Characterisation of structure-to-function relationships in free and protein-linked glycans by computer simulation techniques

Carl Aaron Fogarty B.Sc. (Hons)

A thesis submitted to Maynooth university in fulfilment of the requirement for the degree of.

Doctor of Philosophy

By

Carl Aaron Fogarty B.Sc. (Hons)

Department of Chemistry
Maynooth University

December 2023

Research supervisor:Associate Prof. Elisa Fadda

Head of Department: Prof. Denise Rooney

Table of Contents

Acknowledgements

Firstly, I would like to thank my supervisor Dr. Elisa Fadda for her help and support over these last 5/6 years. From a chance encounter on the train home one day to simulating $\approx$ 700,000 atom systems without your support, this PhD project would not be possible. It has been a pleasure to study with you.

I would also like to thank the other past and present members of the group. Mathew, Aoife, Akash, Silvia, Callum, Ojas and Beatrice. Special thanks to Aoife and Akash for humoring my nonsense for hours at a time during the lockdown, a better research group you will not find.

I would like to thank the Irish Research Council for funding the research presented in this thesis and the Maynooth University Graduate Studies Office and the HEA Cost Extensions for Research Disrupted by COVID-19 for their assistance in finishing off my research.

I also thank the members of staff and researchers from the Chemistry Department and Hamilton Institute (past and present) who have helped me, gave advice, and generally been virtual fountains of knowledge and kindness throughout my stay.

I want to express my gratitude to the postgraduates of the Chemistry Department and the Hamilton Institute. From coffee-break discussions to five-a-side football matches, you immeasurably enriched the overall experience.

I want to express my heartfelt thanks to my parents, Frank and Anita, and my ever supportive dog, Bella (aka Carl #2) who has literally been right behind me for a good portion of the research done. Completing my Ph.D. was a challenging journey, and I couldn't have done it without your unwavering support.

Lastly, Shauna. The love of my life, you deserve a heartfelt acknowledgement for enduring my endless rants and standing by me through the chaotic journey of my Ph.D. Your support, even when the subject matter was incomprehensible nonsense, meant the world to me. Your patience and encouragement made the difficult moments more bearable and I am grateful to have you by my side. Thanks for being the steady force that kept me going.

Declaration

This thesis has not been submitted before, in whole or in part, to this or any other University for any degree, and is, except where otherwise stated, the original work of the author

Carl A. Fogarty

Abbreviations

r_w	Van der Waals Radius
7TM	7 Trans Membrane Regions
ACE2	Angiotensin-Converting Enzyme 2
ADCC	Antibody-Dependent Cellular Cytotoxicity
AI	Artificial Intelligence
Arg	Arginine
Asn	Asparagine
BAI1	Brain-specific Angiogenesis Inhibitor 1
CH	Central Helices
CHARMM	Chemistry at HARvard Macromolecular Mechanics
Cryo-EM	Cryogenic Electron Microscopy
DBSCAN	Density-Based Spatial Clustering of Applications With Noise
EMBL-EBI	European Molecular Biology Labratory, European Bioinformatics Institute
ER	Endoplasmic Reticulum
FA2G2	Biantennary Complex N-Glycan
FcγRIIIa	CD16a Fc γ Receptor
ff14SB9	AMBER Amino Acid Force Field
FFT	Fast Fourier Transform
Fuc	Fucose
GAGs	Glycosaminoglycans
GAIN	Autoproteolysis-Inducing Domain
Gal	Galactose
GLYC	GROMOS 53A6
GLYCAM06j-1	GLYCAM06 Force Field
GNTI	GlcNAc Transferase I
GPCR	G-Protein-Coupled Receptor
HBD	Hormone Binding Domain

HPC	High-Performance Computing
HS	Heparan Sulfate
ICHEC	Irish Centre for High-End Computing
IgG1	Immunoglobulin G1
IRC	Irish Research Council
KDE	Kernel Density Estimate
LeA	Lewis A
LeX	Lewis X
Lys	Lysine
Man	Mannose
Man3	Paucimannose
Man3	Trimannose Core
MD	Molecular Dynamics
ML	Machine Learning
NK	Natural Killer
NMR	Nuclear magnetic Resonance
NPT	Isobaric-Isothermal Ensemble
NVE	Microcanonical Ensemble
NVT	Canonical Ensamble
OST	Oligosaccharyltransferase
PBCs	Periodic Boundary Conditions
PBM	PDZ-Binding Motif
PDB	RCSB Protein Data Bank
PME	Particle Mesh Ewald
PRR	Proline Rich Region
RBD	Receptor-Binding Domain
RBM	Receptor Binding Motif
RGD	Arg-Gly-Asp Integrin-binding Motif
RMSD	Root Mean Square Deviation

ABBREVIATIONS

SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
Ser	Serine
SNFG	Symbol Nomenclature for Glycans
Spike	SARS-CoV-2-S
ST6GAL1	Beta-Galactoside Alpha-2,6-Sialyltransferase
STM	Scanning Tunnelling Microscopy
Thr	Threonine
Trp	Tryptophan
TSR	Thrombospondin Repeat
vdW	Van der Waals
VMD	Visual Molecular Dynamics
Xyl	Xylose

Abstract

Complex carbohydrates (glycans) are the most abundant biopolymers in Nature. They functionalize proteins and lipids and form a thick coating on the surface of cells, which facilitates the cell's movement and its interaction with toxins, viruses, and other cells. This is usually accomplished through a "handshake" recognition between proteins and glycans. Many proteins have their surfaces decorated with glycans (or glycosylated). It is known that the type of glycans present can influence the protein's function and stability. In my extensive research, I employed thorough molecular dynamics (MD) simulations to delve into the subtle yet crucial modifications affecting the N-glycan architecture in various biological contexts.

I investigated modifications involving core α(1-3)-Fuc and β(1-2)-Xyl in standard N-glycoforms found in plants and invertebrates, known to be immunogenic in humans. MD simulations disclosed notable changes in the 3D structure and dynamics of N-glycans, underscoring their pivotal role in selective recognition by lectin receptors and antibodies. The detailed, atomistic-level analysis emphasised that these functionalizations predominantly impact the local spatial vicinity of the modified monosaccharide. Consequently, a novel approach was proposed that employs structural 3D units or glycoblocks to predict the architecture of N-glycans.

My focus shifted to characterising human oligomannose N-glycans free and glycosylated on the CD16a Fc γ Receptor(FcγRIIIa). Through conventional MD simulations, I unraveled a complex architecture shaped by a network of transient hydrogen-bonding interactions. Specific glycoforms exhibited distinct sets of constraints, determining the accessibility for further functionalisation and shedding light on glycoform-specific interactions in modulating antibody-dependent cellular cytotoxicity (ADCC).

Then I investigated the impact of SARS-CoV-2 S N-glycosylation variations on protein function. MD simulations revealed that altering the size of N-glycans at specific sites influenced the stability of receptor binding domain (RBD) conformations, providing insights into the structural dynamics of the virus and potential implications for viral infectivity. Comparative analysis of ancestral sequences suggested the contribution of changes in the topology of the glycan shield to the increased activity of SARS-CoV-2 relative to closely related coronaviruses.

Lastly, I explore the role of C-mannosylation in Thrombospondin Type 1 Repeats (TSR) in proteins, focussing on TSR 1 in BAI1. MD simulations highlighted position-specific effects and the profound influence of the glycan type on stability. The transition between glycan types, including α-mannose, α-rhamnose, α-quinovose, and β-mannose, unveiled nuanced impacts on folding energy and structural dynamics, offering valuable insights for therapeutic protein engineering and drug development.

These findings enhance our understanding of the intricate relationships between

glycan dynamics and protein function, paving the way for novel approaches in glycoscience research and therapeutic interventions.

List of Publications

i Newby, M. L.; **Fogarty, C. A.**; Allen, J. D.; Butler, J.; Fadda, E.; Crispin, M. Variations within the Glycan Shield of SARS-CoV-2 Impact Viral Spike Dynamics. Journal of Molecular Biology 2023, 435, 167928.

ii Harbison, A. M.; **Fogarty, C. A.**; Phung, T. K.; Satheesan, A.; Schulz, B. L.; Fadda, E. Fine-tuning the spike: role of the nature and topology of the glycan shield in the structure and dynamics of the SARS-CoV-2 S. Chemical Science 2022, 13, 386–395.

iii Bagdonas, H.; **Fogarty, C. A.**. A.; Fadda, E.; Agirre, J. The case for post-predictional modifications in the AlphaFold Protein Structure Database. Nature Structural amp; Molecular Biology 2021, 28, 869–870.

iv Zimmerman, M. I. **et al**. SARS-CoV-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome. Nature Chemistry 2021, 13, 651–659.

v **Fogarty, C. A.**. A.; Fadda, E. Oligomannose N-Glycans 3D Architecture and Its Response to the FcRIIIa Structural Landscape. The Journal of Physical Chemistry B 2021, 125, 2607–2616.

vi Casalino, L.; Gaieb, Z.; Goldsmith, J. A.; Hjorth, C. K.; Dommer, A. C.; Harbison, A. M.; **Fogarty, C. A.**. A.; Barros, E. P.; Taylor, B. C.; McLellan, J. S.; Fadda, E.; Amaro, R. E. Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein. ACS Central Science 2020, 6,1722–1734.

vii **Fogarty, C. A.**. A.; Harbison, A. M.; Dugdale, A. R.; Fadda, E. How and why plants and human N-glycans are different: Insight from molecular dynamics into the "glycoblocks" architecture of complex carbohydrates. Beilstein Journal of Organic Chemistry 2020, 16, 2046–2056.2

In BioRχiv

i Ives, C. M.; Nguyen, L.; **Fogarty, C. A.**. A.; Harbison, A. M.; Durocher, Y.; Klassen, J. S.; Fadda, E. GotGlycans: Role of N343 Glycosylation on the SARS-CoV-2 S RBD Structure and Co-Receptor Binding Across Variants of Concern. 2023,

ii Ives, C. M.; Singh, O.; D'Andrea, S.; **Fogarty, C. A.**. A.; Harbison, A. M.; Satheesan, A.; Tropea, B.; Fadda, E. Restoring Protein Glycosylation with GlycoShape. 2023,

1 Chapter 1: Introduction

1.1 Aims and objectives

My Ph.D research, described in this thesis, was aimed at characterising the sequence-to-structure-to-function relationships in glycans through the use of high-performanhelping touting (HPC) simulation. These relationships can be broken down into two main overarching categories:

1. **Sequence to structure**: The relationship between the seshielding. Asnosaccharides and their ensemble of 3D structures of the resulting polysaccharide.

2. **Structure to function**: The biological fin the functional role of N370. to each glycan structure and its direct links to the spatial arrangement and flexibility of the branches and on their conformational flexibility.

Throughout my Ph.D. I performed extensive sampling by molecular dynamics (MD) simulations of both glycans and glycoprotiens. Through my work some of it published in peer-reviewed publications and some of it yet to be published, I have shown that the structure of glycans are not random and that knowing the structure of these glycans gives us insight into the functional role they play when they are linked to glycoproteins. In this chapter, I will introduce the background of my work and describe how the insights I gathered through molecular simulations contributed significantly to our understanding of the relationship between sequence and structure to functional in glycans.

In Chapters 2 to 5 I will present specific case studies that I worked on during my Ph.D. some of which are reproduced verbatim from the published paper. These are.

Sequence to Structure:

Chapter 3: How and why plants and human N-glycans are different. Insight from molecular dynamics into the "glycoblocks" architecture of complex carbohydrates.

Chapter 4: Oligomannose N-Glycan 3D Architecture and Its Response to the FcγRIIIa Structural Landscape.

Structure to Function
Chapter 5: Fine-tuning the spike Role of the nature and topology of the glycan shield in the structure and dynamics of the SARS-CoV-2 S.

In Chapter 6 I present my most recent work on Role of C-Mannosylation in the Folding and Stability of TSR Domains TSR Domains and C-Mannose. We hope to have these results published in the near future.

The Introduction, Computational Methods, and Discussion and Conclusions are original works to this thesis.

1.2   Glycans

Complex carbohydrates, also known as glycans, are the most abundant biomolecule in nature. They are highly diverse and serve a wide range of biological purposes. Monosaccharides are the basic unit of glycans. They have the general formula $(CH_2O)_n$ and are classified formally as polyhydroxyl aldehyde and/or ketones, see Figure 1.1. The differences between natural monosaccharides can be extremely subtle and often consists of only symmetry. For example, sugars can be D or L based on the absolute configuration of the sterogenic carbon centre farthest carbon atom from carbonyl carbon, which is known as anomeric carbon; see Figure 1.1
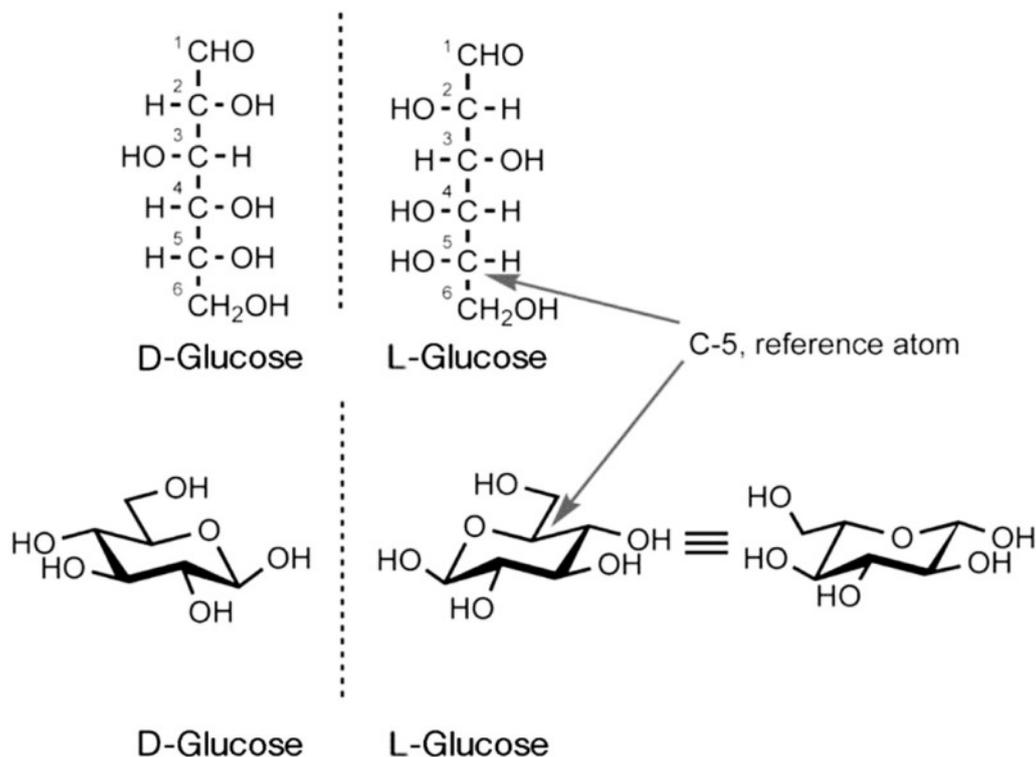
Figure 1.1: The two configurations of glucose, D- and L-, are shown as its acyclic form (illustrated as the Fischer projection above) and its cyclic form (in the $^4C_1$ conformation) [1].

In solution, these monosaccharides are found in cyclic forms, which are five- or six-membered rings shown in Figure 1.1. These rings adopt major conformations known as "Chairs" which can be in $^4C_1$ or $^1C_4$ [2] (depending on which carbon points up / down) and will frequently interchange between other less populated conformations under normal physiological conditions in a complex energy landscape [3]. A Cremer-Pople sphere [4] is a three-dimensional representation of a monosaccharide molecule, which is used to visualise the spatial arrangement of atoms in the molecule. It is a useful tool for understanding the structure of monosaccharides when the Stodart diagram is superimposed on it, where $^4C_1$ and $^1C_4$ at their respective poles; see Figure 1.2 which shows the possible arrangements that a pyranose may take.
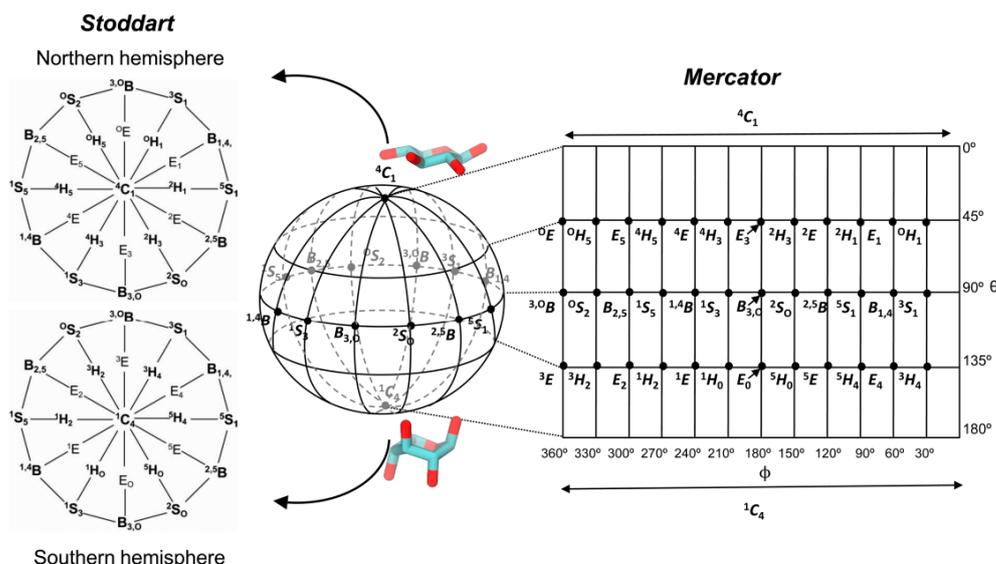
Figure 1.2: Main ring conformations a pyranose may take shown on a Cremer- Pople sphere with the Stoddart diagram projected onto it. Adapted from reference [5]

Cyclisation of pyranose provides an additional asymmetric sterocentre in the C1 anomeric carbon, which means that monosaccharides can be described as (α) or (β) depending on the position of the hydroxyl group of C1 relative to the plane of the ring. Monosaccharides with different stereochemistry in any of their chiral carbon atoms, in a single chiral carbon, are called epimers [6].

The formation of oligosaccharides and polysaccharides is enabled via a condensation reaction between different or identical monosaccharides forming glycosidic bonds. This wide range of possibilities in glycosidic linkages is due not only to the selection of monosaccharides, but also to the way in which they are bonded. If there where only one way to link monosaccharides, the number of potential glycans would be comparable to that of amino/nucleic acids. However, the ability to make glycosidic bonds between the anomeric carbon of one sugar and any other hydroxyl group in another mono or oligosaccharide greatly increases the number of potential glycans, allowing for the formation of both linear and branched products in which more than one hydroxyl group on a given sugar is used to make glycosidic bonds. Furthermore, each anomeric carbon is a stereogenic centre (as described above), which means that each glycosidic linkage can be constructed in configuration α or β. Examples of the branched nature of glycans are shown in Figure 1.3.

Taking into account the two anomeric configurations, the combination of two different hexoses, such as glucose and galactose, leads to sixteen theoretically possible disaccharides. As the number of monosaccharides in an oligo- or polysaccharide increases, the number of possible regioisomers and stereoisomers also increases exponentially [7] resulting in an immense number of distinct glycans. Fortunately, glycan sequences are not randomly synthesised in biological systems, indicating that sequence and branching are essential for glycan functions in biology.
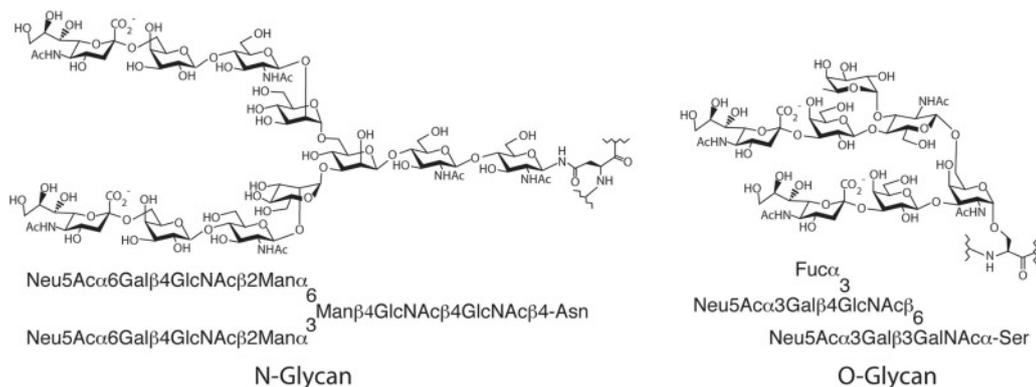
Figure 1.3: Examples showing the branched nature of oligosaccharides / glycans adapted from reference [1]

1.3   Glycosylation

Glycans form covalent bonds with other biomolecules, such as proteins [1], lipids [8, 9], and nucleic acids [10], to form glycoconjugates. In this work I will discuss glycosylation of amino acids and proteins. Depending on the particular amino acid, there are different types of glycosylation that may occur. In this thesis, I will primarily explore N-linked glycosylation, which is so named because the glycan is attached to the amide group of an asparagine amino acid residue (Asn). The topic of glycosylation is described in Chapters 3 and in 4. Furthermore, O-linked glycosylation is where the glycan is chemically bonded to the protein through a hydroxyl group of a serine amino acid residue (Ser) or a threonine amino acid residue (Thr) in a protein. There is also a less common form of glycosylation known as C-linked glycosylation [11], where mannose is linked to the CD1 atom of a tryptophan residue, which is the topic discussed in Chapter 6.

Protein-linked glycans can be modified by extending one monosaccharide at a time via glycosyltransferases or, in the case of N-glycosylation, by a combination of extension via glycosyltransferases and fragment trimming with glycosylhydrolase. The sequence of a glycan is determined by a competition between many glycan processing enzymes that can select the same substrate in a stochastic manner. As a result, glycosylation can be classified as microheterogenous [12]. Consequently, factors such as accessibility of the glycosylation site, expression levels of glycan processing enzymes, availability of sugar donors, and the localisation of enzymes within cellular organelles are all important in determining the glycan sequence, and the glycan sequence can also vary even within the same glycosylation site leading to what is known as macroheterogeneity of the glycoforms. Glycosylation follows an extremely complex order of events tightly regulated by enzymatic reactions [13–16], supporting the notion that there are sequence-structure-function relationships at play here.

N- and O-linked glycosylation are the two most common types of post-translational modification in eukaryotes. Both processes occur in the cell secretory pathway, which involves the movement of newly synthesised glycoproteins through the endoplasmic reticulum and the Golgi apparatus. N- and O- linked glycans may be similar in the sense that they are both complex carbohydrate, they differ in the type of sugar that is attached to the protein; see Figure 1.3. There are considerable distinctions between N- and O-linked glycosylations. To start, O-linked glycosylation forms an O-linked bond to serine (Ser) or threonine (Thr) residues. The synthetic pathway is initiated primarily by enzymes located

in the Golgi apparatus, where the addition of sugar to the newly formed *O*-glycosylation site is an enzymatic stepwise addition from various cores [17] facilitated exclusively by glycosyltransferases. *N*-linked glycosylation is facilitated by the action of glycosidases and glycosyltransferases. A distinctive feature of *N*-linked glycan production is the partial removal of fragments by glycosidase enzymes to make way for glycosyltransferases to further decorate mature *N*-glycans along a nontemplated pathway.[12]. The *N*-linked glycosylation process starts in the endoplasmic reticulum, where a 15 monosaccharide long intermediate glycan is attached, the *N*-glycan is transferred 'en bloc' to a newly translated glycoprotein through the oligosaccharyltransferase (OST) complex and then down by mannosidases, which has a consensus sequence of Asn-X-Ser/Thr (where X is any amino acid except proline) [18]. The eventual transfer to the Golgi apparatus for further and more elaborate processing [19] see, Figure 1.4.
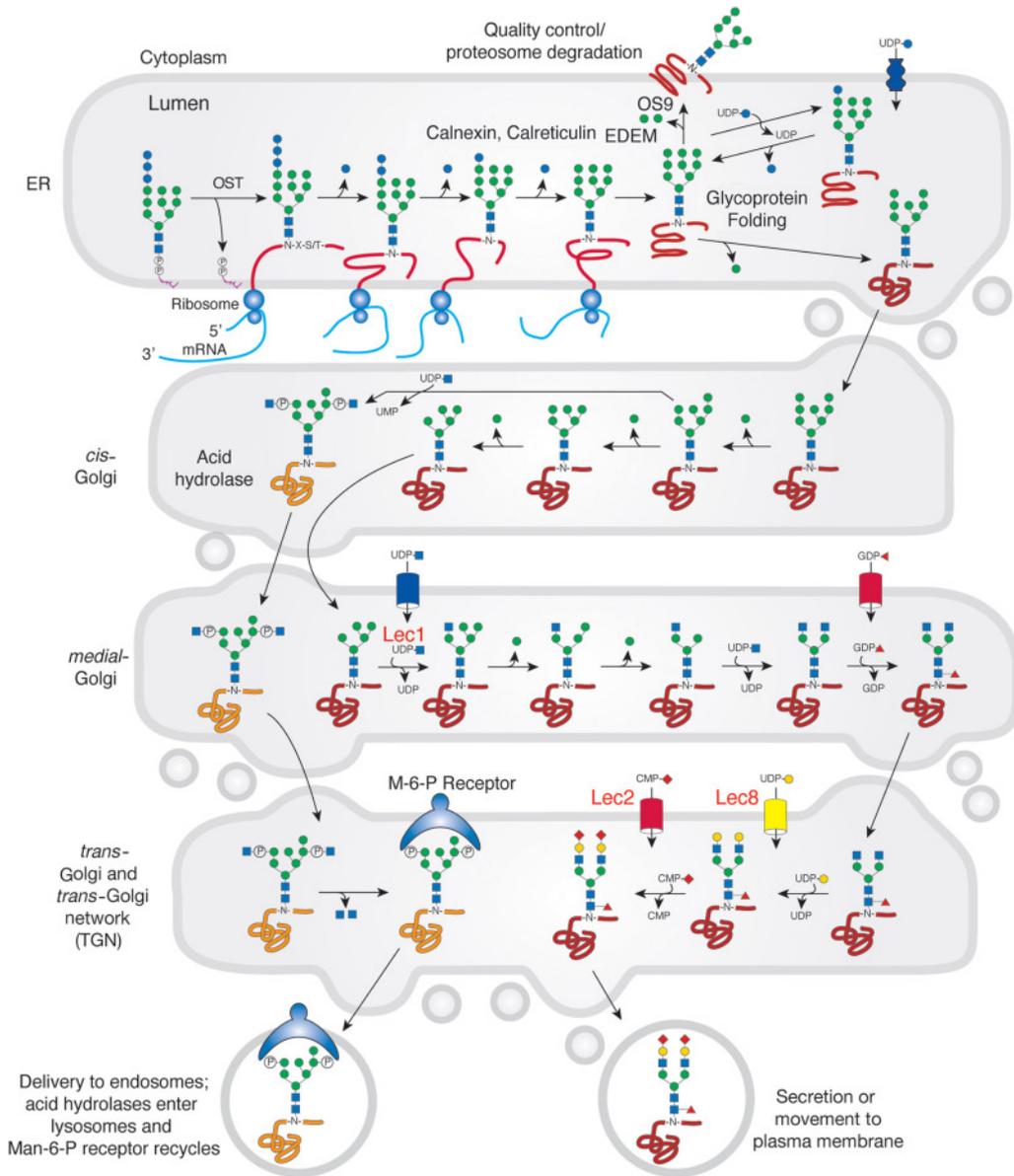
Figure 1.4: Biosynthesis of N-Glycans showing the transfer of a GlcNAc3Man 9 *en bloc* to a translated protein in the rough ER which is then taken into the Golgi apparatus to be further functionalised. Adapted from reference [1]

N-linked glycans have a common pentasaccharide core that is not affected by glycosidase enzymes; see Figure 1.5, unlike *O*-linked glycans that may have multiple core structures [17], due to the exclusive action of glycosyltransferases. *O*-glycosylation does not yet have a known consensus sequence protein sequence.
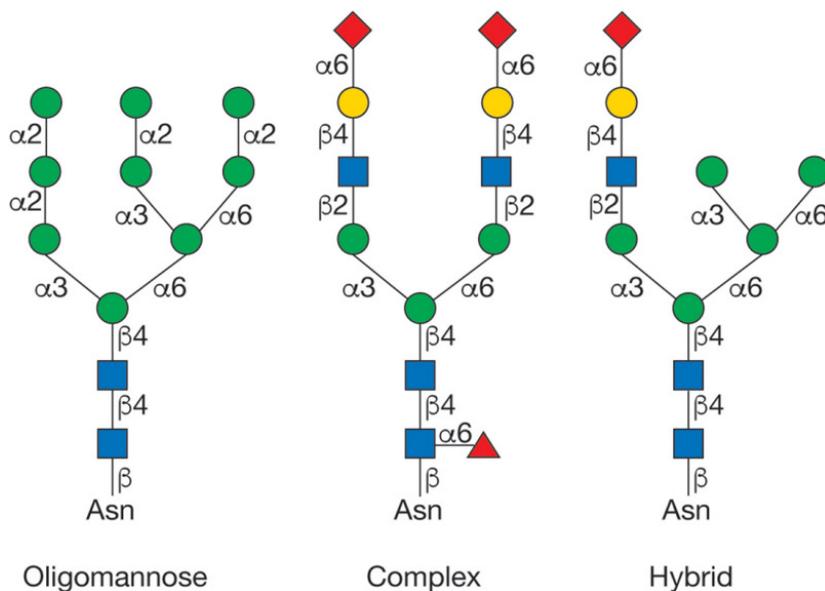
Figure 1.5: Main types of *N*-glycosylation. To note how all types share a common pentasaccharide core linked to the protein through an Asn within an NXS/T sequon where X is any amino acid but Pro. Adapted from reference [1]

1.4 Studying Glycans

The biological pathway that regulates *N*-glycosylation is awesome, complex and much less is known about it than an idealistic diagram may suggest; for example, the 3D structure of the OST complex has only been resolved relatively recently via cryogenic electron microscopy (cryo-EM) [20]. Biological pathways are dependent on intricate networks of molecular interactions that are regulated thermodynamically, with the fit between binding partners and structural complementary components being the primary drivers in recognition. Recent developments in structural biology have been instrumental in helping us gain insight into these dynamic events at the atomic level, providing us with a better understanding of the biological processes that they control. Recent advances in cryo-EM have become a powerful tool for the determination of macromolecular structures [21–24]. These additional advancements have allowed for the visualisation of some of the larger complexes [20, 25, 26], in nature: in addition to X-ray crystallography and nuclear magnetic resonance (NMR), which remain the most commonly used techniques to deposit structures in the RCSB Protein Data Bank (PDB) (PDB; www.rcsb.org).

Recently, a remarkable achievement in biomolecular structure determination was reached with the help of high-performance computing. Advanced machine learning (ML) algorithms were created to accurately predict the three-dimensional structure of proteins from their amino acid sequence [27–29]. Carrying on this momentum Google's DeepMind in collaboration with European Molecular Biology Labratory, European Bioinformatics Institute (EMBL-EBI) created a database of predicted 3D structures, which covers the entirety of the human proteome and the proteomes of chosen model organisms, has been

the result of a great effort. This database currently includes 48 complete proteomes [30].

All of these remarkable breakthroughs have one thing in common: protein structure. Glycans, which are the focus of this thesis, are often not present or partially resolved, often in distorted, peculiar, or even implausible forms [31]. The lack of glycans is not caused by a collective blindness of experimental techniques from the entire structural biology community. But it is due to the fact that glycans are simply very difficult to characterise by traditional structural biology techniques such as X-ray crystalography and cryo-EM [32]. It is difficult to crystallise glycoproteins and in an effort to make crystallisation easier, it is common to remove partial or complete glycans from proteins[32, 33]. In terms of structural characterization by cryo-EM, the rapid conformational dynamics and microheterogeneity of glycans make it difficult to identify them [12], with the resulting structures usually showing only the first or two least flexible monosaccharides directly linked to conformationally stable protein regions. Very recent advances in scanning tunnelling microscopy (STM) have allowed the observation of glycans bonded to proteins and lipids at the single-molecule level [34].

The lack of experimental data on the shapes of glycans can now be remedied by means of library-based 3D structural approaches. Through a proof-of-concept implementation in Privateer [35], equilibrium glycan structures obtained from conformational sampling can be chemically linked to target proteins[36]. Atomistic molecular simulations can provide information on the specific dynamics of different glycan structures [37–39], and how they can be used directly to complement protein structures.

Recent technological advances in high-performance computing and the sophistication of empirical all-atom additive forcefields have enabled classical MD or stochastic simulations to generate informative 3D glycan structural ensembles. Through complete or sufficient sampling based on MD simulations (details of which are discussed in Chapter 2) of free glycans in aqueous solution [38], it has been revealed that glycans are highly flexible biomolecules, yet their conformations are not random. Furthermore, their 3D conformational space is sequence-dependent and thus glycan-specific. This is the first topic that my research has helped to elucidate.

1.5 Sequence-to-Structure Relationship in Glycans

Glycans can be considered to be intrinsically disorderd molecules because of their lack of structural data from the experiments and their dynamic nature. The complexity and variety of their structural arrangements are the consequence of a combination of factors, such as the electronic structure of the monosaccharide rings and the non-covalent intramolecular interactions between glycans, which depend on the sequence and branching [40]. The glycosidic bonds are the most dynamic of the covalent linkages. The complexity of glycans is further increased by the fact that each major conformer has a high level of flexibility, allowing the torsion angles of the two-bond glycosidic linkages to be highly flexible, namely ϕ (O1-C1-Ox-Cx) and ψ (C1-Ox-Cx-Cx + 1), and the three-bond glycosidic linkages, namely ϕ, ψ, and ω (O6-C6-C5-C4), can vary within a range between 10 ° and 20 ° [38, 40] What the research group has uncovered is that we can quantify this degree of conformational disorder by clustering analysis of MD data from exhaustive sampling.

In Chapter 3, I examine the consequences of sequence alterations, such as the addition of β(1-2)-linked Xyl to the central Man of the core pentasaccharide, A modification commonly seen in plants and complex N-glycans of invertebrates. For example, we show through multimicrosecond cumulative MD sampling revealed that the inclusion of Xyl significantly shifts the conformational equilibrium.
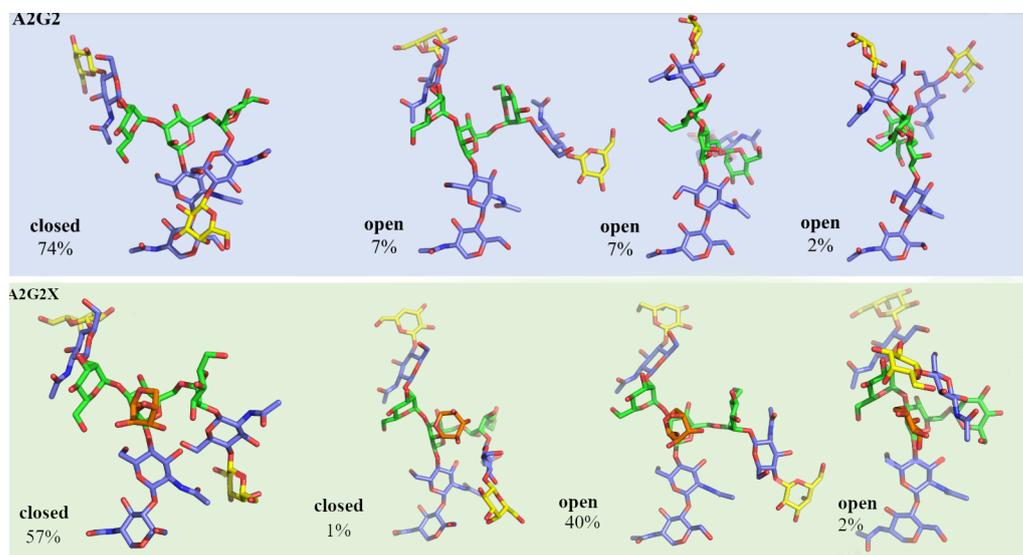
Figure 1.6: The top panel in blue shows the most populated structures of the mammalian biantennary complex N-glycan A2G2. The most common conformation is closed, which is characterised by the 1-6 arm folded over the chitobiose core (1-6 arm torsion angle values: $\phi = 75°(\pm12)$; $\psi= 93°(\pm11)$; $\omega= 48°(\pm8)$; pop 74%). This was determined by three MD trajectories of 250 ns each, for a total of 750 ns, as reported in Ref.[40]. The bottom panel in green displays the highest populated structures of the mammalian xylosilated complex N-glycan. The most open conformations are far more enriched due to xyliose with open torsion angles $\phi = 70°(\pm9)$; ψ= -174°(±15); $\omega= 56°(\pm8)$; pop 40%); and fold arm torsion angle values (1-6):$\phi = 72°(\pm9)$; $\psi= 104°(\pm11)$; $\omega= 56°(\pm7)$, pop 57%).

An illustration of the sequence-to-structure relationship is presented in Figure 1.7 and the lower panel is an introduction to the work done in Chapter 4 of the CD16a molecule. The Man5 N-glycan located at N45 of the human CD16a molecule has a single favoured conformation while being bound to the protein out of the four potential options. The 1-6 arm is able to fit in a space between two structural domains of the CD16a molecule and as a result selects that conformation preferably. Man5 at position N162 is exposed to the solvent and has the same conformational characteristics as when it is not bound to anything. Extensive sampling of Man5 in N45 revealed its ability and inclination to insert the arm 1-6 into accessible clefts, which explains why this site has been discovered to be especially abundant in hybrid N-glycans [41]
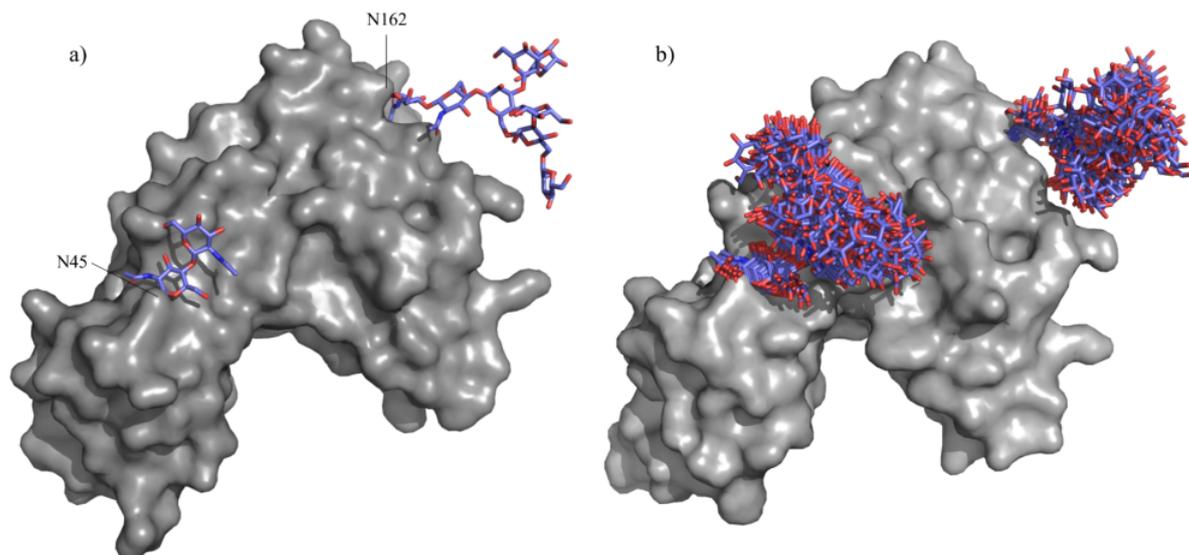
Figure 1.7:   Panel a) The human CD16a (FcγRIIIa, PDB 3SGK) protein is depicted in grey with a solvent-accessible surface, and the *N*-glycan fragment at Asn 45 and a Man6 structure at N162 in a distorted shape (sticks with purple C atoms and red O atoms). Panel b) Structural snapshots taken every 45 ns from a 1 µs MD trajectory of the human CD16a, with *N*-glycosylation at Asn 45 and Asn 162 with Man5 (in sticks), were generated by matching Man5 3D equilibrium structures obtained in solution. The graphical representation was done with PyMOL (www.pymol.org).

1.6   Structure to Function Relationships

An example of structure-to-function relationships in glycans is discussed in Chapter 5, dedicated to the study of the *N*-linked glycosylation of SARS-CoV-2 spike (S) glycoprotein. The SARS-CoV-2 S protein is densely covered by *N*-glycans, also known as a glycan shield. This shield may have evolved for a variety of reasons, such as helping fold the spike, maintaining structural integrity [42, 43], and providing a protective barrier against recognition by the adaptive immune system, also known as shielding. As a key functional change in this shield, which appeared for the first time in SARS-CoV-2, the T372A mutation caused the loss of glycosylation at N370 functional role. MD simulations presented and discussed in Chapter 5 show that the presence of N370 stabilises the receptor-binding domain (RBD) in a closed conformation. The lack of a N370 glycan site in SARS-CoV-2 can be explained by this.

The spatial availability of glycosylation sites has a direct correlation with the local topology of the glycoprotein [12, 44]. MD simulations carried out in the context of the S glycoprotein have revealed that the *N*-glycan at position N234 is among the least accessible in both the closed and the open S conformations, see Figure 1.8. This explains why large oligomannose-type glycans are so abundant at this site [42, 45]. MD simulations presented in Chapter 5 demonstrate that Man 9 stabilises the open conformation and its motion of the RBD [42]. When MD simulations were performed [42], it was observed that the N234 glycan is capable of entering and filling the gap left by the opening of the RBD, thus stabilising it. This is also supported by the N165 and N343 glycans, which

are shown in Figure 1.8. In fact, deleting these glycans through single-point mutations of N234A, N165A, and N343A leads to a significant decrease in binding to the ACE2 receptor [42, 43, 46]. No single-point mutation was able to completely eliminate binding, which implies that multiple residues are involved in the opening of the RBD [42, 43]. MD and cryo-EM studies have revealed that glycosylation in N343 is essential to control the opening and closing of the RBD [43]. Combined affect of *N*-glycans at glycosylation sites N165, N234, and N343 with the surrounding protein help maintain the orientation and movement of the open RBD conformation, allowing for various orientations, and thus facilitates receptor recognition.
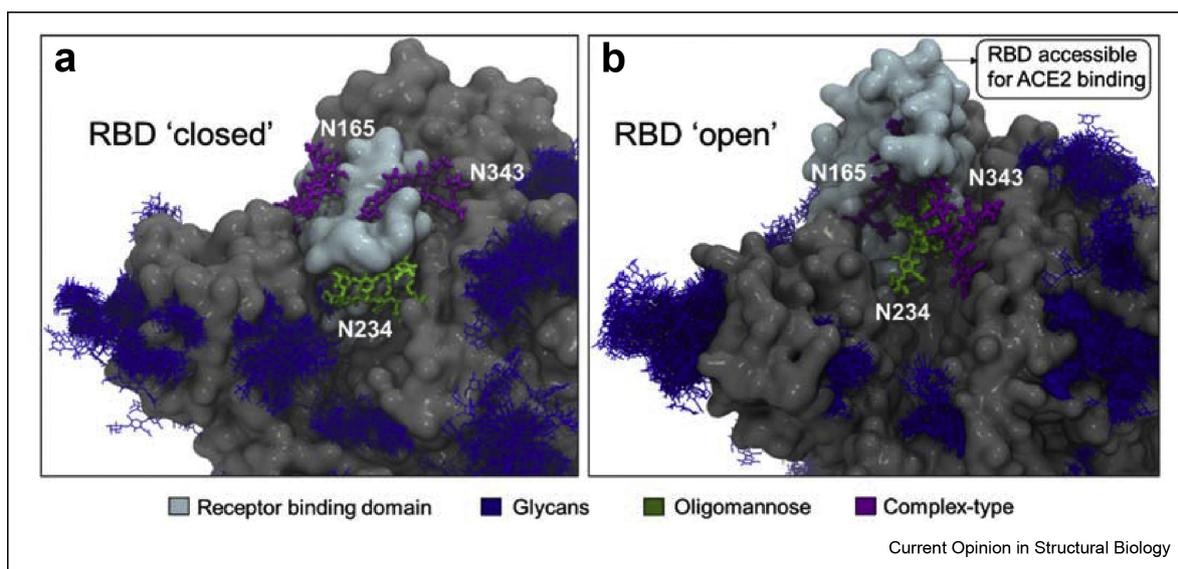
Figure 1.8: Role of glycans in stabilizing dynamics of RBD of spike glycoprotein adapted from reference [47]

Thrombospondin repeat (TSR) proteins are known for their role in cell-matrix interactions and the regulation of cellular activities. In Chapter 6, we present a novel Structure to Function relationship involving C glycosylation patterns. The MD simulations were conducted to investigate the structural and dynamic characteristics of these proteins in the presence of different glycans, such as α-Man, β-Man, α-Rha, α-Qui. The results show that the ideal alpha mannose encourages folding. This knowledge not only enhances our understanding of the biological functions of TSR motifs and their PTMs but also lays the foundation for the design of proteins with tailored glycans and protein folding patterns for various applications, including therapeutic protein engineering.

