; Monaco, S.; Šuligoj, T.; Vaux, L.; Lallement, R.; Khedri, Z.; Yu, H.; Lecointe, K.; Walshaw, J., Unravelling the specificity and mechanism of sialic acid recognition by the gut symbiont Ruminococcus gnavus. *Nature Communications* **2017**, *8* (1), 1-15.
98. Kang, Y.; Gohlke, U.; Engström, O.; Hamark, C.; Scheidt, T.; Kunstmann, S.; Heinemann, U.; Widmalm, G. r.; Santer, M.; Barbirz, S., Bacteriophage Tailspikes and Bacterial O-Antigens as a Model System to Study Weak-Affinity Protein–Polysaccharide Interactions. *Journal of the American Chemical Society* **2016**, *183* (29), 9109-9118.

99. Otting, G.; Liepinsh, E., Protein hydration viewed by high-resolution NMR spectroscopy: implications for magnetic resonance image contrast. *Accounts of Chemical Research* **1995,** *28* (4), 171-177.
100. Ulrich, E.; Akutsu, H.; Doreleijers, J.; Harano, Y.; Ioannidis, Y.; Lin, J.; Livny, M.; Mading, S.; Maziuk, D.; Miller, Z., Biological magnetic resonance data bank. *Nucleic Acids Research* **2007,** *36*, 402-408.
101. Han, B.; Liu, Y.; Ginzinger, S. W.; Wishart, D. S., SHIFTX2: significantly improved protein chemical shift prediction. *Journal of Biomolecular NMR* **2011,** *50* (1), 43.
102. Sastry, G. M.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W., Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *Journal of Computer-Aided Molecular Design* **2013,** *27* (3), 221-234.
103. Mayer, M.; James, T. L., Detecting ligand binding to a small RNA target via saturation transfer difference NMR experiments in D2O and H2O. *Journal of the American Chemical Society* **2002,** *124* (45), 13376-13377.
104. Erlanson, D. Poll results: affiliation, metrics, and fragment-finding methods. http://practicalfragments.blogspot.co.uk/2016/10/poll-results-affiliation-metrics-and.html (accessed March 24th, 2018).
105. Simard, J. R.; Zunszain, P. A.; Hamilton, J. A.; Curry, S., Location of high and low affinity fatty acid binding sites on human serum albumin revealed by NMR drug-competition analysis. *Journal of Molecular Biology* **2006,** *361* (2), 336-351.
106. Honoré, B.; Brodersen, R., Albumin binding of anti-inflammatory drugs. Utility of a site-oriented versus a stoichiometric analysis. *Molecular Pharmacology* **1984,** *25* (1), 137-150.
107. Maruthamuthu, M.; Kishore, S., Binding of naproxen to bovine serum albumin and tryptophan-modified bovine serum albumin. *Journal of Chemical Sciences* **1987,** *99* (4), 273-279.
108. Yan, J.; Kline, A. D.; Mo, H.; Shapiro, M. J.; Zartler, E. R., The effect of relaxation on the epitope mapping by saturation transfer difference NMR. *Journal of Magnetic Resonance* **2003,** *163* (2), 270-276.
109. Asensio, J. L.; Ardá, A.; Cañada, F. J.; Jiménez-Barbero, J. s., Carbohydrate–aromatic interactions. *Accounts of Chemical Research* **2012,** *46* (4), 946-954.
110. Capila, I.; Linhardt, R. J., Heparin–protein interactions. *Angewandte Chemie International Edition* **2002,** *41* (3), 390-412.
111. Crost, E. H.; Tailford, L. E.; Monestier, M.; Swarbreck, D.; Henrissat, B.; Crossman, L. C.; Juge, N., The mucin-degradation strategy of Ruminococcus gnavus: The importance of intramolecular trans-sialidases. *Gut Microbes* **2016,** *7* (4), 302-312.
112. Mandal, S.; Mandal, M. D.; Pal, N. K., Cholera: a great global concern. *Asian Pacific Journal of Tropical Medicine* **2011,** *4* (7), 573-580.
113. Colwell, R. R., A voyage of discovery: cholera, climate and complexity. *Environmental Microbiology* **2002,** *4* (2), 67-69.
114. Publication, W., Cholera vaccines: WHO position paper—Recommendations. *Vaccine* **2010,** *28* (30), 4687-4688.
115. Chinnapen, D. J.-F.; Chinnapen, H.; Saslowsky, D.; Lencer, W. I., Rafting with cholera toxin: endocytosis and trafficking from plasma membrane to ER. *FEMS Microbiology Letters* **2006,** *266* (2), 129-137.
116. Sixma, T. K.; Pronk, S. E.; Kalk, K. H.; Wartna, E. S.; Van Zanten, B. A.; Witholt, B.; Hoi, W. G., Crystal structure of a cholera toxin-related heat-labile enterotoxin from E. coli. *Nature* **1991,** *351* (6325), 371.
117. Basu, I.; Mukhopadhyay, C., Insights into binding of cholera toxin to GM1 containing membrane. *Langmuir* **2014,** *30* (50), 15244-15252.
118. Moens, P. D.; Digman, M. A.; Gratton, E., Modes of diffusion of cholera toxin bound to GM1 on live cell membrane by image mean square displacement analysis. *Biophysical Journal* **2015,** *108* (6), 1448-1458.

119. Nunez, M. F.; Veatch, S. L., Clusters of Cholera Toxin B Subunit on the Outer Leaflet Stabilize Lipid Heterogeneity on the Inner Leaflet of B Cell Membranes. *Biophysical Journal* **2016,** *110* (3), 84a.
120. Zhang, R.-G.; Scott, D. L.; Westbrook, M. L.; Nance, S.; Spangler, B. D.; Shipley, G. G.; Westbrook, E. M., The three-dimensional crystal structure of cholera toxin. *Journal of molecular Biology* **1995,** *251* (4), 563-573.
121. Mhalu, F.; Mmari, P.; Ijumba, J., Rapid emergence of El Tor vibrio cholera resistant to antimicrobial agents during first six months of fourth cholera epidemic in Tanzania. *The Lancet* **1979,** *313* (8112), 345-347.
122. Hochhut, B.; Lotfi, Y.; Mazel, D.; Faruque, S. M.; Woodgate, R.; Waldor, M. K., Molecular analysis of antibiotic resistance gene clusters in Vibrio cholerae O139 and O1 SXT constins. *Antimicrobial Agents and Chemotherapy* **2001,** *45* (11), 2991-3000.
123. Beaber, J. W.; Hochhut, B.; Waldor, M. K., Genomic and functional analyses of SXT, an integrating antibiotic resistance gene transfer element derived from Vibrio cholerae. *Journal of Bacteriology* **2002,** *184* (15), 4259-4269.
124. Lauer, S.; Goldstein, B.; Nolan, R. L.; Nolan, J. P., Analysis of cholera toxin− ganglioside interactions by flow cytometry. *Biochemistry* **2002,** *41* (6), 1742-1751.
125. Zhang, R.-G.; Westbrook, M. L.; Westbrook, E. M.; Scott, D. L.; Otwinowski, Z.; Maulik, P. R.; Reed, R. A.; Shipley, G. G., The 2.4 Å crystal structure of cholera toxin B subunit pentamer: choleragenoid. *Journal of Molecular Biology* **1995,** *251* (4), 550-562.
126. Merritt, E. A.; Kuhn, P.; Sarfaty, S.; Erbe, J. L.; Holmes, R. K.; Hol, W. G., The 1.25 Å resolution refinement of the cholera toxin B-pentamer: evidence of peptide backbone strain at the receptor-binding site. *Journal of Molecular Biology* **1998,** *282* (5), 1043-1059.
127. Turnbull, W. B.; Precious, B. L.; Homans, S. W., Dissecting the cholera toxin− ganglioside gm1 interaction by isothermal titration calorimetry. *Journal of the American Chemical Society* **2004,** *126* (4), 1047-1054.
128. Bernardi, A.; Arosio, D.; Potenza, D.; Sánchez-Medina, I.; Mari, S.; Cañada, F. J.; Jiménez-Barbero, J., Intramolecular carbohydrate–aromatic interactions and intermolecular van der Waals interactions enhance the molecular recognition ability of GM1 glycomimetics for cholera toxin. *Chemistry - A European Journal* **2004,** *10* (18), 4395-4406.
129. Arosio, D.; Baretti, S.; Cattaldo, S.; Potenza, D.; Bernardi, A., Ganglioside GM1 mimics: lipophilic substituents improve affinity for cholera toxin. *Bioorganic & Medicinal Chemistry Letters* **2003,** *13* (21), 3831-3834.