C-Mannose   C-Rhamnose   C-Quinovose
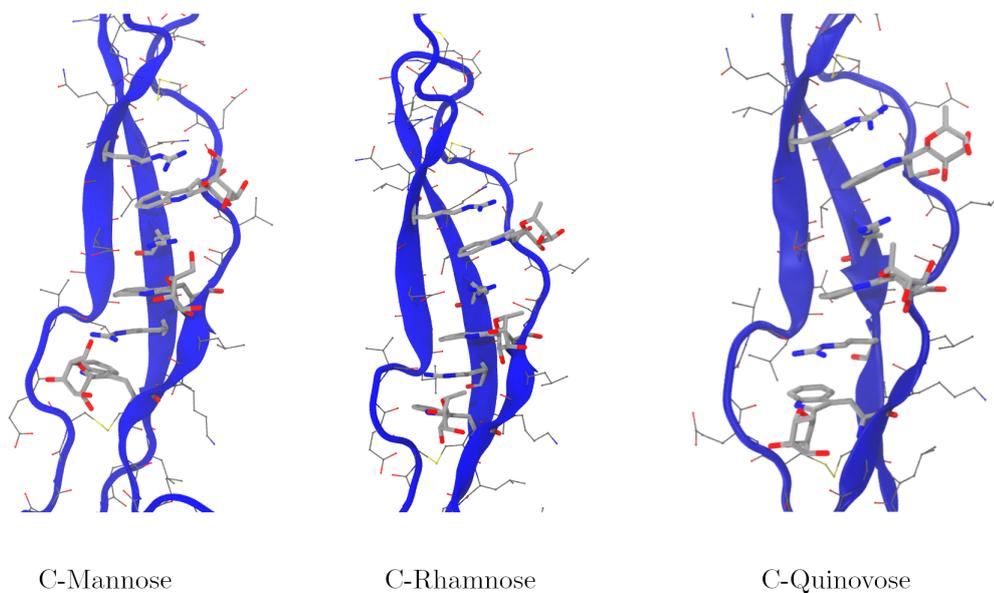
Figure 1.9: The graph shows three distinct TSR domains, each with its own C-glycosylation pattern. The first domain is decorated with C-mannose, the second with C-rhamnose, forming a unique glycan structure, and the third with C-quinovose, demonstrating how the glycan residues stabilise the respective TSR domain.

Bibliography

[1] Stanley, P. In Essentials of Glycobiology, 3rd ed.; Varki, A., Ed.; Cold Spring Harbor Laboratory Press, p 99–112.

[2] Conformational Nomenclature for Five and Six-Membered Ring Forms of Monosaccharides and Their Derivatives. Recommendations 1980. European Journal of Biochemistry **1980**, 111, 295–298.

[3] Mayes, H. B.; Broadbelt, L. J.; Beckham, G. T. How Sugars Pucker: Electronic Structure Calculations Map the Kinetic Landscape of Five Biologically Paramount Monosaccharides and Their Implications for Enzymatic Catalysis. Journal of the American Chemical Society **2014**, 136, 1008–1022.

[4] Cremer, D. t.; Pople, J. General definition of ring puckering coordinates. Journal of the American Chemical Society **1975**, 97, 1354–1358.

[5] Davies, G. J.; Planas, A.; Rovira, C. Conformational Analyses of the Reaction Coordinate of Glycosidases. Accounts of Chemical Research **2011**, 45, 308–316.

[6] Witczak, Z. J. Glycoscience; Springer Berlin Heidelberg, 2008; pp 815–840.

[7] Laine, R. A. Invited Commentary: A calculation of all possible oligosaccharide isomers both branched and linear yields 1.05 x 10(12) structures for a reducing hexasaccharide: the Isomer Barrier to development of single-method saccharide sequencing or synthesis systems. Glycobiology **1994**, 4, 759–767.

[8] Grant, C. W. Model membranes bearing glycolipids and glycoproteins. Chemistry and Physics of Lipids **1986**, 40, 285–302.

[9] Guo, Z. The structural diversity of natural glycosphingolipids (GSLs). Journal of Carbohydrate Chemistry **2022**, 41, 63–154.

[10] Flynn, R. A.; Pedram, K.; Malaker, S. A.; Batista, P. J.; Smith, B. A.; Johnson, A. G.; George, B. M.; Majzoub, K.; Villalta, P. W.; Carette, J. E.; Bertozzi, C. R. Small RNAs are modified with N-glycans and displayed on the surface of living cells. Cell **2021**, 184, 3109–3124.e22.

[11] Crine, S. L.; Acharya, K. R. Molecular basis of C-mannosylation – a structural perspective. The FEBS Journal **2021**, 289, 7670–7687.

[12] Zacchi, L. F.; Schulz, B. L. N-glycoprotein macroheterogeneity: biological implications and proteomic characterization. Glycoconjugate Journal **2015**, 33, 359–376.

[13] Urban, J.; Jin, C.; Thomsson, K. A.; Karlsson, N. G.; Ives, C. M.; Fadda, E.; Bojar, D. Predicting glycan structure from tandem mass spectrometry via deep learning. **2023**,

[14] Lundstrøm, J.; Korhonen, E.; Lisacek, F.; Bojar, D. LectinOracle: A Generalizable Deep Learning Model for Lectin–Glycan Binding Prediction. Advanced Science **2021**, 9.

[15] McDonald, A. G.; Mariethoz, J.; Davey, G. P.; Lisacek, F. In silico analysis of the human milk oligosaccharide glycome reveals key enzymes of their biosynthesis. Scientific Reports **2022**, 12.

[16] Duvaud, S.; Gabella, C.; Lisacek, F.; Stockinger, H.; Ioannidis, V.; Durinx, C. Expasy, the Swiss Bioinformatics Resource Portal, as designed by its users. Nucleic Acids Research **2021**, *49*, W216–W227.

[17] Wilkinson, H.; Saldova, R. Current Methods for the Characterization of O-Glycans. Journal of Proteome Research **2020**, *19*, 3890–3905.

[18] Mellquist, J. L.; Kasturi, L.; Spitalnik, S. L.; Shakin-Eshleman, S. H. The Amino Acid Following an Asn-X-Ser/Thr Sequon Is an Important Determinant of N-Linked Core Glycosylation Efficiency. Biochemistry **1998**, *37*, 6833–6837.

[19] Kornfeld, R.; Kornfeld, S. ASSEMBLY OF ASPARAGINE-LINKED OLIGOSAC-CHARIDES. Annual Review of Biochemistry **1985**, *54*, 631–664.

[20] Ramírez, A. S.; Kowal, J.; Locher, K. P. Cryo–electron microscopy structures of human oligosaccharyltransferase complexes OST-A and OST-B. Science **2019**, *366*, 1372–1375.

[21] Callaway, E. Revolutionary cryo-EM is taking over structural biology. Nature **2020**, *578*, 201–201.

[22] Scheres, S. H. RELION: Implementation of a Bayesian approach to cryo-EM structure determination. Journal of Structural Biology **2012**, *180*, 519–530.

[23] Danev, R.; Yanagisawa, H.; Kikkawa, M. Cryo-Electron Microscopy Methodology: Current Aspects and Future Directions. Trends in Biochemical Sciences **2019**, *44*, 837–848.

[24] Cheng, Y.; Glaeser, R. M.; Nogales, E. How Cryo-EM Became so Hot. Cell **2017**, *171*, 1229–1231.

[25] Guo, H.; Suzuki, T.; Rubinstein, J. L. Structure of a bacterial ATP synthase. eLife **2019**, *8*.

[26] Ognjenović, J.; Grisshammer, R.; Subramaniam, S. Frontiers in Cryo Electron Microscopy of Complex Macromolecular Assemblies. Annual Review of Biomedical Engineering **2019**, *21*, 395–415.

[27] AlQuraishi, M. AlphaFold at CASP13. Bioinformatics **2019**, *35*, 4862–4865.

[28] Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature **2021**, *596*, 583–589.

[29] Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science **2021**, *373*, 871–876.

[30] Tunyasuvunakool, K. et al. Highly accurate protein structure prediction for the human proteome. Nature **2021**, *596*, 590–596.

[31] Agirre, J.; Davies, G.; Wilson, K.; Cowtan, K. Carbohydrate anomalies in the PDB. Nature Chemical Biology **2015**, *11*, 303–303.

[32] Prestegard, J. H. A perspective on the PDB's impact on the field of glycobiology. Journal of Biological Chemistry **2021**, *296*, 100556.

[33] Chang, V. T.; Crispin, M.; Aricescu, A. R.; Harvey, D. J.; Nettleship, J. E.; Fennelly, J. A.; Yu, C.; Boles, K. S.; Evans, E. J.; Stuart, D. I.; Dwek, R. A.; Jones, E. Y.; Owens, R. J.; Davis, S. J. Glycoprotein Structural Genomics: Solving the Glycosylation Problem. Structure 2007, 15, 267–273.

[34] Anggara, K.; Sršan, L.; Jaroentomeechai, T.; Wu, X.; Rauschenbach, S.; Narimatsu, Y.; Clausen, H.; Ziegler, T.; Miller, R. L.; Kern, K. Direct observation of glycans bonded to proteins and lipids at the single-molecule level. Science 2023, 382, 219–223.

[35] Agirre, J.; Iglesias-Fernández, J.; Rovira, C.; Davies, G. J.; Wilson, K. S.; Cowtan, K. D. Privateer: software for the conformational validation of carbohydrate structures. Nature Structural & Molecular Biology 2015, 22, 833–834.

[36] Bagdonas, H.; Fogarty, C. A.; Fadda, E.; Agirre, J. The case for post-predictional modifications in the AlphaFold Protein Structure Database. Nature Structural & Molecular Biology 2021, 28, 869–870.

[37] Woods, R. J. Predicting the Structures of Glycans, Glycoproteins, and Their Complexes. Chemical Reviews 2018, 118, 8005–8024.

[38] Fadda, E. Molecular simulations of complex carbohydrates and glycoconjugates. Current Opinion in Chemical Biology 2022, 69, 102175.

[39] Mallajosyula, S. S.; Jo, S.; Im, W.; MacKerell, A. D. Methods in Molecular Biology; Springer New York, 2015; pp 407–429.

[40] Harbison, A. M.; Brosnan, L. P.; Fenlon, K.; Fadda, E. Sequence-to-structure dependence of isolated IgG Fc complex biantennary N-glycans: a molecular dynamics study. Glycobiology 2018, 29, 94–103.

[41] Patel, K. R.; Roberts, J. T.; Barb, A. W. Allotype-specific processing of the CD16a N45-glycan from primary human natural killer cells and monocytes. Glycobiology 2020, 30, 427–432.

[42] Casalino, L.; Gaieb, Z.; Goldsmith, J. A.; Hjorth, C. K.; Dommer, A. C.; Harbison, A. M.; Fogarty, C. A.; Barros, E. P.; Taylor, B. C.; McLellan, J. S., et al. Beyond shielding: the roles of glycans in the SARS-CoV-2 spike protein. ACS central science 2020, 6, 1722–1734.

[43] Sztain, T.; Ahn, S.-H.; Bogetti, A. T.; Casalino, L.; Goldsmith, J. A.; Seitz, E.; McCool, R. S.; Kearns, F. L.; Acosta-Reyes, F.; Maji, S., et al. A glycan gate controls opening of the SARS-CoV-2 spike protein. Nature chemistry 2021, 13, 963–968.

[44] Mathew, C.; Weiß, R. G.; Giese, C.; Lin, C.-w.; Losfeld, M.-E.; Glockshuber, R.; Riniker, S.; Aebi, M. Glycan–protein interactions determine kinetics of N-glycan remodeling. RSC Chemical Biology 2021, 2, 917–931.

[45] Watanabe, Y.; Allen, J. D.; Wrapp, D.; McLellan, J. S.; Crispin, M. Site-specific glycan analysis of the SARS-CoV-2 spike. Science 2020, 369, 330–333.

[46] Henderson, R.; Edwards, R. J.; Mansouri, K.; Janowska, K.; Stalls, V.; Kopp, M.; Haynes, B. F.; Acharya, P. Glycans on the SARS-CoV-2 spike control the receptor binding domain conformation. BioRxiv 2020,

[47] Chawla, H.; Fadda, E.; Crispin, M. Principles of SARS-CoV-2 glycosylation. Current Opinion in Structural Biology 2022, 75, 102402.

2   Chapter 2: Computational Methods

In this chapter, I will present and discuss the techniques used to run all the computer simulations in this thesis, starting with free glycans from Chapters 3 and 4 and concluding with complete glycoprotein systems in Chapters 5 and 6. I will explain the foundations of these computational methods, starting with classical mechanics, which is the driving force behind these simulations, and then moving on to the empirical force fields I used and the methods used to analyse the results of these simulations.

2.1   Molecular dynamics

Molecular dynamics (MD) simulations [1–3] are an approach based on classical mechanics used to describe the motion of molecules.

$$F = ma \tag{1}$$

Eq 1 can be written as Eq 2 and writing it in terms of coordinates $\mathbb{X}$, the potential energy of the system (represented by an empirical force field) $\mathcal{V}(\mathbb{X}_i)$ and the masses m_i of the atoms. Equation 2 can be used to determine the motion of a classical mechanical system.

$$
\begin{aligned}
F_i(t) =& m_i \frac{d^2 \mathbb{X}_i(t)}{dt^2} \\
=& -\frac{\partial \mathcal{V}(\mathbb{X}_i)}{\partial \mathbb{X}_i(t)}
\end{aligned}
\tag{2}
$$

It is essential to note that these equations are applicable to finite steps. To obtain a trajectory, the second law must be numerically integrated with a timestep that must be smaller than the fastest movement of the system. Typically, the vibrations of hydrogen atoms bonded together take place on a femtosecond timescale, which are the fastest vibrations that occur in our molecular simulations; however, the timestep can be increased by using techniques such as the SHAKE algorithm [4]. This algorithm involves restraining the length of hydrogen-covalent bonds to the equilibrium position, which can expand the timestep from 1 to 2 femtoseconds.

The Verlet algorithm [5] shown in Equation 3, one of the most basic numerical integrators. This algorithm is based on the Taylor expansion of the position of a particle at a certain time t: [5] , The system makes use of three main components: the coordinates at time t $\mathbb{X}_i(t)$, the coordinates before a given time step dt $\mathbb{X}_i(t)(t - dt)$, and the acceleration at that given t $a_i(t)dt^2$. Additionally, the Taylor expansion for $\mathbb{X}_i(t)$ is included in $t - dt$ and $t + dt$.

$$
\begin{aligned}
\mathbb{X}_i(t + dt) =& \mathbb{X}_i(t) + v(t)dt + \frac{1}{2}a(t)dt^2 \\
\mathbb{X}_i(t - dt) =& \mathbb{X}_i(t) - v(t)dt + \frac{1}{2}a(t)dt^2 \\
\mathbb{X}_i(t + dt) =& 2\mathbb{X}_i(t) - \mathbb{X}_i(t - dt) + a(t)dt^2
\end{aligned}
\tag{3}
$$

It should be noted that the velocities $v(t) = \frac{d\mathbb{X}_i(t)}{dt}$, are not explicitly included in this setup. They are cancelled out in the Taylor expansion of $\mathbb{X}_i(t + dt)$ and $\mathbb{X}_i(t - dt)$. Although they are not necessary to describe the trajectory, their evaluation is a necessary step to compute the kinetic energy of the system. The velocity Verlet integrator solves this problem [6]. The leap frog is another and is described in the following. It is described

in [7] and serves as an introduction to the Langevin integrator scheme, described later in the simulation procedure section.

The leap frog integrator uses the velocities calculated at $t + \frac{1}{2}dt$ to calculate the acceleration at time t, which in turn calculates the positions as follows. The leap frog integrator is a second-order symplectic integrator, which means that it conserves important dynamical properties of the system, such as energy, provided an appropriate time step is chosen [8].

$$
\begin{aligned}
v_i(t + \frac{1}{2}dt) =& v_i(t - \frac{1}{2}dt) + a(t)dt \\
\mathbb{X}_i(t + dt) =& \mathbb{X}_i(t) + v(t + \frac{1}{2}dt)dt
\end{aligned}
\tag{4}
$$

In AMBER a version of the leap frog integrator called the Langevin leap frog integrator is used. This integrator introduces a friction parameter γ, which indicates the rate of friction or the frequency of collisions. This adds a random element to the dynamics; for example, when γ is set to 0, it approximates Brownian motion. This scheme is divided into four steps, which are outlined in Equation 5.

$$
\begin{aligned}
&1)v = v(t - \frac{1}{2}dt) + a(t)dt \\
&2)\Delta v = -\gamma v + \sqrt{\gamma(2 - \gamma)(\frac{k_b T_{ref}}{m})}\epsilon \\
&3)\mathbb{X}_i(t + dt) = \mathbb{X}_i(t) + (v + \frac{1}{2}\Delta v)dt \\
&4)v(t + \frac{1}{2}dt) = v + \Delta v \\
&\quad \epsilon \sim N(0, 1)
\end{aligned}
\tag{5}
$$

Step 1 of the leapfrog scheme is the same as the MD velocity update from the standard leapfrog scheme. Step 2 involves the application of the friction coefficient γ and the random noise ϵ to adjust the velocities to the reference temperature T_{ref}, thus acting as a thermostat for the MD simulation. This makes it a stochastic leapfrog integrator, unlike Equation (4). Step 3 can be divided into two parts, which update the velocities and coordinates at the same time.

$$
\begin{aligned}
&3a)\mathbb{X}_i(t + \frac{1}{2}dt) = \mathbb{X}_i(t) + v\frac{1}{2}dt \\
&3b)\mathbb{X}_i(t + dt) = \mathbb{X}_i(t + \frac{1}{2}dt) + (v + \Delta v)\frac{1}{2}dt
\end{aligned}
\tag{6}
$$

Finally, step 4) normalizes the velocities as per step 2) to ensure that the temperature remains constant. The entire process is depicted in Figure 2.1.
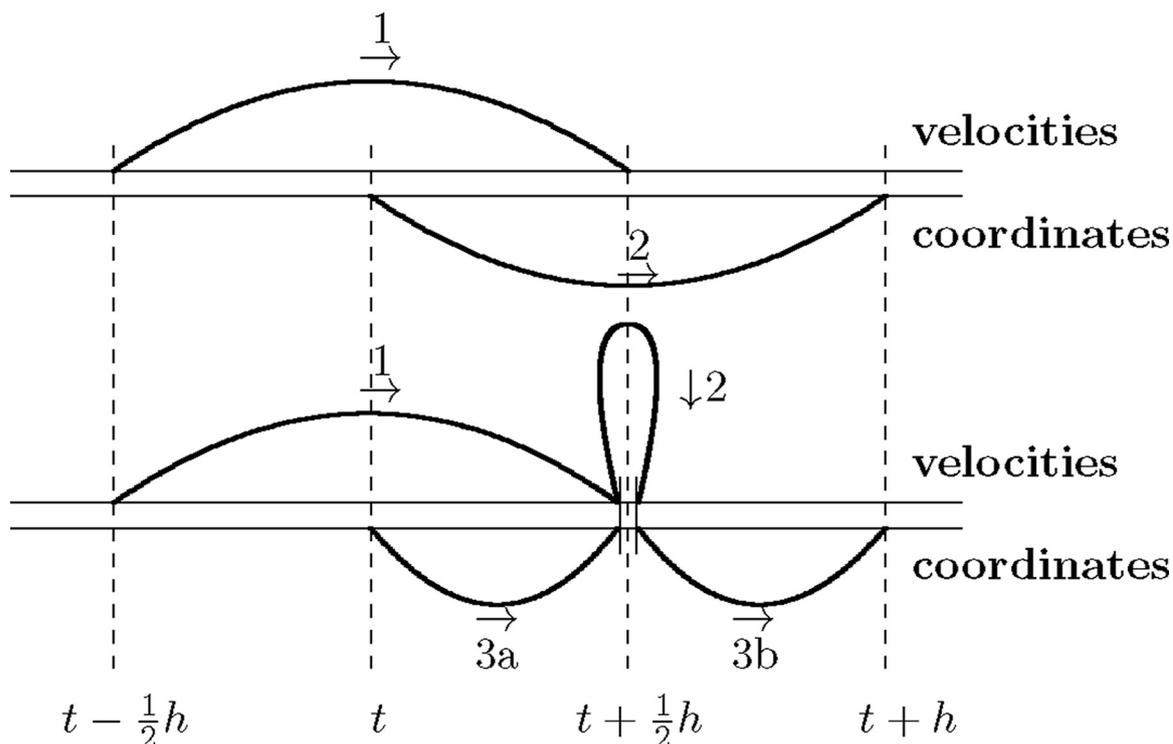
Figure 2.1: Traditional Leap frog scheme described in Eq. 4 (Top) v.s The Leap-frog scheme with the Temperature correction. Adaped from [9]

2.2   Classical Force-field

Throughout the course of my work, all my molecular simulations have been run within a classical formalism, explained in Equations 1 and 2 where the potential energy is central to the description of the systems dynamics. Using a classical mechanical approximation, atoms are represented as solid, hard spheres with known masses (m_i), size (Van der Waals radius), and electrostatic point charge for each atom type. The Born-Oppenheimer approximation allows us to separate the motion of electrons from the motion of nuclei.

The potential energy of a molecule is described by an empirical force field. The success of any MD simulation is contingent upon the selection of a force field that is appropriately tested and specifically developed for the system at hand. When it comes to glycans, the choice of all-atom empirical force fields lies between between several parameter sets, namely Glycam06-J [10–12] which is extensively used in this work and Chemistry at HARvard Macromolecular Mechanics (CHARMM) [13–17]. The majority of common monosaccharides can be adequately represented by these force-fields together with the corresponding glycosidic bonds and the intricacy of branching. [18]. Some other parameter sets of note are GROMOS 53A6 (GLYC) parameter set is used for the simulations of carbohydrates within a CHARMM formalism and that employs united-atoms [19–21] and also the availability of Drude polarization models [22–24] the differences of which are summarized in Figure 2.2 .
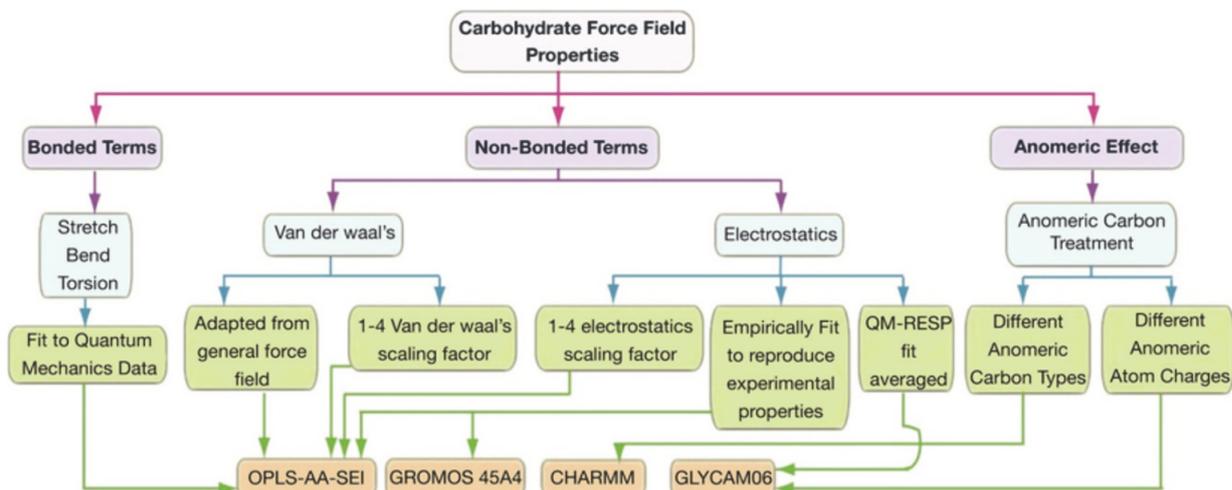
Figure 2.2: Diagram comparing the parameterisation protocol between the various carbohydrate force fields. Adapted from [25]

The parameter sets described above are defined as force fields, where the corresponding potential energy can be described by the general equation below, see Equation 7.

$$
\begin{aligned}
\mathcal{V}(\mathbb{X}_i) = &\sum_{\text{bonds}} \frac{k_r}{2}(r - r_0)^2 + \\
&\sum_{\text{angles}} \frac{k_\theta}{2}(\theta - \theta_0)^2 + \\
&\sum_{\text{dihhedrals}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \\
&\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}} \right]
\end{aligned}
\tag{7}
$$

Equation 7 is an example of an additive force field where partial charges (q_i) of atoms approximate the effect that electrons would have on the structure and dynamics of molecules. All of these interactions have parameters $(k_r, r_0, k_\theta, \theta_0, V_n\phi, \gamma, A_{ij}, B_i j)$ that differ in different force fields. The bonding effects are modelled like a spring using Hook's law to approximate bond stretching and bending, while torsion's are treated as a sinusoidal function. Meanwhile, nonbonding interactions are approximated via a 12/6 Lenard Jones potential, and electrostatic interactions are represented by a Coulomb potential.

2.3   Statistical Thermodynamics

Ergodic Hypothesis

The ergodic hypothesis is a fundamental concept in MD and is a cornerstone of Boltzmann's [26] theory of statistical mechanics. The ergodic hypothesis states that, given enough time, a trajectory will explore all possible states of a system. Assuming that the ergodic hypothesis is correct, if one were to observe a single system over an infinite time period, the result would be the same to that of examining a large collection of systems at one time. It is commonly stated that the average of the small-scale behaviour ($\langle A \rangle_{\text{time}}$) is the same as the average of the large-scale behaviour ($\langle A \rangle_{\text{ensemble}}$) over a period of time.

$$\langle A \rangle_{\text{time}} = \langle A \rangle_{\text{ensemble}} \tag{8}$$

The ergodic hypothesis allows us to analyse the thermodynamic characteristics of a system on the basis of its trajectory through time.

Statistical Ensembles

Running MD without any temperature and pressure, the resulting simulation will have constant energy. If the volume is maintained, the resulting ensemble would be microcanonical, i.e. (N,V,E). The NVE ensemble is advantageous when studying the intrinsic dynamics of a system and its energy conservation, since the number of particles (N), volume (V) and total energy (E) are all maintained throughout the simulation.

The NVT Ensemble, also known as the Canonical Ensemble, is a simulation environment in which the number of particles (N), the volume (V), and the temperature (T) are kept constant. This type of simulation is ideal for situations where temperature control is necessary. Temperature control can be achieved through velocity rescaling, thermostat algorithms such as the Nose-Hoover thermostat [27, 28] or the Berendsen thermostat [29], or Langevin dynamics [2, 30].

The Isobaric-Isothermal Ensemble, also known as the NPT Ensemble, is a simulation environment where x y and z are kept constant. the number of particles (N), the pressure (P), and the temperature (T). This method is ideal for simulations that require both temperature and pressure control, such as those conducted in solution phase.

Within a standard simulation protocol, the structure of a molecular system would be sourced from experiments, such as cryo-EM, X-ray crystallography or NMR, or from computational predictions, such as homology modelling or AI (AlphaFold [31], RoseTTAFold [32]), and the system will start in the crystaline state , *i.e.*, 0 K and 0 atm pressure. The temperature is then adjusted from 0 to 300 K in the microcanonical environment (N,V,T). Then transitions to (N,P,T) when equilibrating to 1 atm all with the solute constrained. Finally, the solute is unconstraned and equilibrated at (N,P,T) before production can be run in NPT with both the temperature and pressure controlled.

Thermostats

A thermostat is a computational tool used to control and maintain the temperature of a simulated system. The thermostat mimics the behaviour of a heat bath. The heat bath acts as a source or sink of energy, allowing the simulated system to exchange energy with it. This exchange helps maintain the desired temperature, ensuring that the system remains at a specific temperature or follows a particular temperature profile throughout the simulation. Temperature control is essential in MD simulations because it helps replicate real-world conditions and enables researchers to study the behaviour of molecules under specific temperature regimes. Different algorithms have been developed over the years to achieve temperature control in a simulation. The most widely used are outliened below.

$$
\begin{aligned}
\lambda =& \sqrt{1 + \frac{\Delta t}{\tau} \left(\frac{T_{\text{target}}}{T_{\text{current}}} - 1 \right)} \\
v'(t) =& \lambda \cdot v(t)
\end{aligned}
\tag{9}
$$

Berendsen Thermostat: The Berendsen thermostat [33] is a simple and commonly used thermostat in MD simulations. That scales the velocities of the particles in the simulation by a scaling factor λ to achieve the desired temperature T_{target} with a given relaxation time τ. It is a weak-coupling thermostat, meaning that it does not mimic a true heat bath, but is effective for temperature control.

$$
\begin{aligned}
\frac{d^2 \mathbb{X}_i(t)}{dt^2} =& F_i - \xi p_i \\
\frac{d\xi}{dt} =& \frac{1}{Q} \left(\sum_i \frac{p_i^2}{m_i} - N k_B T \right) \\
s(t) =& e^{-\frac{\xi(t) \Delta t}{2}} \\
v'(t) =& s(t) \cdot v(t)
\end{aligned}
\tag{10}
$$

Nose-Hoover Thermostat: The Nose-Hoover thermostat [27, 28] is a more sophisticated thermostat that provides better temperature control. It does this by adding degress of freedom to the system (ξ) which are represented by a set of auxiliary variables (Q : Thermostat mass), which evolve along with the system and help it reach the desired temperature (T) while conserving energy. This thermostat is particularly useful in NVT (constant temperature) ensemble simulations.

$$
\begin{aligned}
\Delta v =& -\gamma v + \sqrt{\gamma (2 - \gamma) (\frac{k_b T_{ref}}{m})} \cdot \epsilon \\
\epsilon \sim& N(0, 1) \\
v'(t) =& v(t) + \Delta v
\end{aligned}
\tag{11}
$$

Langevin dynamics [2, 30]: In addition to controlling temperature using the friction coefficient γ, the Langevin thermostat also introduces a stochastic term ϵ into the equations of motion of the particles. This term mimics the effect of random collisions with a heat bath, effectively modelling the system's interaction with its surroundings. Langevin dynamics is often used to simulate systems in contact with a solvent or to model Brownian motion. This particular thermostat is described in the Simulation Procedures section.

Barostats

Barostats in the context of MD simulations are tools used to control and maintain the pressure of a simulated system. Pressure is a crucial thermodynamic variable that influences the density and compressibility of the system, and controlling it is essential to reproduce realistic conditions in simulations. Barostats simulate the effect of an external pressure or pressure bath on the system to ensure that it stays at the desired pressure or follows a specific pressure profile throughout the simulation.

$$
\begin{aligned}
\mu &= \sqrt[3]{1 + \frac{\Delta t}{\tau_p}(P_i - P_0)} \\
\mathbb{X}_i'(t) &= \mu \cdot \mathbb{X}_i(t) \\
l'(t) &= \mu \cdot l(t) \\
V'(t) &= \mu^3 \cdot V
\end{aligned}
\tag{12}
$$

Berendsen Barostat: Similarly to the Berendsen thermostat, the Berendsen barostat [33] scales the volume of the simulation box to maintain the desired pressure. It is a weak-coupling barostat and is often used in NPT (isothermal-isobaric) ensemble simulations.

Periodic Boundary Conditions

Periodic Boundary Conditions (PBCs) are a key element in MD simulations, especially when modelling systems of atoms and molecules in a finite simulation box. PBCs are employed to replicate an infinite, bulk-like environment and prevent the emergence of edge effects that can occur in finite-sized systems.

PBCs assume that the simulation box is surrounded by identical replicas of itself in all three dimensions (x, y, and z). This creates an infinite lattice of simulation boxes that are repeated at regular intervals. When a particle passes through one side of the simulation box, PBCs cause it to reappear on the opposite side of the box [34]. This allows the particle to keep moving as if it had never left the simulation area. Simulating Bulk PBCs can be used to simulate the bulk characteristics of materials without having to simulate an extremely large number of particles. By creating a system that is effectively infinite, it is possible to obtain thermodynamic and structural data as if the system were actually infinite. In order to take into account the long-range interactions, such as electrostatic forces in periodic systems, special methods must be used (e.g., Ewald summation or Particle Mesh Ewald).
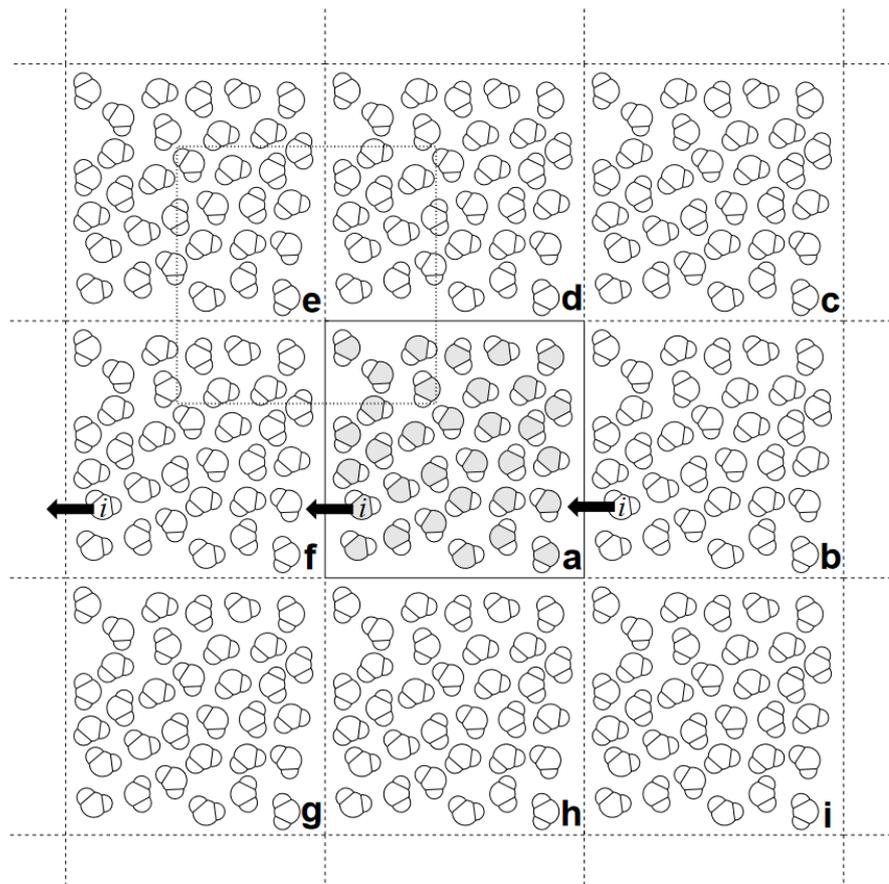
Figure 2.3:   A two-dimensional representation of a simulation cell that is replicated in three directions of space is shown. Utilizing periodic boundary conditions (PBC), when molecule i exits the central box a, its replicas in the adjacent ghost boxes move in a similar manner. The cell delineated by a dotted line, which overlaps with cells a, d, e and f, symbolizes the so-called "minimum image" convention. Adapted from [35]

Particle Mesh Ewald (PME) is a particular implementation of Ewald summation that combines the benefits of Ewald summation with computational efficiency. PME utilizes a three-dimensional mesh to Fourier transform the charge distribution of the system, thus enabling a faster calculation of the reciprocal space component.

The PME algorithm involves the following steps: Assigning Charges to the Mesh: The charges of the system are allocated to a three-dimensional mesh, and a fast Fourier transform (FFT) is used to calculate the charge distribution in reciprocal space. Computing the Fourier-Transformed Potential: The potential in reciprocal space is calculated by utilising the Fourier-transformed charge distribution. Inverse Fourier Transformation: The potential in reciprocal space is transformed into an inverse Fourier transform to obtain the real-space potential. Direct Space Summation: The short-range part of the electrostatic potential is calculated by direct summation. Combining contributions: The contributions from the reciprocal and direct space are combined to generate the total electrostatic potential, and a full description of the PME summation could take up a chapter for itself [34].

2.4   Simulation procedure

In this section, a general simulation procedure is outlined, while the specifics of each individual simulation are discussed in the relevant chapter. The initial phase of any simulation requires the selection of an appropriate starting structure, such as a crystal structure from the PDB [36] or, more recently, a prediction from AlphaFold [31].The initial glycan structures from the GLYCAM Builder (www.glycam.org). Additionally, CHARMM-GUI [37] can also been used. The resulting 3D structures are then simulated via a schema similar to that described in Figure 2.4.

System Prep
- Build Co-Ordinates (PDB / Glycam / PyMOL)
- Build Parameters (tLEaP)

Minimisation
- Steepest Descent (500,000 cycles)

Equilibration
1. Temperature: 0 − 200 K (NVE)
2. Temperature: 100 − 300K (NVE)
3. Pressure: 0 − 1 atm (NPT)
4. Relaxation: 15 ns (NPT)

Production
- RMSD is tracked until converged → previous frames discarded as equilibration
- Multiple replicas (Different initial co-ordinates) → sufficient sampling

Figure 2.4: This protocol outlines a generalized approach to molecular dynamics simulations, which includes the steps of system preparation, minimization, equilibration, and production. The process begins with the construction of molecular coordinates and parameterization. Then, energy minimization is achieved through steepest descent. Temperature and pressure equilibration follows, and the production run is conducted with RMSD tracking to ensure convergence.

All MD simulations employed the AMBER ff14SB9 set of parameters to represent the protein and counterions (200 mM), while the glycans were represented by a version of the GLYCAM06 force field called GLYCAM06j-1 [10]. The water molecules were modelled using the TIP3P model [38–40]. All simulations were conducted using the AMBER software package, version v18. The following protocol was employed for all MD simulations: the energy is minimised 50,000 cycles of the steepest descent algorithm. During the initial minimization, all heavy atoms were restrained with a potential weight of 5 kcal mol^{-1} Å^{2}, while the solvent, counterions, and hydrogen atoms were left unrestrained. After the energy minimization, the system was equilibrated in the NVT ensemble with the same restraints scheme, with heating occurring in two stages over a period of 1 ns, from 0 to 100 K (stage 1) and then 100 K to 300 K (stage 2). The SHAKE algorithm was employed to constrain all bonds to hydrogen atoms during the equilibration process. Van der Waals interactions were cut off at 11 Å and Particle Mesh Ewald (PME) was used to handle long-range electrostatics with B-spline interpolation of order 4. To regulate

the temperature, Langevin dynamics with a collision frequency of 1.0 ps^{-1} was used, and a pseudorandom variable seed was employed to avoid any synchronisation artefacts. After the system was brought to 300 K, an equilibration phase in the NPT ensemble with a time step of 2 ps was employed to set the pressure to 1 atmosphere. An isotropic pressure scaling and a pressure relaxation time of 2.0 ps were used to keep the pressure constant. After this, all limitations on the heavy atoms of the protein were taken away, allowing the system to adjust for 15 ns before the production phase. MD simulations were conducted with resources from ICHEC kay or cineca marconi and some of the later simulation was conducted on ORACLE's cloud platform via GROMACS using the same settings.

2.5   Sampling in molecular dynamics

The most crucial feature of a molecular simulation experiment is its reproducibility, and thus the statistical accuracy of the results derived from it.A crucial feature that ensures reproducibility of the results is the exhaustiveness of the conformational sampling, this is exhaustive (or sufficient) conformational sampling, which can be summed up in the following question: How long is enough?

When it comes to Molecular Dynamics simulations, it may require a lot of computational power to achieve sufficient sampling. The chapters in Section 3 and 4 (illustrated in Figure 2.5) demonstrate an approach that maps the free energy of various conformations of two glycosidic bonds, a β (1–4) and an α (1–6) linkage. MD simulations using a conventional approach enable us to explore the conformational potential energy surface as defined by the force field through a continuous progression in time. Suppose we use conventional MD to identify the structure(s) that correspond to the energy maps in Figure 2.5. For the β (1–4) linkage on the left side, if the initial structure of our MD trajectory is the one represented by the red dot, a relatively short MD will be enough to characterize the most stable structure and estimate its degree of flexibility in terms of standard deviations.
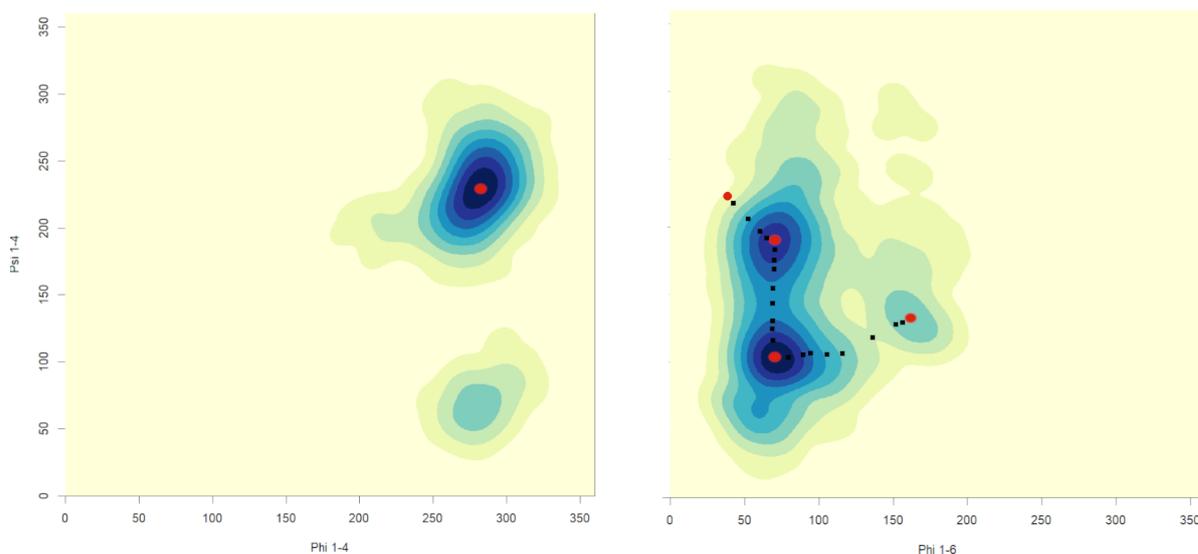
Figure 2.5: . Heat maps showing the conformations of a *N*-glycan with a) a β(1-4) linkage and b) an α(1-6) linkage were generated through a series of parallel MD simulations. The red dots and black dashed lines indicate a hypothetical journey through time and space as determined by the MD simulations. Adapted from [18]

In Figure 2.5, the α(1–6) linkage on the right-hand side has multiple minima, as is typical for a flexible linkage. Some of these minima are linked by low-energy paths, while others are not. The dashed line in the figure illustrates a potential MD trajectory that can be used to identify some, but not all, of the stable structures (minima). It is highly unlikely that a conventional MD trajectory will cross a high energy barrier at 300 K, thus making it necessary to extend the MD simulation. To explore the conformational space more efficiently, it is recommended to initiate uncorrelated MD runs from different structures. For instance, the red dots in Figure 2.5 panel b can be used as an example. This system allows for shorter MD runs and the amount of structural conversion between the runs can be used to measure the sampling completeness. This approach is especially beneficial for the analysis of free carbohydrates [41] as well as Chapter 3, 4, since conformational interconversion does not usually require crossing high energy barriers and can be achieved within reasonable trajectory lengths, except in certain complex cases.