130. Bernardi, A.; Arosio, D.; Manzoni, L.; Monti, D.; Posteri, H.; Potenza, D.; Mari, S.; Jiménez-Barbero, J., Mimics of ganglioside GM1 as cholera toxin ligands: replacement of the GalNAc residue. *Organic & Biomolecular Chemistry* **2003,** *1* (5), 785-792.
131. Vrasidas, I.; Kemmink, J.; Liskamp, R. M.; Pieters, R. J., Synthesis and cholera toxin binding properties of a lactose-2-aminothiazoline conjugate. *Organic Letters* **2002,** *4* (10), 1807-1808.
132. Arosio, D.; Vrasidas, I.; Valentini, P.; Liskamp, R. M.; Pieters, R. J.; Bernardi, A., Synthesis and cholera toxin binding properties of multivalent GM1 mimics. *Organic & Biomolecular Chemistry* **2004,** *2* (14), 2113-2124.
133. Ermolinsky, B.; Peredelchuk, M.; Provenzano, D., α-Cyclodextrin decreases cholera toxin binding to GM1-gangliosides. *Journal of Medical Microbiology* **2013,** *62* (7), 1011-1014.
134. Leaver, D. J.; Hughes, A. B.; Dawson, R. M.; Postma, A.; Malic, N.; Polyzos, A., Synthesis of RAFT polymers as bivalent inhibitors of cholera toxin. *RSC Advances* **2014,** *4* (29), 14868-14871.
135. Richards, S. J.; Jones, M. W.; Hunaban, M.; Haddleton, D. M.; Gibson, M. I., Probing Bacterial-Toxin Inhibition with Synthetic Glycopolymers Prepared by Tandem Post-Polymerization Modification: Role of Linker Length and Carbohydrate Density. *Angewandte Chemie International Edition* **2012,** *51* (31), 7812-7816.

136. Jones, M.; Otten, L.; Richards, S.-J.; Lowery, R.; Phillips, D.; Haddleton, D.; Gibson, M., Glycopolymers with secondary binding motifs mimic glycan branching and display bacterial lectin selectivity in addition to affinity. *Chemical Science* **2014,** *5* (4), 1611-1616.
137. Branson, T. R.; McAllister, T. E.; Garcia-Hartjes, J.; Fascione, M. A.; Ross, J. F.; Warriner, S. L.; Wennekes, T.; Zuilhof, H.; Turnbull, W. B., A Protein-Based Pentavalent Inhibitor of the Cholera Toxin B-Subunit. *Angewandte Chemie International Edition* **2014,** *53* (32), 8323-8327.
138. Oleinikovas, V.; Saladino, G.; Cossins, B. P.; Gervasio, F. L., Understanding cryptic pocket formation in protein targets by enhanced sampling simulations. *Journal of the American Chemical Society* **2016,** *138* (43), 14257-14263.
139. Bouguet-Bonnet, S.; Yemloul, M.; Canet, D., New application of proton nuclear spin relaxation unraveling the intermolecular structural features of low-molecular-weight organogel fibers. *Journal of the American Chemical Society* **2012,** *134* (25), 10621-10627.
140. Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T., Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein–ligand complexes. *Journal of Medicinal Chemistry* **2006,** *49* (21), 6177-6196.
141. Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L., Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *Journal of Medicinal Chemistry* **2004,** *47* (7), 1750-1759.
142. Sender, R.; Fuchs, S.; Milo, R., Revised estimates for the number of human and bacteria cells in the body. *PLoS Biology* **2016,** *14* (8), e1002533.
143. Peterson, J.; Garges, S.; Giovanni, M.; McInnes, P.; Wang, L.; Schloss, J. A.; Bonazzi, V.; McEwen, J. E.; Wetterstrand, K. A.; Deal, C., The NIH human microbiome project. *Genome Research* **2009,** *19* (12), 2317-2323.
144. Power, S. E.; O'Toole, P. W.; Stanton, C.; Ross, R. P.; Fitzgerald, G. F., Intestinal microbiota, diet and health. *British Journal of Nutrition* **2014,** *111* (3), 387-402.
145. Arumugam, M.; Raes, J.; Pelletier, E.; Le Paslier, D.; Yamada, T.; Mende, D. R.; Fernandes, G. R.; Tap, J.; Bruls, T.; Batto, J.-M., Enterotypes of the human gut microbiome. *Nature* **2011,** *473* (7346), 174.
146. El Kaoutari, A.; Armougom, F.; Gordon, J. I.; Raoult, D.; Henrissat, B., The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nature Reviews Microbiology* **2013,** *11* (7), 497-504.
147. Icaza-Chávez, M., Gut microbiota in health and disease. *Revista de Gastroenterología de México (English Edition)* **2013,** *78* (4), 240-248.
148. Blaser, M. J., The theory of disappearing microbiota and the epidemics of chronic diseases. *Nature Reviews Immunology* **2017,** *17* (8), nri. 2017.77.
149. Sekirov, I.; Russell, S. L.; Antunes, L. C. M.; Finlay, B. B., Gut microbiota in health and disease. *Physiological Reviews* **2010,** *90* (3), 859-904.
150. Johansson, M. E.; Larsson, J. M. H.; Hansson, G. C., The two mucus layers of colon are organized by the MUC2 mucin, whereas the outer layer is a legislator of host–microbial interactions. *Proceedings of the National Academy of Sciences* **2011,** *108* (Supplement 1), 4659-4665.
151. Ottman, N.; Smidt, H.; De Vos, W. M.; Belzer, C., The function of our microbiota: who is out there and what do they do? *Frontiers in Cellular and Infection Microbiology* **2012,** *2*.
152. Donaldson, G. P.; Lee, S. M.; Mazmanian, S. K., Gut biogeography of the bacterial microbiota. *Nature Reviews Microbiology* **2016,** *14* (1), 20-32.
153. Martens, E. C.; Chiang, H. C.; Gordon, J. I., Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont. *Cell host & Microbe* **2008,** *4* (5), 447-457.
154. Ouwerkerk, J. P.; de Vos, W. M.; Belzer, C., Glycobiome: bacteria and mucus at the epithelial interface. *Best practice & research Clinical gastroenterology* **2013,** *27* (1), 25-38.
155. Juge, N., Microbial adhesins to gastrointestinal mucus. *Trends in Microbiology* **2012,** *20* (1), 30-39.

156. Etzold, S.; Juge, N., Structural insights into bacterial recognition of intestinal mucins. *Current Opinion in Structural Biology* **2014,** *28*, 23-31.
157. Ng, K. M.; Ferreyra, J. A.; Higginbottom, S. K.; Lynch, J. B.; Kashyap, P. C.; Gopinath, S.; Naidu, N.; Choudhury, B.; Weimer, B. C.; Monack, D. M., Microbiota-liberated host sugars facilitate post-antibiotic expansion of enteric pathogens. *Nature* **2013,** *502* (7469), 96-99.
158. Lewis, A. L.; Lewis, W. G., Host sialoglycans and bacterial sialidases: a mucosal perspective. *Cellular Microbiology* **2012,** *14* (8), 1174-1182.
159. Tailford, L. E.; Crost, E. H.; Kavanaugh, D.; Juge, N., Mucin glycan foraging in the human gut microbiome. *Frontiers in Genetics* **2015,** *6*.
160. Vimr, E. R.; Kalivoda, K. A.; Deszo, E. L.; Steenbergen, S. M., Diversity of microbial sialic acid metabolism. *Microbiology and Molecular Biology Reviews* **2004,** *68* (1), 132-153.
161. Moustafa, I.; Connaris, H.; Taylor, M.; Zaitsev, V.; Wilson, J. C.; Kiefel, M. J.; Von Itzstein, M.; Taylor, G., Sialic acid recognition by Vibrio cholerae neuraminidase. *Journal of Biological Chemistry* **2004,** *279* (39), 40819-40826.
162. Boraston, A. B.; Ficko-Blean, E.; Healey, M., Carbohydrate recognition by a large sialidase toxin from Clostridium perfringens. *Biochemistry* **2007,** *46* (40), 11352-11360.
163. Cantarel, B. L.; Coutinho, P. M.; Rancurel, C.; Bernard, T.; Lombard, V.; Henrissat, B., The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Research* **2008,** *37* (suppl_1), D233-D238.
164. Qin, J.; Li, R.; Raes, J.; Arumugam, M.; Burgdorf, K. S.; Manichanh, C.; Nielsen, T.; Pons, N.; Levenez, F.; Yamada, T., A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **2010,** *464* (7285), 59.