Bibliography

[1] Frenkel, D.; Smit, B. Understanding molecular simulation from algorithms to applications; Academic Press, an imprint of Elsevier, 2012.

[2] Brooks, C. L. Computer simulation of liquids. Journal of Solution Chemistry **1989**, 18, 99–99.

[3] Leach, A. R. Molecular modelling: Principles and applications; Prentice Hall, 2001.

[4] Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. Journal of Computational Physics **1977**, 23, 327–341.

[5] Verlet, L. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. Physical Review **1967**, 159, 98–103.

[6] Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. The Journal of Chemical Physics **1982**, 76, 637–649.

[7] University of California: San Francisco, U., CA AMBER 2018. 2018.

[8] Fass, J.; Sivak, D. A.; Crooks, G. E.; Beauchamp, K. A.; Leimkuhler, B.; Chodera, J. D. Quantifying configuration-sampling error in Langevin simulations of complex molecular systems. Entropy **2018**, 20, 318.

[9] Goga, N.; Rzepiela, A. J.; de Vries, A. H.; Marrink, S. J.; Berendsen, H. J. C. Efficient Algorithms for Langevin and DPD Dynamics. Journal of Chemical Theory and Computation **2012**, 8, 3637–3649, PMID: 26593009.

[10] Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; González-Outeiriño, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. GLYCAM06: A generalizable biomolecular force field. Carbohydrates. Journal of Computational Chemistry **2007**, 29, 622–655.

[11] Tessier, M. B.; DeMarco, M. L.; Yongye, A. B.; Woods, R. J. Extension of the GLYCAM06 biomolecular force field to lipids, lipid bilayers and glycolipids. Molecular Simulation **2008**, 34, 349–364.

[12] Singh, A.; Tessier, M. B.; Pederson, K.; Wang, X.; Venot, A. P.; Boons, G.-J.; Prestegard, J. H.; Woods, R. J. Extension and validation of the GLYCAM force field parameters for modeling glycosaminoglycans. Canadian journal of chemistry **2016**, 94, 927–935.

[13] Guvench, O.; Mallajosyula, S. S.; Raman, E. P.; Hatcher, E.; Vanommeslaeghe, K.; Foster, T. J.; Jamison, F. W.; MacKerell Jr, A. D. CHARMM additive all-atom force field for carbohydrate derivatives and its utility in polysaccharide and carbohydrate–protein modeling. Journal of chemical theory and computation **2011**, 7, 3162–3180.

[14] Guvench, O.; Hatcher, E.; Venable, R. M.; Pastor, R. W.; MacKerell Jr, A. D. CHARMM additive all-atom force field for glycosidic linkages between hexopyranoses. Journal of chemical theory and computation **2009**, 5, 2353–2370.

[15] Mallajosyula, S. S.; Guvench, O.; Hatcher, E.; MacKerell Jr, A. D. CHARMM additive all-atom force field for phosphate and sulfate linked to carbohydrates. Journal of Chemical Theory and Computation **2012**, _8_, 759–776.

[16] Raman, E. P.; Guvench, O.; MacKerell Jr, A. D. CHARMM additive all-atom force field for glycosidic linkages in carbohydrates involving furanoses. The Journal of Physical Chemistry B **2010**, _114_, 12981–12994.

[17] Mallajosyula, S. S.; Jo, S.; Im, W.; MacKerell, A. D. Molecular dynamics simulations of glycoproteins using CHARMM. Glycoinformatics **2015**, 407–429.

[18] Fadda, E. Molecular simulations of complex carbohydrates and glycoconjugates. Current Opinion in Chemical Biology **2022**, _69_, 102175.

[19] Nester, K.; Gaweda, K.; Plazinski, W. A GROMOS Force Field for Furanose-Based Carbohydrates. Journal of Chemical Theory and Computation **2019**, _15_, 1168–1186.

[20] Pol-Fachin, L.; Rusu, V. H.; Verli, H.; Lins, R. D. GROMOS 53A6 GLYC, an Improved GROMOS Force Field for Hexopyranose-Based Carbohydrates. Journal of Chemical Theory and Computation **2012**, _8_, 4681–4690.

[21] Pol-Fachin, L.; Verli, H.; Lins, R. D. Extension and validation of the GROMOS 53A6 - glyc parameter set for glycoproteins. Journal of Computational Chemistry **2014**, _35_, 2087–2095.

[22] Aytenfisu, A. H.; Yang, M.; MacKerell, A. D. CHARMM Drude Polarizable Force Field for Glycosidic Linkages Involving Pyranoses and Furanoses. Journal of Chemical Theory and Computation **2018**, _14_, 3132–3143.

[23] Yang, M.; Aytenfisu, A. H.; MacKerell, A. D. Proper balance of solvent-solute and solute-solute interactions in the treatment of the diffusion of glucose using the Drude polarizable force field. Carbohydrate Research **2018**, _457_, 41–50.

[24] Kognole, A. A.; Aytenfisu, A. H.; MacKerell, A. D. Extension of the CHARMM Classical Drude Polarizable Force Field to N- and O-Linked Glycopeptides and Glycoproteins. The Journal of Physical Chemistry B **2022**, _126_, 6642–6653.

[25] Perez, S.; Fadda, E.; Makshakova, O. Comprehensive Glycoscience; Elsevier, 2021; pp 374–404.

[26] Boltzmann, L. Vorlesungen über Gastheorie: 2. Teil; BoD–Books on Demand, 2017.

[27] Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. The Journal of Chemical Physics **1984**, _81_, 511–519.

[28] Hoover, W. G.; Holian, B. L. Kinetic moments method for the canonical ensemble distribution. Physics Letters A **1996**, _211_, 253–257.

[29] van Gunsteren, W. F.; Berendsen, H. J. C. Computer Simulation of Molecular Dynamics: Methodology, Applications, and Perspectives in Chemistry. Angewandte Chemie International Edition in English **1990**, _29_, 992–1023.

[30] Brooks, C. L.; Berkowitz, M.; Adelman, S. A. Generalized Langevin theory for many-body problems in chemical dynamics: Gas-surface collisions, vibrational energy relaxation in solids, and recombination reactions in liquids. The Journal of Chemical Physics **1980**, _73_, 4353–4364.

[31] Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature **2021**, _596_, 583–589.

[32] Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science **2021**, _373_, 871–876.

[33] Berendsen, H. J.; Postma, J. v.; Van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. The Journal of chemical physics **1984**, _81_, 3684–3690.

[34] Allen, M. P.; Tildesley, D. J. Computer Simulation of Liquids; Oxford University PressOxford, 2017.

[35] Chipot, C. Nanoscience; Springer Berlin Heidelberg, 2009; pp 803–838.

[36] Armstrong, D. R. et al. PDBe: improved findability of macromolecular structure data in the PDB. Nucleic Acids Research **2019**,

[37] Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: A web-based graphical user interface for CHARMM. Journal of Computational Chemistry **2008**, _29_, 1859–1865.

[38] MacKerell, A. D. et al. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. The Journal of Physical Chemistry B **1998**, _102_, 3586–3616.

[39] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. The Journal of Chemical Physics **1983**, _79_, 926–935.

[40] Price, D. J.; Brooks, C. L. A modified TIP3P water potential for simulation with Ewald summation. The Journal of Chemical Physics **2004**, _121_, 10096–10103.

[41] Harbison, A. M.; Brosnan, L. P.; Fenlon, K.; Fadda, E. Sequence-to-structure dependence of isolated IgG Fc complex biantennary N-glycans: a molecular dynamics study. Glycobiology **2019**, _29_, 94–103.

3   Chapter 3: How and why plants and human *N*-glycans are different.

Insight from molecular dynamics into the "glycoblocks" architecture of complex carbohydrates

3.1   Abstract

The *N*-glycosylation is one of the most abundant and diverse post-translational modifications of proteins, implicated in protein folding and structural stability, and mediating interactions with receptors and with the environment. All *N*-glycans share a common core from which linear or branched arms stem from, with functionalization specific to different species and to the cells' health and disease state. This diversity generates a rich collection of structures, all diversely able to trigger molecular cascades and to activate pathways, which also include adverse immunogenic responses. These events are inherently linked to the *N*-glycans' 3D architecture and dynamics, which remain for the large part unresolved and undetected because of their intrinsic structural disorder. In this work we use molecular dynamics (MD) simulations to provide insight into *N*-glycans' 3D structure by analysing the effects of a set of very specific modifications found in plants and invertebrate *N*-glycans, which are immunogenic in humans. We also compare these structural motifs and combine them with mammalian *N*-glycan motifs to devise strategies for the control of the *N*-glycan 3D structure through sequence. Our results suggest that the *N*-glycans' architecture can be described in terms of the local spatial environment of groups of monosaccharides. We define these "glycoblocks" as self-contained 3D units, uniquely identified by the nature of the residues they comprise, their linkages and structural/dynamic features. This alternative description of glycans' 3D architecture can potentially lead to an easier prediction of sequence-to-structure relationships in complex carbohydrates, with important implications in glycoengineering design.

Keywords: complex carbohydrates; fucose; glycoblocks; molecular dynamics; molecular recognition; *N*-glycans; xylose
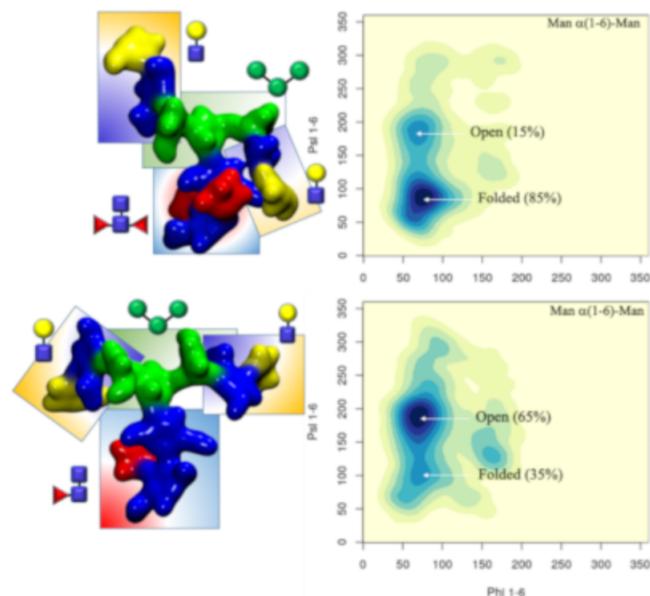
3.2   Introduction

Complex carbohydrates (or glycans) are an essential class of biomolecules, directly implicated in the cell's interactions with its environment, facilitating communication and infection [1, 2]. These processes are often initiated by molecular recognition involving carbohydrate-binding proteins (lectins) or by glycan–glycan interactions [1, 3–5], all events that hinge on specific structural and dynamic features of the glycans. This makes the 3D complementarity of the glycans architecture key towards the success of these processes and an essential piece of information for us to have in order to understand glycan recognition. Because of their chemical nature, glycans are intrinsically flexible and highly dynamic at room temperature, thus their characterization through experimental structural biology methods is hardly straightforward even in cryogenic environments [6] As an additive layer of difficulty, glycosylation is only indirectly dependent on the genome, which often results in a micro- (or macro-)heterogeneity of glycan sequences at specific sites [7] These complexities are very difficult to resolve, requiring high levels of expertise and multi-layered orthogonal approaches [7–10] .Within this framework, the contribution of glycoinformatics tools and databases represents an essential resource to advance glycomics [11–15] while molecular simulations fit in very well as complementary and orthogonal techniques to support and advance structural glycobiology research. Indeed, current high performance computing (HPC) technology allows us to study realistic model systems [16, 17] and to reach experimental timescales [18], so that computing can now contribute as one of the leading research methods in structural glycobiology.

One of the most interesting and remarkably challenging areas in glycoscience research that HPC simulations can address is the study of the links between glycans' sequence and the 3D structure. This direct relationship is a well-recognized and broadly accepted concept in proteins' structural biology, according to which the amino acid sequence dictates the functional 3D fold and its stability. However, the same notion is not generally invoked when discussing other biopolymers or complex carbohydrates. In the specific case of glycans, the structural complexity, in terms of the diversity of monosaccharides, the linkages' stereochemistry and the branched scaffolds, makes the already difficult case even more intricate. Nevertheless, the fact that glycoforms follow recurrent sequence patterns, clearly suggests that the glycans 3D structure is also non-random and very likely sequence-determined. We use computer modelling to gain insight into these relationships

and to define a framework to understand how subtle modifications to the glycans sequence can alter their 3D structure and conformational dynamics, ultimately regulating recognition [19].In this work we use molecular dynamics (MD) simulations to analyse the effects of the inclusion of motifs typically found in plants and invertebrates N-glycans and immunogenic in mammals [20–23] More specifically, we investigate how core $\alpha(1\text{-}3)$-linked fucose (Fuc) and β (1-2)-linked xylose (Xyl) affect the structure and dynamics of plants N-glycoforms [23] and of hybrid constructs with mammalian N-glycoforms [24]

At first glance plants protein N-glycosylation [23] is quite similar to the one of higher species [25],carrying the distinctive trimannose core (Man3), which can be further functionalised with $\beta(1\text{-}2)$-linked GlcNAc residues on the arms. As a trademark feature, shown in Figure 1, plants N-glycans can also have a $\beta(1\text{-}2)$-Xyl linked to the central mannose and core $\alpha(1\text{-}3)$-Fuc, instead of the $\alpha(1\text{-}6)$-Fuc commonly found in mammalian complex N-glycans. Additionally, the arms can be further functionalised with terminal galactose (Gal) in $\beta(1\text{-}3)$ instead of $\beta(1\text{-}4)$ [23] commonly found in vertebrates, which forces the addition of fucose in the $\alpha(1\text{-}4)$ position of the GlcNAc and results in the occurrence of Lewis A (LeA) instead of Lewis X (LeX) terminal motifs on the arms [23, 26] In a previous study, we characterized through extensive sampling the structure and dynamics of complex biantennary N-glycans commonly found in the human IgGs Fc region [24]. The results of this study indicated a clear sequence-to-structure relationship, especially in the context of the dynamics of the (1-6) arm. More specifically, we found that the outstretched (open) conformation of the (1-6) arm gets progressively less populated as the functionalization of the arm grows, i.e., from 85% in Man3, to 52% in (F)A2, (F)A2[3]G1, and (F)A2[3]G1S1, where the (F) indicates the presence or absence of $\alpha(1\text{-}6)$ core fucosylation, to 24% in all structures with (1-6) arm terminating with Gal-$\beta(1\text{-}4)$-GlcNAc or Sia-$\alpha(2\text{-}6)$-Gal-$\beta(1\text{-}4)$-GlcNAc, irrespective of the functionalization of the (1-3) arm [24]. As a practical implication of these results, positional isomers, such as (F)A2[3]G1 and (F)A2[6]G1, have different conformational propensities, the latter with a much lower population of outstretched (1-6) arm and therefore quite different 3D average structures, which ultimately explains their differential recognition in glycan arrays [27]. Additionally, the different conformation of the arms explains the known difficulties in sialylating the (1-6) arm by ST6-Gal1, relatively to the (1-3) arm [28] Also, the different 3D conformational propensity of the arms in function of sequence can have important implications in terms of the N-glycans' biosynthesis and biodegradation [29] As an additional interesting point, we found that the folding of the (1-6) arm over the chitobiose region is completely independent of core $\alpha(1\text{-}6)$ fucosylation [24] with the result that core-fucosylated and non-core fucosylated N-glycans with the same sequence in the (1-6) arm correspond to the same structural ensemble. In this work we discuss how core $\alpha(1\text{-}3)$-Fuc and $\beta(1\text{-}2)$-Xyl regulate the conformational propensity of the (1-6) arm to push a predominantly outstretched (open) conformation when the arms are functionalized with terminal $\beta(1\text{-}3)$-Gal. Within this framework, we explored the possibility of integrating these motifs in the context of mammalian sequences as an exploratory strategy towards the design of N-glycans with the desired 3D structure. For simplicity in the presentation and discussion of the results, we refer to N-glycans as either "plant" or "hybrid" separately. Nevertheless, it is important to underline that some of these motifs, such as $\beta(1\text{-}2)$ xylosylation and difucosylated core are also found in invertebrate N- glycosylation [30]. Finally, we discuss these findings within a framework where the different N-glycoforms can be represented as a combination of spatial self-contained units, named "glycoblocks", rather than in terms of monosaccharides and linkages. We find that this approach helps our understanding of N-glycans architecture in terms of equilibrium structures and relative populations and also of how specific modifications affect molecular recognition.
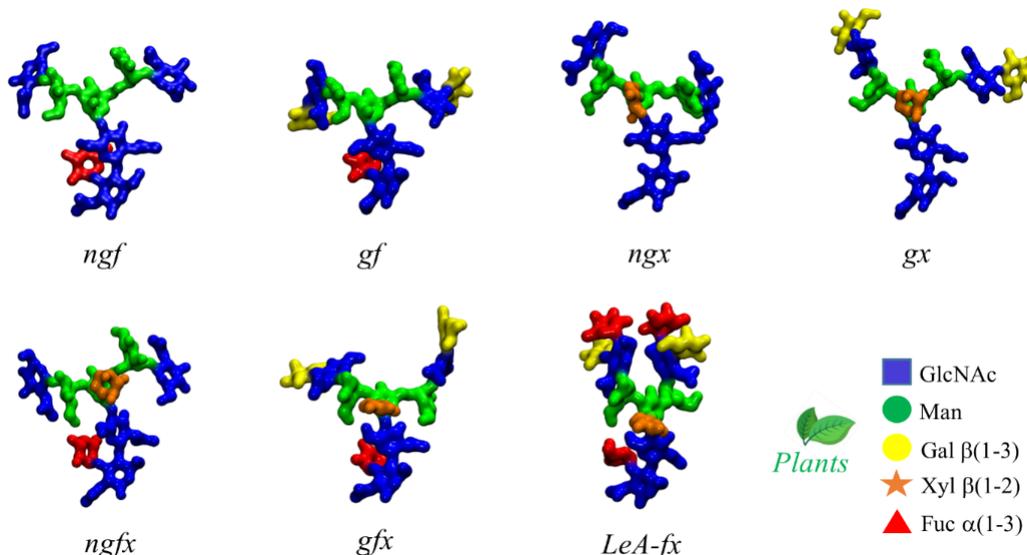
Figure 3.1: Representative structures of the plant *N*-glycans studied in this work with corresponding nomenclature. The letters f, x, and g indicate the presence of Fuc, Xyl and β(1-3) Gal, respectively, and ng the absence of β(1-3) Gal. LeA stands for Lewis A antigen. The *N*-glycans structures are shown with the (1-3) and (1-6) arms on the left and on the right, respectively. The monosaccharides colouring follows the SFNG nomenclature. The plants *N*-glycan characteristic linkages are indicated in the legend. Rendering was done with VMD (https://www.ks.uiuc.edu/Research/vmd/).

3.3   Computational Methods

All starting structures were generated with the GLYCAM Carbohydrate Builder (http://www.glycam.org). For each sequence we selected the complete set of torsion angle values obtained by variation of the 1-6 dihedrals, namely the three *gg*, *gt* and *tg* conformations for each 1-6 torsion. The topology file for each structure was obtained using *tleap* [31] with parameters from the GLYCAM06-j1 [32] for the carbohydrate atoms and with TIP3P for water molecules [33] All calculations were run with the AMBER18 software package [31] on NVIDIA Tesla V100 16GB PCIe (Volta architecture) GPUs installed on the HPC infrastructure kay at the Irish Centre for High-End Computing (ICHEC). Separate production steps of 500 ns each were run for each rotamer (starting system) and convergence was assessed based on conformational and clustering analysis, see Appendix A for all relevant Tables. Simulations were extended, if the sampling was not deemed sufficient, i.e., in case standard deviation values measured were significantly larger than 15° for each cluster in each trajectory. All trajectories were processed using *cpptraj* [31] and visually analysed with the Visual Molecular Dynamics (VMD) software package [34]. Root mean square deviation (RMSD) and torsion angles values were measured using VMD. A density-based clustering method was used to calculate the populations of occupied conformations for each torsion angle in a trajectory and heat maps for each dihedral were generated with a kernel density estimate (KDE) function. Statistical and clustering analysis was done with the R package and data were plotted with RStudio (https://www.rstudio.com). Further details on the simulation set-up and running protocol are included in Appendix A.

3.4 Results

Core α(1-3) fucose in plant *N*-glycans: One distinctive feature of plants *N*-glycans is the occurrence of core fucosylation in α(1-3), rather than α(1-6)-Fuc, normally found in mammalian *N*-glycans [23, 24]. To understand the effects on the 3D structure of this modification, we have considered two biantennary systems, one terminating with β(1-2)-GlcNAc on both arms (*ngf*) and the other with terminal β(1-3)-Gal on both arms (*gf*), shown in Figure 3.1. In both glycoforms core α(1-3)-Fuc occupies a stable position, with one single conformer populated (100%), see Table A.1 and Table A.2 in Appendix A This conformation is supported by a stacking interaction between the core α(1-3)-Fuc and β(1-4) GlcNAc of the chitobiose in a "closed" conformation, which resembles the stable conformation of LeX [35] This spatial arrangement imposes a 20° rotation of the GlcNAc-β(1-4)-GlcNAc linkage, see Table A.1 and Table A.2 in Appendix A, relative to the α(1-6) core fucosylated or non-fucosylated chitobiose [24] where the average psi value is -127.8° (14.8) [24], but doesn't affect the structure of the linkage to the central mannose. As shown by the low standard deviation values and by the lack of multiple minima (clusters), the *N*-glycan core remains relatively rigid throughout the trajectories. The slight torsion of the GlcNAc-β(1-4)-GlcNAc linkage imposed by the α(1-3)-Fuc has a dramatic effect on the conformational dynamics of the (1-6) arm, which is found predominantly in an outstretched (66%, cluster 1) conformation,rather than folded over (34%, clusters 1 and 2), see Table A.1 in Appendix A The addition of a terminal β(1-3)-Gal in the gf *N*-glycan pushes the equilibrium towards an outstretched (1-6) arm even further, with the open conformation populated at 72%, see Table A.2 in Appendix A. Interestingly, in the case of α(1-6) core fucosylated *N*-glycans, and with double fucosylation as discussed later on, the equilibrium of the (1-6) arm was the exact opposite, with a predominance of the folded conformation, especially in the presence of terminal β(1-4) Gal [24]. To note, the folded (1-6) arm conformation can be either a 'front fold', see Figure 3.2 panel a, where the torsion around the α(1-6) linkage brings the arm towards the reader, or a 'back fold' where the (1-6) arm interacts with the α(1-3)-Fuc, away from the reader. As shown in Table A.1 and Table A.2 in Appendix A, the equilibrium of the (1-3) arm is not affected by core α(1-3)-Fuc.
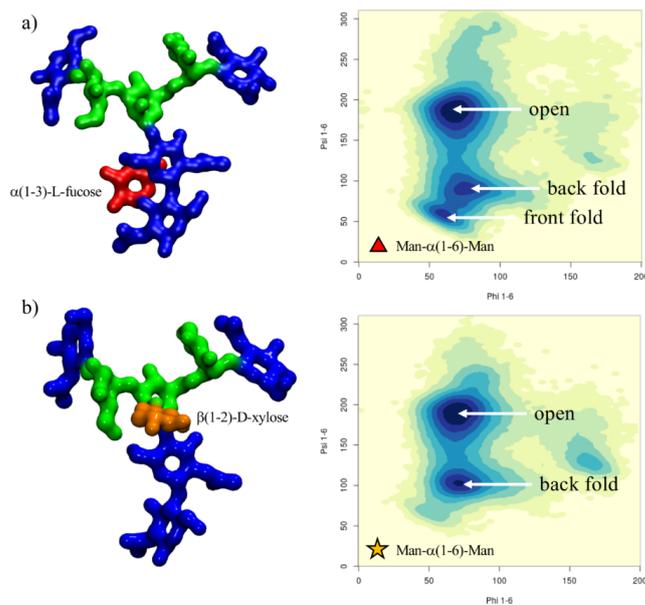
Figure 3.2: A representative structure of the non-galactosylated *N*-glycan with α(1-3)-linked core fucose (*ngf*) is shown in panel a), with on the right-hand side the corresponding heat map showing the conformations accessible to the (1-6) arm in terms of the phi/psi torsion angles. A representative structure of the non-galactosylated *N*-glycan with β(1-2)-linked xylose (*ngx*) is shown in panel b), with on the right-hand side a heat map showing the conformations accessible to the (1-6) arm in terms of the phi/psi torsion angles. The *N*-glycans structures are shown with the (1-3) and (1-6) arms on the left and on the right, respectively. The monosaccharides colouring follows the SFNG nomenclature. The structure rendering was done with VMD (https://www.ks.uiuc.edu/Research/vmd/) and the graphical statistical analysis with RStudio (http://www.glycam.org).

β**(1-2) xylose in plant *N*-glycans:** Because the β(1-2)-Xyl sits in front of the two arms, it greatly affects their dynamics. Because of steric hindrance, the (1-3) arm is much more rigid relative to non-xylosylated species, see Table A.3 in Appendix A losing its "two conformer" dynamics characteristic of the biantennary mammalian *N*-glycans [24], also retained in the plant *N*-glycans with only α(1-3)-Fuc discussed above, see also Table A.1 and Table A.2 in Appendix A. In regards to the (1-6) arm, as shown in 3.2 panel b, the presence of β(1-2)-Xyl has a very similar effect as the α(1-3)-Fuc, pushing the equilibrium towards an open conformation. To note, in the presence β(1-2)-Xyl, the (1-6) arm cannot fold over the chitobiose core in a 'front fold' either, because of steric hindrance. Also, similarly to the α(1-3) fucosylated glycans, the stability of the open structure is slightly increased when the arm is further functionalized with terminal β(1-3)-Gal, see Table A.4 in Appendix A. As an additional interesting feature, through the cumulative 3 μs MD sampling, the xylose ring repeatedly inverts its conformation from the all equatorial ${}^4C_1$ chair, to the ${}^1C_4$ chair, where all hydroxy groups are axial, see Figure 3.3 This transition may be energetically facilitated by the hydrogen bonding interaction xylose is able to form when in a ${}^1C_4$ chair with the α(1-6)-Man, which may compensate for the steric compression, making the ${}^1C_4$ chair the highest populated conformer at 76% within an *N*-glycan scaffold. Both experimental and ab-initio theoretical studies [36–38] have shown that the ${}^1C_4$ chair is energetically accessible in isolated β-D-Xyl at room temperature in different dielectric conditions.
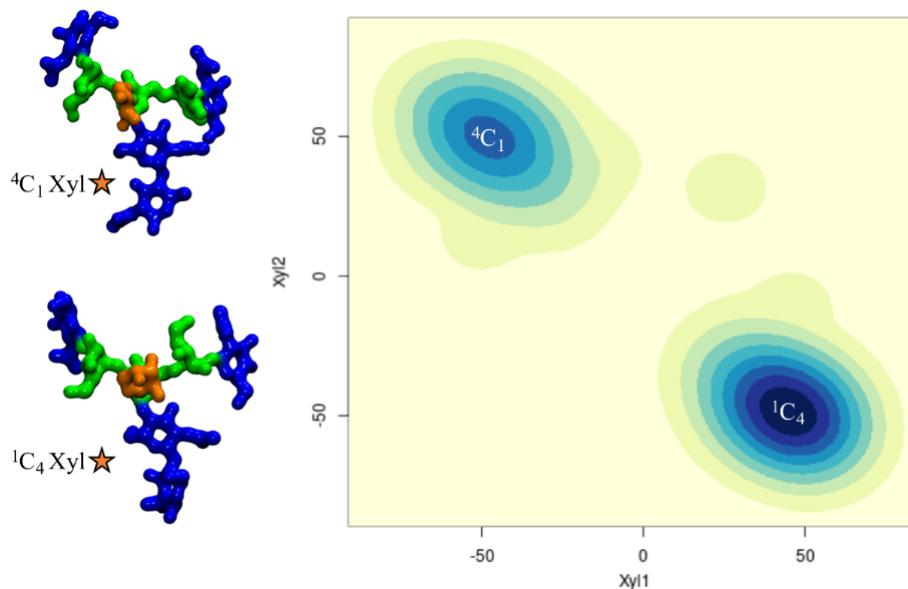
Figure 3.3: β-D-xylose ring pucker analysis over 3 μs of cumulative MD sampling of the *ngx* N-glycan. The two snapshots on the right-hand side are representative *ngx* conformations corresponding to the two different ring puckers. The Xyl1 and Xyl2 axis labels refer to the torsion angles C1–C2–C3–C4 and C2–C3–C4–C5, respectively. The N-glycan structures are shown with the (1-3) and (1-6) arms on the left and on the right, respectively. The monosaccharides colouring follows the SFNG nomenclature. The structure rendering was done with VMD (https://www.ks.uiuc.edu/Research/vmd/) and the graphical statistical analysis with RStudio (http://www.glycam.org).

Core α(1-3) fucose and β(1-2) xylose in plant N-glycans: The presence of both α(1-3)-Fuc and β(1-2)-Xyl brings in the characteristic features highlighted earlier in the analysis of the structures with either α(1-3)-Fuc or β(1-2)-Xyl. Indeed, we see here again the 20° rotation of the chitobiose GlcNAc-β(1-4)-GlcNAc psi angle caused by the stacking of the α(1-3)-Fuc to the chitobiose β(1-4)-GlcNAc and the conformational restraints imposed by the β(1-2)-Xyl on the (1-3) arm, see Table S5 in Supporting Information File 1. We also observed that both α(1-3)-Fuc and β(1-2)-Xyl push the (1-6) arm equilibrium towards an open conformation, which is also the case when both are present in the ngfx N-glycan and to an even higher degree, i.e. 87%, in the gfx N-glycan, when both arms are functionalized with terminal β(1-3)-Gal, see Table A.6 in Appendix A One feature specific to the ngfx N-glycan is the higher flexibility of the core Man-β(1-4)-GlcNAc linkage, which allows for the rotation of the trimannose group relative to the chitobiose core. This conformation was accessible, but only populated around 2% when either β(1-2)-Xyl or α(1-3)-Fuc are present, see Tables A.1 to A.4 in Appendix A When both fucose and xylose are present, the population of the rotated trimannose reaches above 20%, see Table A.5 in Appendix A which can be considered a synergistic effect as this conformation is stabilized by a hydrogen bonding network involving the core fucose, the GlcNAc on the (1-6) arm and the xylose, as shown in Figure A.1 in Appendix A. Such folding event has been observed as a stable conformation in two independent simulations. To note, the functionalization of the arms to include terminal β(1-3)-Gal reduces the occurrence of this event down to around 5%, see Table A.6 in Appendix A.

Terminal LeA and LeX motifs in plant *N*-glycans: To understand how an increased complexity on the arms would affect the dynamics of the $\alpha(1\text{-}3)$ fucosylated and $\beta(1\text{-}2)$ xylosylated *N*-glycans, we considered the functionalization with terminal LeA antigens present in plants *N*-glycans [24] and with LeX for comparison. As expected [35] the LeA and LeX structures are quite rigid, see Table A.7 and Table A.15 in Appendix A, and remain in what is known as the "closed" conformation throughout the 1.5 μs cumulative sampling time for each system. One interesting point is that the branching introduced by functionalizing the terminal GlcNAc residues with $\alpha(1\text{-}4)$-Fuc and $\beta(1\text{-}3)$-Gal, i.e. LeA, promotes the interaction between the two arms, which is not observed when the arms are linear, neither here for plants *N*-glycans, nor for mammalian IgG-type complex biantennary *N*-glycans [24] The interaction between the arms is promoted by the ability to form complex hydrogen bonding networks, which in this specific case, may also involve the central xylose. As outcomes of the complex interaction the branched arms can establish, the equilibrium of the (1-6) arm is restrained in conformations previously not significantly populated, see 3.4 and Table A.15 in Appendix A, within this framework the dynamics of the (1-6) arm is completely different. Contrary to the *N*-glycan with terminal LeA groups, the two arms with LeX are not interacting and the (1-6) arm is predominantly (90%) in an extended (open) conformation, while the closed conformation, which accounts for the remaining 10% is achieved through a rotation around the core Man-$\beta(1\text{-}4)$-GlcNAc. The lack of interaction between the arms is due to the inability to establish the same stable hydrogen bonding network due to the non-complementary position of the deoxy-C6 of the fucose in LeX relative to LeA.
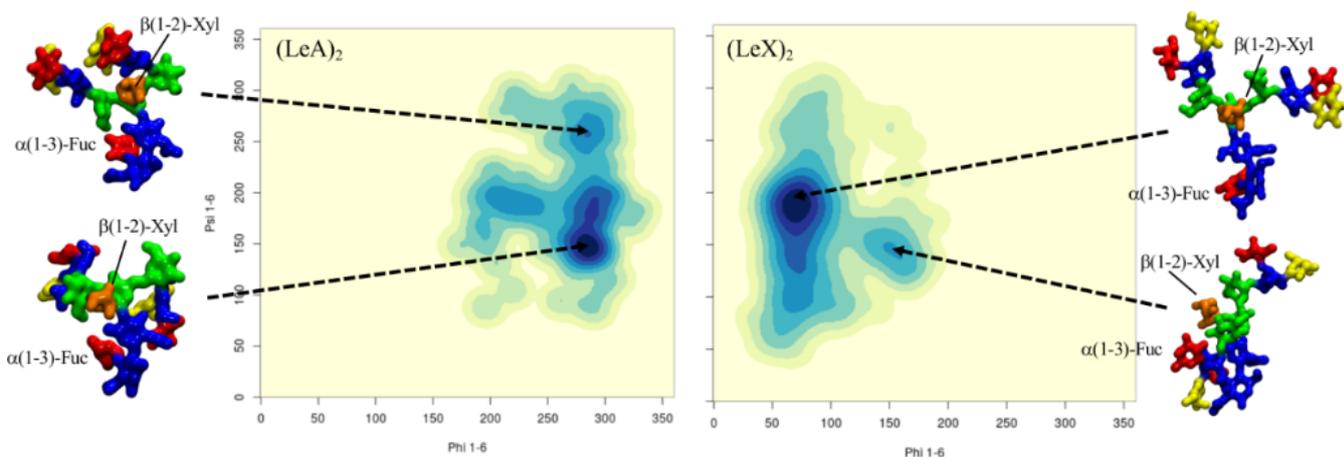
Figure 3.4: Comparison of the different conformational equilibria of the (1-6) arm in a core $\alpha(1\text{-}3)$-Fuc $\beta(1\text{-}2)$-Xyl A2 *N*-glycan with terminal LeA and LeX groups on the left- and right-hand side, respectively. Representative structures from 1.5 μs MD sampling of each system are shown to illustrate the conformations corresponding to the different minima. The *N*-glycans structures are shown with the (1-3) and (1-6) arms on the left and on the right, respectively. The monosaccharides colouring follows the SFNG nomenclature. The structure rendering was done with VMD (https://www.ks.uiuc.edu/Research/vmd/) and the graphical statistical analysis with RStudio (http://www.glycam.org).

Hybrid *N*-glycans. To understand how characteristic plant *N*-glycan motifs can affect the structure of mammalian *N*-glycoforms, we have designed and analysed the dynamics of a set of hybrid systems. In particular, we were interested in the effect of the addition

of β(1-2)-Xyl and α(1-3)-Fuc to (F)A2G2 N-glycans scaffolds in terms of potential alteration of the (1-6) arm dynamics.

β(1-2)-xylosylated mammalian N-glycans. Unlike the case of plants N-glycans, the presence of β(1-2)-Xyl hinders but does not completely prevent the (1-6) arm from folding over when the terminal galactose is β(1-4)-linked, as folding over the chitobiose can be stabilized by stacking, see Figure 3.5 and Table A.8 in Appendix A. The folded conformation with a median psi value of 103.5° ($\pm$ 11.3) is 20° from the average value of 82.9° calculated for the non-xylosylated (mammalian) counterpart [24], so slightly distorted, and its population reduced from 74% to 57%. Nevertheless, the closed conformation is still the predominant form, even with β(1-2)-Xyl. The presence of α(1-6)-linked core fucose to create a β(1-2)-xylosylated FA2G2, which is actually a type of N-glycosylation found in schistosoma [30]. brings in yet another change. As shown in 3.5 and Table A.9 in Appendix A. α(1-6)-Fuc and β(1-2)-Xyl are in an optimal conformation to support the closed (folded) (1-6) arm, by stacking of the terminal galactose by fucose and hydrogen bonding by xylose. Within this context the closed (1-6) arm is the highest populated conformer at 70.0% over 4.5 μs of cumulative sampling of this system. To note that the conformation of the α(1-6)-linked core fucose is the same as the one seen in mammalian N-glycans [24], which on its own we have seen is not enough to affect the (1-6) arm equilibrium, see Table A.9 in Appendix A The interaction of the α(1-6)-Fuc with the terminal β(1-4)-Gal is essential to promote the closed conformation of the (1-6) arm as demonstrated by the results obtained for the xylosylated FA2 systems, which recovers a conformational propensity similar to the non-fucosylated, xylosylated A2G2, see Figure 3.5 and Tables A.8 and A.10 in Appendix A

α(1-3)-fucosylated mammalian N-glycans Because of its orientation tucked "behind" the chitobiose core defined in the context of plants N-glycans earlier, the effect of core α(1-3)-Fuc on the (1-6) arm equilibrium within an A2G2-xylosylated scaffold is not as significant as α(1-6)-Fuc. As shown in Figure 3.5 and Table A.11 in Appendix A, this lack of direct effect is demonstrated by the recovery of the same equilibrium as the non-fucosylated A2G2-xylosylated system. The dynamics of the chitobiose core is very similar to the one determined for the corresponding plant N-glycan. To analyse the effect of core α(1-3) fucosylation without β(1-2)-Xyl, we have looked at two A2G2 hybrid systems, one with only α(1-3)-linked fucose and one with both core α(1-3)- and α(1-6)-linked fucose, a characteristic "double-fucose" glycosylation found in worm and fly cells [30] as shown in Table A.12 in Appendix A unlike in plants N-glycans, the α(1-3)-Fuc alone does not affect the A2G2 (1-6) arm equilibrium [24] as the folding of the (1-6) arm with terminal β(1-4)-Gal is not obstructed by the rotation of the chitobiose core imposed by the α(1-3)-Fuc position. When both α(1-3)- and α(1-6)-linked fucoses are present the (1-6) arm with terminal β(1-4)-Gal is predominantly folded (closed) at 85%, see Figure 3.6 and Table A.11 in Appendix A, this lack of direct effect is demonstrated by the recovery of the same equilibrium as the non-fucosylated A2G2-xylosylated system. The dynamics of the chitobiose core is very similar to the one determined for the corresponding plant N-glycan. To analyse the effect of core α(1-3) fucosylation without β(1-2)-Xyl, we have looked at two A2G2 hybrid systems, one with only α(1-3)-linked fucose and one with both core α(1-3)- and α(1-6)-linked fucose, a characteristic "double-fucose" glycosylation found in worm and fly cells [30]. As shown in Table A.12 in Appendix A unlike in plants N-glycans, the α(1-3)-Fuc alone does not affect the A2G2 (1-6) arm equilibrium [24] as the folding of the (1-6) arm with terminal β(1-4)-Gal is not obstructed by the rotation of the chitobiose core imposed by the α(1-3)-Fuc position. When both α(1-3)- and α(1-6)-linked fucoses are present the (1-6) arm with terminal β(1-4)-Gal is predominantly folded (closed) at

Figure 3.5:   Conformational analysis of the (1-6) arm in four hybrid *N*-glycoforms, β(1-2)-xylosylated A2G2 (top-left), β(1-2)-xylosylated FA2G2 (bottom-left), β(1-2)-xylosylated α(1-3)-core fucosylated A2G2 (top-right) and β(1-2)-xylosylated FA2 (bottom-right). The predominant conformations are indicated in the top- and bottom-left heat maps for simplicity. The simulation time relative to each system is indicated in the top right corner of each heat map. The *N*-glycans structures are shown with the (1-3) and (1-6) arms on the left and on the right, respectively. The monosaccharides colouring follows the SFNG nomenclature. The structure rendering was done with VMD (https://www.ks.uiuc.edu/Research/vmd/) and the graphical statistical analysis with RStudio (http://www.glycam.org).

85%, see Figure 3.6 and Table A.13 in Appendix A, which is higher than in the absence of α(1-3)-Fuc [24]. Indeed, the latter can actively contribute in stabilizing the interaction with the terminal β(1-4)-Gal of the folded (1-6) arm. We also observed interesting events, one representing 10% of 2 μs as indicated by the values of the GlcNAc-β(1-4)-GlcNAc torsion, where the GlcNAc is stacked in between the two fucose residues and another one, contributing to 18% of the simulation time, 14% when the system is also xylosylated, in which the GlcNAc ring transitions from $^4C_1$ to $^1C_4$ allowing the two fucose to stack, see Table A.13, Table A.14 and Figure A.2 in Appendix A.
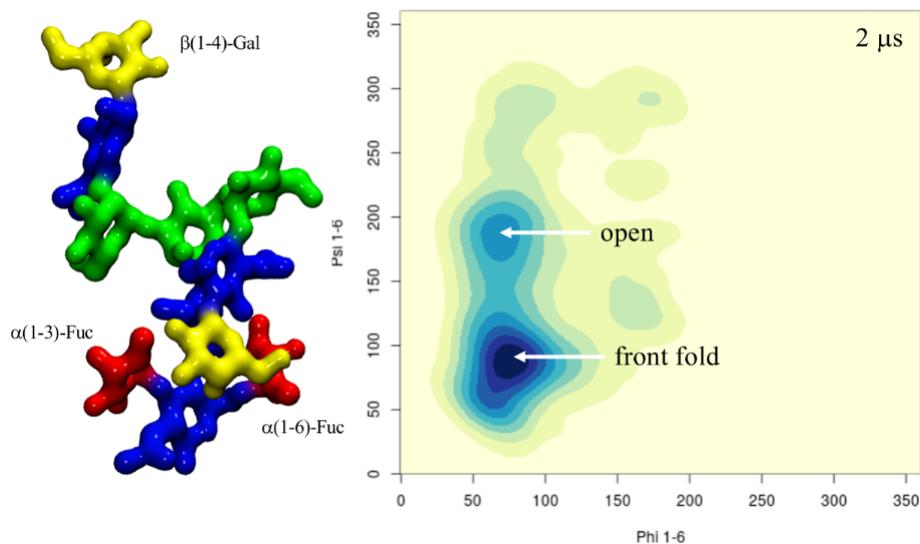
Figure 3.6: Conformational equilibrium of the (1-6) arm in terms of phi/psi torsion angle values for the α(1-3)-fucosylated FA2G2 *N*-glycoform. The structure with the folded (1-6) arm where the terminal β(1-4)-Gal interacts with both fucose residues is shown on the left-hand side. The *N*-glycans structures are shown with the (1-3) and (1-6) arms on the left and on the right, respectively. The monosaccharides colouring follows the SFNG nomenclature. The structure rendering was done with VMD (https://www.ks.uiuc.edu/Research/vmd/) and the graphical statistical analysis with RStudio (http://www.glycam.org).

3.5 Discussion

Differences and similarities in *N*-glycan sequences are highly cell-specific as well as important indicators of health and disease states [1, 39]. Exogenous *N*-glycans motifs can be quite subtle, yet trigger profound differences in terms of molecular recognition [19, 27] and dangerous immunogenic responses [20–22] In this work we have analysed the effects on the *N*-glycans structure and dynamics of two motifs in particular, namely β(1-2)-Xyl and core α(1-3)-Fuc, common in plants [23] and invertebrates [30] but completely absent in mammalian *N*-glycans. Within the context of plant-type *N*-glycans, which have a terminal β(1-3)-Gal, rather than β(1-4)-Gal, both β(1-2)-Xyl and α(1-3)-Fuc contribute independently in promoting an outstretched (open) conformation of the (1-6) arm because of steric hindrance of the xylose and of the rotation forced upon the chitobiose core by the α(1-3)-Fuc. The latter is not an obstruction for the folding of a β(1-4)-Gal terminated (1-6) arm, as we have seen in the hybrid *N*-glycans constructs. Therefore, in β(1-2) xylosylated *N*-glycans terminating with β(1-3)-Gal, both arms should be more available for recognition, binding and further functionalization [30] unlike in mammalian *N*-glycans where the β(1-4)-Gal determines a prevalently closed and inaccessible (1-6) arm [24, 28] Also, the analysis of the structure and dynamics of the LeA terminating plant *N*-glycans showed that the specific branching and spatial orientation of the motif allowed for a stable interaction between the arms, which is not observed in complex *N*-glycans with a linear functionalization of the arms [24]. Notably, the same hydrogen bonding network between the arms cannot be established when the same *N*-glycan terminates with LeX, because of the non-complementary position of the α(1-3)-Fuc deoxy-C6.

The analysis of all these different complex *N*-glycoforms clearly shows that every modification, addition or removal of a specific motif, can greatly affect the 3D architecture of

the *N*-glycan, thus its accessibility and complementarity to a receptor. However, these effects are rather complex to understand or to predict, if we think of the *N*-glycans 3D structure in terms of sequence of monosaccharides, a view that stems from the way we think about proteins. Our results show that the main effect of all functionalizations is actually local. For example, the core α(1-3)-Fuc forces a rotation of the chitobiose, a degree of freedom very lowly populated otherwise; meanwhile, β(1-2)-Xyl restricts the flexibility of the trimannose core and occupies its centre. Within this framework, the 3D structural and dynamics features of the *N*-glycoforms can be rationalized by discretizing their architecture in terms 3D units, or "glycoblocks", that group monosaccharides and their linkages within their immediate spatial vicinity, e.g., the core α(1-3)-Fuc and the chitobiose which structure it has modified. A list of the glycoblocks that we have identified with the corresponding descriptors of their 3D features are listed in Figure 3.7. The whole *N*-glycan 3D architecture, in terms of the structures accessible and their conformational propensity, can be then described through the combination of these glycoblocks, together with the knowledge of their dynamic properties and flexibility. Also, consideration of these glycoblocks as spatial units can be useful to understand recognition by lectins and antibodies, which is often affected primarily by the targeted monosaccharide's immediate vicinity and by its accessibility within a specific glycoform. For example, if we consider the 3D structure of the β(1-2)-Xyl Man3 glycoblock vs the Man3 without Xyl, we can understand how the β(1-2)-Xyl position within that unit negates binding to DC-SIGN lectins [19] see Appendix A, Figure A.3 panels a and b. Additionally, we can see that the slight rotation on the chitobiose imposed by the core α(1-3)-Fuc does not prevent recognition and binding, Appendix A, Figure A.3 panel c.

Glycoblock	Unit 3D structure	Dominant motif descriptor	Characteristic glycosidic linkage		
		Very stable conformation, with a low degree of flexibility around the equilibrium structure	**GlcNAc β(1-4) GlcNAc** phi = -78.7 (11.1)/100 psi = -130.8 (15.7)/99		
		The α(1-3)-Fuc stacking interaction imposes a 20° rotation around the chitobiose glycosidic linkage and further stabilizes the chitobiose structure	**GlcNAc β(1-4) GlcNAc** phi = -72.1 (8.3)/100 psi = -107.1 (7.6)/100		
		The longer glycosydic linkage provides the α(1-6)-Fuc with more degrees of freedom and does not alter the chitobiose structure	**GlcNAc β(1-4) GlcNAc** phi = -77.8 (11.2)/100 psi = -127.8 (14.8)/100		
		This motif is structurally similar to the one with the α(1-3)-Fuc, which acts as the dominant restraint. The presence of the α(1-6)-Fuc increases flexibility	**GlcNAc β(1-4) GlcNAc** phi = -73.9 (9.5)/100 psi = -106.4 (14.1)/90		
		Joint with differently flexible arms. The α(1-6) torsion has two dominant, different conformations accessible that can open and close the (1-6) arm. Populations depend on the attached glycoblock units	**Man α(1-6) Man** (with terminal β(1-4)-Gal) open phi = 76.3 (15.0)/100 psi = -185.1 (22.1)/23 closed phi = 76.3 (15.0)/100 psi = 85.3 (15.8)/76		
		The addition of the β(1-2)-Xyl further restricts the (1-3) arm and hinders the closing forward of the (1-6) arm, which is only favored in (F)A2G2 hybrids	**Man α(1-6) Man** (with terminal β(1-3)-Gal) open phi = 70.5 (10.4)/100 psi = -173.6 (16.3)/70 closed phi = 70.5 (10.4)/100 psi = 103.8 (12.7)/26		
β3		This unit is rather rigid, while the orientation of the terminal Gal in β(1-3/4) is determinant for the interaction with the chitobiose in the closed (1-6) arm. These units in the two arms do not interact	**Gal β(1-3) GlcNAc** phi = -72.9 (11.2)/100 psi = -124.0 (18.2)/96		
β3 α4		The Lewis A epitope is stable in its closed conformation and its presence greatly affects the (1-6) arm conformation because branching promotes association of the arms	**Man α(1-6) Man** (with terminal LeA and α(1-3)-Fuc) open phi = -70.1 (5.7)/100 psi = -177.1 (16.3)/28 closed phi = -70.1 (5.7)/100 psi = 148.2 (7.4)/54		
β4		This unit is rather rigid, while the orientation of the terminal Gal in β(1-3/4) is determinant for the interaction with the chitobiose in the closed (1-6) arm. These units in the two arms do not interact	**Gal β(1-4) GlcNAc** phi = -76.2 (15.1)/100 psi = -125.7 (15.4)/97		
β4 α3		The Lewis X epitope is stable in its closed conformation. Contrary to Lewis A its branching does not favor association of the arms because of the relative positions of the deoxy-C_6.	**Man α(1-6) Man** (with terminal LeX, core α(1-3)-Fuc and β(1-2)-Xyl) open phi = 70.7 (9.2)/90 psi = -173.6 (13.4)/90 closed phi = 152.4 (12.9)/10 psi = 145.6 (12.3)/10		

Figure 3.7: List of 3D structural units of monosaccharides (glycoblocks) that regulate the 3D architecture and dynamics of complex biantennary *N*-glycans from plants and invertebrate sources and hybrid mammalian constructs. The SFNG representation of each glycoblock is indicated in the first column from the left, 3D structures from the highest populated conformers are shown in the second column, rendered with VMD (https://www.ks.uiuc.edu/Research/vmd/). A brief summary of the conformational features of each glycoblock and the characteristic linkage or its effect on the (1-6) arm conformation are indicates in the last two columns, respectively.

3.6   Conclusion

In this work we used extensive sampling through MD simulations to study the effects on the N-glycan architecture of subtle, yet highly consequential modifications, namely core α(1-3)-Fuc and β(1-2)-Xyl [19] These are part of standard N-glycoforms found in plants [23] and invertebrates [30], but immunogenic in humans [21, 22, 26]. Our results show that these modifications can greatly affect the 3D structure of the N-glycan and its structural dynamics, therefore its selective recognition by lectin receptors and antibodies. The atomistic-level of detail information that the MD simulations provide us with, highlights that the effects of different functionalizations, in terms of monosaccharide types and linkages, are primarily local, affecting the immediate spatial vicinity of the monosaccharide within the N-glycan structure. Within this framework, we propose an alternative approach that can help to describe and predict the architecture of N-glycans based on the combination of structural 3D units, or glycoblocks. Unlike a description based on the monosaccharide sequence and linkages as two separate features, the transition to well-defined and self-contained units, integrating information on both monosaccharides and linkages, can help us rationalize and deconvolute the glycans structural disorder and ultimately understand more clearly the relationships between sequence and structure in complex carbohydrates.

3.7   Acknowledgements

3.8   Funding

Bibliography

[1] Varki, A. Biological roles of glycans. Glycobiology **2016**, 27, 3–49.

[2] Stanley, P. In Essentials of Glycobiology, 3rd ed.; Varki, A., Ed.; Cold Spring Harbor Laboratory Press, p 99–112.

[3] Cobb, B. A. The history of IgG glycosylation and where we are now. Glycobiology **2020**, 30, 202–213.

[4] Ferrara, C.; Grau, S.; Jäger, C.; Sondermann, P.; Brünker, P.; Waldhauer, I.; Hennig, M.; Ruf, A.; Rufer, A. C.; Stihle, M.; Umaña, P.; Benz, J. Unique carbohydrate-carbohydrate interactions are required for high affinity binding between FcγRIII and antibodies lacking core fucose. Proc. Natl. Acad. Sci. U. S. A. **2011**, 108, 12669–12674.

[5] Day, C. J.; Tran, E. N.; Semchenko, E. A.; Tram, G.; Hartley-Tassell, L. E.; Ng, P. S. K.; King, R. M.; Ulanovsky, R.; McAtamney, S.; Apicella, M. A.; Tiralongo, J.; Morona, R.; Korolik, V.; Jennings, M. P. Glycan:glycan interactions: High affinity biomolecular interactions that can mediate binding of pathogenic bacteria to host cells. Proc. Natl. Acad. Sci. U. S. A. **2015**, 112, E7266–75.

[6] Nagae, M.; Yamaguchi, Y. Function and 3D Structure of the N-Glycans on Glycoproteins. International Journal of Molecular Sciences **2012**, 13, 8398–8429.

[7] Zacchi, L. F.; Schulz, B. L. N-glycoprotein macroheterogeneity: biological implications and proteomic characterization. Glycoconjugate Journal **2015**, 33, 359–376.

[8] Leoz, M. L. A. D. et al. NIST Interlaboratory Study on Glycosylation Analysis of Monoclonal Antibodies: Comparison of Results from Diverse Analytical Methods. Molecular & Cellular Proteomics **2020**, 19, 11–30.

[9] Mimura, Y.; Katoh, T.; Saldova, R.; O'Flaherty, R.; Izumi, T.; Mimura-Kimura, Y.; Utsunomiya, T.; Mizukami, Y.; Yamamoto, K.; Matsumoto, T.; Rudd, P. M. Glycosylation engineering of therapeutic IgG antibodies: challenges for the safety, functionality and efficacy. Protein & Cell **2017**, 9, 47–62.

[10] Thaysen-Andersen, M.; Packer, N. H. Advances in LC–MS/MS-based glycoproteomics: Getting closer to system-wide site-specific mapping of the N- and O-glycoproteome. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics **2014**, 1844, 1437–1452.

[11] Alocci, D.; Mariethoz, J.; Gastaldello, A.; Gasteiger, E.; Karlsson, N. G.; Kolarich, D.; Packer, N. H.; Lisacek, F. GlyConnect: Glycoproteomics Goes Visual, Interactive, and Analytical. Journal of Proteome Research **2018**, 18, 664–677.

[12] Lisacek, F.; Mariethoz, J.; Alocci, D.; Rudd, P. M.; Abrahams, J. L.; Campbell, M. P.; Packer, N. H.; Ståhle, J.; Widmalm, G.; Mullen, E.; Adamczyk, B.; Rojas-Macias, M. A.; Jin, C.; Karlsson, N. G. In High-Throughput Glycomics and Glycoproteomics: Methods and Protocols; Lauc, G., Wuhrer, M., Eds.; Springer New York: New York, NY, 2017; pp 235–264.

[13] Mariethoz, J.; Alocci, D.; Gastaldello, A.; Horlacher, O.; Gasteiger, E.; Rojas-Macias, M.; Karlsson, N. G.; Packer, N. H.; Lisacek, F. Glycomics@ExPASy: Bridging the Gap. Molecular & Cellular Proteomics **2018**, 17, 2164–2176.

[14] Aoki-Kinoshita, K. et al. GlyTouCan 1.0 – The international glycan structure repository. Nucleic Acids Research **2015**, 44, D1237–D1242.

[15] Rojas-Macias, M. A. et al. Towards a standardized bioinformatics infrastructure for N- and O-glycomics. Nature Communications **2019**, 10.

[16] Durrant, J. D.; Kochanek, S. E.; Casalino, L.; Ieong, P. U.; Dommer, A. C.; Amaro, R. E. Mesoscale All-Atom Influenza Virus Simulations Suggest New Substrate Binding Mechanism. ACS Central Science **2020**, 6, 189–196.

[17] Yu, I.; Mori, T.; Ando, T.; Harada, R.; Jung, J.; Sugita, Y.; Feig, M. Biomolecular interactions modulate macromolecular structure and dynamics in atomistic model of a bacterial cytoplasm. eLife **2016**, 5.

[18] Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Shaw, D. E. Picosecond to Millisecond Structural Dynamics in Human Ubiquitin. The Journal of Physical Chemistry B **2016**, 120, 8313–8320.

[19] Brzezicka, K.; Echeverria, B.; Serna, S.; van Diepen, A.; Hokke, C. H.; Reichardt, N.-C. Synthesis and Microarray-Assisted Binding Studies of Core Xylose and Fucose Containing N-Glycans. ACS Chemical Biology **2015**, 10, 1290–1302.

[20] Wilson, I. B. H.; Harthill, J. E.; Mullin, N. P.; Ashford, D. A.; Altmann, F. Core 1, 3-fucose is a key part of the epitope recognized by antibodies reacting against plant N-linked oligosaccharides and is present in a wide variety of plant extracts. Glycobiology **1998**, 8, 651–661.

[21] van Ree, R.; Cabanes-Macheteau, M.; Akkerdaas, J.; Milazzo, J.-P.; Loutelier-Bourhis, C.; Rayon, C.; Villalba, M.; Koppelman, S.; Aalberse, R.; Rodriguez, R.; Faye, L.; Lerouge, P. $\beta(1, 2)$-Xylose and $\alpha(1, 3)$-Fucose Residues Have a Strong Contribution in IgE Binding to Plant Glycoallergens. Journal of Biological Chemistry **2000**, 275, 11451–11458.

[22] Bardor, M. Immunoreactivity in mammals of two typical plant glyco-epitopes, core alpha(1, 3)-fucose and core xylose. Glycobiology **2002**, 13, 427–434.

[23] Strasser, R. Plant protein glycosylation. Glycobiology **2016**, 26, 926–939.

[24] Harbison, A. M.; Brosnan, L. P.; Fenlon, K.; Fadda, E. Sequence-to-structure dependence of isolated IgG Fc complex biantennary N-glycans: a molecular dynamics study. Glycobiology **2018**, 29, 94–103.

[25] Varki, A., Ed. Essentials of glycobiology, third edition; Cold Spring Harbor Laboratory Press: New York, NY, 2017.

[26] Fitchette-Laine, A.-C.; Gomord, V.; Cabanes, M.; Michalski, J.-C.; Macary, M.; Foucher, B.; Cavelier, B.; Hawes, C.; Lerouge, P.; Faye, L. N-glycans harboring the Lewis a epitope are expressed at the surface of plant cells. The Plant Journal **1997**, 12, 1411–1417.

[27] Echeverria, B.; Serna, S.; Achilli, S.; Vivès, C.; Pham, J.; Thépaut, M.; Hokke, C. H.; Fieschi, F.; Reichardt, N.-C. Chemoenzymatic Synthesis of N-glycan Positional Isomers and Evidence for Branch Selective Binding by Monoclonal Antibodies and Human C-type Lectin Receptors. ACS Chemical Biology **2018**, 13, 2269–2279.

[28] Barb, A. W.; Brady, E. K.; Prestegard, J. H. Branch-Specific Sialylation of IgG-Fc Glycans by ST6Gal-I. Biochemistry **2009**, <u>48</u>, 9705–9707.

[29] Möginger, U.; Grunewald, S.; Hennig, R.; Kuo, C.-W.; Schirmeister, F.; Voth, H.; Rapp, E.; Khoo, K.-H.; Seeberger, P. H.; Simon, J. C.; Kolarich, D. Alterations of the Human Skin N- and O-Glycome in Basal Cell Carcinoma and Squamous Cell Carcinoma. Frontiers in Oncology **2018**, <u>8</u>.

[30] Paschinger, K.; Wilson, I. B. H. Comparisons of N-glycans across invertebrate phyla. Parasitology **2019**, <u>146</u>, 1733–1742.

[31] University of California: San Francisco, U., CA AMBER 2018. 2018.

[32] Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; González-Outeiriño, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. GLYCAM06: A generalizable biomolecular force field. Carbohydrates. Journal of Computational Chemistry **2007**, <u>29</u>, 622–655.

[33] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. The Journal of Chemical Physics **1983**, <u>79</u>, 926–935.

[34] Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. Journal of Molecular Graphics **1996**, <u>14</u>, 33–38.

[35] Alibay, I.; Burusco, K. K.; Bruce, N. J.; Bryce, R. A. Identification of Rare Lewis Oligosaccharide Conformers in Aqueous Solution Using Enhanced Sampling Molecular Dynamics. The Journal of Physical Chemistry B **2018**, <u>122</u>, 2462–2474.

[36] Mayes, H. B.; Broadbelt, L. J.; Beckham, G. T. How Sugars Pucker: Electronic Structure Calculations Map the Kinetic Landscape of Five Biologically Paramount Monosaccharides and Their Implications for Enzymatic Catalysis. Journal of the American Chemical Society **2014**, <u>136</u>, 1008–1022.

[37] Rönnols, J.; Manner, S.; Siegbahn, A.; Ellervik, U.; Widmalm, G. Exploration of conformational flexibility and hydrogen bonding of xylosides in different solvents, as a model system for enzyme active site interactions. Organic & Biomolecular Chemistry **2013**, <u>11</u>, 5465.

[38] Iglesias-Fernández, J.; Raich, L.; Ardèvol, A.; Rovira, C. The complete conformational free energy landscape of β-xylose reveals a two-fold catalytic itinerary for β-xylanases. Chemical Science **2015**, <u>6</u>, 1167–1177.

[39] Reily, C.; Stewart, T. J.; Renfrow, M. B.; Novak, J. Glycosylation in health and disease. Nature Reviews Nephrology **2019**, <u>15</u>, 346–366.

4 Chapter 4: Oligomannose *N*-Glycans 3D Architecture and Its Response to the FcγRIIIa Structural Landscape

The following text is reproduced verbatim from the paper.

4.1 Abstract

Oligomannoses are evolutionarily the oldest class of *N*-glycans, where the arms of the common pentasaccharide unit, *i.e.*, Manα(1–6)-[Manα(1–3)]-Manβ(1–4)-GlcNAcβ(1–4)-GlcNAcβ1-Asn, are functionalized exclusively with branched arrangements of mannose (Man) monosaccharide units. In mammalian species oligomannose *N*-glycans can have up to 9 Man; meanwhile structures can grow to over 200 units in yeast mannan. The highly dynamic nature, branching complexity, and 3D structure of oligomannoses have been recently highlighted for their roles in immune escape and infectivity of enveloped viruses, such as HIV-1 and SARS-CoV2. The architectural features that allow these *N*-glycans to perform their functions are yet unclear, due to their intrinsically disordered nature that hinders their structural characterization. In this work we will discuss the results of over 54 μs of cumulative sampling by molecular dynamics (MD) simulations of differently processed, free (not protein-linked) oligomannose *N*-glycans common in vertebrates. We then discuss the effects of a protein surface on their structural equilibria based on over 4 μs cumulative MD sampling of the fully glycosylated CD16a Fc γ receptor (FcγRIIIa), where the type of glycosylation is known to modulate its binding affinity for IgG1s, regulating the antibody-dependent cellular cytotoxicity (ADCC). Our results show that the protein's structural constraints shift the oligomannoses conformational ensemble to promote conformers that satisfy the steric requirements and hydrogen bonding networks demanded by the protein's surface landscape. More importantly, we find that the protein does not actively distort the *N*-glycans into structures not populated in the unlinked forms in solution. Ultimately, the highly populated conformations of the Man5 linked glycans support experimental evidence of high levels of hybrid complex forms at N45 and show a specific presentation of the arms at N162, which may be involved in mediating binding affinity to the IgG1 Fc.

4.2 Introduction

Complex carbohydrates (or glycans) are the most abundant biomolecules in nature. Within a human biology context, glycans coat cell membranes and protein surfaces, mediating a myriad of essential biological processes in health and disease states.[1–6] *N*-glycosylation is one of the most abundant and diverse type of post-translational modification that can affect protein trafficking and structural stability and mediate interactions with different receptors [6–11]. *N*-glycan recognition and binding affinities are often highly specific to their sequence, intended as the types of monosaccharides, their stereochemistry, and branching patterns, [12] a principle that has been successfully exploited in the development of glycan microarray technology [13].

Molecular recognition is fundamentally dependent, among other considerations, on structural and electrostatic complementarity between the ligand and the receptor's binding site.
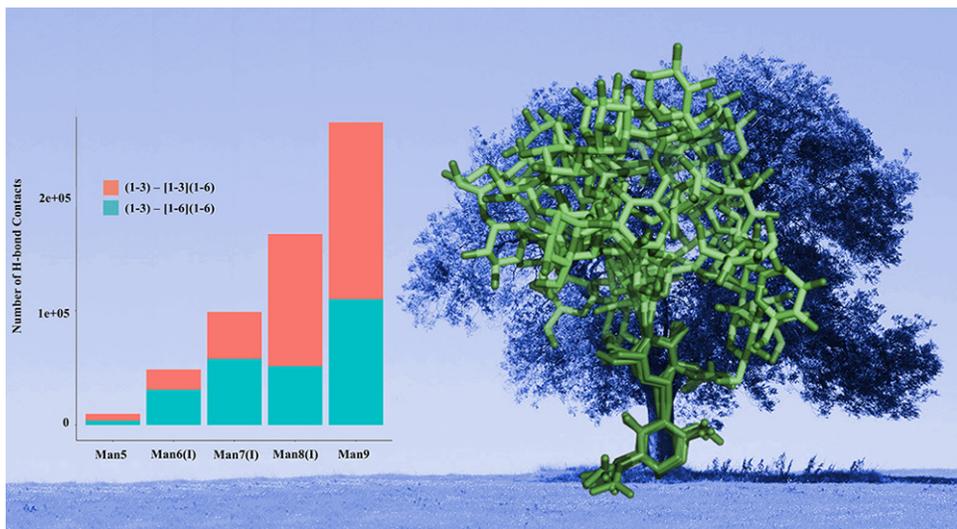
Within this framework, the prediction and characterization of glycan binding specificity are an extremely difficult task, due to their high degree of flexibility or intrinsic disorder, which hinders our ability to determine their 3D structure by means of experimental techniques. Indeed, glycans can only be structurally resolved in their entirety only when tightly bound to a receptor, thus when their conformational degrees of freedom are heavily restrained. Because of their inherent flexibility, free glycans can adopt different 3D structures within a weighted conformational ensemble, which cannot be determined with currently available experimental methods, although very promising steps forward have been recently made in advancing imaging techniques for single glycans[14, 15].

High performance computing (HPC) molecular simulations can contribute a great deal toward our understanding of the relationships between glycans' sequence, structure, and function. Indeed, conformational sampling through conventional and/or enhanced molecular dynamics (MD) schemes allows us to characterize the dynamic behavior of different glycoforms at the atomistic level of details. Within this context, for the past few years our lab contributed to the knowledge of *N*-glycans dynamics by providing information on their 3D architecture and relative flexibility from extensive MD-based conformational sampling. [16, 17] As an example, we have shown how the sequence (and branching) of complex *N*-glycans determines the 3D structure, which in turn drives their recognition [16, 17].In this work we extend our data set of free (unlinked) *N*-glycans structures to the vertebrate oligomannose type, where, as shown in 4.1 , the common pentasaccharide unit, i.e., Manα(1–6)-[Manα(1–3)]-Manβ(1–4)-GlcNAcβ(1–4)-GlcNAcβ1-Asn, is functionalized by a branched arrangement of only Man units. In addition, we also determine how the protein surface landscape affects their conformational dynamics, which is a very important question in terms of its impact on molecular recognition and function while challenging to answer in absolute terms because of the site-specific character.

Oligomannoses are often defined as "immature" *N*-glycans, as they are processed toward complex functionalization in the Golgi [6] and are not abundant in vertebrates. Nevertheless, these *N*-glycans are a common post-translational modification of viral envelope proteins expressed in human cell lines; [20, 21].for example, it is the prevalent type of glycosylation of the HIV-1 fusion trimer. [22–25].Furthermore, an increase in large oligomannose-type *N*-glycosylation in humans has been linked to breast cancer progression [26–28].and can occur where the protein landscape at the *N*-glycan site does not allow easy access to the required glycohydrolases and glycotransferases for further functionalization [6, 29, 30]Interestingly, recent work has shown that oligomannose *N*-glycans functionalizing CD16a low-affinity Fc γ receptors (FcγIIIa) determine an increase in IgG1-binding
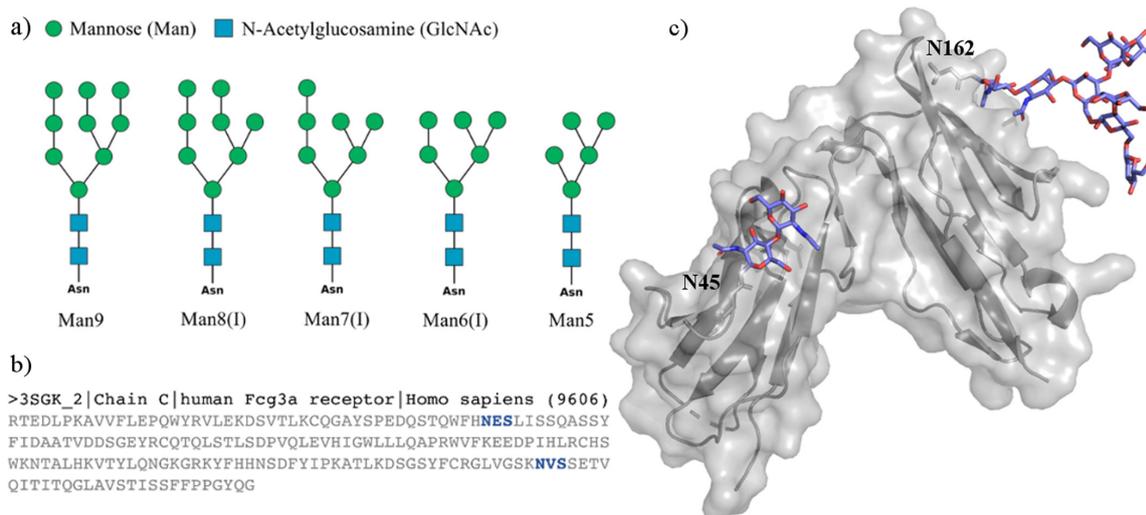
Figure 4.1: (a) SNFG representation [18] of a subset of the oligomannose *N*-glycans discussed in this work. Man6/7/8(I) indicate specific positional isomers. For the complete list of isomers see B.1 (b) Sequence of the human CD16a (FcγIIIa) of the PDB entry 3SGK [19] with the occupied sequons highlighted in blue. (c) Structure of the human CD16a (FcγIIIa) from PDB entry 3SGK with the resolved *N*-glycans at positions N45 and N162 represented with sticks. Image was rendered with VMD (http://www.ks.uiuc.edu/Research/vmd/). *N*-Glycan sketches were rendered with DrawGlycan (http://www.virtualglycome.org/DrawGlycan/).

affinity by 51-fold,[31]relative to the more common complex *N*-glycans,[32] although the *N*-glycosylation composition varies depending on the glycosylation site. [32] In this work we have studied the effect of the FcγIIIa protein surface landscape on the intrinsic conformational propensity of different oligomannose *N*-glycans we determined for the unlinked forms. Our results show that the two FcγIIIa *N*-glycosylation sites, N45 and N162, affect the oligomannose dynamics rather differently, in function of the structural constrains of the sites and of the 3D architecture of the glycan. More specifically, we find that the protein landscape affects the glycans conformational equilibrium by promoting structures that are complementary to it and not by actively changing their intrinsic architecture. Indeed, all the 3D conformers observed in the analysis of the bound oligomannoses are always identified in the simulations of the corresponding unlinked forms in solution, although in different populations. Interestingly, we also determined that the progressive elongation of the arms/branches promotes interarm contacts, where the Man9 3D architecture is almost entirely structured with interacting arms. Finally, these findings fit very well within the framework of our recently proposed "glycoblocks" glycans structure representation,[16] whereby groups of specifically linked monosaccharides within *N*-glycans represent independent structural elements (or glycoblocks), whose exposure, or presentation in function of the particular protein landscape, drives molecular recognition.

4.3 Computational Method

All 12 oligomannose starting structures, shown in 4.1 and B.1 were obtained with the GLY-CAM carbohydrate builder online tool (https://glycam.org/). For each of these oligomannoses, we built nine structures characterized by different combinations of the two $\alpha(1-6)$ torsions values. Complete topology and parameter files were generated with the tLEaP

tool from version 18 of the AMBER software package,[33] with the GLYCAM06-j1 parameter set [34] to represent the carbohydrates and the TIP3P model [35] for water molecules. Because our simulations do not involve the calculation of hydration or of binding free energies [36, 37] and also because of consistency with our previous work, [16, 17] we consider the choice of GLYCAM06-j1/TIP3P parameter set as appropriate. All simulations were run in 200 mM NaCl salt concentration, with counterions represented by AMBER parameters[38] in a cubic simulation box of 16 Å sides. Long range electrostatic were treated by particle mesh Ewald (PME) with cutoff set at 11 Å and a B-spline interpolation for mapping particles to and from the mesh of order of 4. van der Waals (vdW) interactions were cut off at 11 Å. The MD trajectories were generated by Langevin dynamics with collision frequency of 1.0 ps^{-1}. Pressure was kept constant by isotropic pressure scaling with a pressure relaxation time of 2.0 ps. After an initial 500.000 cycles of steepest descent energy minimization, with all protein/glycans heavy atoms restrained by a harmonic potential with a force constant of 5 $kcal\ mol^{-1}$Å^{-2}, the system was heated in two stages, i.e., from 0 to 100 K over 500 ps at constant volume and then from 100 to 300 K over 500 ps at constant pressure. After the heating phase, all restraints were removed and the system was allowed to equilibrate for 5 ns at 300 K and at 1 atm of pressure. Production and subsequent analysis were done on 500 ns trajectories run in parallel for each uncorrelated starting structure, i.e., each conformer generated with GLYCAM-Web. Analysis was done using the cpptraj tool and with VMD [39] ((https://www.ks.uiuc.edu/Research/vmd/). The dihedral distributions from the trajectories were obtained in terms of kernel density estimates (KDE), with a smoothing parameter of [1000,1000], with the ks package in R and rendered with heat maps with RStudio (http://www.rstudio.com/) in conjunction with the DBSCAN clustering algorithm. The highest populated conformers resulting from the analysis of Man5 and Man9 were then grafted in positions N45 and N162 of the FcγRIIIa (PDB code 3SGK) by structural alignment to the resolved chitobiose at N45 and at N162, see 4.1 c,to obtain two systems, one with only Man5 and the other with only Man9 at both positions. As a note, the structure of the N-glycan at N165 from the PDB structure is quite distorted with uncommon ring conformations of some of the monosaccharides, probably resulting from the fitting to the electron density; therefore it was disregarded and only the chitobiose was used for structural alignment. These systems were run in duplicates from uncorrelated starting structures with the same simulation protocol used for the free glycans. Production runs were extended to 1 μs for each trajectory for a total of 4 μs of cumulative sampling time. All simulations were run on NVIDIA Tesla V100 16GB PCIe (Volta architecture) GPUs on resources from the Irish Centre for High-End Computing (ICHEC) (www.ichec.ie).

4.4 Results

We used conventional MD simulations, run in parallel for 500 ns from nine uncorrelated starting points, [16, 17] to characterize the 3D structure and dynamics of human oligomannose N-glycans, when unlinked; see 4.1 amd B.1. The effects of the protein on their intrinsic dynamics were studied on two models with Man5 and Man9 linked to the human FcγIIIa protein on the two N-glycosylation sites, namely, N45 and N162; see 4.1 c. This section is organized as follows, first we present the results obtained for the unlinked oligomannoses, starting with Man5 that we used as a reference to describe sequence-to-structure changes in the larger forms. The subset of representative isomers shown in 4.1 is presented here for simplicity, while the complete analysis of all positional isomers with heat maps and tables is included as B.The section concludes with the results obtained for Man5 and Man9 when linked to the FcγRIIIa.

Man5 is the simplest oligomannose found in vertebrates and the substrate of GlcNAc

transferase I (GnTI), responsible for starting the *N*-glycan complex functionalization in the Golgi. [6] As found for complex biantennary *N*-glycans, [16, 17] the Man5 chitobiose core and the following Manβ(1–4)-GlcNAc linkage are rigid with only one conformation significantly occupied, see 4.2 and B.1 while the (1–3) arm adopts an outstretched conformation with flexibility in a range of 40° around the ψ torsion angle; see 4.2 and 4.4. The Man5 (1–6) arm has a relatively more complex dynamics, hinging around the preferential "open" conformation, [16, 17] populated at 82%, where the Manα(1–3)-Man branch can be orientated toward the front of the page and the Manα(1–6)-Man branch toward the back of the page or vice versa. We also identified two alternative, less populated conformers, namely, a "front fold" ($\phi = 79°$, $\psi = 87°$) with a relative population of 12% and a "back fold" ($\phi = 83°$, $\psi = $ -76°) with a relative population of 6%; see 4.2 and 4.4. In the front fold the terminal Manα(1–6)-Man interacts through hydrogen bonds with the *N*-acetyl group of the second core GlcNAc, pushing Manα(1–3)-Man upward, while the back fold is stabilized by hydrogen bonds between the terminal Manα(1–6)-Man and both GlcNAc residues of the chitobiose; see 4.2 b. The open conformation can be further analyzed in terms of the α(1–6) linkage ω torsion angle that determines different orientations of the Manα(1–6)-Man and Manα(1–3)-Man branches relative to the core. As shown in 4.3, there are two dominant conformers contributing to the open structure, open (cluster 1) populated at 48% with $\omega = 56°$ and open (cluster 2) populated at 33% with $\omega = $ -175°; see 4.4. The dynamics of the branches on the Man5 (1–6) arm follows the same pattern observed in the (1–3) and (1–6) arms, with only small differences dictated by their immediate environment. For example, the terminal Manα(1–6)-Man is found predominantly (80%) in the open (cluster 1) conformation, see 4.4, and does not show a back fold orientation.
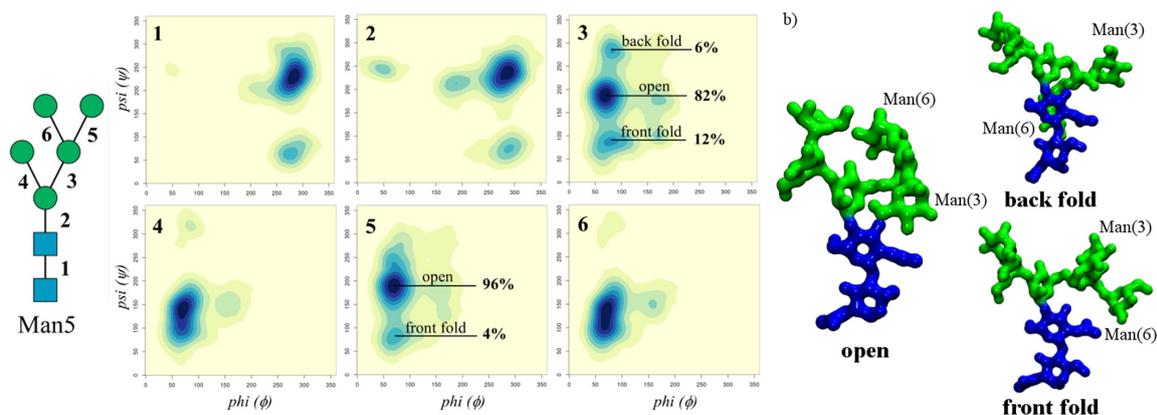
Figure 4.2:   Figure 2. (a, Left) Man5 conformational analysis in terms of the ϕ and ψ torsion angles, with axes ranging from 0° to 360°. Each torsion is numbered as indicated on the left-hand side, and the heat maps are labeled in the top-left corner accordingly. (b) 3D structures of the dominant conformers determined by the flexibility of the (1–6) arm. The Man(3/6) labels indicate the position of the Man on the 3/6 branch on the (1–6) arm. Heat maps were made with RStudio (www.rstudio.com), and molecular models were rendered with VMD (http://www.ks.uiuc.edu/Research/vmd/). *N*-glycans are colored according to the SNFG convention.

Table 4.1: **Torsion Angle Median Values of the Linkages in the Man5 (1–3/6) Arms:** Standard deviation values are shown in parentheses, with relative populations obtained from clustering analysis. Angle values are in degrees. The number in parentheses in the first column indicates the linkages, as shown on the Man5 sketch in Figure 4.2

Manα(1–6)-Man arm (3)	ϕ	ψ	ω	population (%)
open (cluster 1)	71 (11)	–172 (17)	56 (11)	49
open (cluster 2)	68 (10)	–175 (14)	–175 (13)	33
front fold	79 (16)	87 (13)	51 (10)	12
back fold	83 (9)	–76 (11)	–150 (10)	6
Manα(1–6)-Man branch (5)	ϕ	ψ	ω	population (%)
open (cluster 1)	70 (10)	–171 (16)	55 (10)	80
open (cluster 2)	70 (8)	–173 (13)	–81 (13)	6
open (cluster 3)	69 (8)	–120 (14)	–64 (10)	6
open (cluster 4)	70 (8)	–169 (17)	–164 (9)	4
front fold	71 (10)	83 (12)	48 (9)	4
Manα(1–3)-Man branch (6)	ϕ	ψ	ω	population (%)
cluster 1	72 (9)	139 (5)	-	62
cluster 2	68 (10)	100 (10)	-	38
Manα(1–3)-Man arm (4)	ϕ	ψ	ω	population (%)
cluster 1	72 (8)	142 (14)	-	74
cluster 2	69 (9)	100 (10)	-	26
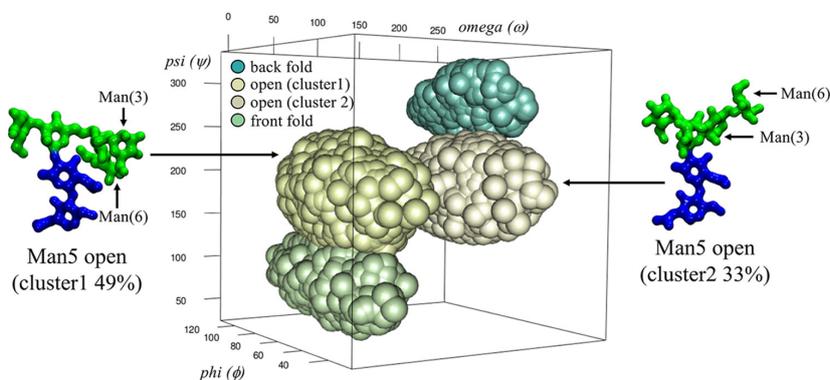
Figure 4.3: 3D representation of the clustering analysis of the Man5 (1–6) arm torsion angles populations. While the back and front fold conformers have one set of values each for the ϕ, ψ, and ω torsions, the open conformation adopts two distinct orientations of the biantennary branch, namely, open (cluster 1) with a representative structure shown on the left-hand side and open (cluster 2) with a representative structure shown on the right-hand side. The relative positions of the Man(3)- and Man(6)-linked units are also indicated. The rgl package in RStudio (www.rstudio.com) was used to make the graphics. Molecular models were rendered with VMD (http://www.ks.uiuc.edu/Research/vmd/). *N*-glycans are colored according to the SNFG convention.

Man6(I) and Man7(I) both have a longer (1–3) arm relative to Man5 with one and two Manα(1–2)-Man additional linkages, respectively. Note that different Man6/7 positional isomers exist, where the terminating Man can functionalize either branch on the (1–6)

arm. We decided to highlight the Man6/7(I) positional isomers to present the effect of the elongation of the (1–3) arm in combination with a shorter (1–6) arm on the dynamics of the system and their role in enhancing contacts between the arms. A full set of all positional isomers is presented in the B for completeness. As shown in 4.4 and Tables B.2 and B.6 both Manα(1–2)-Man linkages occupy two conformers, one at ($\phi = 74°$, $\psi = 151°$) and the other at ($\phi = 70°$, $\psi = 107°$) with a relative population of 73% and 27% for Man 6, respectively, and one at ($\phi = 74°$, $\psi = 151°$) and the other at ($\phi = 70°$, $\psi = 106°$) with population of 76% and 24% for Man 7, respectively. As shown by the population analysis in B.2 and B.7 the elongation of the (1–3) arm with Manα(1–2)-Man linkages does not affect the conformational propensity of the (1–6) arm relative to Man5, yet it slightly enhances the flexibility of the (1–3) arm, decreasing the population of the dominant conformer ($\phi = 72°$, $\psi = 142°$) at 74% in Man5 down to 63% in Man7. Notably, the progressive elongation of the (1–3) arm with rigid Manα(1–2)-Man linkages determines an increase of the interarm contacts with both (1–6) branches relative to Man5, as discussed in the next subsection. These contacts are stabilized by a complex network of short-lived and interchanging hydrogen bonds that mostly involve the terminal residues of the arms. The open (cluster 1) conformation with α(1–6) torsion values ($\phi = 71°$, $\psi = -172°$, $\omega = 56°$) populated at 41% and 46% in Man6 and Man7, respectively, favors the formation of these arm–arm interactions; see Figure 6b. Notably, elongation of the (1–3) branch on the (1–6) arm in Man6 (II) and Man 7 (II) determines an increase of these arm–arm interactions that contributes to increasing the population of a previously negligeably populated "cluster 3" conformer, see Figures B.4 and B.7 and Tables B.3 and B.6
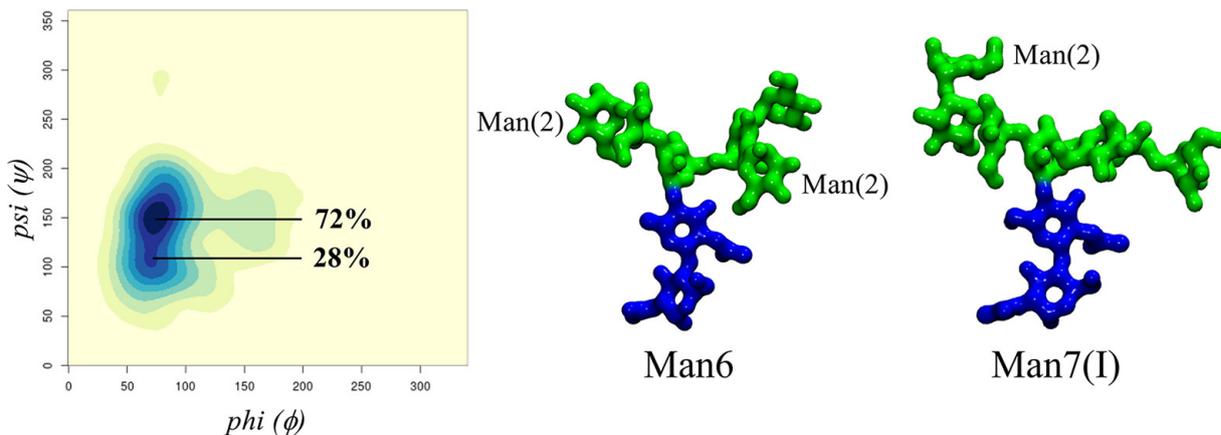
Figure 4.4: On the left-hand side, heat map representing the conformational analysis with corresponding populations of the first Manα(1–2)-Man linkage on the Man7(I) (1–3) arm, also representative of the corresponding linkage in the Man6 (1–3) arm. On the right-hand side, representative structures of Man6(I) and Man7(I) corresponding to the highest populated Manα(1–2)-Man linkage rotamers in the open (cluster 2) ($\phi = 71°$, $\psi = -172°$, $\omega = -176°$) conformation where the arms do not interact and the conformation of the Manα(1–2)-Man is more clearly visible. Heat maps were made with RStudio (www.rstudio.com) was used to make the graphics. Molecular models were rendered with VMD (http://www.ks.uiuc.edu/Research/vmd/). *N*-glycans are colored according to the SNFG convention.