165. Crost, E. H.; Tailford, L. E.; Le Gall, G.; Fons, M.; Henrissat, B.; Juge, N., Utilisation of mucin glycans by the human gut symbiont Ruminococcus gnavus is strain-dependent. *PLoS One* **2013,** *8* (10), e76341.
166. Luo, Y.; Li, S.-C.; Li, Y.-T.; Luo, M., The 1.8 Å structures of leech intramolecular trans-sialidase complexes: evidence of its enzymatic mechanism. *Journal of Molecular Biology* **1999,** *285* (1), 323-332.
167. Luo, Y.; Li, S.-C.; Chou, M.-Y.; Li, Y.-T.; Luo, M., The crystal structure of an intramolecular trans-sialidase with a NeuAcα2→ 3Gal specificity. *Structure* **1998,** *6* (4), 521-530.
168. Gut, H.; King, S. J.; Walsh, M. A., Structural and functional studies of Streptococcus pneumoniae neuraminidase B: An intramolecular trans-sialidase. *FEBS Letters* **2008,** *582* (23-24), 3348-3352.
169. Xu, G.; Potter, J. A.; Russell, R. J.; Oggioni, M. R.; Andrew, P. W.; Taylor, G. L., Crystal structure of the NanB sialidase from Streptococcus pneumoniae. *Journal of Molecular Biology* **2008,** *384* (2), 436-449.
170. Xu, G.; Kiefel, M. J.; Wilson, J. C.; Andrew, P. W.; Oggioni, M. R.; Taylor, G. L., Three Streptococcus pneumoniae sialidases: three different products. *Journal of the American Chemical Society* **2011,** *133* (6), 1718-1721.
171. Khan, A.; Leśniak, R.; Brem, J.; Rydzik, A.; Choi, H.; Leung, I.; McDonough, M.; Schofield, C.; Claridge, T., Development and application of ligand-based NMR screening assays for γ-butyrobetaine hydroxylase. *MedChemComm* **2016,** *7* (5), 873-880.
172. Etzold, S.; Kober, O. I.; MacKenzie, D. A.; Tailford, L. E.; Gunning, A. P.; Walshaw, J.; Hemmings, A. M.; Juge, N., Structural basis for adaptation of lactobacilli to gastrointestinal mucus. *Environmental Microbiology* **2014,** *16* (3), 888-903.
173. Gunning, A. P.; Kavanaugh, D.; Thursby, E.; Etzold, S.; MacKenzie, D. A.; Juge, N., Use of atomic force microscopy to study the multi-modular interaction of bacterial adhesins to mucins. *International Journal of Molecular Sciences* **2016,** *17* (11), 1854.
174. Ribeiro, J. P.; Pau, W. K.; Pifferi, C.; Renaudet, O.; Varrot, A.; Mahal, L. K.; Imberty, A., Characterization of a high-affinity sialic acid specific CBM40 from Clostridium perfringens and engineering of divalent form. *Biochemical Journal* **2016**, BCJ20160340.

175. Amaya, M. a. F.; Watts, A. G.; Damager, I.; Wehenkel, A.; Nguyen, T.; Buschiazzo, A.; Paris, G.; Frasch, A. C.; Withers, S. G.; Alzari, P. M., Structural insights into the catalytic mechanism of Trypanosoma cruzi trans-sialidase. *Structure* **2004,** *12* (5), 775-784.
176. Buschiazzo, A.; Amaya, M. a. F.; Cremona, M. a. L.; Frasch, A. C.; Alzari, P. M., The crystal structure and mode of action of trans-sialidase, a key enzyme in Trypanosoma cruzi pathogenesis. *Molecular Cell* **2002,** *10* (4), 757-768.
177. Padler-Karavani, V.; Song, X.; Yu, H.; Hurtado-Ziola, N.; Huang, S.; Muthana, S.; Chokhawala, H. A.; Cheng, J.; Verhagen, A.; Langereis, M. A., Cross-comparison of protein recognition of sialic acid diversity on two novel sialoglycan microarrays. *Journal of Biological Chemistry* **2012,** *287* (27), 22593-22608.
178. Hajduk, P. J.; Mack, J. C.; Olejniczak, E. T.; Park, C.; Dandliker, P. J.; Beutel, B. A., SOS-NMR: a saturation transfer NMR-based method for determining the structures of protein–ligand complexes. *Journal of the American Chemical Society* **2004,** *126* (8), 2390-2398.