Man8(I) and Man9 have further functionalizations of the (1–6) arm with one Manα(1–2)-Man linkage on the (1–3) branch for the Man8(I) positional isomer and an additional one on the (1–6) branch in Man9; see 4.1. As seen for the other oligomannoses, in Man8(I) and Man9 the dominant conformation is with an open (clusters 1, 2, and 3) (1–6) arm, see Table 4.4 and Tables B.9 and B.13 with a slightly more pronounced preference for the back vs front fold in Man9, due to the interactions of the longer (1–6) branch with the chitobiose, see Figure 4.5 and Table 4.4.

Table 4.2: **Torsion Angle Median Values of the Linkages in the Man9 (1–3/6) Arms:** Standard deviation values are shown in parentheses, with relative populations obtained from clustering analysis. Angle values are in degrees. The number in parentheses in the first column indicates the linkages, as shown on the Man5 sketch in Figure 4.5

Manα(1–6)-Man arm (3)	ϕ	ψ	ω	population (%)
open (cluster 1)	79 (13)	−173 (16)	55 (12)	38
open (cluster 2)	66 (10)	−179 (13)	−177 (12)	37
open (cluster 3)	73 (9)	−157 (21)	−65 (11)	7
cluster 4	147 (9)	−172 (8)	47 (7)	1
front fold	78 (11)	87 (11)	46 (9)	3
back fold (cluster 1)	82 (8)	−74 (10)	−151 (10)	12
back fold (cluster 2)	76 (9)	−96 (10)	−75 (9)	2
Manα(1–6)-Man branch (5)	ϕ	ψ	ω	population (%)
open (cluster 1)	71 (10)	−169 (17)	55 (10)	57
open (cluster 2)	69 (10)	−179 (23)	−166 (11)	8
open (cluster 3)	71 (7)	−180 (15)	−73 (8)	4
front fold (cluster 1)	73 (12)	83 (12)	−164 (18)	10
front fold (cluster 2)	79 (16)	86 (12)	51 (10)	10
back fold	70 (7)	−118 (12)	−67 (10)	12
Manα(1–3)-Man branch (6)	ϕ	ψ	ω	population (%)
cluster 1	76 (12)	139 (5)	-	55
cluster 2	70 (11)	100 (10)	-	26
cluster 3	148 (11)	151 (10)	-	18
Manα(1–3)-Man arm (4)	ϕ	ψ	ω	population (%)
cluster 1	72 (9)	142 (15)	-	59
cluster 2	71 (11)	97 (11)	-	41

The structures of all Manα(1–2)-Man linkages are the same as described for Man6(I) and Man7(I), yet the elongation of both branches on the (1–6) arm with relatively rigid Manα(1–2)-Man linkages determines structures with a high number of contacts between the two arms. Indeed, as shown in 4.6, inter-arm contacts only occur within one conformational cluster in Man6(I) and Man7(I), namely, open (cluster 1); meanwhile in Man9, interactions between the arms are a feature of virtually all structural populations. These contacts are stabilized by complex networks of rapidly interchanging hydrogen bonds involving mainly the terminal monosaccharides on the arms and branches.
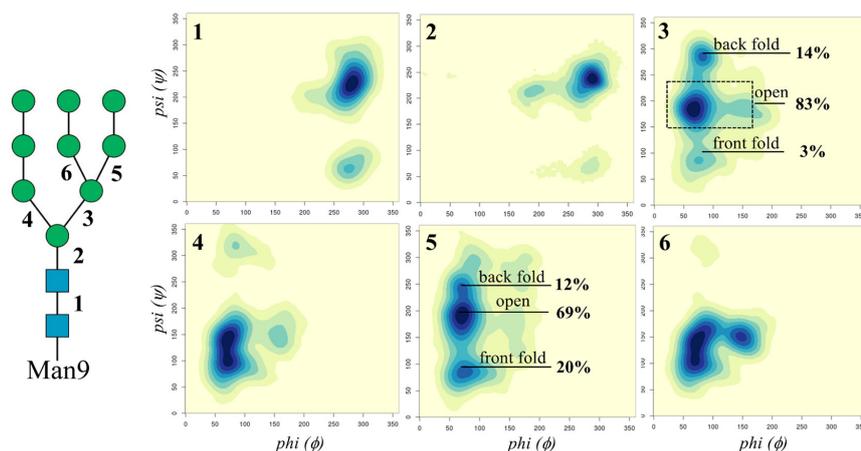
Figure 4.5: Man9 conformational analysis in terms of the ϕ and ψ torsion angles (ranging between 0° and 360°) explored during the 4.5 μs of cumulative MD sampling. Each torsion was numbered as indicated on the left-hand side sketch, and the heat maps have been labeled accordingly in the top left corner. Heat maps were made with RStudio (www.rstudio.com).
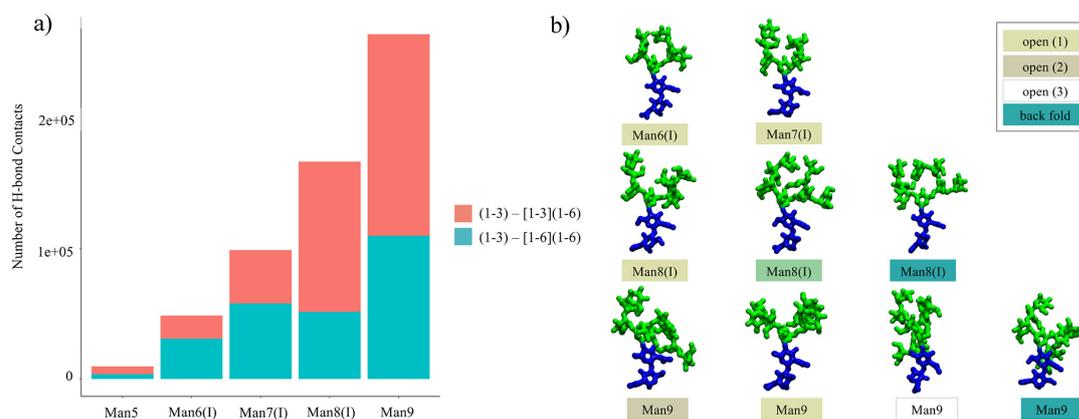
Figure 4.6: (a) Number of hydrogen bond contacts (distance threshold 4 Å between donor and acceptor atoms) counted over the 4.5 μs cumulative sampling for each oligomannose indicated on the x-axis. Contacts between the (1–3) arm and the (1–6) branch of the (1–6) arm are shown in cyan. (b) Representative snapshots from the MD simulations illustrating examples of the interarm contacts occurring within each conformational cluster. Different clusters are indicated by the colors in the legend on the top right-hand side, in agreement with the coloring scheme used in Figure 4.3. Histograms were made with RStudio (www.rstudio.com), and molecular models were rendered with VMD (http://www.ks.uiuc.edu/Research/vmd/). *N*-glycans are colored according to the SNFG convention.

4.5 FcγRIIIa-Linked Man5/9

The FcγRIIIa (CD16a) is a cell-bound receptor responsible for modulating antibody-dependent cellular cytotoxicity (ADCC) through its interaction with the IgG1 Fc region. [2] Recent studies have shown that the FcγRIIIa glycosylation contributes to the binding to IgG1s by stabilizing the interaction to a degree that is highly dependent on the type of the *N*-glycans present. [31, 40, 41] Human FcγRIIIa is glycosylated on two sites, namely, N45 and N162; see4.1. These two sites are very different in terms of their surrounding protein

landscape; while N162 is highly exposed to the solvent, N45 is located in the core of one of the two structural domains. To understand the effect of the protein surface landscape on the oligomannoses structure and dynamics, we studied two FcγRIIIa glycoforms, one with Man5 at N45 and N162 and the other with Man9 at N45 and N162.

As shown in Figure 4.7, results obtained from 2 μs of cumulative sampling from two independent runs show that the conformational dynamics of the Man5 at N45 is significantly restrained compared to the unlinked form. Indeed, a network of hydrogen bonds connects the terminal Man on the (1–3) branch of the (1–6) arm within a protein's cleft located between the two domains. These interactions result in shifting the Man5 intrinsic conformational equilibrium so that at N45 the Man5 (1–6) arm is mainly allowed in the open (cluster 2) conformation; see Figure 4.7 and Table 4.3. The flexibility of the (1–6) branch and of the (1–3) arm, not interacting with the protein, is the same as found for the unlinked Man5; see also Figure 4.2.
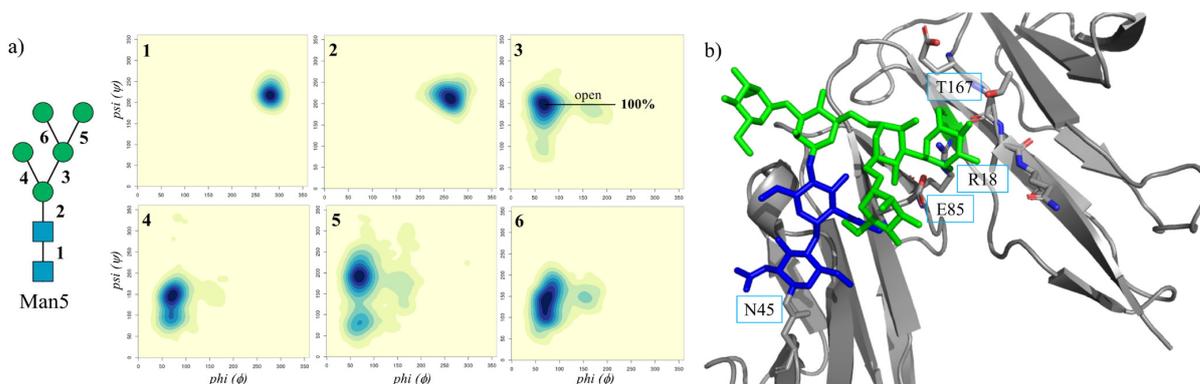
Figure 4.7: (a) Conformational analysis in terms of the ϕ and ψ torsion angles of the N40-linked Man5 linked explored during the 2 μs of cumulative MD sampling of the Man5 glycosylated FcγRIIIa. Each torsion is numbered as indicated on the left-hand side, and the corresponding heat maps are labeled in the top-left corner accordingly. (b) Dominant conformation of the N45-linked Man5, see 4.3, with the terminal Man on the (1–3) branch of the (1–6) arm restrained by hydrogen bonds to residues T167, R18 and E85, labeled in the figure. Heat maps were made with RStudio (www.rstudio.com), and structure was rendered with PyMol (www.pymol.org). *N*-glycan is colored according to the SNFG convention.

Man9 has two Manα(1–2)-Man linkages elongating both branches on the (1–6) arm, denying the pose found for Man5, which indeed disappears; see Tables 4.3 and B.13 Despite a higher flexibility relative to Man5, the N45-linked Man9 is less dynamic relative to the unlinked form due to the protein's landscape. Indeed, as shown in Table 4.3, only three out of the seven populated conformers are accessible.

As shown in Figure 4.1,the N162 position is much more exposed to the solvent relative to N45. Consequently, the intrinsic dynamics of the N162-Man5 is almost entirely retained, with a shift promoting the open (cluster 2) relative to the open (cluster 1) as the dominant conformer; see Table 4.4. Meanwhile in the case of a N162-linked Man9, the dynamics of the longer arms is limited due to the proximity to the protein's surface, see Figure 4.8, and in particular due to the presence of Lys 128, which because of its position denies a number of conformers due to steric hindrance and also potentially stabilizes the open (cluster 1) conformation through a hydrogen bonding interaction with the α(1–6)-linked Man on the (1–6) arm.

Table 4.3: **Torsion Angle Median Values of the Linkages in the N45-Linked Man5 and Man9 (1–6) Arm:** Standard deviation values are shown in parentheses, with relative populations obtained from clustering analysis. Angle values are in degrees. The number in parentheses in the first column indicates the linkages, as shown on the Man5 and Man9 sketches in Figures 4.2 amd 4.5

Manα(1–6)-Man arm (3)	ϕ	ψ	ω	population (%)
N45-Man5				
open (cluster 2)	69 (10)	−162 (13)	−172 (11)	100
N45-Man9				
open (cluster 1)	71 (12)	174 (13)	59 (15)	57
open (cluster 3)	72 (9)	−174 (15)	−75 (11)	15
back fold (cluster 2)	67 (7)	−112 (12)	−69 (12)	28

Table 4.4: **Torsion Angle Median Values of the Linkages in the N162-Linked Man5 and Man9 (1–6) Arm:** Standard deviation values are shown in parentheses, with relative populations obtained from clustering analysis. Angle values are in degrees. The number in parentheses in the first column indicates the linkages, as shown on the Man5 and Man9 sketches in Figures 4.2 amd 4.5

Manα(1–6)-Man arm (3)	ϕ	ψ	ω	population (%)
N162-Man5				
open (cluster 2)	72 (12)	−173 (17)	−169 (15)	83
open (cluster 1)	69 (11)	−175 (16)	52 (12)	10
front fold	74 (13)	88 (13)	53 (12)	7
N162-Man9				
open (cluster 1)	70 (11)	−171 (16)	55 (11)	83
front fold	83 (14)	88 (12)	48 (9)	17

4.6 Discussion

In this work we analyzed the 3D structure and dynamics of human oligomannose *N*-glycans, from Man5 to Man9, when free (unlinked) in solution and also determined how the effect of FcγRIIIa (CD16a) surface landscape modulates their structural equilibria. Despite similarities with complex *N*-glycans,[16, 17]. in terms of the core chitobiose rigidity and of the relatively low degree of flexibility of the (1–3) arm, oligomannoses have a very unique architecture, which changes with the progressive functionalization of the arms. More specifically, Man5 shows a clear propensity for an "open" structure, where the (1–6) arm is outstretched orientating the two branches on either side of the (1–3) arm; see Figures 4.2 and 4.3. Small variations of the open structure, determined by two accessible values of the (1–6) arm ω torsion, are also populated, see Figure 4.3,and of the additional degrees of freedom of both (1–3/6) branches, which closely reflect the dynamics of the arms, see Table 4.4 In larger oligomannoses the (1–3) arm and (1–3/6) branches on the (1–6) arm are terminated with Manα(1–2)-Man groups, giving raise to different Man6 up to Man8 isoforms and then to Man9. We analyzed 10 different positional isomers of Man6 to Man8, see Figure B.1, and focused our attention on the isoforms named (I) as representative examples, shown in Figure 4.1, while including all others for completeness as Appendex B. Sampling results show that the Manα(1–2)-Man linkages are rigid and do not significantly affect the intrinsic dynamics of the other linkages. The most interesting and unique aspect of the arms elongation in oligomannoses is that the orientation
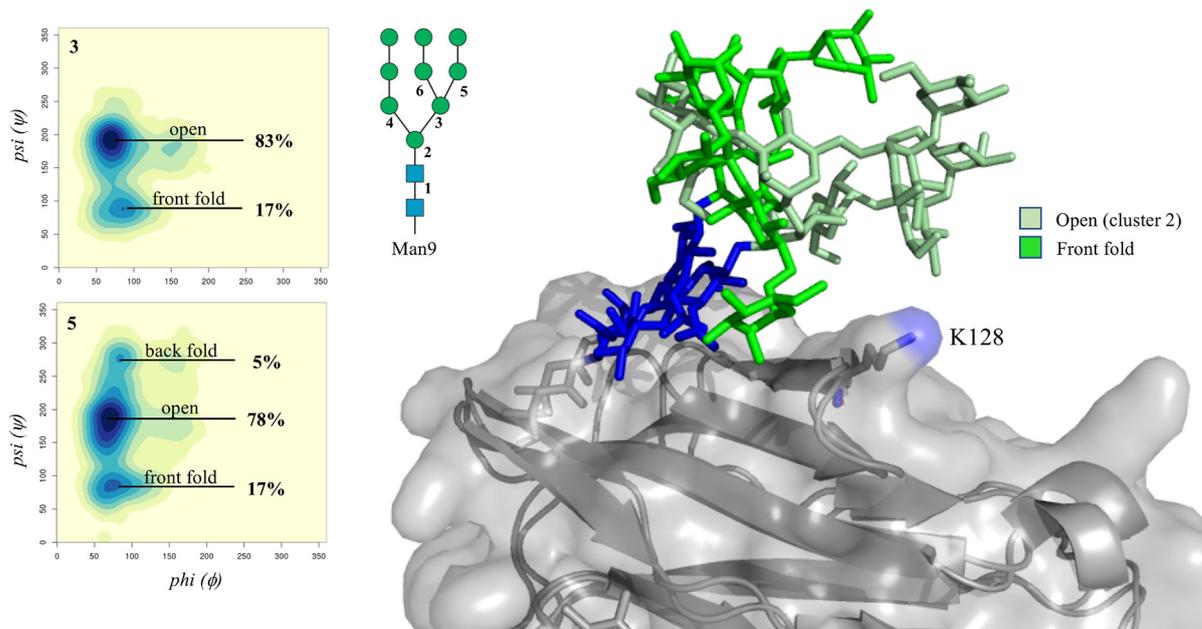
Figure 4.8: (a) Conformational analysis Conformational analysis of the (1–6) arm and (1–6) branch of the N162-linked Man9 in terms of the ϕ and ψ torsion angles, obtained from the 2 μs of cumulative MD sampling of the Man9 glycosylated FcγRIIIa. Heat maps are labeled in the top-left corner according to the Man9 numbering in the sketch. The two dominant conformations of the N162-linked Man9 are shown on the right-hand side, with the protein represented by the solvent accessible surface and underlying cartoons in gray and the mannose residues with different shades of green as described in the legend. Heat maps were made with RStudio (www.rstudio.com), and structure was rendered with PyMol (www.pymol.org). *N*-glycan is colored according to the SNFG convention.

of the additional Manα(1–2)-Man linkages within the underlying architecture of Man5 determines a progressive increase of interarm contacts, see Figure 4.6; so the structure of Man9 is quite compact, or more "tree-like", relative to smaller oligomannoses, where the arms are shorter but characterized by a more independent dynamics. As a further step in the analysis, a direct comparison of the results we obtained for Man9 and Man8(II) with NMR-validated REMD analysis [42] shows a very good agreement, supporting that the trimming of terminal residues allows for more extended arm structures, which expose embedded glycotopes; see Figure B.18.

The results obtained for the unlinked oligomannoses also confirm an earlier observation we made in the context of complex *N*-glycans, [16]. whereby the overall 3D architecture is determined by the local spatial arrangement of independent groups of monosaccharides we named "glycoblocks". The oligomannoses dynamics can be also discretized in terms of these structural units, [16]. With the addition of a unique Manα(1–2)-Man glycoblock that can be added to the arms with, as we have seen, minimal effect to the dynamics of the underlying units it builds on. This observation can offer a practical advantage to the study of glycan recognition through molecular docking, for example, where the receptor binds a specific glycoblock unit and recognition depends only on its accessibility within a specific glycoform.

To understand how the protein affects the presentation of the glycans to potential receptors, we have looked at the human FcγRIIIa (CD16a). Human FcγRIIIa has two

N-glycosylation sites, namely, N45 and N162, where the type of glycosylation affects the receptor's binding affinity to IgG1s [31, 32, 43]. The surface landscape around these two sites is quite different, with N162 exposed to the solvent while N45 is located in the core of one of the two structural domains; see Figure 4.1. Conformational sampling of a Man5 at N45 shows that the (1–6) arm dynamics is heavily restrained to one of its two open conformations accessible in solution; see Figure 4.7. More specifically, we found that the terminal Man on the (1–3) branch is engaged in a network of hydrogen bonding interactions involving a number of residues near the glycosylation site, namely, Arg 18, Glu 85, and Thr 167. The stabilization of this glycoform by the FcγRIIIa surface landscape renders the (1–3) branch on the (1–6) arm virtually inaccessible for further functionalization. This result agrees with recent work highlighting the unique prevalence of hybrid and oligomannose type *N*-glycans at N45. [32, 43] The N162 position determines very little steric hindrance to the dynamics of Man5, which retains most of the degrees of freedom characterized for the glycan free in solution. Meanwhile, the dynamics of the larger Man9 is greatly affected by the presence of Lys 128, which forces the glycan to adopt only two of the conformations accessible to the unlinked form; see Table 4.4 and Figure 4.8. Ultimately, the comparison between the conformational propensity of the unlinked Man5 and Man9 oligomannoses relative to their FcγRIIIa-linked counterparts suggests that the protein landscape affects the glycans structure by shifting their intrinsic conformational equilibria toward forms that complement it, yet it does not actively morph the glycan into unnatural conformers.

4.7   Conclusions

In this work we have characterized the 3D structure and dynamics of human oligomannose *N*-glycans unlinked and linked to FcγRIIIa through extensive sampling based on conventional MD simulations. The simulations of the unlinked oligomannose *N*-glycans show a complex architecture that is derived from a progressively intricate network of transient hydrogen bonding interactions involving the terminal residues on the arms, all linked through rigid Manα(1–2)-Man glycoblocks. The protein landscape affects the conformational equilibrium of the *N*-glycans favoring conformations that complement it, but it does not actively distort the oligomannoses' structure. Indeed, the two FcγRIIIa glycosylation sites studied in this work present different sets of constraints to different glycoforms and accordingly shift each conformational equilibrium specifically. This determines a diverse degree of accessibility of the arms for further functionalization by glycotransferases and glycohydrolases at N45, [32, 43] which has been found to have an unusually high degree of hybrid *N*-glycoforms, and ultimately exposure of the arms at N162 for contact with the IgG1 Fc *N*-glycans, which is implicated in modulating ADCC. [19, 31, 44, 45]. Work in this direction is currently ongoing in our lab.

4.8   Acknowledgements

Bibliography

[1] Varki, A. Biological roles of glycans. Glycobiology **2016**, 27, 3–49.

[2] Strasser, R. Plant protein glycosylation. Glycobiology **2016**, 26, 926–939.

[3] Christiansen, M. N.; Chik, J.; Lee, L.; Anugraham, M.; Abrahams, J. L.; Packer, N. H. Cell surface protein glycosylation in cancer. PROTEOMICS **2014**, 14, 525–546.

[4] Cobb, B. A. The history of IgG glycosylation and where we are now. Glycobiology **2019**, 30, 202–213.

[5] Moremen, K. W.; Tiemeyer, M.; Nairn, A. V. Vertebrate protein glycosylation: diversity, synthesis and function. Nature Reviews Molecular Cell Biology **2012**, 13, 448–462.

[6] Schjoldager, K. T.; Narimatsu, Y.; Joshi, H. J.; Clausen, H. Global view of human protein glycosylation pathways and functions. Nature Reviews Molecular Cell Biology **2020**, 21, 729–749.

[7] Gu, J.; Taniguchi, N. Regulation of integrin functions by N-glycans. Glycoconjugate Journal **2004**, 21, 9–15.

[8] Strasser, R. Biological significance of complex N-glycans in plants and their impact on plant physiology. Frontiers in Plant Science **2014**, 5.

[9] Paschinger, K.; Wilson, I. B. H. Comparisons of N-glycans across invertebrate phyla. Parasitology **2019**, 146, 1733–1742.

[10] Deshpande, N.; Wilkins, M. R.; Packer, N.; Nevalainen, H. Protein glycosylation pathways in filamentous fungi. Glycobiology **2008**, 18, 626–637.

[11] Thompson, A. J.; de Vries, R. P.; Paulson, J. C. Virus recognition of glycan receptors. Current Opinion in Virology **2019**, 34, 117–129.

[12] Aebi, M.; Bernasconi, R.; Clerc, S.; Molinari, M. N-glycan structures: recognition and processing in the ER. Trends in biochemical sciences **2010**, 35, 74–82.

[13] Rillahan, C. D.; Paulson, J. C. Glycan microarrays for decoding the glycome. Annual review of biochemistry **2011**, 80, 797–823.

[14] Wu, X.; Delbianco, M.; Anggara, K.; Michnowicz, T.; Pardo-Vargas, A.; Bharate, P.; Sen, S.; Pristl, M.; Rauschenbach, S.; Schlickum, U., et al. Imaging single glycans. Nature **2020**, 582, 375–378.

[15] Anggara, K.; Zhu, Y.; Delbianco, M.; Rauschenbach, S.; Abb, S.; Seeberger, P. H.; Kern, K. Exploring the molecular conformation space by soft molecule–surface collision. Journal of the American Chemical Society **2020**, 142, 21420–21427.

[16] Fogarty, C. A.; Harbison, A. M.; Dugdale, A. R.; Fadda, E. How and why plants and human N-glycans are different: Insight from molecular dynamics into the "glycoblocks" architecture of complex carbohydrates. Beilstein journal of organic chemistry **2020**, 16, 2046–2056.

[17] Harbison, A. M.; Brosnan, L. P.; Fenlon, K.; Fadda, E. Sequence-to-structure dependence of isolated IgG Fc complex biantennary N-glycans: a molecular dynamics study. Glycobiology **2019**, 29, 94–103.

[18] Neelamegham, S.; Aoki-Kinoshita, K.; Bolton, E.; Frank, M.; Lisacek, F.; Lütteke, T.; O'Boyle, N.; Packer, N. H.; Stanley, P.; Toukach, P., et al. Updates to the symbol nomenclature for glycans guidelines. Glycobiology **2019**, 29, 620–624.

[19] Ferrara, C.; Grau, S.; Jäger, C.; Sondermann, P.; Brünker, P.; Waldhauer, I.; Hennig, M.; Ruf, A.; Rufer, A. C.; Stihle, M., et al. Unique carbohydrate–carbohydrate interactions are required for high affinity binding between FcγRIII and antibodies lacking core fucose. Proceedings of the National Academy of Sciences **2011**, 108, 12669–12674.

[20] Watanabe, Y.; Bowden, T. A.; Wilson, I. A.; Crispin, M. Exploitation of glycosylation in enveloped virus pathobiology. Biochimica et Biophysica Acta (BBA)-General Subjects **2019**, 1863, 1480–1497.

[21] Watanabe, Y.; Allen, J. D.; Wrapp, D.; McLellan, J. S.; Crispin, M. Site-specific glycan analysis of the SARS-CoV-2 spike. Science **2020**, 369, 330–333.

[22] Bonomelli, C.; Doores, K. J.; Dunlop, D. C.; Thaney, V.; Dwek, R. A.; Burton, D. R.; Crispin, M.; Scanlan, C. N. The glycan shield of HIV is predominantly oligomannose independently of production system or viral clade. PloS one **2011**, 6, e23521.

[23] Doores, K. J.; Bonomelli, C.; Harvey, D. J.; Vasiljevic, S.; Dwek, R. A.; Burton, D. R.; Crispin, M.; Scanlan, C. N. Envelope glycans of immunodeficiency virions are almost entirely oligomannose antigens. Proceedings of the National Academy of Sciences **2010**, 107, 13800–13805.

[24] Stewart-Jones, G. B.; Soto, C.; Lemmin, T.; Chuang, G.-Y.; Druz, A.; Kong, R.; Thomas, P. V.; Wagh, K.; Zhou, T.; Behrens, A.-J., et al. Trimeric HIV-1-Env structures define glycan shields from clades A, B, and G. Cell **2016**, 165, 813–826.

[25] Struwe, W. B.; Chertova, E.; Allen, J. D.; Seabright, G. E.; Watanabe, Y.; Harvey, D. J.; Medina-Ramirez, M.; Roser, J. D.; Smith, R.; Westcott, D., et al. Site-specific glycosylation of virion-derived HIV-1 Env is mimicked by a soluble trimeric immunogen. Cell reports **2018**, 24, 1958–1966.

[26] De Leoz, M. L. A.; Young, L. J.; An, H. J.; Kronewitter, S. R.; Kim, J.; Miyamoto, S.; Borowsky, A. D.; Chew, H. K.; Lebrilla, C. B. High-mannose glycans are elevated during breast cancer progression. Molecular & Cellular Proteomics **2011**, 10.

[27] Li, Q.; Li, G.; Zhou, Y.; Zhang, X.; Sun, M.; Jiang, H.; Yu, G. Comprehensive N-glycome profiling of cells and tissues for breast cancer diagnosis. Journal of proteome research **2019**, 18, 2559–2570.

[28] Liu, X.; Nie, H.; Zhang, Y.; Yao, Y.; Maitikabili, A.; Qu, Y.; Shi, S.; Chen, C.; Li, Y. Cell surface-specific N-glycan profiling in breast cancer. PLoS One **2013**, 8, e72704.

[29] Thaysen-Andersen, M.; Packer, N. H. Site-specific glycoproteomics confirms that protein structure dictates formation of N-glycan type, core fucosylation and branching. Glycobiology **2012**, 22, 1440–1452.

[30] Zacchi, L. F.; Schulz, B. L. N-glycoprotein macroheterogeneity: biological implications and proteomic characterization. Glycoconjugate journal **2016**, 33, 359–376.

[31] Subedi, G. P.; Barb, A. W. CD16a with oligomannose-type N-glycans is the only "low-affinity" Fc γ receptor that binds the IgG crystallizable fragment with high affinity in vitro. Journal of Biological Chemistry **2018**, 293, 16842–16850.

[32] Roberts, J. T.; Patel, K. R.; Barb, A. W. Site-specific N-glycan analysis of antibody-binding Fc γ receptors from primary human monocytes. Molecular & Cellular Proteomics **2020**, 19, 362–374.

[33] Case, D. A.; Aktulga, H. M.; Belfon, K.; Ben-Shalom, I.; Brozell, S. R.; Cerutti, D. S.; Cheatham III, T. E.; Cruzeiro, V. W. D.; Darden, T. A.; Duke, R. E., et al. Amber 2021; University of California, San Francisco, 2021.

[34] Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; González-Outeiriño, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. GLYCAM06: a generalizable biomolecular force field. Carbohydrates. Journal of computational chemistry **2008**, 29, 622–655.

[35] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. The Journal of chemical physics **1983**, 79, 926–935.

[36] Fadda, E.; Woods, R. J. On the role of water models in quantifying the binding free energy of highly conserved water molecules in proteins: The case of concanavalin a. Journal of Chemical Theory and Computation **2011**, 7, 3391–3398.

[37] Sauter, J.; Grafmüller, A. Solution properties of hemicellulose polysaccharides with four common carbohydrate force fields. Journal of Chemical Theory and Computation **2015**, 11, 1765–1774.

[38] Joung, I. S.; Cheatham III, T. E. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. The journal of physical chemistry B **2008**, 112, 9020–9041.

[39] Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. Journal of molecular graphics **1996**, 14, 33–38.

[40] Hayes, J. M.; Frostell, A.; Karlsson, R.; Müller, S.; Martín, S. M.; Pauers, M.; Reuss, F.; Cosgrave, E. F.; Anneren, C.; Davey, G. P., et al. Identification of Fc gamma receptor glycoforms that produce differential binding kinetics for rituximab. Molecular & Cellular Proteomics **2017**, 16, 1770–1788.

[41] Subedi, G. P.; Barb, A. W. The structural role of antibody N-glycosylation in receptor interactions. Structure **2015**, 23, 1573–1583.

[42] Yamaguchi, T.; Sakae, Y.; Zhang, Y.; Yamamoto, S.; Okamoto, Y.; Kato, K. Exploration of Conformational Spaces of High-Mannose-Type Oligosaccharides by an NMR-Validated Simulation. Angewandte Chemie International Edition **2014**, 53, 10941–10944.

[43] Patel, K. R.; Roberts, J. T.; Barb, A. W. Allotype-specific processing of the CD16a N45-glycan from primary human natural killer cells and monocytes. Glycobiology **2020**, 30, 427–432.

[44] Falconer, D. J.; Subedi, G. P.; Marcella, A. M.; Barb, A. W. Antibody fucosylation lowers the FcγRIIIa/CD16a affinity by limiting the conformations sampled by the N162-glycan. ACS chemical biology **2018**, 13, 2179–2189.

[45] Day, C. J.; Tran, E. N.; Semchenko, E. A.; Tram, G.; Hartley-Tassell, L. E.; Ng, P. S.; King, R. M.; Ulanovsky, R.; McAtamney, S.; Apicella, M. A., et al. Glycan: glycan interactions: High affinity biomolecular interactions that can mediate binding of pathogenic bacteria to host cells. Proceedings of the National Academy of Sciences **2015**, 112, E7266–E7275.

5   Chapter 5: Fine-tuning the spike

Role of the nature and topology of the glycan shield in the structure and dynamics of the SARS-CoV-2 S

The following text is reproduced verbatim from the paper.
Harbison, A. M.‡; **Fogarty, C. A.**‡; Phung, T. K.; Satheesan, A.; Schulz, B. L.; Fadda, E. Fine-tuning the spike: role of the nature and topology of the glycan shield in the structure and dynamics of the SARS-CoV-2 S. Chemical Science 2022, 13, 386–395.
‡: These authors contributed equally to the work.

5.1   Abstract

The dense glycan shield is an essential feature of the SARS-CoV-2 spike (S) architecture, key to immune evasion and to the activation of the prefusion conformation. Recent studies indicate that the occupancy and structures of the SARS-CoV-2 S glycans depend not only on the nature of the host cell, but also on the structural stability of the trimer; a point that raises important questions about the relative competence of different glycoforms. Moreover, the functional role of the glycan shield in the SARS-CoV-2 pathogenesis suggests that the evolution of the sites of glycosylation is potentially intertwined with the evolution of the protein sequence to affect optimal activity. Our results from multi-microsecond molecular dynamics simulations indicate that the type of glycosylation at N234, N165 and N343 greatly affects the stability of the receptor binding domain (RBD) open conformation, and thus its exposure and accessibility. Furthermore, our results suggest that the loss of glycosylation at N370, a newly acquired modification in the SARS-CoV-2 S glycan shield's topology, may have contributed to increase the SARS-CoV-2 infectivity as we find that N-glycosylation at N370 stabilizes the closed RBD conformation by binding a specific cleft on the RBD surface. We discuss how the absence of the N370 glycan in the SARS-CoV-2 S frees the RBD glycan binding cleft, which becomes available to bind cell-surface glycans, and potentially increases host cell surface localization.
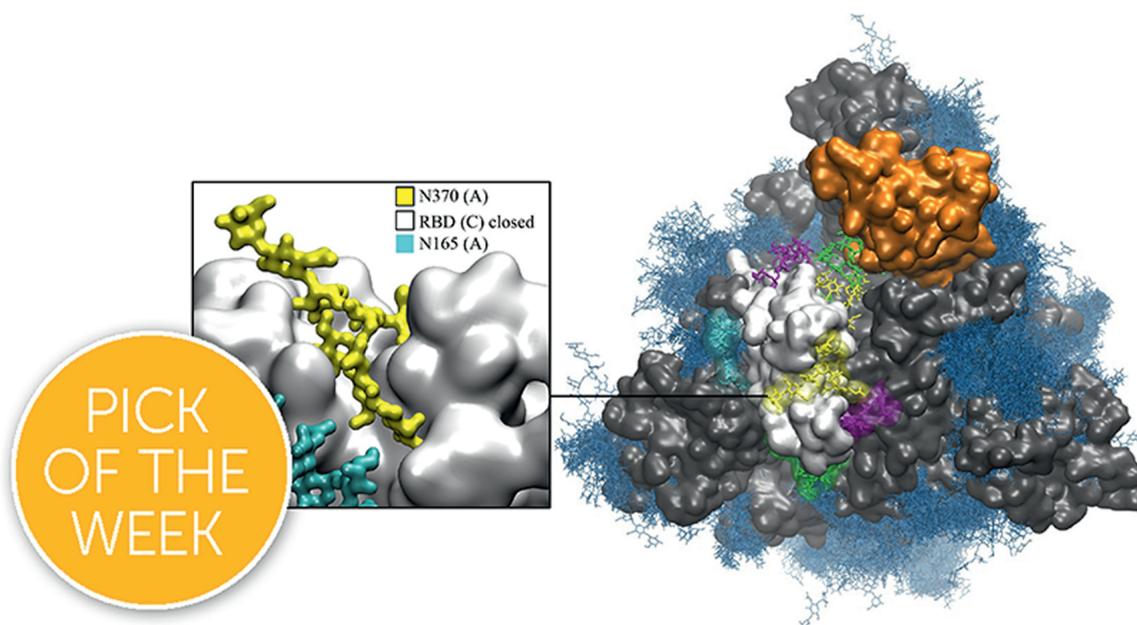
5.2   Introduction

Spike (S) glycoproteins mediate the adhesion and fusion of enveloped viruses to the host cell, initiating viral infection [1–3]. The interaction with the cell-bound receptor leads to a complex conformational change, dependent on the S architecture,[4, 5] that terminates with the fusion of the viral envelope with the host cell's membrane, giving the virus access to the cellular machinery for replication [3]. Viral envelope S are heavily coated with a dense layer of complex carbohydrates, also known as a glycan shield, that performs different intrinsic and extrinsic functions, from modulating protein folding, stability and trafficking, to masking the virus from the immune system and mediating contacts with lectins and antibodies [1]. The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has a trimeric S glycoprotein [6, 7] protruding from the viral envelope surface.[8, 9]. The SARS-CoV-2 S has 22 N-glycosylation sequons per protomer, of which at least 18 appear to be consistently occupied in different constructs,[8, 10–15] and O-glycosylation sites with significantly lower occupancy. [11, 12, 16]

A unique feature of the SARS-CoV-2 S glycoprotein's architecture is the key role of the glycan shield in its activation mechanism. [17, 18] Binding of SARS-CoV-2 S to its primary receptor, namely the angiotensin-converting enzyme 2 (ACE2),[19, 20]requires the opening of one (or more) receptor binding domains (RBDs), which need to emerge from the glycan shield to become accessible [9, 17, 18, 21–23]. The RBD opening creates a cavity within the SARS-CoV-2 S prefusion trimer's structure, [6, 7] see Fig 5.1 , which extends deeply into the trimer's core. In the absence of strategically positioned N-glycans as a support [17], upon RBD opening this large pocket would be filled by water molecules, likely weakening the S prefusion structural integrity, especially considering the SARS-CoV-2 S pre-cleaved polybasic furin site at the S1/S2 boundary [24, 25]. Multi-microsecond molecular dynamics (MD) simulations supported by biolayer interferometry experiments [17] have shown that this structural weakness is effectively recovered by the N-glycan at position N234, where a site-specific large oligomannose [8, 10, 14, 15] is able to fill the cavity, supporting the RBD open conformation [17]. Furthermore, MD simulations have also shown that the N-glycans at positions N165 and N343, see Fig. 5.1 are directly involved in important interactions with residues of the open RBD, supporting its open conformation [17] and mediating (or gating) its transition from open to closed, [18] respectively.
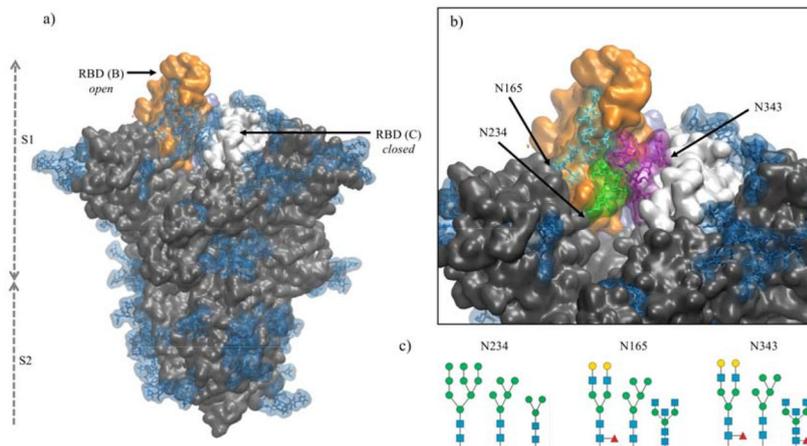
Figure 5.1: (panel a) Structure of the fully glycosylated SARS-CoV-2 S (PDBid 6VYB) ectodomain [7] The protein is shown in grey with the RBDs of chain B and C highlighted in orange and white, respectively; the glycan shield is highlighted in blue. (panel b) Close up on the open pockets with the *N*-glycans at the strategic positions N234, N165, and N343 highlighted in green, cyan and purple, respectively. (panel c) *N*-glycans considered in the different models studied in this work, represented through the SNFG [26] and drawn with DrawGlycan [27] (http://www.virtualglycome.org/DrawGlycan/). Molecular rendering done with VMD [28].

The crucial role of the *N*-glycans at N234, N165 and N343 is exerted through the contacts these structures can make with protein residues, both in the open RBD (of chain B, following the PDBid 6VYB nomenclature) and in the adjacent closed RBD (chain C) at either side of the empty cleft [17], see Fig. 5.1 Their ability to engage in effective interactions is intrinsically linked to the nature, size, sequence and branching of the *N*-glycans at these sites, opening the floor to a broader discussion on the relative structural stability and dynamics of different S glycoforms. This is a very important point to explore, especially in view of the design of specific antiviral therapeutic strategies targeting glycosylation [29],yet a very difficult (if not impossible) one to systematically address experimentally.

In this work, we present the results of a set of multi-microsecond MD simulations of different SARS-CoV-2 S glycoforms aimed at characterizing the effect of changes in the type of glycosylation at positions N234, N165 and N343, see Fig. 5.1, while the rest of the glycan shield is represented consistently with a stable recombinant S prefusion trimer (S_{trimer}). [10, 17] More specifically, we investigated if shorter oligomannoses structures at N234, such as Man5 and paucimannose (Man3), see Fig. 5.1, could occupy the empty cleft in the open SARS-CoV-2 S as effectively as larger oligomannoses such as Man9, characteristic of a highly stable prefusion S_{trimer} [10]. As an important note, Man5, rather than Man7/8/9 is found to be present, or even to be the dominant glycosylation type at N234 in the virus, vaccine epitopes and in some recombinant S constructs [8, 11, 30]. We also explored how the role of oligomannose structures at N234 is supplemented by the complex N-glycans at N165 and N343, which can form a tight network of glycan–glycan and glycan–protein interactions that stabilize the orientation and dynamics of the open RBD across different possible orientations.

Finally, we also investigated the effect of a mutation, unique to the SARS-CoV-2 Wuhan-Hu-1 (NCBI reference sequence: NC_045512.2) and derived strains[31],that changes the RBD glycan shield's topology. More specifically, in SARS-CoV and MERS [32],as well as in the bat RaTG13 and pangolin CoV SARS-CoV-2 variants[25, 33, 34],position N370 on the RBD is part of an occupied NST sequon [32], which is lost in SARS-CoV-2 due to a T372A mutation. We performed ancestral sequence reconstruction of selected SARS S sequences to investigate the gain and loss of glycosylation sequons during S evolution. To address the effect of this change in the glycan shield's topology, we re-introduced the N-glycan at N370 in the SARS-CoV-2 S native sequence and ran MD simulations to study its effect on the stability of both, the open and closed RBDs. Our simulations indicate that within the SARS-CoV-2 S architecture, glycosylation at N370 does not interfere with the glycan network at N234, N165 and N343, but actively contributes to it by stabilizing very effectively the RBD open conformation.

Interestingly, analysis of the closed protomers shows that the N370 glycan from RBD (A) is firmly bound to the RBD (C) surface, where it occupies a specific cleft. This interaction results in tying the closed RBD (C) to the adjacent closed RBD (A), very much like the laces in a shoe, thus potentially hindering the opening of the RBDs. Based on these findings, we propose that the recent loss of N-glycosylation at N370 allows for a higher availability of open S conformations by lowering the energetic cost of the opening reaction, which is likely to be beneficial by providing higher infectivity of SARS-CoV-2 relative to closely related variants carrying this sequon. We also discuss how the cleft on the closed RBD surface in SARS-CoV-2 S, which is occupied by the N370 glycan in S glycoforms with the sequon, may be used by other glycans found on the cell surface, such as glycosaminoglycans[35–40], sialogangliosides, and blood group antigens [39] , where these interactions may contribute to increasing the S cell-surface localization.

5.3   Results and Discussion

In this section we will present the results of extensive conformational sampling based on multi-microsecond MD simulations of SARS-CoV-2 S models with different glycosylation at N234, N165 and N343, see Table5.1. Specifically, we determined the effects of site-specific glycan structure on the SARS-CoV-2 S ectodomain's structure and dynamics by systematically reducing the size of the oligomannose at N234; from Man9 (N234-Man9), found in highly stable prefusion SARS-CoV-2 S constructs [10, 12, 30] and also studied in previous work [17, 41], we progressed to a shorter Man5 (N234-Man5), found on the

virus and on other SARS-CoV-2 S constructs[11, 30], and to paucimannose (N234-Man3) as a hypothetical size limit. In these models the glycosylation at N165 and N343 is biantennary complex (FA2G2), and in one case we have also considered bisecting GlcNAc F/A2B forms at N343 and N165, respectively, see Fig. 5.1 and TableC.1 The results obtained for the N234-Man5 models were compared to a SARS-CoV-2 S model with a uniform immature glycosylation[42], *i.e* in which all *N*-glycans are Man5 (all-Man5). Finally, to assess the effect of loss of glycosylation at N370 on SARS-CoV-2 S, we added a complex *N*-glycan (FA2G2) at N370 in the N234-Man9 model with bisecting GlcNAc (F/A2B) *N*-glycans at N165 and N343, respectively. In all of these models the glycosylation at all sites other than the ones listed above is consistent and the same as the profile experimentally determined for the stable prefusion S_{trimer} [10] see Table C.1 Results are based on the analysis of multiple uncorrelated MD trajectories (replicas) run in parallel for each model (see details in the Computational methods section in Appendix C

Table 5.1: Orientation of the RBD in different SARS-CoV-2 S glycoforms in terms of average lateral and axial angle values relative to the SARS-CoV-2 S ectodomain, as described in ref. [17] Standard deviation values are indicated in parentheses. Total simulation time for each replica is indicated in parentheses. Note: the first 300 ns of each trajectory were considered as part of conformational equilibration and were omitted from the analysis

Model system	Replicas	Lateral angle (°)	Axial angle (°)
N234-Man5	R1 (2.0 μs)	0.0 (2.4)	2.9 (1.1)
	R2 (2.0 μs)	-6.7 (2.9)	4.8 (0.9)
	R3 (1.5 μs)	2.7 (3.7)	5.9 (0.9)
N234-Man3	R1 (2.0 μs)	4.4 (3.1)	1.8 (1.1)
	R2 (1.9 μs)	1.3 (2.5)	-0.1 (1.0)
	R3 (1.4 μs)	2.5 (2.5)	6.7 (1.1)
N234-Man9	R1 (2.0 μs)	3.4 (2.0)	-0.5 (0.7)
	R2 (1.4 μs)	-0.7 (2.0)	4.8 (0.7)
N234-Man9 (+N370)	R1 (1.4 μs)	-3.4 (2.4)	6.0 (0.7)
	R2 (1.3 μs)	-0.6 (2.1)	-1.0 (0.7)
All-Man5	R1 (1.4 μs)	-14.1 (2.6)	3.2 (0.9)
	R2 (1.1 μs)	-3.5 (3.6)	6.9 (1.0)
	R3 (1.9 μs)	-1.9 (6.4)	-0.2 (3.8)
N234-Man9* (Man5 N165/N343)	R1 (2.1 μs)	0.2 (2.5)	7.3 (1.0)

5.3.1   N234-Man5

The structure and dynamics of the RBD (residues 330 to 530) from a model with a Man5 at N234 and FA2G2 at N165 and N343 was analysed based on three replicas with sampling times indicated in Table 5.1. Note that the first 300 ns of each replica were omitted from the analysis to allow conformational equilibration, see Fig. C.1 The dynamics of the RBD is rather complex and for consistency with previous work [17] we defined it here in terms of lateral and axial angle values populations. The lateral angle indicates an in-plane motion of the open RBD along a hypothetical circle centred on the central helices of the spike. This angle is defined by three points, corresponding to (1) the centre of mass of the open RBD core β-sheets at frame 0, i.e. the orientation of the RBD in PDB 6VYB, (2) the centre of mass of the top section of the central helices (CH), and (3) the centre

of mass of the open RBD core β-sheets at each frame along the trajectory. [17] The axial-angle defines a tilting motion of the open RBD either away from (negative values) or toward (positive values) the central helices of the spike. The axial angle is defined by (1) the centre of mass of the open RBD core β-sheets, (2) the centre of mass of the central helices, and (3) the centre of mass of the top section of the central helices. [17] A graphical representation of the axial and lateral angle coordinate frame is shown in Fig. 5.2
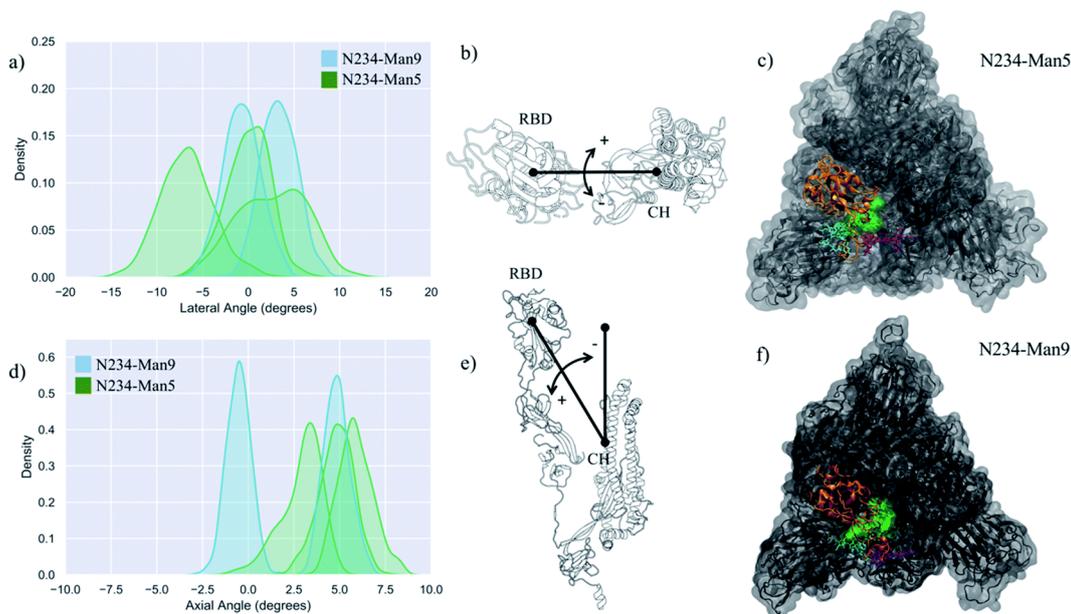
Figure 5.2: (panels a and d) Kernel density estimation (KDE) analysis of the lateral and axial angles distributions calculated through the uncorrelated MD trajectories obtained for N234-Man5 (green), replicas R1–3, and for N234-Man9 (cyan), replicas R1, 2 for comparison. (panels b and e) Graphical representation of the lateral (b) and axial (e) angles. Displacements relative to the initial trajectory frame are measured in terms of positive and negative values. The centre of mass of the RBD and of the central helices (CH) are used as reference points. (panels e and f) Top view on an equilibrated snapshot of the N234-Man5 (R3) S and N234-Man9 (R2) S, respectively. Man5/9 at N234 are shown in green, while the N-glycans at N165 and N343 are shown in cyan and purple, respectively. All other glycans are not represented for clarity. The RBD of chain B is shown in orange, while the rest of the protein is in grey. Data analysis and graphs were done with seaborn (www.seaborn.pydata.org) and molecular rendering with PyMol (www.pymol.org) and VMD[28].

Results shown in Fig. 5.2 and Table 5.1 indicate that in the N234-Man5 S model the open RBD can access a larger conformational space relative to the N243-Man9 model along the lateral displacement coordinate, while it is still able to adopt a stable open conformation, indicated by the axial tilt values. To note, the results from the MD simulations of the N243-Man9 ectodomain model presented here agree with the results obtained for the whole N243-Man9 S glycoprotein model [17]. The N234-Man5 S open conformations obtained from the conformational sampling scheme we used are slightly diverse, encompassing different degrees of opening, as shown in Fig. C.1 and C.2 More specifically, the RBD can be 'wide-open', see in Fig. 5.2 (panel b), i.e. in a similar orientation seen in the cryo-EM structures [6], stabilized by interactions involving different RBD residues and the N165 and N234 N-glycans, while the N343 N-glycan

is relatively free on the opposite side of the pocket, or alternatively the RBD can be in an intermediate conformation between the wide-open and the closed. In the latter the *N*-glycans at all three positions N165, N234 and N343 are involved in complex interactions with protein residues in the RBD, namely with the loop from L460 and F489, and in and around the pocket, and with each other, see Fig. C.3. Within this framework, the higher lateral degree of flexibility of the open RBD in the N234-Man5 S model is due to the smaller size of the Man5 glycan and its specific conformational propensity relative to Man9, which adopts a more 'tree-like' conformation, supported by extensive inter-arm interactions [43] , that fills the pocket much more effectively, see Fig 5.2 (panels e and f). More specifically, the Man5 at N234 appears to be less competent than Man9 at forming interactions that bridge both sides of the pocket as it accesses the cavity left open by the RBD, while it mainly interacts with residues at the base of the open RBD, leaving it 'unhinged' when in the "wide-open" conformation, relative to the N234-Man9 model. As an interesting point to note, in all the trajectories the core fucose of the N343 FA2G2 glycan is exposed to the solvent, potentially allowing for its recognition. This is in agreement with cryo-EM studies reporting interactions involving the core fucose of the *N*-glycan at N343 and human neutralizing antibodies[44].

To further assess the role of the type of glycosylation around the N234-Man5 in the RBD dynamics, we ran three MD simulations of a model with uniform Man5 glycosylation (all Man5 model), see Table 5.1 and Fig. C.4 The results indicate that the replacement of the complex glycans with immature Man5 structures is problematic for the stability of the open RBD. More specifically, the shorter size and 3D architecture of the Man5 at N343 does not allow it to engage effectively with residues in the RBD, while the Man5 at N234, as seen for the N234-Man5 model, is not able to effectively engage with the closed RBD (chain C) that flank the opposite side of the empty pocket. These results agree with recent work indicating that immature glycosylation, achieved in GnTI-/- mutant cells, leads to a less competent S glycoprotein relative to the fully glycosylated variant [42]. For comparison we tested a N234-Man9 model where only the *N*-glycans at N165 and N343 were Man5. The results shown in Table 5.1 and Fig. C.5 indicate that the presence of a large oligomannose structure at N243 helps recover the instability by more effectively occupying the empty pocket. In the equilibrium conformation sampled through 2.1 μs, the RBD conformation is more closed, shown also by the positive value of the axial angle, see Table 5.1 where both Man5 at N165 and N343 can interact with each other and with the RBD residues, see Fig. C.5.

5.3.2   N234-Man3

To gauge the implications of the size of the oligomannose at N243 for the orientation and dynamics of the open RBD, we studied a model were N243 is modified with a paucimannose (Man3). Note, to our knowledge, the presence of Man3 at this or at any sequon in the SARS-CoV-2 S has not been detected to date; [8, 10–14, 30, 45] nevertheless, as Man3 is one of the smallest oligomannose structures, this S glycoform represents a case study to account for the effect of the reduction in the size of the *N*-glycan to an extreme at this strategic position. The results shown in Table 5.1 and Fig. 5.3 in addition to the visual analysis of all trajectories, show that Man3 at N234 is less competent than Man 5 and Man9 in supporting a fully wide-open RBD. In only one of the three replicas, namely R3 shown in Fig. 5.3 (panels b and d), the Man3 accesses the interior of the open pocket, although its size does not allow for it to form any interactions that contribute to stability. In both R1 and R2, Man3 interacts with residues outside or at the edge of the pocket, as shown in Fig. C.6.
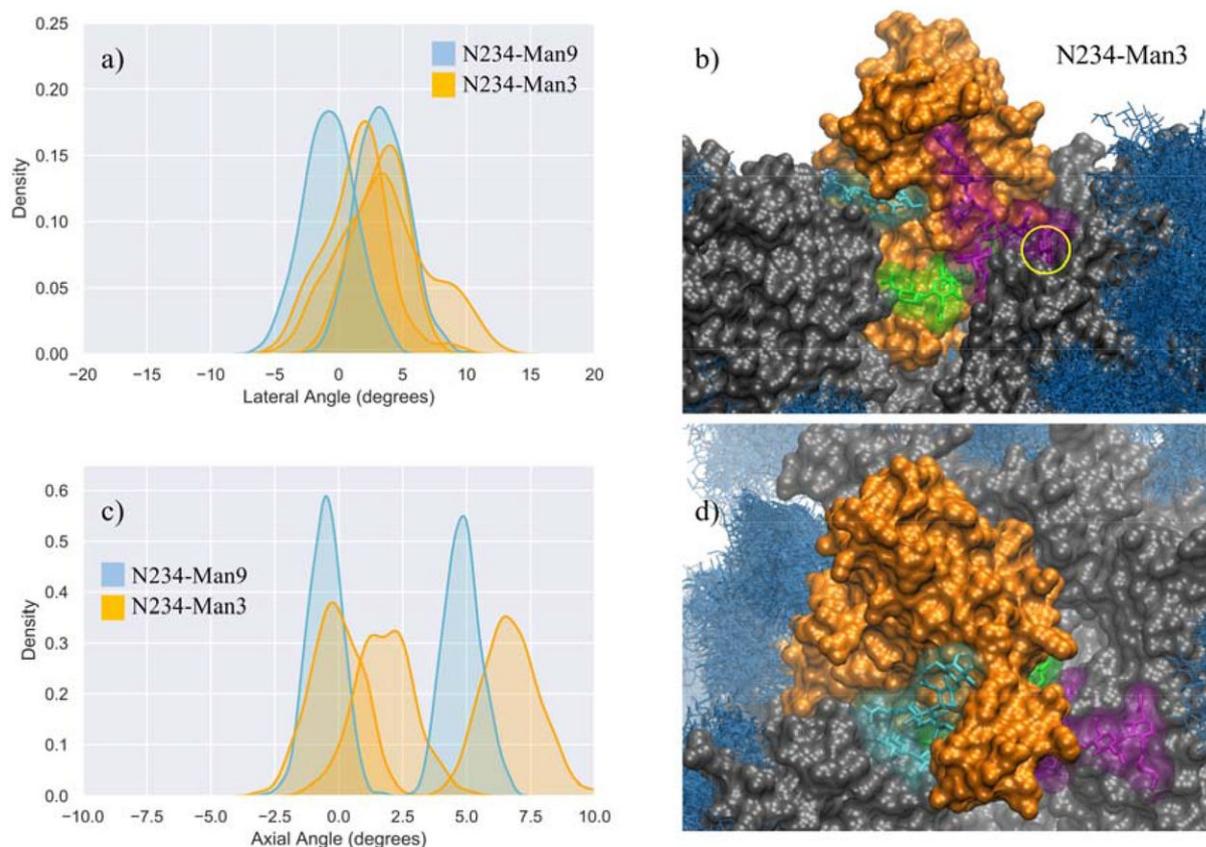
Figure 5.3:   (panels a and c) Kernel density estimation (KDE) analysis of the lateral and axial angle distributions calculated through the uncorrelated MD trajectories obtained for N234-Man3 (orange), replicas R1–3, and for N234-Man9 (cyan), replicas R1, 2 for comparison. (panels b and d) Close-ups on a representative snapshot of the N234-Man3 (R3) simulation from the side and top, respectively. Man3 at N234 is shown in green, while the *N*-glycans at N165 and N343 are shown in cyan and purple, respectively. The solvent accessible surface of the open RBD is shown in orange, while the rest of the protein is shown in grey. All other glycans are shown in blue, as an overlay of snapshots collected every 10 frames. The position of the core fucose in the FA2G2 *N*-glycans at N343 is highlighted within a yellow circle. Data analysis and graphs were done with seaborn (www.seaborn.pydata.org) and molecular rendering with VMD[28].

In the N234-Man3 model, the open RBD does not show any significant lateral excursions, see Fig. 5.3 (panel a), as its position is firmly held in place by interactions with the complex *N*-glycans at N165 and N343, which contribute to pulling it towards a conformation intermediate near to a closed RBD, shown in Fig.C.2 and C.6 In this conformation the N165 and N343 *N*-glycans interact extensively with residues in the receptor binding motif (RBM), possibly preluding to a final closing, as reported in earlier work, showing the gating activity of the glycans at N343.[17] When the RBD is near closing there is a very limited degree of freedom in the lateral angle coordinate. As an interesting point, when the FA2G2 *N*-glycan at N343 interacts with the RBD, its core fucose becomes less accessible to a potential recognition, see Fig. 5.3 (panel b).

5.3.3   N234-Man9 with FA2G2 at N370

Analysis of the glycan shield topology through the ancestral sequence reconstruction of select SARS S sequences, shown in Fig. 5.4 (panel d), indicates the loss of a sequon at N370 in the SARS-CoV-2 Wuhan-Hu-1 strain, due to a mutation to NSA of the conserved NST present in closely related strains of current interest, namely SARS-CoV, the bat RaTG13G S and the pangolin CoV S. To study the effect on SARS-CoV-2 S structure and dynamics of this additional ancestral glycan at N370, we restored the sequon at N370 and glycosylated this position with a complex FA2G2 N-glycan, consistently with data reported in the literature for SARS-CoV [32]. In this model, position N234 is modified with Man9, while positions N343 and N165 are glycosylated with complex N-glycans with a bisecting GlcNAc and core fucosylation at N343 (FA2B), and without core-fucosylation at N165 (A2B). Our results, shown in Fig. 5.4 and Table 5.1, indicate that the wide-open RBD conformation in the N370-glycosylated SARS-CoV-2 S is as stable as the one observed for the corresponding N234-Man9, where the N-glycan at N370 fills the interior of the empty pocket together with the Man9. In this model, the Man9 at N234 is only able to access the entrance of the pocket due to steric hindrance with the N370 glycan that occupies the core. This suggest that once the RBD opens, an S glycoform with glycosylated N370 would be highly competent in exposing the RBD to the ACE2 receptor, as the SARS-CoV-2 S with a large oligomannose at N234.

The analysis of the dynamics of the closed RBDs of protomers A and C (PDB 6VYB numbering) shows that the N370 N-glycan of RBD (A) is tightly bound to the surface of the adjacent closed RBD (C), threading the two closed RBDs together, see Fig. 5.4 (panel b, e and f) and Fig. C.7 The 3D architecture of the complex FA2G2 N-glycan at N370, characterized by independent dynamics of the arms, [46] allows for stable interactions of the (1–6) arm within a cleft in the RBD, flanked by residues between N448 and Y453 on one side and F490 and Y495 on the other, see Fig. 5.4 (panel e and f) and Fig. C.7 that support binding through hydrogen bonding and hydrophobic interactions with different monosaccharide units along the (1–6) arm. Because of the stability of the N370 glycan–protein interaction in the closed RBD, N370 glycosylation may hinder the opening mechanism. Thus, the loss of glycosylation at N370 likely contributes to enhancing the binding activity of S and infectivity of SARS-CoV-2 relative to other variants with an N-glycosylation sequon at this position.

5.4   Discussion

The SARS-CoV-2 S glycosylation profile with expression in (or infection of) mammalian cells has been reported by several studies [8, 10–14, 30, 45], almost all of which find a large oligomannose N-glycan, such as Man9–7, as the most common structure at N234. This is especially true in highly stable prefusion SARS-CoV-2 S trimer glycoforms[10, 30] ,which bear the 2P mutation.[6, 47] Meanwhile, a shorter Man5 appears to be present or even the dominant structure at N234 in the virus [8, 30] and in the secreted ChAdOx1 nCoV-19 (AZD1222) vaccine epitope [14, 15] , suggesting a higher degree of accessibility to that site by alpha-mannosidases in the ER. The results of multi-microsecond simulations presented in this work indicate that a reduced degree of filling of the pocket left empty by the opening of the RBD by a smaller N-glycan at N243 leads progressively higher degree of instability of the wide-open conformation of the RBD. More specifically, while the N234-Man5 model appears competent in exposing the open RBD despite its higher dynamics relative to the N234-Man9 model, the N234-Man3 model leads to a dominant "more closed" conformation, see Fig. C.2 where the paucimannose cannot form stable interactions within the pocket, interacting only with residues at the pocket's gate or outside

it, see Fig. 5.3 and C.6 The progressive destabilization upon reduction of the N-glycan size at N234 is in agreement with the results obtained for the N234A/N165A mutant, [17] designed to account for the complete removal of glycosylation at N234, which leads to 60% less binding to the ACE2 receptor as determined through biolayer interferometry assays [17]. It should be noted that the compact architecture of the S trimer does not allow for very large excursions in terms of lateral and especially of axial angles, nevertheless changes in these parameters in the same numerical range observed for the N234A/N165A mutant [17] have been shown to correspond to dramatic difference in ACE2 binding.

In agreement with recent work that discusses the roles of the glycan at N343 in supporting the RBD intermediate dynamics between open and closed conformations,[41] and of the glycan at N165 in supporting the open RBD [17], we observed that the stability of the N234-Man5 is significantly reduced when all the N-glycans in the shield are reduced to Man5 (All-Man5 model). In agreement with recent work [41], this destabilization is due primarily to the lack of interactions that the shorter Man5 at N343 can make with the open RBD, and in particular with residues in the disordered loop within the RBM (400 to 508), in addition to the Man5 at N234 allowing the RBD to be relatively unhinged. We have shown that this destabilization is partially recovered when the N234 is occupied by a larger oligomannose, such as Man9. However, within this framework, the RBD adopts a more closed conformation, where the Man5 at both N165 and N343 can interact with the open RBD and with each other, see Fig.C.5

The introduction of an additional glycosylation site at N370, which SARS-CoV-2 S has lost due to the T372A mutation in the Wuhan-Hu-1 and derived strains, illustrates the importance of an effective filling of the cavity left empty by the opening of the RBD. Indeed, the complex FA2G2 glycan at N370 can easily access the pocket in addition to Man9 at N234, which in this specific case can only partially fill it due to steric hindrance. We have shown that the presence of an N370 glycan contributes effectively to the stability of the wide-open RBD state, which nevertheless needs to be achieved starting from a closed S conformation [41]. To this end, we find that the N370 N-glycan on the SARS-CoV-2 S throughout all our simulations occupies a specific cleft on the surface of the closed RBDs. This binding mode is stabilized by a network of hydrogen bonds and hydrophobic interactions between the protein and the monosaccharides of the FA2G2 N370 glycan (1–6) arms, see Fig. 5.4 and C.7 Because the N370 glycan involved in this interaction originates from the adjacent closed RBD, its binding results in tying the closed RBDs together, likely hindering opening in the first place. This is in agreement with recent work [48] suggesting that the introduction of a N370 sequon in SARS-CoV-2 S negatively affects binding to human ACE2, contributing to increased replication of SARS-CoV-2 S in human cells relative to its putative ancestral variant. Furthermore, it is also interesting to note that cryo-EM S structures from the bat RatG13 and pangolin CoV variants, both carrying the N370 sequon, have been only solved in their closed states [23, 33, 34], possibly also suggesting opening is less favoured in these S glycoproteins relative to the SARS-CoV-2 S.
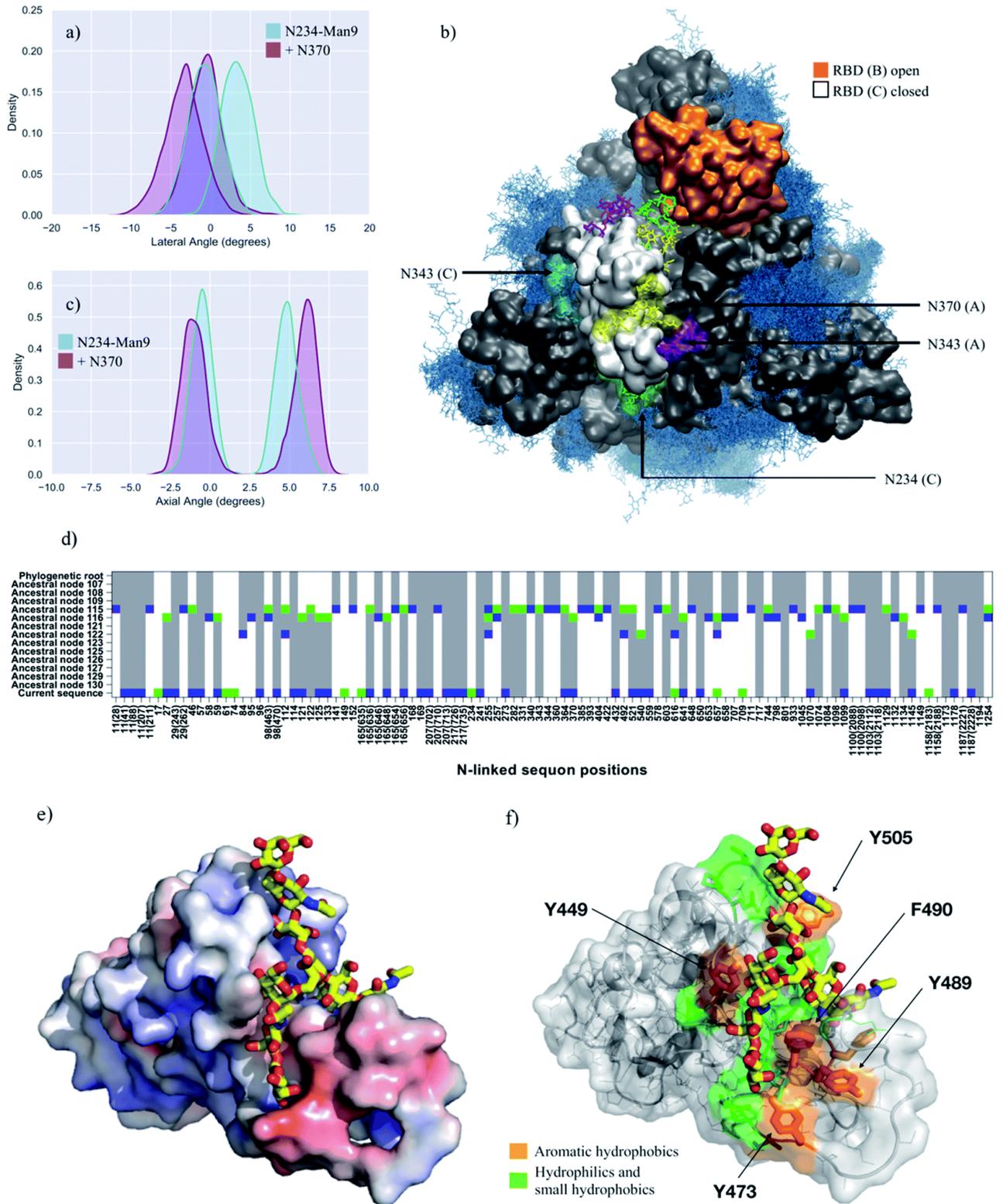
Figure 5.4: see following page.

Figure 5.4 (cont.)(panels a and c) Kernel density estimation (KDE) analysis of the lateral and axial angle distributions calculated through the uncorrelated MD trajectories obtained for N370-glycosylated N234-Man9 (purple), replicas R1, 2, and for N234-Man9 (cyan), replicas R1, 2 for comparison. (panel b) Close-ups on a representative snapshot of the N370-glycosylated N234-Man3 (R2) simulation from top. Man9 at N234 (B, C) is shown in green, while the complex *N*-glycans at N165 (B, C), N343 (C, A) and N370 (C, A) are shown in cyan, purple and yellow, respectively. The surface of the open RBD (B) is shown in orange and of the closed RBD (C) is shown in white; the rest of the protein is shown in grey. All other glycans are shown in blue, as an overlay of snapshots collected every 10 frames. Data analysis and graphs were done with seaborn (www.seaborn.pydata.org) and molecular rendering with VMD[28]. (panel d) Gain, loss, and retention of *N*-glycosylation sequons through the evolution of SARS-CoV-2 S as inferred from ancestral sequence reconstruction based on selected coronavirus spike proteins. The predicted ancestral sequence is at the top and the current SARS-CoV-2 S sequence at the base. Colours show gain (green), retention (grey), or loss (blue) of a sequon at a specific amino acid position at each phylogenetic node. *N*-glycosylation sequon positions are numbered as in the current SARS-CoV-2 sequence. Positions in the current SARS-CoV-2 sequence that are aligned with multiple sequons in reconstructed ancestral sequences due to insertion/deletion events also include their position in the multiple sequence alignment in parentheses. (panel e) Close-up of the RBD C bound to the N370 FA2G2 *N*-glycan represented with sticks and yellow C atoms. The RBD C is represented through a solvent accessible surface colorised based on the electrostatic potential calculated with the APBS plugin in PyMol (www.pymol.org). Darker shades of blue indicate increasingly positive charge, white indicate neutral charge and increasingly red shades indicate negatively charged regions. (panel f) Representation of the RBD C bound to the N370 FA2G2 *N*-glycan in the same orientation as in (panel e) highlighting potentially critical residues for binding. Molecular rendering done with PyMol.

Ultimately, the tight binding we observed between the N370 *N*-glycan and the surface of the SARS-CoV-2 S closed RBD surface is not only interesting in terms of the implications for higher ACE2-binding activity through loss of the corresponding sequon suggests, but also indicates the presence of a glycan binding site in that cleft, which is occupied in CoV variants that retain that sequon. In this context, recent work [35] provided evidence that heparan sulfate (HS) binds the SARS-CoV-2 S RBD in a ternary complex with ACE2, as an essential interaction for cell infection, meanwhile unfractionated heparin, non-anti-coagulant heparin, heparin lyases, and lung HS are found to potently block SARS-CoV-2 S binding to ACE2 and infection. Notably, the same study provides evidence that SARS-CoV S binding to heparin–BSA is significantly reduced, yet not completely negated[35]. Furthermore, recent work has also shown evidence that the RBD of SARS-CoV-2 S specifically binds sialogangliosides, such as GM1 and GM2, with the same affinity observed for glycosaminoglycans (GAGs), as well as blood group antigens with a lower affinity[39]. Our results suggest that in SARS-CoV2 S a glycan-binding cleft on the RBD surface is available and broadly accessible to be occupied by GAGs, sialogangliosides as well as blood group antigens, provided that they fit the structural and electronic constraints that the site imposes. Meanwhile, in SARS-CoV S binding of these species may be disfavoured, because the cleft is occupied by the N370 glycan from the adjacent protomer, further stabilizing the closed conformation. Furthermore, because of the high density of GAGs and sialogangliosides displayed on the surface of mammalian cells, we can speculate that this recently acquired topological change of the glycan shield may be advantageous for the virus towards cell surface localization and increased affinity to ACE2, where these glycans act as co-receptors [35, 39]. Further investigation on these topics is underway.

Analysis of reconstructed SARS S ancestral sequences indicates that while the N370 sequon was recently lost in SARS-CoV-2 S, this sequon was only quite recently acquired within the phylogeny. However, proximal N-glycosylation sequons for example at position 364 (D364-YS in CoV2) have been gained and lost alternatively, see Fig. 5.4 (pannel d). Based on our results, it is reasonable to think that glycosylation within this S topological region may have been evolutionarily conserved, because of its role in effectively stabilizing the open RBD, despite the higher energetic cost involved in the transition from the RBD down-to-up state. In SARS-CoV-2, glycosylation at N234, which according to our analysis, shown in Fig. 5.4 (pannel d), also appeared recently, would then functionally take the role of a glycan at N370. This evolutionary flexibility in the precise positions of glycosylation sites would make the presence of any specific glycan dispensable, with the consequent advantage of ensuring easier RBD opening reaction, and thus a more active S.

5.5   Conclusions

In this work we have used multi-microsecond MD simulations to determine the effect of changes in the nature and topology of the SARS-CoV-2 S N-glycosylation at sites known to be involved in its function. Our results indicate that reducing the size of the N-glycans at N234 led to the instability of the "wide-open" RBD conformation, with a consequent increase in RBD dynamics and a progressive stabilization of conformations favouring the closed protomer. Additionally, the structure of the N-glycans at N165 and N343 also affects the stability of the open RBD with shorter structures unable to effectively interact with the RBD disordered loop within the RBM. This effect is especially dramatic when a shorter N-glycan, such as Man5, is also present at N234. To account for changes in the glycan shield topology, we explored the effect of re-introducing N-glycosylation to a recently lost sequon at N370. Our results indicate that while the N-glycan at N370 is highly effective in stabilizing the open RBD in conjunction with the N-glycan at N234, it tightly binds a specific cleft on the surface of the closed RBD, tying the closed protomers together and likely increasing the energetic cost of the RBD opening. Because the architecture of this RBD cleft is particularly able to bind multiple monosaccharides through a network of hydrogen bonds and dispersion interactions, we suggest that in SARS-CoV-2 it can be occupied by other diverse glycan structures, such as glycosaminoglycans, proteoglycans or sialylated species, which have also been shown to bind S. This flexibility in glycan-binding preference would provide an additional advantage in terms of increasing localization at the host cell surface. Finally, comparative analysis of reconstructed SARS ancestral sequences suggests that specific changes in the glycan shield topology at and around N370, in conjunction with the gain of N-glycosylation at N234, may have contributed to an increase in S activity, and thus of the infectivity of the SARS-CoV-2 relative to closely related coronaviruses.

Bibliography

[1] Watanabe, Y.; Bowden, T. A.; Wilson, I. A.; Crispin, M. Exploitation of glycosylation in enveloped virus pathobiology. Biochimica et Biophysica Acta (BBA) - General Subjects **2019**, 1863, 1480–1497.

[2] Li, F. Structure, Function, and Evolution of Coronavirus Spike Proteins. Annual Review of Virology **2016**, 3, 237–261.

[3] Plemper, R. K. Cell entry of enveloped viruses. Current Opinion in Virology **2011**, 1, 92–100.

[4] Tortorici, M. A.; Veesler, D. Advances in Virus Research; Elsevier, 2019; pp 93–116.

[5] Walls, A. C.; Tortorici, M. A.; Snijder, J.; Xiong, X.; Bosch, B.-J.; Rey, F. A.; Veesler, D. Tectonic conformational changes of a coronavirus spike glycoprotein promote membrane fusion. Proceedings of the National Academy of Sciences **2017**, 114, 11157–11162.

[6] Wrapp, D.; Wang, N.; Corbett, K. S.; Goldsmith, J. A.; Hsieh, C.-L.; Abiona, O.; Graham, B. S.; McLellan, J. S. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science **2020**, 367, 1260–1263.

[7] Walls, A. C.; Park, Y.-J.; Tortorici, M. A.; Wall, A.; McGuire, A. T.; Veesler, D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell **2020**, 181, 281–292.e6.

[8] Yao, H. et al. Molecular Architecture of the SARS-CoV-2 Virus. Cell **2020**, 183, 730–738.e13.

[9] Turoňová, B. et al. In situ structural analysis of SARS-CoV-2 spike reveals flexibility mediated by three hinges. Science **2020**, 370, 203–208.

[10] Watanabe, Y.; Allen, J. D.; Wrapp, D.; McLellan, J. S.; Crispin, M. Site-specific glycan analysis of the SARS-CoV-2 spike. Science **2020**, 369, 330–333.

[11] Shajahan, A.; Supekar, N. T.; Gleinich, A. S.; Azadi, P. Deducing the N- and O-glycosylation profile of the spike protein of novel coronavirus SARS-CoV-2. Glycobiology **2020**, 30, 981–988.

[12] Sanda, M.; Morrison, L.; Goldman, R. N- and O-Glycosylation of the SARS-CoV-2 Spike Protein. Analytical Chemistry **2021**, 93, 2003–2009.

[13] Zhao, P. et al. Virus-Receptor Interactions of Glycosylated SARS-CoV-2 Spike and Human ACE2 Receptor. Cell Host & Microbe **2020**, 28, 586–601.e6.

[14] Watanabe, Y. et al. Native-like SARS-CoV-2 Spike Glycoprotein Expressed by ChAdOx1 nCoV-19/AZD1222 Vaccine. ACS Central Science **2021**, 7, 594–602.

[15] Brun, J.; Vasiljevic, S.; Gangadharan, B.; Hensen, M.; Chandran, A. V.; Hill, M. L.; Kiappes, J.; Dwek, R. A.; Alonzi, D. S.; Struwe, W. B.; Zitzmann, N. Assessing Antigen Structural Integrity through Glycosylation Analysis of the SARS-CoV-2 Viral Spike. ACS Central Science **2021**, 7, 586–593.

[16] Bagdonaite, I.; Thompson, A. J.; Wang, X.; Søgaard, M.; Fougeroux, C.; Frank, M.; Diedrich, J. K.; Yates, J. R.; Salanti, A.; Vakhrushev, S. Y.; Paulson, J. C.; Wandall, H. H. Site-Specific O-Glycosylation Analysis of SARS-CoV-2 Spike Protein Produced in Insect and Human Cells. Viruses **2021**, <u>13</u>, 551.

[17] Casalino, L.; Gaieb, Z.; Goldsmith, J. A.; Hjorth, C. K.; Dommer, A. C.; Harbison, A. M.; Fogarty, C. A.; Barros, E. P.; Taylor, B. C.; McLellan, J. S.; Fadda, E.; Amaro, R. E. Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein. ACS Central Science **2020**, <u>6</u>, 1722–1734.

[18] Sztain, T. et al. A glycan gate controls opening of the SARS-CoV-2 spike protein. Nature Chemistry **2021**, <u>13</u>, 963–968.

[19] Shang, J.; Wan, Y.; Luo, C.; Ye, G.; Geng, Q.; Auerbach, A.; Li, F. Cell entry mechanisms of SARS-CoV-2. Proceedings of the National Academy of Sciences **2020**, <u>117</u>, 11727–11734.

[20] Letko, M.; Marzi, A.; Munster, V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. Nature Microbiology **2020**, <u>5</u>, 562–569.

[21] Grant, O. C.; Montgomery, D.; Ito, K.; Woods, R. J. Analysis of the SARS-CoV-2 spike protein glycan shield reveals implications for immune recognition. Scientific Reports **2020**, <u>10</u>.

[22] Zimmerman, M. I. et al. SARS-CoV-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome. Nature Chemistry **2021**, <u>13</u>, 651–659.

[23] Choi, Y. K.; Cao, Y.; Frank, M.; Woo, H.; Park, S.-J.; Yeom, M. S.; Croll, T. I.; Seok, C.; Im, W. Structure, Dynamics, Receptor Binding, and Antibody Binding of the Fully Glycosylated Full-Length SARS-CoV-2 Spike Protein in a Viral Membrane. Journal of Chemical Theory and Computation **2021**, <u>17</u>, 2479–2487.

[24] Coutard, B.; Valle, C.; de Lamballerie, X.; Canard, B.; Seidah, N.; Decroly, E. The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. Antiviral Research **2020**, <u>176</u>, 104742.

[25] Wrobel, A. G.; Benton, D. J.; Xu, P.; Roustan, C.; Martin, S. R.; Rosenthal, P. B.; Skehel, J. J.; Gamblin, S. J. SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects. Nature Structural & Molecular Biology **2020**, <u>27</u>, 763–767.

[26] Neelamegham, S. et al. Updates to the Symbol Nomenclature for Glycans guidelines. Glycobiology **2019**, <u>29</u>, 620–624.

[27] Cheng, H. D.; Stöckmann, H.; Adamczyk, B.; McManus, C. A.; Ercan, A.; Holm, I. A.; Rudd, P. M.; Ackerman, M. E.; Nigrovic, P. A. High-throughput characterization of the functional impact of IgG Fc glycan aberrancy in juvenile idiopathic arthritis. Glycobiology **2017**, <u>27</u>, 1099–1108.

[28] Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. Journal of molecular graphics **1996**, <u>14</u>, 33–38.

[29] Williams, S. J.; Goddard-Borger, E. D. α-glucosidase inhibitors as host-directed an-
tiviral agents with potential for the treatment of COVID-19. Biochemical Society
Transactions **2020**, 48, 1287–1295.

[30] Brun, J.; Vasiljevic, S.; Gangadharan, B.; Hensen, M.; Chandran, A. V.; Hill, M. L.;
Kiappes, J.; Dwek, R. A.; Alonzi, D. S.; Struwe, W. B.; Zitzmann, N. Assessing Anti-
gen Structural Integrity through Glycosylation Analysis of the SARS-CoV-2 Viral
Spike. ACS Central Science **2021**, 7, 586–593.

[31] Harvey, W. T.; Carabelli, A. M.; Jackson, B.; Gupta, R. K.; Thomson, E. C.; Harri-
son, E. M.; Ludden, C.; Reeve, R.; Rambaut, A.; Peacock, S. J.; and, D. L. R. SARS-
CoV-2 variants, spike mutations and immune escape. Nature Reviews Microbiology
2021, 19, 409–424.

[32] Watanabe, Y.; Berndsen, Z. T.; Raghwani, J.; Seabright, G. E.; Allen, J. D.; Py-
bus, O. G.; McLellan, J. S.; Wilson, I. A.; Bowden, T. A.; Ward, A. B.; Crispin, M.
Vulnerabilities in coronavirus glycan shields despite extensive glycosylation. Nature
Communications **2020**, 11.

[33] Wrobel, A. G.; Benton, D. J.; Xu, P.; Calder, L. J.; Borg, A.; Roustan, C.; Mar-
tin, S. R.; Rosenthal, P. B.; Skehel, J. J.; Gamblin, S. J. Structure and binding
properties of Pangolin-CoV spike glycoprotein inform the evolution of SARS-CoV-2.
Nature Communications **2021**, 12.

[34] Zhang, S.; Qiao, S.; Yu, J.; Zeng, J.; Shan, S.; Tian, L.; Lan, J.; Zhang, L.; Wang, X.
Bat and pangolin coronavirus spike glycoprotein structures provide insights into
SARS-CoV-2 evolution. Nature Communications **2021**, 12.

[35] Clausen, T. M. et al. SARS-CoV-2 Infection Depends on Cellular Heparan Sulfate
and ACE2. Cell **2020**, 183, 1043–1057.e15.

[36] Tandon, R.; Sharp, J. S.; Zhang, F.; Pomin, V. H.; Ashpole, N. M.; Mitra, D.;
McCandless, M. G.; Jin, W.; Liu, H.; Sharma, P.; Linhardt, R. J. Effective Inhibition
of SARS-CoV-2 Entry by Heparin and Enoxaparin Derivatives. Journal of Virology
2021, 95.

[37] Mycroft-West, C. J. et al. Heparin Inhibits Cellular Invasion by SARS-CoV-2: Struc-
tural Dependence of the Interaction of the Spike S1 Receptor-Binding Domain with
Heparin. Thrombosis and Haemostasis **2020**, 120, 1700–1715.

[38] Liu, L.; Chopra, P.; Li, X.; Bouwman, K. M.; Tompkins, S. M.; Wolfert, M. A.;
de Vries, R. P.; Boons, G.-J. Heparan Sulfate Proteoglycans as Attachment Factor
for SARS-CoV-2. ACS Central Science **2021**, 7, 1009–1018.

[39] Nguyen, T.-D.; Dang, T.-T. T. Demystifying the momilactone pathway. Nature
Chemical Biology **2021**, 17, 126–128.

[40] Baker, A. N. et al. The SARS-COV-2 Spike Protein Binds Sialic Acids and Enables
Rapid Detection in a Lateral Flow Point of Care Diagnostic Device. ACS Central
Science **2020**, 6, 2046–2052.

[41] Sztain, T.; Ahn, S.-H.; Bogetti, A. T.; Casalino, L.; Goldsmith, J. A.; Seitz, E.;
McCool, R. S.; Kearns, F. L.; Acosta-Reyes, F.; Maji, S., et al. A glycan gate controls
opening of the SARS-CoV-2 spike protein. Nature chemistry **2021**, 13, 963–968.

[42] Bouwman, K. M.; Tomris, I.; Turner, H. L.; van der Woude, R.; Shamorkina, T. M.; Bosman, G. P.; Rockx, B.; Herfst, S.; Snijder, J.; Haagmans, B. L.; Ward, A. B.; Boons, G.-J.; de Vries, R. P. Multimerization- and glycosylation-dependent receptor binding of SARS-CoV-2 spike proteins. PLOS Pathogens **2021**, 17, e1009282.

[43] Fogarty, C. A.; Fadda, E. Oligomannose N-Glycans 3D Architecture and Its Response to the FcγRIIIa Structural Landscape. The Journal of Physical Chemistry B **2021**, 125, 2607–2616.

[44] Barnes, C. O.; Jette, C. A.; Abernathy, M. E.; Dam, K.-M. A.; Esswein, S. R.; Gristick, H. B.; Malyutin, A. G.; Sharaf, N. G.; Huey-Tubman, K. E.; Lee, Y. E.; Robbiani, D. F.; Nussenzweig, M. C.; West, A. P.; Bjorkman, P. J. SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. Nature **2020**, 588, 682–687.

[45] Parker, R. et al. Mapping the SARS-CoV-2 spike glycoprotein-derived peptidome presented by HLA class II on dendritic cells. Cell Reports **2021**, 35, 109179.

[46] Harbison, A. M.; Brosnan, L. P.; Fenlon, K.; Fadda, E. Sequence-to-structure dependence of isolated IgG Fc complex biantennary N-glycans: a molecular dynamics study. Glycobiology **2018**, 29, 94–103.

[47] Pallesen, J. et al. Immunogenicity and structures of a rationally designed prefusion MERS-CoV spike antigen. Proceedings of the National Academy of Sciences **2017**, 114.

[48] Kang, L.; He, G.; Sharp, A. K.; Wang, X.; Brown, A. M.; Michalak, P.; Weger-Lucarelli, J. A selective sweep in the Spike gene has driven SARS-CoV-2 human adaptation. Cell **2021**, 184, 4392–4400.e4.

6   Chapter 6 Role of C-Mannosylation in the Folding and Stability of TSR Domains

6.1   Introduction

Proteins can be modified by the addition of sugar molecules through a process known as glycosylation. This post-translational modification has a major impact on the structure, function, and stability of the proteins. glycosylation is an essential part of many cellular activities, such as protein folding, trafficking, and interactions. N-linked and O-linked glycosylation are the two main forms of glycosylation, which differ in the point of attachement of the glycan to the protein, namely an Asn for N-glycans and a Ser or a Thr for O-glycans where the attachment is made. The addition of glycans to proteins can have an effect on their 3D conformation as described in Chapter 5, solubility and susceptibility to enzymatic breakdown[1].

For example, ADCC is a mechanism of the immune system by which certain cells, such as natural killer (NK) cells, recognise and destroy target cells that are coated with antibodies. Antibodies mediate ADCC by binding to the target cell through their antigen-binding fragments (Fabs) and engaging with Fc receptors (such as the fcγRIIIa described in chapter 4) on the surface of the effector cells. The fucosylation of the core significantly decreases ADCC [2, 3].

Another less common and less studied form of glycosylation is tryptophan C-mannosylation, which is the only known type of protein C-glycosylation, This glycosylation takes place in the Endoplasmic Reticulum (ER) [4, 5] C-glycosylation is an essential post-translational modification for the trafficking, folding, and function of secretory transmembrane protiens. Here, a C-C bond is formed between CD1 of tryptophan residue proteins that harbour the $WX_2WX_2WX_2C$ consensus sequences by and the anomeric position C1 of a mannose monosaccaride, resulting in the post-translational modification Trp (Man). Mannose adopts the $^1C_4$ chair conformation with all hydroxyl groups except C2 in the axial position instead of the $^4C_1$ conformation that is more stable in solution. This is due to the proposed mechanism that results in a reclining of the chair in the final step [6].

The Thrombospondin Type 1 Repeats (TSRs) are structural domains found in proteins, TSR's which are known for their role in cell-matrix interactions and the regulation of cellular activities [7–9]. specifically the TSR 1 in Brain-specific angiogenesis inhibitor 1 (BAI1) (UniProt: O14514) in this chapter. The TSR domain is composed of a conserved sequence of approximately 60 amino acids that form a small and stable three-stranded antiparallel beta-sheet. The TSR motif is composed of six conserved cysteine residues (C16 C52, C20 C57 and C31 C42) which make up 3 disulphide bonds, and 6 residues across the repeat are consistent which make up a Trp-Arg backbone (W7 , R27 ,W10 , R29 , W13 and R50) all of which are described in Figure 6.1 .
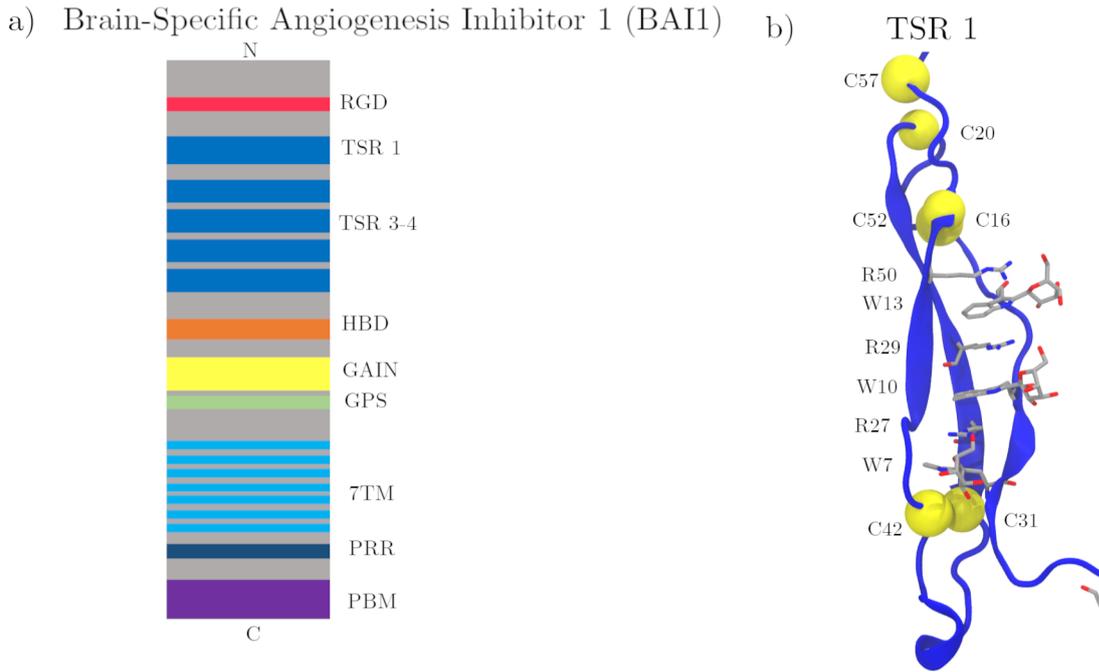
Figure 6.1:   Panel a): A diagram of the BAI1 protein structure ((UniProt: O14514))
is shown, featuring several domains, such as the Arg-Gly-Asp integrin-binding motif
(RGD), TSR 1-4, Hormone Binding Domain (HBD), G-protein-coupled receptor
(GPCR) autoproteolysis-inducing domain (GAIN), seven transmembrane domains,
a proline-rich region (PRR), and the PDZ-binding motif (PBM). Panel b): A more
detailed view of the TSR 1 domain within BAI1 is shown, which includes three α-
mannose molecules attached to the C-manosylation sites W7, W10, and W13. The
figure emphasizes the three disulfide bonds that contribute to the stability of TSR,
as well as the core 'Trp-Arg' backbone that is essential to the structure.

Our collaborator, Prof Ethan Goddard Borger and his research team at WEHI in
Melbourne have devised a new technique to synthesise thrombospondin repeats [10].
Initially, we were interested in the effect of individual C-mannosylation on the stability on
an indivudal TSR domain. Our colaborators showed that there is a $\Delta\Delta G_f \approx 2.4 kJ/mol$
for α Mannose but not for β Mannose that translates to a 10 degree stabilisation. Which
posed the question why was this?

Due to the fact that our collaborators domains are synthesized rather than expressed
[10] the incorporation of a variety of glycosylation patterns are possible, such as alpha
mannose, alpha rhamnose, alpha quinovose, and beta mannose. Figure 6.2 shows the
monosaccharides that were linked to the TSR domains. Thrombospondin repeats, can
now be synthesized with a range of glycosylation modifications. By using the subtle
changes in stereochemistry from monosacharide to monosacharide in this chaper I aim to
explain the stabalisation that α- mannose gives to TSR domains.

To achieve this,molecular dynamics simulations of TSR 1 from the BAI1 alpha-fold model
were perfomed with each of these monosacharides C linked to W7, W10, W13 and also ran
a simulation without C-linked glycans. These simulations involved a detailed examination
of the protein structures in the presence of different glycosylation moieties. Specifically,
we investigated the impact of alpha mannose, alpha rhamnose, alpha quinovose, and

| α-Mannose | α-Rhamnose | α-Quinovose | β-Mannose |
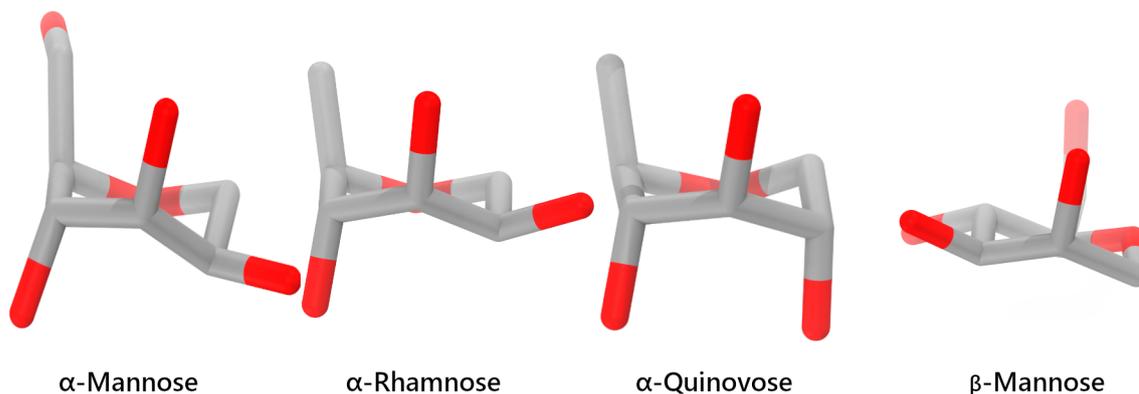
Figure 6.2: 3D structures of $^1C_4$ α-mannose,$^1C_4$ α-rhamnose, $^1C_4$ α-quinovose and $^4C_1$ β-mannose. The O6 3D position is lost in the transition to α-rhamnose, and the O2 position is altered when transitioning to $^1C_4$ α-quinovose. Due to the ring flip that occurs when moving from α-mannose to β-mannose, β-mannose shares none of these positions with α-mannose.

beta mannose on the stability of the TSR. The simulations aimed to unravel the molecular factors that contribute to the stabilisation of the protein fold, ultimately leading to a better understanding of the protein's behaviour in different glycosylation contexts.

The outcomes of these simulations provide crucial information about the structural dynamics and energetics of the glycosylated TSRs. This knowledge not only enhances our understanding of the biological functions of these proteins but also lays the foundation for the design of proteins with tailored glycosylation patterns for various applications, including therapeutic protein engineering and drug development. The ability to modulate protein folding energy through specific glycosylation strategies opens up new possibilities to advance both basic science and applied research in the fields of protein chemistry and molecular biology.

6.2 Computational Methods

MD simulations: Set-up and protocol of the system. TSR domains were generated from Alpha-fold using human BAI1 (UniProt: O14514)[11] To build our starting simulation model. Three C-linked mannoses in $^1C_4$ chair conformations were glycosidically added to W7,W12 and W13, based on structural alignment to the mouse BAI1 (ADGRB1) TSR3 domain (PDB 7R84) [12] the mouse data was used for the co-ordinates for the $^1C_4$ carbohydrates.

In all MD simulations, the protein and counterions (200 mM) were represented by the set of parameters of AMBER ff14SB9 [13, 14], while the glycans were represented by a version of the GLYCAM06 force field called GLYCAM06j-1 [15], the parameters of the C-Mannose bond were taken from the GAFF force field. The water molecules were represented by the TIP3P model [16–18]. All simulations were run with version v18 of the AMBER software package. The following running protocol was used for all MD simulations. The energy of the TSR domains was minimised through 50,000 cycles of the steepest descent algorithm. During the first minimisation, all heavy atoms were

harmonically restricted using a potential weight of 5 kcal mol^{-1} Å^2, while the solvent, counterions, and hydrogen atoms were left unrestrained. After energy minimisation, the system was equilibrated in the NVT ensemble with the same restraints scheme, where heating was performed in two stages over a total time of 1 ns, from 0 to 100 K (stage 1) and then 100 K to 300 K (stage 2). During equilibration, the SHAKE algorithm was used to constrain all bonds to hydrogen atoms.The Van der Waals interactions were truncated at 11 Å and Particle Mesh Ewald (PME) was used to treat long-range electrostatics with B-spline interpolation of order 4. Langevin dynamics with a collision frequency of 1.0 ps^{-1} was used to control the temperature, which was a pseudorandom variable seed to ensure there are no synchronisation artefacts. Once the system was brought to 300 K, an equilibration phase in the NPT ensemble of 1 ns was used to set the pressure at 1 atm. The pressure was held constant with an isotropic pressure scaling and a pressure relaxation time of 2.0 ps. At this point, all restrictions on the heavy atoms of the protein were removed, allowing the system to evolve for 15 ns of conformational equilibration prior to production. MD simulations were performed with resources allocated on ICHEC kay, using 2 Tesla V100 GPUs per simulation, with a benchmark standard of approximately 320 ns day^{-1}. The total simulation time for each model was 1.2 μs.

A similar method was employed for the simulations of the TSRs in complex with a glycosaminoglycan (GAG) octamer. The starting structures for these simulations were generated by coupling with AutoDock Vina [19, 20] using rigid body docking with a search box of 60x60x60 Å with a centroid of (35.8,48.7,24.5). Using MD-equilibrated structures of GAGs from our GlycoShape database (https://glycoshape.org available OA at the end of November 20203) with the TSRs. The same MD simulation protocol described above was performed to assess the stability of the complexes.

6.3 Results and Discussion

6.3.1 C-Mannosylation Sites
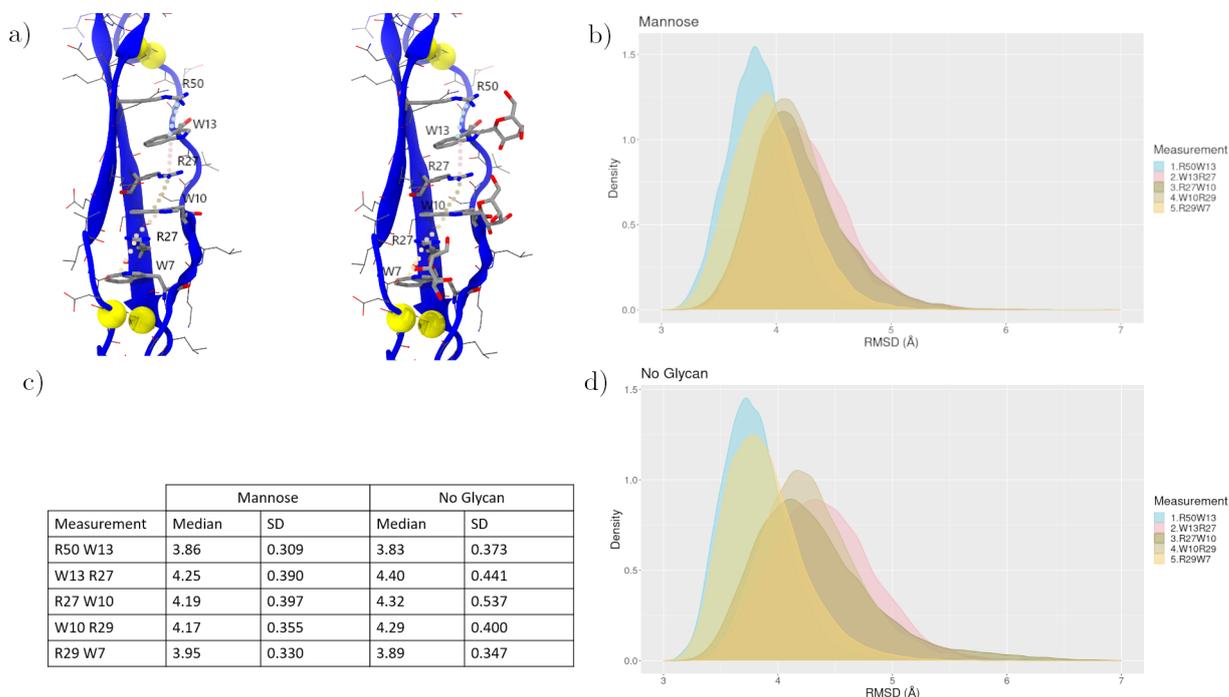
Figure 6.3: Panel a) illustrates the C_g Distance between the amino acids on the Trp-Arg ladder along two models of TSR's with and without C-glycosylation. Panel b illustrates the distribution of each of these distances for the glycosylated TSR and panel d) is the same plot for the non glycosylated TSR and panel c) is a summary table for each.

C-mannosylation at positions W10 and W13 in thrombospondin repeats (TSR) induces a significant decrease in folding energy. In particular, specific distances along the Trp-Arg ladder, such as W13-R27 and R27-W10, exhibit decreased standard deviations in glycosylated TSR, emphasising the stabilising effect of glycosylation at these positions. On the contrary, the R29-W7 interaction shows minimal change. These findings underscore the position-specific impact of C-mannosylation on TSR structural dynamics

A closer look at the Trp-Arg ladder in TSRs reveals the impact of C-mannosylation at positions W10 and W13. The protein structure wraps around W7 stabilising its position; meanwhile, this is not the case for W10 and W13, where you can see the C-Man connecting through O2 O4 and O6, suggesting that the C-Man of W10 and W13 is determinant for folding. This effect is visible in Figure 6.3 of C-mannosylation at positions 10 and W13 on this essential part of the protein structure.

An analysis of the gap between Arg27 (R27) and Trp10 (W10) showed a considerable increase in standard deviation, from 0.397 in the unglycosylated TSR to 0.537 in the glycosylated TSR. This alteration suggests that glycosylation at positions W10 and W13 is essential to reduce structural fluctuations within the Trp-Arg ladder, particularly between R27 and W10. In contrast, the distance between R29 and W7 showed a minor change in standard deviation from 0.330 to 0.347, implying a minor effect on the stability

of this particular interaction.

These nuanced changes in standard deviations along the Trp-Arg ladder further support
the hypothesis that C-mannosylation, particularly at positions and 3, contributes
significantly to the folding free energy by imparting structural stability, as it seems to
guide the formation of the disulphide bridges, so it may be essential to form contacts
that guide the correct assembly of the TSR by interlocking the structure with disulfide
bridges. The specific and position-dependent nature of these effects provides valuable
insight into the intricate relationship between glycosylation and protein folding dynamics.

6.3.2   Different Glycans
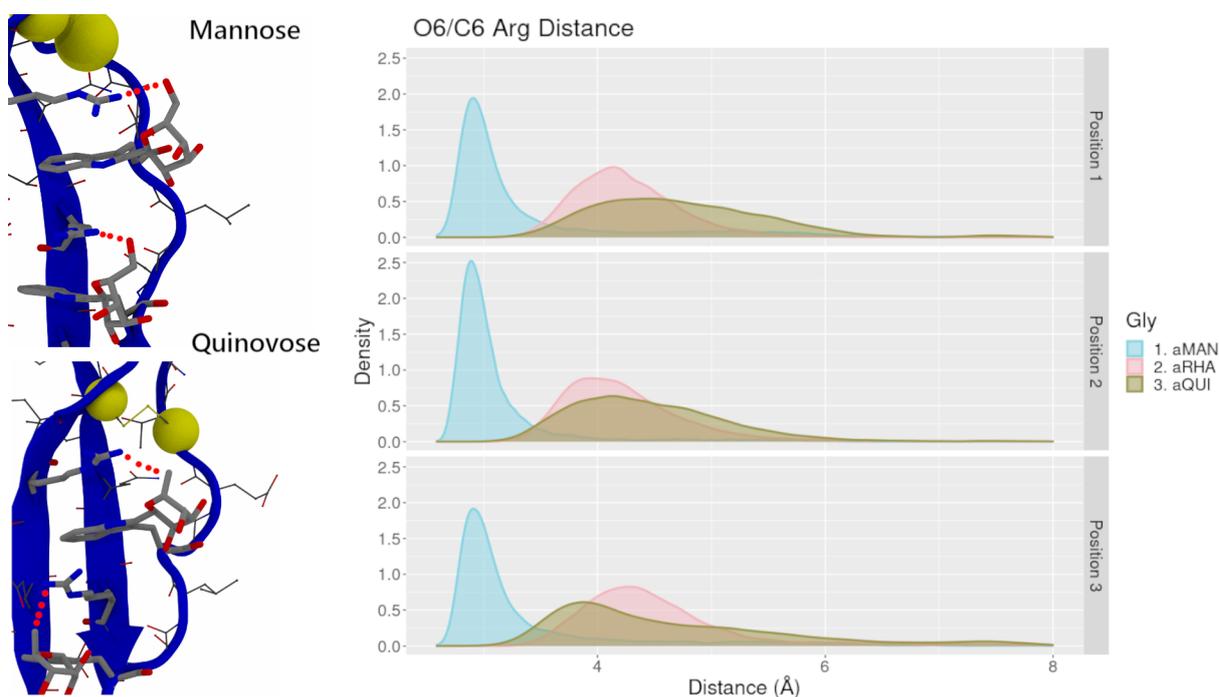
Figure 6.4: Figure showing the distances between the O6/C6 of the C-Linked sugar
and the Arg of the TSR.

The transition from mannose to rhamnose or quinovose in glycosylation patterns brings
about more than just structural changes. A key difference is the presence or absence
of the hydroxyl group in position C6, see Figure 6.2. This is especially relevant when
considering C-mannosylation, as the hydroxyl group at C6 is lost when mannose is
replaced by rhamnose or quinovose. This alteration results in the loss of a key hydrogen
bond with the relevant Arg resude in the Trp-Arg Ladder. Which has results in the
stability as the O6 position in mannose is responsible for hydrogen bonding with nearby
arginine residues see Figure 6.4.

The hydroxyl group on C6 is absent in rhamnose and quinovose, leading to a loss of the
hydrogen bonding capacity seen in α mannose. This lack of a potential hydrogen bond
creates a distinguishing factor between α C-mannosylation and functionalisation with
sugars such as α-rhamnose or α-quinovose, where the main difference is at position C6.
Specific molecular interactions between the glycan and adjacent amino acid residues, such

as Arg, play a crucial role in shaping the structural stability of TSRs.
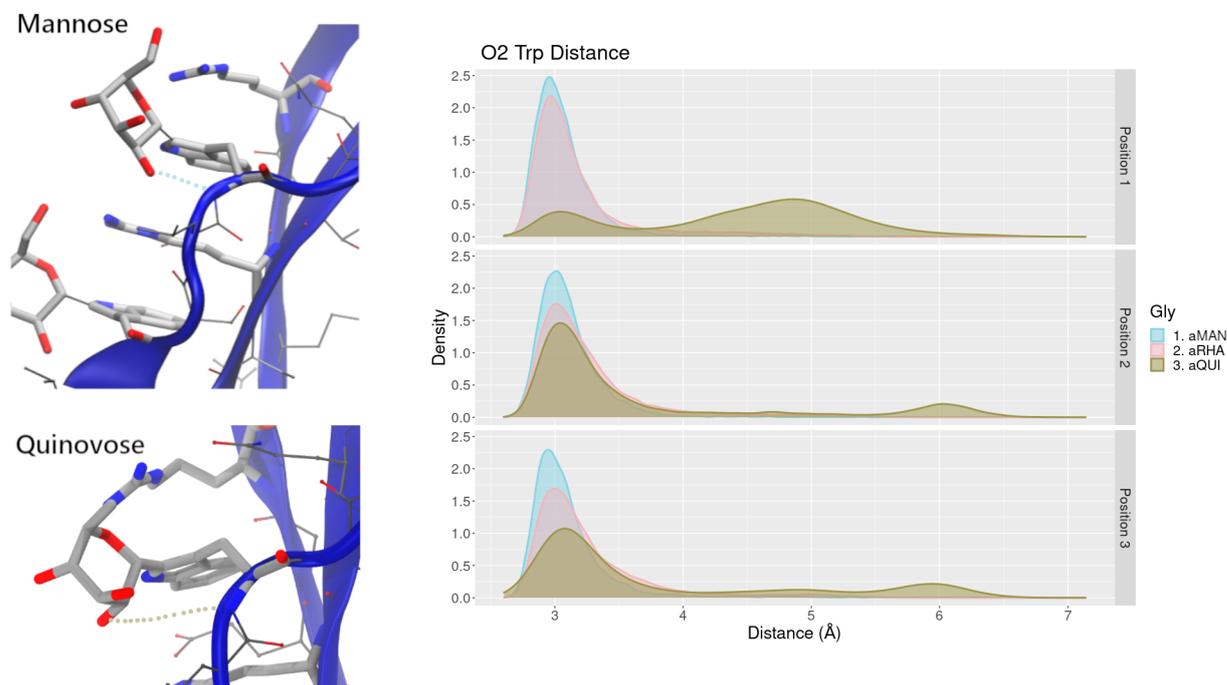
Figure 6.5: Figure showing the distances between the O2 of the C-Linked sugar and the backbone N of the Tryptophan of the TSR

The change from α-mannose to α-quinovose in glycosylation patterns is a significant transformation as it changes the orientation of the hydroxyl group at the C2 position. This structural change has far-reaching consequences for the intricate hydrogen-bonding interactions found in glycoproteins. In the case of α-mannose and α-rhamnose, the equatorial position of O2 provides a platform for potential hydrogen bonding, which stabilises the monosaccharide in a similar position, promoting critical interactions within the protein structure. However, compared to α-mannose and α-rhamnose, α-quinovose as well as loss in C6, at C2 the hydroxl adopts an axial orientation in the $^1C_4$ conformation, leading to loss of the potential hydrogen bonding capacity observed in α-mannose and α-rhamnose. This is visible in the KDE's in Figure 6.5

A particularly significant hydrogen bond found in α-mannose and α-rhamnose involves equatorial O2 and the nitrogen backbone of bound tryptophan. This interaction contributes substantially to the stabilisation of the glycosylation site within the protein structure. However, with the axial orientation of O2 in α-quinovose, this specific hydrogen bond is notably absent. The absence of this interaction significantly destabilises the TSR domain.

The transition from α-mannose to β-mannose within glycosylation patterns introduces a distinct modification, characterised by the absence of key hydrogen bonding interactions described above . Unlike α-mannose, β-mannose in the $^4C_1$ conformation lacks both the O6 Arg interactions and the O2 Trp N stabilisation. The absence of O6 interactions signifies a departure from potential hydrogen bonding with nearby arginine (Arg) residues, a feature observed in alpha mannose. This structural modification in beta mannose, i.e.
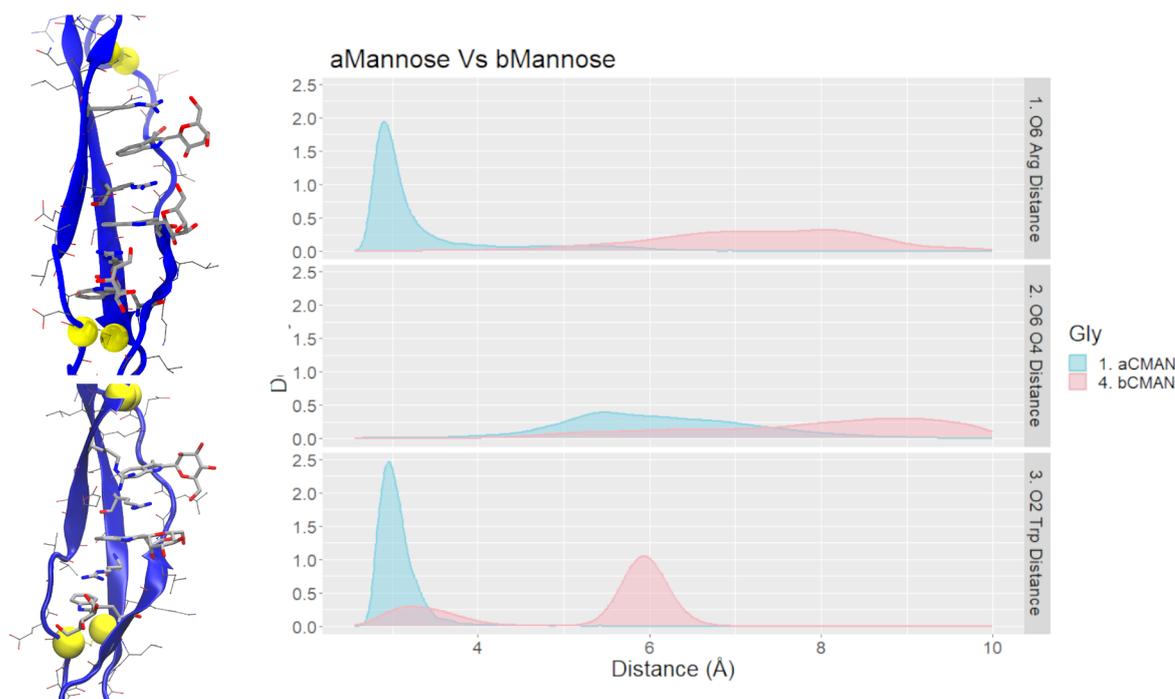
Figure 6.6: Figure illustrates that β-mannose posseses none of the stabalisation effects which are present in α mannose.

the ring flip, seems to discount both structural stabalisations present in the other two monosachardies.

The lack of stabilisation of O2 Trp N in β-mannose is due to the inherent ring flip in its structure. This dynamic alteration in the orientation of the sugar ring precludes the stabilisation of the hydrogen bond between the O2 position of the sugar and the nitrogen in the backbone of bound tryptophan (Trp). This distinctive feature distinguishes beta mannose from its alpha counterpart, highlighting the nuanced impact of glycosylation patterns on the landscape of hydrogen bonding within glycoproteins. Understanding these specific changes is crucial to unraveling the complex interaction between sugar moieties and protein structures, providing valuable insight into the molecular intricacies of glycoprotein function.

6.4   Conclusions

In conclusion, C-mannosylation plays a crucial role in the folding and stability of Thrombospondin Type 1 Repeats (TSRs) in proteins, with a specific focus on TSR 1 in Brain-specific angiogenesis inhibitor 1 (BAI1). The impact of C-mannosylation at different positions on the Trp-Arg ladder within TSRs has been investigated through molecular dynamics simulations. The results reveal position-specific effects, with glycosylation at W10 and W13 significantly reducing the folding energy and stabilising the Trp-Arg ladder.

Furthermore, the type of glycan attached to TSRs has a profound influence on their stability. The transition from α-mannose to α-rhamnose or α-quinovose alters the hydrogen bonding capacity, leading to changes in the structural dynamics. The loss of the hydroxyl group at C6 in rhamnose and quinovose results in the destabilisation of key

interactions, emphasising the importance of this specific moiety in glycosylation-induced stabilisation.

Furthermore, the comparison between α-mannose and β-mannose highlights the drastic impact of the anomeric configuration on stabilising interactions. β-mannose lacks the stabilising effects observed in α-mannose, with the absence of critical hydrogen bonding interactions at the positions O6 and O2. Ring flip in β-mannose disrupts the hydrogen bonding network, further underscoring the nuanced effects of glycosylation patterns on protein stability.

Bibliography

[1] Hang, I.; Lin, C.-w.; Grant, O. C.; Fleurkens, S.; Villiger, T. K.; Soos, M.; Morbidelli, M.; Woods, R. J.; Gauss, R.; Aebi, M. Analysis of site-specificN-glycan remodeling in the endoplasmic reticulum and the Golgi. Glycobiology **2015**, 25, 1335–1349.

[2] Li, T.; DiLillo, D. J.; Bournazos, S.; Giddens, J. P.; Ravetch, J. V.; Wang, L.-X. Modulating IgG effector function by Fc glycan engineering. Proceedings of the National Academy of Sciences **2017**, 114, 3485–3490.

[3] Harbison, A.; Fadda, E. An atomistic perspective on antibody-dependent cellular cytotoxicity quenching by core-fucosylation of IgG1 Fc N-glycans from enhanced sampling molecular dynamics. Glycobiology **2019**, 30, 407–414.

[4] Doucey, M. A.; Hess, D.; Cacan, R.; Hofsteenge, J. Protein C-mannosylation is enzyme-catalysed and uses dolichyl-phosphate-mannose as a precursor. Mol. Biol. Cell **1998**, 9, 291–300.

[5] Krieg, J.; Hartmann, S.; Vicentini, A.; Gläsner, W.; Hess, D.; Hofsteenge, J. Recognition signal for C-mannosylation of Trp-7 in RNase 2 consists of sequence Trp-x-x-Trp. Mol. Biol. Cell **1998**, 9, 301–309.

[6] Bloch, J. S.; John, A.; Mao, R.; Mukherjee, S.; Boilevin, J.; Irobalieva, R. N.; Darbre, T.; Scott, N. E.; Reymond, J.-L.; Kossiakoff, A. A.; Goddard-Borger, E. D.; Locher, K. P. Structure, sequon recognition and mechanism of tryptophan C-mannosyltransferase. Nature Chemical Biology **2023**, 19, 575–584.

[7] Albuquerque-Wendt, A.; Jacot, D.; Dos Santos Pacheco, N.; Seegers, C.; Zarnovican, P.; Buettner, F. F.; Bakker, H.; Soldati-Favre, D.; Routier, F. H. C-Mannosylation of Toxoplasma gondii proteins promotes attachment to host cells and parasite virulence. Journal of Biological Chemistry **2020**, 295, 1066–1076.

[8] Schultz-Cherry, S.; Lawler, J.; Murphy-Ullrich, J. E. The type 1 repeats of thrombospondin 1 activate latent transforming growth factor-beta. Journal of Biological Chemistry **1994**, 269, 26783–26788.

[9] Adams, J. C.; Tucker, R. P. The thrombospondin type 1 repeat (TSR) superfamily: diverse proteins with related roles in neuronal development. Developmental dynamics: an official publication of the American Association of Anatomists **2000**, 218, 280–299.

[10] Mao, R.; Xi, S.; Shah, S.; Roy, M. J.; John, A.; Lingford, J. P.; Gäde, G.; Scott, N. E.; Goddard-Borger, E. D. Synthesis of C-Mannosylated Glycopeptides Enabled by Ni-Catalyzed Photoreductive Cross-Coupling Reactions. Journal of the American Chemical Society **2021**, 143, 12699–12707.

[11] Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature **2021**, 596, 583–589.

[12] Miao, Y.; Jude, K.; Garcia, K. Structure of mouse BAI1 (ADGRB1) TSR3 domain in P21 space group. 2021; https://doi.org/10.2210/pdb7r84/pdb.

[13] Tsai, H.-C.; Lee, T.-S.; Ganguly, A.; Giese, T. J.; Ebert, M. C.; Labute, P.; Merz, K. M.; York, D. M. AMBER Free Energy Tools: A New Framework for the Design of Optimized Alchemical Transformation Pathways. Journal of Chemical Theory and Computation **2023**, 19, 640–658.

[14] Case, D. et al. AMBER 2018. *University of California: San Francisco, CA, USA* **2018**,

[15] Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; González-Outeiriño, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. GLYCAM06: A generalizable biomolecular force field. Carbohydrates. *Journal of Computational Chemistry* **2007**, *29*, 622–655.

[16] MacKerell, A. D. et al. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *The Journal of Physical Chemistry B* **1998**, *102*, 3586–3616.

[17] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **1983**, *79*, 926–935.

[18] Price, D. J.; Brooks, C. L. A modified TIP3P water potential for simulation with Ewald summation. *The Journal of Chemical Physics* **2004**, *121*, 10096–10103.

[19] Eberhardt, J.; Santos-Martins, D.; Tillack, A. F.; Forli, S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *Journal of Chemical Information and Modeling* **2021**, *61*, 3891–3898.

[20] Trott, O.; Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry* **2009**, *31*, 455–461.

7   Discussions and Conclusions

My PhD research focused on the structural complexity of glycans, highlighting their intrinsically disordered nature and uncovering ensembles of 3D structures. Using high-performance computing simulations, I have sought to uncover the connections between sequence, structure, and function in free glycans and changes occurring when the saccharides are linked to proteins. Through extensive molecular dynamics simulations that included both glycans and glycoproteins, I have discovered that these biomolecules have a highly complex, flexible structures, with conformational degrees of freedom that are far from random, yet strictly dictated by sequence and branching symmetries.

My work not only has unveiled the non-random nature of the glycan and glycoprotein structures, but also has shown the profound implications of understanding these structures. Specifically, through my work we gained insight into some of the effects of glycosylation of proteins and into the structure-to-function relationships of glycans in biological processes. In this thesis, I presented the structural foundation for those relationships and explored their effects in the context of biological system of high interest and impact in health and disease.

Starting in **Chapter 3** I looked into the sequence-to-structure relationship of the N-glycans particularly exploring the difference between plant / invertebrate and mammalian N-glycans. Unraveling the profound impact of modifications on their structure and dynamics. Specifically in this work, we explored alterations in the N-glycan sequence originating from plants and invertebrates as well as non-natural modifications to mammalian sequences. In particular, changes such as shifting linkages (e.g., from α1-6 to α1-3 linked core-fucose) or the addition of a single monosaccharide (e.g., β1-2 xylose) were identified as potential strategies to significantly reshape the conformational space occupied by adjacent monosaccharides. This, in turn, exerted a discernible influence on the overall conformational landscape of the N-glycans.

Moving onto **Chapter 4** I continued to investigate the sequence-to-structure relationship in the dynamics of human oligomannose N-glycans, ranging from Man5 to Man9, both in their unlinked state in solution and when linked to the FcγRIIIa protein. Despite the core chitobiose rigidity and moderate flexibility shared with complex N-glycans, oligomannoses exhibit a distinctive architecture that changes with the functionalisation of their arms. Man5, in particular, demonstrates a notable preference for an "open" structure, with the (1–6) arm outstretched and variations in the open structure influenced by the ω torsion of the (1–6) arm. Larger oligomannoses, from Man6 to Man9, bear additional Manα (1–2) -Man links, contributing to a more compact, 'tree-like' structure with increased interarm contacts. Examining the influence of the Fcγ RIIIa protein on the presentation of N-glycans reveals different effects at the glycosylation sites N45 and N162. The protein landscape at N45 restricts Man5 dynamics, limiting the accessibility of the (1–3) branch on the (1–6) arm. In contrast, N162 has a distinct structural landscape, preserving most of the intrinsic dynamics of the free glycan. Overall showing that these sequence-to-structure relationships carry over to glycoproteins, we show that the protien's surface selects from an already existing structures in conformational equilibrium, which is determined by the glycan sequence and branching.

In **Chapter 5** I explored the effects of glycosylation through extensive multi-microsecond MD simulations of SARS-CoV-2 Spike (S) protein. Specifically, I modulated the glycosyatlion sites N234, N165, N343 in order to understand on the intricate dynamics

of the receptor binding domain (RBD) and its susceptibility to glycan modifications, reducing the size of N-glycans at N234 was found to induce instability in the wide-open RBD conformation, leading to increased dynamics and a progressive shift toward a closed trimer state. The interplay of glycan sizes at multiple sites, particularly the presence of a shorter Man5 at N234, further influenced the stability of the open RBD. Reintroducing N-glycosylation into the recently lost sequon N370 provided information on its role in stabilising the open RBD. However, this glycan tightly bound a specific cleft on the surface of the closed RBD, potentially hindering the opening process. The observed cleft, capable of binding various glycan structures, suggests flexibility in glycan-binding preferences, including glycosaminoglycans, proteoglycans, and sialylated species. This adaptability may contribute to an increase in the localisation on the surface of the host cell. Overall, it shows that glycans are not mere decorations on these massive glycoprotiens but that they play a role in the formation of protien specifically the SARS CoV2 S in this case.

Lastly in **Chapter 6** of this thesis I moved my focus from N-glycans to a unique and less common type of glycosylation, namely C-mannosylation, found in thrombosponding repeats. The primary focus of this investigation was the study of TSR domain folding dynamics while systematically altering the C-mannosylation from mannose to rhamnose to quinovose. Our findings, revealing that each successive alteration in monosaccharide the stability of critical interactions—specifically, the O6-Arg and O2-Trp stabilizing non-covalent contacts. As a consequence, these structural modifications induced dramatic changes in the folding kinetics of TSR domains.

In summary, my Ph.D. work has provided valuable insights into the structural dependencies and dynamic aspects of N-glycans, extending our understanding of molecular recognition and functional roles. The key insight comes from viewing the structure of the glycan as free to choose from an ensamble of already existing structures, modulating the function of glycoproteins, exemplified by the functional role of glycans in the SARS-CoV-2 Spike proteins.

Appendices

A   Supporting Information for Chapter 3

A.1   Computational Methods

System preparation. All N-glycan starting structures for the MD simulations were generated with the GLYCAM Carbohydrate Builder (http://www.glycam.org). For each sequence we selected the complete set of rotamers obtained by variation of the 1-6 torsion angles, namely gg, gt and tg conformations for each 1-6 torsion. The topology file corresponding to each structure was obtained using tleap [1] with parameters from the GLYCAM06-j1 [2] or the carbohydrate atoms and with TIP3P for water molecules[3]. ach N-glycan was placed in the centre of a cubic simulation box of 16 Åides with no counterions to be consistent with the simulations run in earlier work [4]. Long range electrostatic were treated by Particle Mesh Ewald (PME) with cut-off set at 11 ÅLong range electrostatic were treated by Particle Mesh Ewald (PME) with cut-off set at 11 Å. The MD trajectories were generated by Langevin dynamics [5, 6] with collision frequency of 1.0 ps^{-1} Pressure was kept constant by isotropic pressure scaling [7] with a pressure relaxation time of 2.0 ps. Integration was done with a time step of 0.002 ps for all simulations, with bonds to hydrogen atoms restrained with the SHAKE algorithm [8] All calculations were run with the AMBER18 software package [1] on NVIDIA Tesla V100 16GB PCIe (Volta architecture) GPUs installed on the HPC infrastructure *kay* at the Irish Centre for High-End Computing (ICHEC).

Simulation Protocol. The energy of the hydrated systems was initially minimized through 500,000 cycles of steepest descent, with all heavy atoms restrained with a harmonic potential with a force constant of 5 $kcal\ mol^{-1}Å^{-2}$. After minimization, the system was heated in two stages. During the first stage the temperature was raised from 0 to 100 K over 500 ps at constant volume and in the second stage from 100 K to 300 K over 500 ps at constant pressure. Through the heating process all heavy atoms were kept restrained. After heating phase all restraints were removed and the system was allowed to equilibrate for 5 ns at 300 K and at 1 atm of pressure. Separate production steps of 500 ns each were run for each rotamer (starting system) and convergence was assessed based on conformational and clustering analysis. Simulations were extended, if the sampling was not deemed as fully converged.

Data analysis. All trajectories were processed using *cpptraj* [1] nd visually analysed with the Visual Molecular Dynamics (VMD) software package [9]. Backbone Root Mean Square Deviation (RMSD) and torsion angles values were measured using VMD. A density-based clustering method was used to calculate the populations of occupied conformations for each torsion angle in a trajectory and heat maps for each dihedral were generated with a kernel density estimate (KDE) function. Statistical and clustering analysis was done with the R package and data were plotted with RStudio (https://www.rstudio.com).

A.2   Supplementary tables and figures

.

Table A.1: Results of the clustering analysis showing the median and standard deviation values (in parenthesis) for the torsion angles (°) measured through a cumulative 1.5 μs MD sampling of the α(1-3) core fucosylated *ngf* glycan.

Fuc α(1-3) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-71.1 (8.9)	141.1 (6.3)	100
GlcNAcβ(1-4) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-72.1 (8.3)	-107.1 (7.6)	100
Manβ(1-4) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-73.3 (12.3)	-122.1 (14.9)	90.0
Cluster 2	-166.4 (16.7)	-145.6 (9.8)	8.3
Cluster 3	-64.0 (10.0)	75.3 (9.3)	1.7
Manα(1-6)Man	Phi	Psi	Pop(%)
Cluster 1	69.2 (10.2)	-176.6 (17.5)	65.7
Cluster 2	74.2 (9.4)	89.2 (10.8)	25.2
Cluster 3	60.4 (6.4)	60.0 (5.8)	9.1
GlcNAc β(1-2) Man (1-6)	Phi	Psi	Pop(%)
Cluster 1	-78.8 (14.3)	160.7 (20.6)	96.2
Cluster 2	69.6 (9.3)	154.8 (10.3)	2.5
Man α(1-3) Man	Phi	Psi	Pop(%)
Cluster 1	70.9 (8.5)	140.3 (13.5)	66.8
Cluster 2	70.6 (8.7)	99.2 (9.5)	33.2
GlcNAc β(1-2) Man (1-3)	Phi	Psi	Pop(%)
Cluster 1	-77.8 (13.7)	161.5 (13.5)	88.8
Cluster 2	-78.5 (8.0)	109.2 (7.2)	9.7

.

Table A.2: Results of the clustering analysis showing the median and standard deviation values (in parenthesis) for the torsion angles (°) measured through a cumulative 1.5 μs MD sampling of the α(1-3) core fucosylated *ngf* glycan.

Fuc α(1-3) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-72.5 (10.7)	125.2 (17.4)	100
GlcNAc β(1-4) GlcNAc	**Phi**	**Psi**	**Pop(%)**
Cluster 1	-72.9 (8.6)	-105.7 (11.7)	100
Man β(1-4) GlcNAc	**Phi**	**Psi**	**Pop(%)**
Cluster 1	-74.0 (12.3)	-123.9 (15.7)	88.5
Cluster 2	-165.7 (14.7)	-145.7 (8.8)	11.5
Man α(1-6) Man	**Phi**	**Psi**	**Pop(%)**
Cluster 1	69.3 (10.6)	-177.6 (19.4)	72.5
Cluster 2	71.5 (6.2)	79.2 (7.1)	16.2
Cluster 3	60.7 (5.7)	60.4 (5.1)	1.1
GlcNAc β(1-2) Man (1-6)	**Phi**	**Psi**	**Pop(%)**
Cluster 1	-76.7 (13.8)	162.5 (12.4)	88.4
Cluster 2	-76.0 (6.2)	113.5 (6.6)	5.9
Cluster 3	-147.6 (9.5)	97.9 (9.4)	1.9
Gal β(1-3) GlcNAc (1-6)	**Phi**	**Psi**	**Pop(%)**
Cluster 1	-72.9 (11.2)	124.0 (18.2)	96.6
Cluster 2	-82.1 (12.2)	-64.0 (8.4)	3.0
Cluster 3	-150.0 (11.3)	-101.0 (7.3)	1.4
Man α(1-3) Man	**Phi**	**Psi**	**Pop(%)**
Cluster 1	71.7 (8.9)	141.5 (14.5)	66.9
Cluster 2	70.4 (8.6)	99.12 (9.5)	33.1
GlcNAc β(1-2) Man (1-3)	**Phi**	**Psi**	**Pop(%)**
Cluster 1	-77.9 (14.2)	160.9 (14.5)	83.3
Cluster 2	67.2 (9.2)	152.4 (10.4)	8.6
Cluster 3	-77.7 (7.7)	152.4 (6.2)	8.1
Gal β(1-3) GlcNAc (1-3)	**Phi**	**Psi**	**Pop(%)**
Cluster 1	-72.4 (9.8)	125.0 (15.8)	100

.

Table A.3: Results of the clustering analysis showing the median and standard deviation values (in parenthesis) for the torsion angles (°) measured through a cumulative 3 μs MD sampling of the β(1-2) core xylosylated *ngx* glycan.

GlcNAc β(1-4) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-78.1 (10.2)	-129.6 (15.67)	96.3
Cluster 2	-81.8 (9.8)	63.8 (8.11)	3.7
Man β(1-4) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-75.5 (12.6)	-123.5 (14.6)	94.4
Cluster 2	-64.2 (6.8)	74.2 (9.7)	5.6
Xyl β(1-2) Man	Phi	Psi	Pop(%)
Cluster 1	-80.2 (12.0)	133.2 (17.0)	100
Man α(1-6) Man	Phi	Psi	Pop(%)
Cluster 1	69.7 (10.3)	-174.4 (16.8)	77.2
Cluster 2	73.5 (10.1)	106.1 (10.6)	22.8
GlcNAc β(1-2) Man (1-6)	Phi	Psi	Pop(%)
Cluster 1	-80.9 (15.5)	161.7 (12.0)	90.1
Cluster 2	-73.1 (7.9)	114.2 (7.8)	9.3
Man α(1-3) Man	Phi	Psi	Pop(%)
Cluster 1	78.4 (7.3)	114.8 (14.8)	100
GlcNAc β(1-2) Man (1-3)	Phi	Psi	Pop(%)
Cluster 1	-77.4 (13.3)	161.5 (12.2)	88.2
Cluster 2	-78.9 (6.4)	109.1 (7.18)	8.2
Cluster 3	-66.9 (9.0)	149.1 (12.3)	3.6

.

Table A.4: Results of the clustering analysis showing the median and standard deviation values (in parenthesis) for the torsion angles (°) measured through a cumulative 1.5 μs MD sampling of the β(1-2) core xylosylated *gx* glycan.

GlcNAc β(1-4) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-78.0 (10.1)	-130.2(15.7)	97.5
Cluster 2	-84.0 (5.4)	-64.8 (5.6)	2.5
Man β(1-4) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-75.2 (13.5)	-124.7 (15.0)	87.6
Cluster 2	-67.9 (14.7)	72.9 (11.7)	11.3
Cluster 3	-178.3 (6.9)	-175.5 (7.3)	1.0
Xyl β(1-2) Man	Phi	Psi	Pop(%)
Cluster 1	-78.2 (9.3)	138.4 (15.4)	100
Man α(1-6) Man	Phi	Psi	Pop(%)
Cluster 1	70.5 (10.4)	-173.6 (19.4)	70.0
Cluster 2	71.8 (9.6)	103.8 (12.9)	25.7
Cluster 3	161.6 (8.0)	132.5 (9.2)	2.3
Cluster 4	79.0 (6.7)	-80.2 (8.05)	1.59
GlcNAc β(1-2) Man (1-6)	Phi	Psi	Pop(%)
Cluster 1	-82.1 (15.0)	161.3 (12.2)	88.2
Cluster 2	-77.7 (7.47)	112.5 (7.0)	10.0
Cluster 3	-147.3 (9.0)	98.9 (9.4)	1.9
Gal β(1-3) GlcNAc (1-6)	Phi	Psi	Pop(%)
Cluster 1	-72.1 (10.0)	126.1 (15.5)	80.8
Cluster 2	-83.6 (11.4)	-63.0 (7.57)	19.2
Man α(1-3) Man	Phi	Psi	Pop(%)
Cluster 1	72.5 (10.7)	125.2 (17.4)	100
GlcNAc β(1-2) Man (1-3)	Phi	Psi	Pop(%)
Cluster 1	-78.5 (14.2)	162.4 (13.0)	91.2
Cluster 2	-77.8 (7.4)	111.1 (7.0)	8.8
Gal β(1-3) GlcNAc (1-3)	Phi	Psi	Pop(%)
Cluster 1	-71.1 (7.2)	125.5 (11.8)	100

.

Table A.5: Results of the clustering analysis showing the median and standard deviation values (in parenthesis) for the torsion angles (°) measured through a cumulative 1.5 μs MD sampling of the β(1-2) core xylosylated and α(1-3) core fucosylated *ngxf* glycan.

Fuc α(1-3) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-71.3 (9.0)	140.5 (6.0)	100
GlcNAc β(1-4) GlcNAc	**Phi**	**Psi**	**Pop(%)**
Cluster 1	-73.2 (11.0)	-106.4 (21.0)	100
Man β(1-4) GlcNAc	**Phi**	**Psi**	**Pop(%)**
Cluster 1	-75.9 (14.4)	-124.2 (15.3)	75.4
Cluster 2	-66.5 (11.3)	73.3 (10.8)	21.1
Cluster 3	179.0 (7.6)	-174.7 (8.2)	2.3
Cluster 4	-151.9 (8.2)	-146.5 (6.7)	1.2
Xyl β(1-2) Man	**Phi**	**Psi**	**Pop(%)**
Cluster 1	-78.8 (10.8)	135.9 (17.4)	100
Man α(1-6) Man	**Phi**	**Psi**	**Pop(%)**
Cluster 1	70.6 (9.0)	175.2 (17.6)	68.5
Cluster 2	71.7 (10.0)	106.5 (12.3)	24.5
Cluster 3	81.6 (8.0)	-75.0 (10.0)	4.4
Cluster 4	158.0 (8.9)	135.9 (10.4)	2.4
GlcNAc β(1-2) Man (1-6)	**Phi**	**Psi**	**Pop(%)**
Cluster 1	-79.2 (14.4)	158.4 (24.9)	90.1
Cluster 2	-80.0 (6.33)	113.0 (7.3)	9.3
Man α(1-3) Man	**Phi**	**Psi**	**Pop(%)**
Cluster 1	69.1 (9.4)	113.8 (16.8)	100
GlcNAc β(1-2) Man (1-3)	**Phi**	**Psi**	**Pop(%)**
Cluster 1	-77.4 (13.3)	161.5 (12.2)	88.2
Cluster 2	-78.9 (6.4)	109.1 (7.18)	8.2
Cluster 3	-66.9 (9.0)	149.1 (12.3)	3.6

.
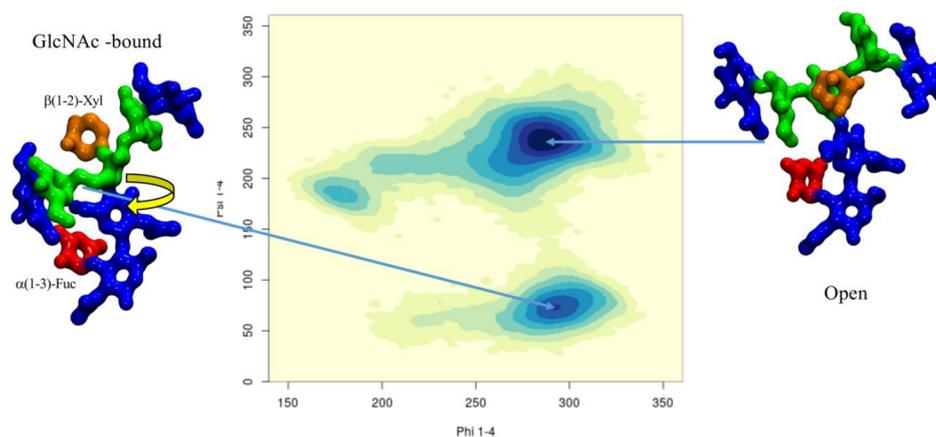
Figure A.1: Conformational analysis of the ngxf glycoform in terms of phi/psi torsion values, shown on the x and y axis, respectively, of the core Man-β(1-4)-GlcNAc linkage. Representative structures selected from MD sampling are shown on the left- and right-hand side of the heat map. The GlcNAc-bound conformation is obtained through a torsion of the Man3 "glycoblock" relative to the chitobiose and it is stabilized by hydrogen bonding interactions between the β(1-2)-Xyl and α(1- 3)-Fuc. The monosaccharides colouring follows the SFNG nomenclature. The structure rendering was done with VMD and the graphical statistical analysis with RStudio (https://www.rstudio.com).

Table A.7: Results of the clustering analysis showing the median and standard deviation values (in parenthesis) for the torsion angles (°) measured through a cumulative 1.5 μs MD sampling of the α(1-3) core fucosylated LeA glycan.

Fuc α(1-3) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-71.7 (9.8)	141.8 (6.6)	82.8
Cluster 2	-156.3 (7.02)	88.8 (17.6)	17.2
Man β(1-4) GlcNAc	**Phi**	**Psi**	**Pop(%)**
Cluster 1	-72.6 (8.7)	-107.4 (7.4)	100
Man β(1-4) GlcNAc	**Phi**	**Psi**	**Pop(%)**
Cluster 1	-72.2 (12.0)	-124.5 (14.0)	97.1
Cluster 2	179.7 (6.1)	-178.2 (6.6)	2.9
Man α(1-6) Man	**Phi**	**Psi**	**Pop(%)**
Cluster 1	-75.5 (7.4)	148.1 (7.4)	53.6
Cluster 2	-70.1 (5.8)	-177.0 (11.2)	28.2
Cluster 3	-73.2 (6.7)	-101.9 (6.0)	9.4
Cluster 4	-148.0 (6.3)	-165.3 (5.8)	8.8
Xyl β(1-2) Man	**Phi**	**Psi**	**Pop(%)**
Cluster 1	-83.9 (10.5)	132.8 (14.0)	100
GlcNAc β(1-2) Man (1-6)	**Phi**	**Psi**	**Pop(%)**
Cluster 1	-150.0(9.3)	98.5 (7.9)	40.9
Cluster 2	-91.5 (8.5)	151.6 (8.6)	29.9
Cluster 3	-62.7 (8.5)	161.4 (9.9)	29.2
Gal β(1-3) GlcNAc (1-6)	**Phi**	**Psi**	**Pop(%)**
Cluster 1	-70.1 (7.6)	131.5 (7.6)	100

Table A.6: Results of the clustering analysis showing the median and standard deviation values (in parenthesis) for the torsion angles (°) measured through a cumulative 1.5 μs MD sampling of the β(1-2) core xylosylated and α(1-3) core fucosylated *gxf* glycan.

Fuc α(1-3) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-71.1 (9.8)	140.5 (7.1)	88.9
Cluster 2	-156.8 (5.6)	90.7 (6.7)	11.1
GlcNAc β(1-4) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-72.2 (8.5)	-107.0 (8.0)	100
Man β(1-4) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-76.7 (15.1)	-124.5 (17.0)	85.2
Cluster 2	-179.5 (8.5)	-174.7 (9.5)	8.3
Cluster 3	-63.0 (9.9)	-75.6 (10.3)	5.5
Cluster 4	-153.1 (6.2)	-147.4 (5.31)	1.0
Xyl β(1-2) Man	Phi	Psi	Pop(%)
Cluster 1	-78.2 (9.9)	135.9 (17.4)	100
Man α(1-6) Man	Phi	Psi	Pop(%)
Cluster 1	70.73 (10.3)	-174.2 (19.0)	87.2
Cluster 2	70.8 (8.1)	101.5 (10.7)	9.3
Cluster 3	161.6 (8.5)	131.9 (10.6)	3.5
GlcNAc β(1-2) Man (1-6)	Phi	Psi	Pop(%)
Cluster 1	-78.1 (14.3)	161.9 (12.9)	84.7
Cluster 2	-76.1 (7.6)	111.6 (7.2)	10.5
Cluster 3	-147.1 (10.9)	99.5 (10.1)	3.2
Cluster 4	67.4 (7.4)	152.4 (8.4)	1.6
Gal β(1-3) GlcNAc (1-6)	Phi	Psi	Pop(%)
Cluster 1	-70.7 (10.4)	125.0 (16.8)	96.7
Cluster 2	-81.6 (7.07)	-64.6 (6.4)	3.3
Cluster 3	-81.6 (7.07)	-64.6 (6.4)	1.3
Man α(1-3) Man	Phi	Psi	Pop(%)
Cluster 1	68.7 (9.1)	114.1 (16.6)	100
GlcNAc β(1-2) Man (1-3)	Phi	Psi	Pop(%)
Cluster 1	-78.2 (14.7)	162.0 (13.0)	90.1
Cluster 2	-78.1 (7.2)	112.1 (6.8)	9.9
Gal β(1-3) GlcNAc (1-3)	Phi	Psi	Pop(%)
Cluster 1	-71.9 (6.9)	126.0 (11.9)	100

Table A.8: Results of the clustering analysis showing the median and standard deviation values (in parenthesis) for the torsion angles (°) measured through a cumulative 1.5 μs MD sampling of the β(1-2) core xylosylated *mgx* glycan. Note: *mg* refers to the mammalian terminal β(1-4)-Gal.

GlcNAc β(1-4) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-78.2 (10.9)	-131.1 (15.8)	97.5
Cluster 2	-79.6 (11.3)	66.6 (11.5)	2.5
Man β(1-4) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-75.7 (17.1)	-123.7 (14.7)	91.3
Cluster 2	-68.1 (12.6)	72.1 (11.9)	8.7
Xyl β(1-2) Man	Phi	Psi	Pop(%)
Cluster 1	-81.7 (19.1)	133.6 (20.1)	100
Man α(1-6) Man	Phi	Psi	Pop(%)
Cluster 1	72.2 (9.4)	103.5 (11.3)	56.9
Cluster 2	69.9 (8.3)	-173.8 (15.7)	39.8
Cluster 3	162.2 (9.1)	131.2 (9.3)	1.7
Cluster 4	58.5 (3.2)	59.98 (4.3)	1.6
GlcNAc β(1-2) Man (1-6)	Phi	Psi	Pop(%)
Cluster 1	-78.1 (14.3)	161.9 (12.9)	84.7
Cluster 2	-76.1 (7.6)	111.6 (7.2)	10.5
Cluster 3	-147.1 (10.9)	99.5 (10.1)	3.2
Cluster 4	67.4 (7.4)	152.4 (8.4)	1.6
Gal β(1-4) GlcNAc (1-6)	Phi	Psi	Pop(%)
Cluster 1	-72.9 (15.8)	-119.9 (16.0)	98.9
Cluster 2	-73.5 (12.6)	-73.24 (12.5)	0.6
Cluster 3	63.6 (10.7)	-117.8 (6.8)	0.8
Man α(1-3) Man	Phi	Psi	Pop(%)
Cluster 1	68.9 (9.7)	114.5 (16.7)	100
GlcNAc β(1-2) Man (1-3)	Phi	Psi	Pop(%)
Cluster 1	-79.0 (13.8)	162.5 (12.2)	87.5
Cluster 2	-80.9 (9.12)	109.4 (9.0)	12.5
Gal β(1-4) GlcNAc (1-3)	Phi	Psi	Pop(%)
Cluster 1	-72.41 (10.9)	-118.4 (15.1)	100

.

Table A.9:   Results of the clustering analysis showing the median and standard deviation values (in parenthesis) for the torsion angles (°) measured through a cumulative 4.5 μs MD sampling of the β(1-2) xylosylated and α(1-6) core fucosylated *mgmfx* glycan. Note: *mg* refers to the mammalian terminal β(1-4)-Gal and **mf** to the mammalian core α(1-6)-Fuc.

Fuc α(1-6) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-74.5 (9.6)	172.4 (14.5)	92.1
Cluster 2	-95.9 (4.5)	71.78 (6.1)	6.9
Cluster 3	-75.6 (2.5)	1113.8 (2.17)	1.0
GlcNAc β(1-4) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-77.2 (9.5)	-126.0 (14.3)	100
Man β(1-4) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-74.3 (15.0)	-122.7 (14.2)	97.1
Cluster 2	-67.4 (10.1)	73.6 (10.7)	2.9
Man α(1-6) Man	Phi	Psi	Pop(%)
Cluster 1	71.8 (9.4)	103.3 (10.98)	70.0
Cluster 2	70.4 (9.0)	-177.0 (11.2)	26.9
Cluster 3	67.5 (5.24)	-62.5 (5.7)	3.1
Xyl β(1-2) Man	Phi	Psi	Pop(%)
Cluster 1	-83.9 (10.5)	132.8 (14.0)	100
GlcNAc β(1-2) Man (1-6)	Phi	Psi	Pop(%)
Cluster 1	-91.5 (12.5)	159.9 (10.3)	96.3
Cluster 2	-75.5 (3.9)	113.6 (3.8)	2.9
Cluster 3	66.0 (6.8)	154.3 (5.8)	0.8
Gal β(1-4) GlcNAc (1-6)	Phi	Psi	Pop(%)
Cluster 1	-74.9 (13.56)	-122.1 (15.5)	97.5
Cluster 2	-83.6 (18.6)	65.3 (13.5)	2.5
Man α(1-3) Man	Phi	Psi	Pop(%)
Cluster 1	68.7 (9.9)	115.6 (16.9)	100
GlcNAc β(1-2) Man (1-3)	Phi	Psi	Pop(%)
Cluster 1	-79.1 (13.2)	162.3 (11.4)	90.9
Cluster 2	-77.9 (6.8)	110.4 (7.6)	9.1
Gal β(1-4) GlcNAc (1-3)	Phi	Psi	Pop(%)
Cluster 1	-72.7 (15.4)	-118.6 (15.7)	97.3
Cluster 2	-74.0 (12.4)	70.12 (13.1)	2.7

Table A.10:   Results of the clustering analysis showing the median and standard deviation values (in parenthesis) for the torsion angles (°) measured through a cumulative 4.5 μs MD sampling of the β(1-2) xylosylated and α(1-6) core fucosylated *nmgmfx* glycan. Note: *nmg* refers to the absence of mammalian terminal β(1-4)-Gal and *mf* to the mammalian core α(1-6)-Fuc.

Fuc α(1-6) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-73.0 (9.6)	177.0 (14.8)	93.1
Cluster 2	-95.6 (4.0)	75.0 (5.4)	6.9
GlcNAc β(1-4) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-77.2 3(9.7)	-126.0 (14.4)	100
Man β(1-4) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-76.6 (18.5)	-124.8 (16.9)	91.1
Cluster 2	-66.2 (12.5)	73.0 (11.9)	8.2
Man α(1-6) Man	Phi	Psi	Pop(%)
Cluster 1	70.1 (9.2)	-173.6 (14.6)	71.5
Cluster 2	72.1 (8.3)	104.25 (9.7)	28.5
Xyl β(1-2) Man	Phi	Psi	Pop(%)
Cluster 1	-80.6 (16.5)	135.8 (14.0)	100
GlcNAc β(1-2) GlcNAc (1-6)	Phi	Psi	Pop(%)
Cluster 1	-82.8 (14.94)	161.1 (11.9)	90.6
Cluster 2	-77.8 (7.3)	111.3 (6.8)	7.4
Cluster 3	66.0 (7.3)	152.9 (8.7)	2.0
Man α(1-3) Man	Phi	Psi	Pop(%)
Cluster 1	68.9 (9.4)	114.4 (16.9)	100
GlcNAc β(1-2) GlcNAc (1-3)	Phi	Psi	Pop(%)
Cluster 1	-78.4 (13.4)	162.6 (11.4)	88.8
Cluster 2	-78.9 (7.2)	110.4 (7.0)	9.0
Cluster 3	66.45 (10.9)	152.6 (10.7)	2.2

Table A.11:   Results of the clustering analysis showing the median and standard deviation values (in parenthesis) for the torsion angles (°) measured through a cumulative 1.5 μs MD sampling of the β(1-2) xylosylated and α(1-3) core fucosylated *mgpfx* glycan. Note: *mg* refers to mammalian terminal β(1-4)-Gal and *pf* to the plant core α(1-3)-Fuc.

Fuc α(1-3) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-70.8 (10.7)	141.3 (8.9)	93.8
Cluster 2	-156.8 (7.7)	91.3 (9.3)	6.2
GlcNAc β(1-4) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-72.5 (8.8)	-107.3 (9.1)	100
Man β(1-4) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-76.4 (15.1)	-123.6 (17.3)	87.0
Cluster 2	-68.2 (12.8)	70.9 (11.2)	13.0
Xyl β(1-2) Man	Phi	Psi	Pop(%)
Cluster 1	-80.7 (15.3)	134.5 (16.5)	100
Man α(1-6) Man	Phi	Psi	Pop(%)
Cluster 1	69.9 (9.0)	-175.5 (15.0)	49.1
Cluster 2	72.4 (9.7)	105.1 (12.5)	46.7
Cluster 3	158.6 (11.8)	135.8 (13.6)	4.2
GlcNAc β(1-2) GlcNAc (1-6)	Phi	Psi	Pop(%)
Cluster 1	-86.2 (15.0)	160.9 (11.4)	96.4
Cluster 2	-78.5 (6.1)	113.5 (4.9)	3.6
Gal β(1-4) GlcNAc (1-6)	Phi	Psi	Pop(%)
Cluster 1	-72.9 (15.0)	-119.9 (15.6)	97.2
Cluster 2	-74.1 (13.1)	70.1 (13.8)	2.8
Man α(1-3) Man	Phi	Psi	Pop(%)
Cluster 1	68.8 (9.6)	114.5 (16.7)	100
GlcNAc β(1-2) GlcNAc (1-3)	Phi	Psi	Pop(%)
Cluster 1	-78.8 (13.8)	162.6 (12.6)	88.3
Cluster 2	-80.7 (8.0)	109.5 (7.6)	9.6
Cluster 3	65.2 (9.9)	150.8 (10.8)	2.1
Gal β(1-4) GlcNAc (1-3)	Phi	Psi	Pop(%)
Cluster 1	-72.5 (11.5)	-118.6 (15.5)	100

Table A.12:   Results of the clustering analysis showing the median and standard deviation values (in parenthesis) for the torsion angles (°) measured through a cumulative 1.5 μs MD sampling of the α(1-3) core fucosylated *mgpf* glycan. Note: *mg* refers to the mammalian terminal β(1-4)-Gal and *pf* to the plant core α(1-3)-Fuc

Fuc α(1-3) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-71.1 (10.6)	140.3 (14.8)	98.1
Cluster 2	-156.5 (6.2)	91.4 (9.3)	1.9
GlcNAc β(1-4) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-73.0 (14.8)	-121.5 (15.7)	95.9
Cluster 2	-80.9 (15.3)	62.5 (13.4)	4.1
Man β(1-4) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-76.6 (14.1)	-124.9 (16.1)	77.9
Cluster 2	-153.7 (13.3)	-139.7 (8.5)	12.8
Cluster 3	-71.1 (12.5)	69.6 (11.4)	9.3
Man α(1-6) Man	Phi	Psi	Pop(%)
Cluster 1	74.2 (13.0)	86.5 (14.7)	74.8
Cluster 2	70.3 (9.1)	-176.5 (14.7)	25.2
GlcNAc β(1-2) GlcNAc (1-6)	Phi	Psi	Pop(%)
Cluster 1	-80.2 (14.6)	163.2 (12.4)	100
Gal β(1-4) GlcNAc (1-6)	Phi	Psi	Pop(%)
Cluster 1	-73.0 (14.88)	-121.58 (15.8)	95.9
Cluster 2	-80.9 (15.3)	62.5 (13.4)	4.1
Man α(1-3) Man	Phi	Psi	Pop(%)
Cluster 1	70.95 (9.6)	140.95 (15.2)	73.0
Cluster 2	70.1(8.21)	101.2 (8.8)	27.0
GlcNAc β(1-2) GlcNAc (1-3)	Phi	Psi	Pop(%)
Cluster 1	-78.5 (13.3)	162.1 (11.7)	91.2
Cluster 2	-77.9 (6.4)	110.7 (6.62)	8.8
Gal β(1-4) GlcNAc (1-3)	Phi	Psi	Pop(%)
Cluster 1	-72.2 (10.9)	-118.4 (15.0)	100

Table A.13:   Results of the clustering analysis showing the median and standard deviation values (in parenthesis) for the torsion angles (°) measured through a cumulative 2 μs MD sampling of the α(1-3) and α(1-6) core fucosylated *mgmfpf* glycan. Note: *mg* refers to the mammalian terminal β(1-4)-Gal, *pf* to the plant core α(1-3)-Fuc and *mf* to the mammalian core α(1-6)-Fuc.

Fuc α(1-6) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-72.0 (9.6)	-179.5 (14.8)	74.3
Cluster 2	-76.2 (4.0)	117.36 (12.1)	12.9
Cluster 3	-144.4 (7.7)	171.0 (5.5)	12.8
Fuc α(1-6) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-70.2 (10.3)	140.1 (9.7)	88.6
Cluster 2	-157.3 (7.7)	91.4 (8.9)	11.4
GlcNAc β(1-4) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-73.8 (9.5)	-106.4 (14.1)	91.7
Cluster 2	-154.9 (10.6)	-147.8 (7.4)	8.3
Man β(1-4) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-72.1 (10.9)	-120.7 (12.6)	74.8
Cluster 2	-153.0 (13.0)	-139.9 (8.3)	25.2
Man α(1-6) Man	Phi	Psi	Pop(%)
Cluster 1	73.5 (11.7)	86.6 (17.3)	85.1
Cluster 2	69.5 (9.9)	-176.3 (15.8)	14.9
GlcNAc β(1-2) GlcNAc (1-6)	Phi	Psi	Pop(%)
Cluster 1	-85.1 (14.9)	161.7 (13.4)	100
Gal β(1-4) GlcNAc (1-6)	Phi	Psi	Pop(%)
Cluster 1	-73.9 (11.5)	-123.5 (16.5)	98.7
Cluster 2	-146.2 (8.5)	-142.1 (6.9)	1.3
Man α(1-3) Man	Phi	Psi	Pop(%)
Cluster 1	71.1 (9.3)	140.5 (15.6)	72.9
Cluster 2	69.9 (8.8)	100.63(9.8)	27.1
GlcNAc β(1-2) GlcNAc (1-3)	Phi	Psi	Pop(%)
Cluster 1	-78.7 (13.8)	161.1 (12.8)	90.4
Cluster 2	-79.1 (7.7)	110.7 (7.2)	9.6
Gal β(1-4) GlcNAc (1-3)	Phi	Psi	Pop(%)
Cluster 1	-72.4 (10.9)	-118.6 (15.2)	95.0
Cluster 1	-143.2 (10.4)	-144.4 (5.9)	2.9
Cluster 1	-74.2 (9.3)	67.6 (9.4)	2.1
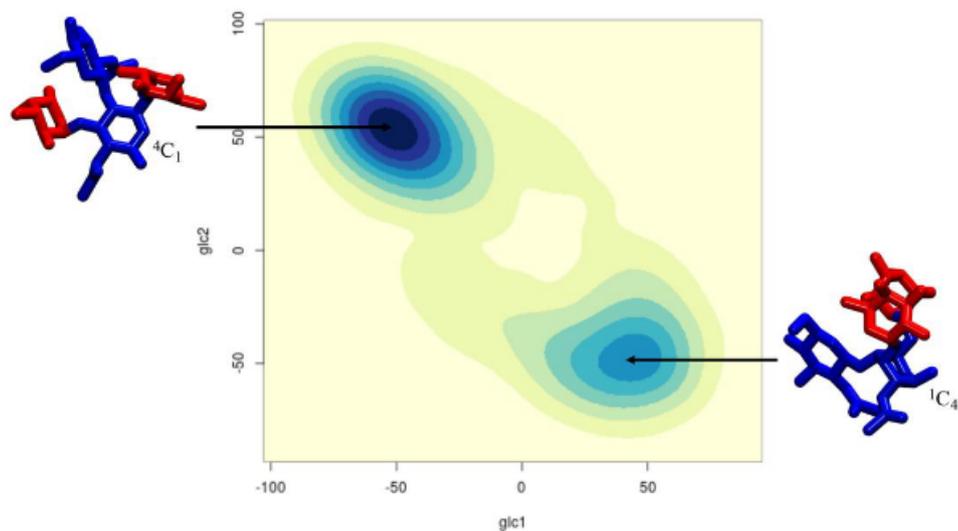
Figure A.2: Conformational analysis N-linked GlcNAc pucker along the 1.5 μs cumulative sampling of the α(1-3) and α(1-6) core fucosylated A2G2 (*mgmfpf*) N-glycan. Representative structures of the fucosylated chitobiose selected from MD sampling of the whole N-glycans are shown on the left- and right-hand side of the heat map where the ring pucker is also indicated. The x and y axis are labelled with the torsion angled measured, namely C1C2C3C4 (glc1) and C2C3C4C5 (glc2), respectively. The monosaccharides colouring follows the SFNG nomenclature. The structure rendering was done with VMD and the graphical statistical analysis with RStudio (https://www.rstudio.com).

Table A.14: Results of the clustering analysis showing the median and standard deviation values (in parenthesis) for the torsion angles (°) measured through a cumulative 2 μs MD sampling of the β(1-2) xylosylated and α(1-3) and α(1-6) core fucosylated *mgxmfpf* glycan. Note: *mg* refers to the mammalian terminal β(1-4)-Gal, pf to the plant core α(1-3)-Fuc and *mf* to the mammalian core α(1-6)-Fuc.

Fuc α(1-6) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-75.9 (11.25)	175.65 (14.0)	79.3
Cluster 2	-142.7 (9.8)	171.3(6.2)	16.9
Cluster 3	-95.48 (4.4)	71.95(4.8)	3.8
Fuc α(1-3) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-70.6 (10.3)	141.4 (8.3)	88.8
Cluster 2	-156.0 (8.1)	89.5 (8.7)	11.2
GlcNAc β(1-4) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-72.8 (8.9)	-106.6 (11.0)	96.7
Cluster 2	-82.1 (7.1)	-154.8 (6.8)	3.3
Man β(1-4) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-75.5 (11.0)	-123.5 (12.6)	94.6
Cluster 2	-62.42 (7.9)	72.2 (9.5)	4.4
Man α(1-6) Man	Phi	Psi	Pop(%)
Cluster 1	73.4 (10.5)	102.8 (10.6)	55.3
Cluster 2	70.4(9.4)	-176.8 (16.7)	26.7
Cluster 3	102.5 (9.3)	55.9 (7.7)	17.9
GlcNAc β(1-2) GlcNAc (1-6)	Phi	Psi	Pop(%)
Cluster 1	-91.8 (13.5)	159.4 (11.1)	100
Gal β(1-4) GlcNAc (1-6)	Phi	Psi	Pop(%)
Cluster 1	-74.2 (11.1)	-121.1 (15.3)	99.2
Cluster 2	-146.6 (7.5)	-143.6 (4.9)	0.8
Man α(1-3) Man	Phi	Psi	Pop(%)
Cluster 1	68.7 (9.7)	115.4 (16.8)	100
GlcNAc β(1-2) GlcNAc (1-3)	Phi	Psi	Pop(%)
Cluster 1	-79.5 (13.8)	162.3 (12.4)	92.4
Cluster 2	-79.3 (6.9)	110.3 (6.9)	7.6
Gal β(1-4) GlcNAc (1-3)	Phi	Psi	Pop(%)
Cluster 1	-72.6 (11.0)	-119.1 (15.2)	93.4
Cluster 2	-143.2 (10.4)	-144.4 (5.9)	4.9
Cluster 3	-74.2 (11.6)	69.7 (12.1)	1.7

Table A.15: Results of the clustering analysis showing the median and standard deviation values (in parenthesis) for the torsion angles (°) measured through a cumulative 2 μs MD sampling of the β(1-2) xylosylated and α(1-3) fucosylated A2 glycan terminating with LeX on both arms.

Fuc α(1-3) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-70.2 (11.0)	141.9 (8.4)	90.8
Cluster 2	-156.7 (7.6)	90.2 (8.6)	9.2
GlcNAc β(1-4) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-72.2 (8.6)	-107.5 (7.7)	100
Man β(1-4) GlcNAc	Phi	Psi	Pop(%)
Cluster 1	-76.4 (15.7)	-124.6 (16.7)	75.0
Cluster 2	-68.1 (9.7)	71.0 (10.1)	25.0
Man α(1-6) Man	Phi	Psi	Pop(%)
Cluster 1	70.7 (9.2)	-173.6 (13.4)	89.7
Cluster 2	152.4 (12.9)	145.6 (12.3)	10.3
Xyl β(1-2) Man	Phi	Psi	Pop(%)
Cluster 1	-77.6 (9.6)	140.1 (18.1)	100
GlcNAc β(1-2) GlcNAc (1-6)	Phi	Psi	Pop(%)
Cluster 1	-78.2 (14.3)	162.3 (13.1)	91.8
Cluster 2	-76.9 (7.4)	111.1 (6.8)	8.2
Fuc α(1-4) GlcNAc (1-6)	Phi	Psi	Pop(%)
Cluster 1	-69.7 (8.9)	142.1 (7.0)	100
Gal β(1-3) GlcNAc (1-6)	Phi	Psi	Pop(%)
Cluster 1	-67.7 (8.0)	-107.9 (7.2)	100
Man α(1-3) Man	Phi	Psi	Pop(%)
Cluster 1	61.5 (8.5)	111.0 (15.1)	100
GlcNAc β(1-2) GlcNAc (1-3)	Phi	Psi	Pop(%)
Cluster 1	-78.8 (13.9)	162.6 (11.7)	91.6
Cluster 2	-79.1 (6.2)	112.5 (6.2)	8.3
Gal β(1-3) GlcNAc (1-3)	Phi	Psi	Pop(%)
Cluster 1	-69.1 (9.45)	142.2 (8.0)	100
Fuc α(1-4) GlcNAc (1-3)	Phi	Psi	Pop(%)
Cluster 1	-67.8 (7.9)	-108.1 (7.4)	94.8
Cluster 2	-153.9 (8.8)	-141.5 (7.7)	5.2
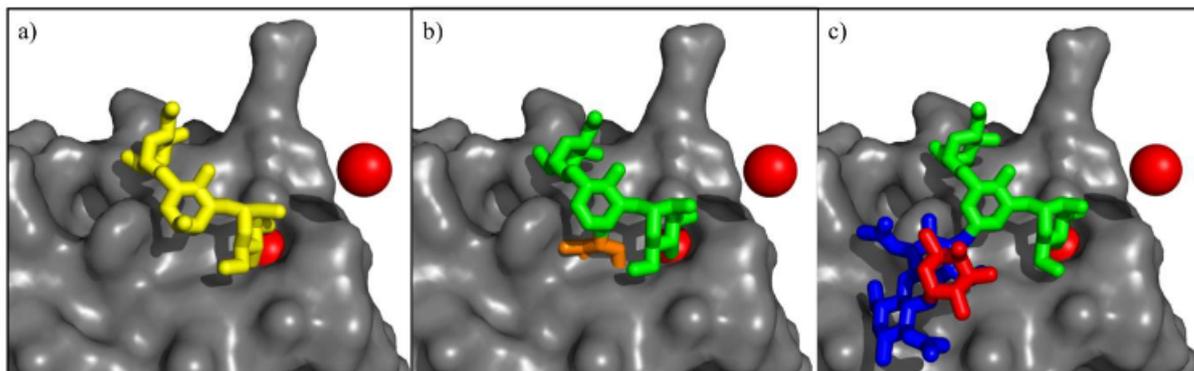
Figure A.3: Structural alignment of different plant N-glycoforms from our MD simulations on to the DC-SIGN/GlcNAc2Man3 complex resolved at 2.5 Åresolution (PDBid 1k9i). Panel a) The water accessible surface of the DC-SIGN (chain C) binding site is shown in grey and the Man3 region of the co-crystallized glycan in 1k9i is rendered as yellow sticks. Ca2+ ions are shown as red spheres. Panel b) Structural alignment of representative structure from our MD simulation of the *ngx* plant N-glycan shows that the β(1-2) xylose sterically hinders binding by clashing with the surface of the binding site. Only the Xyl-Man3 glycoblock from the whole N-glycan is represented. Panel c) Structural alignment of representative structure from our MD simulation of the ngf plant N-glycan shows that the α(1-3) fucose does not hinders recognition or binding by DC-SIGN.Only the Man3 and α(1-3)-Fuc chitobiose glycoblocks from the whole N-glycan is represented. The monosaccharides colouring, aside from panel a), follows the SFNG nomenclature. The structure rendering was done with VMD and the graphical statistical analysis with RStudio (https://www.rstudio.com).

.

B   Supporting Information for Chapter 4

B.1   Oligomannose Isomers
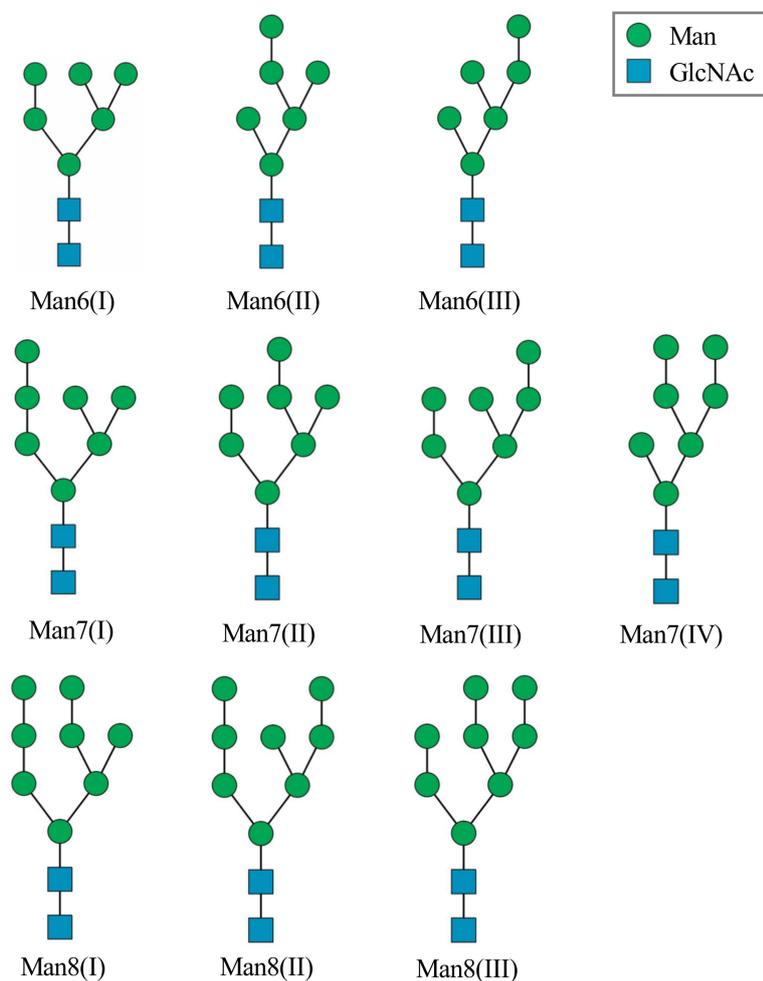
Figure B.1: SNFG representation of all the Man6-8 oligomannose positional isomers studied in this work in addition to Man5/9

B.2   Free Glycan Tables and Figures.

Man 5
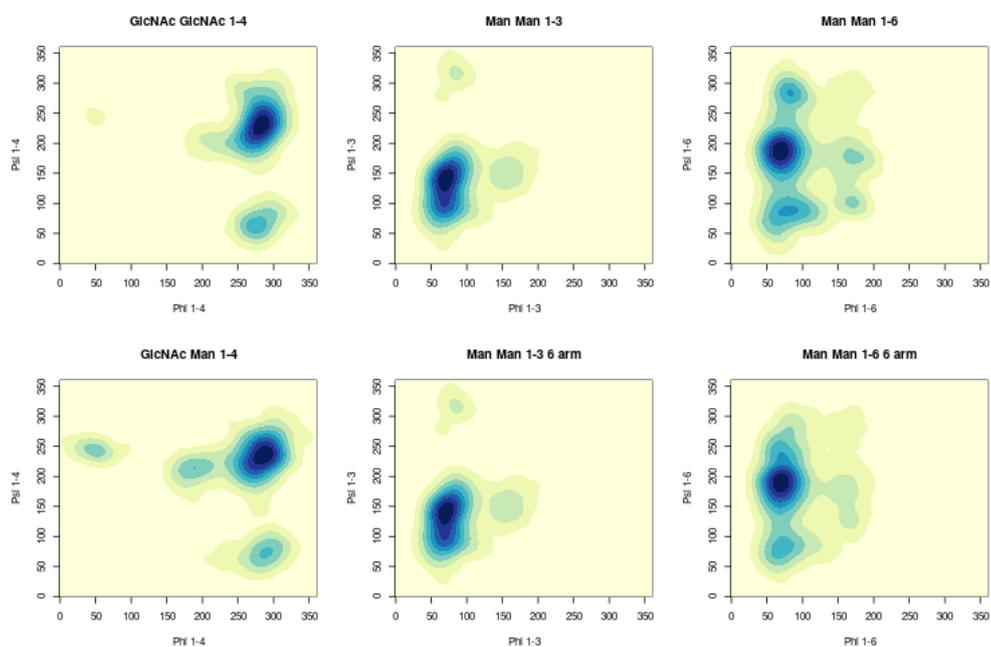
Figure B.2: 2D Kernel density estimates for the ϕ / ψ angle distributions for Man 5.

Table B.1: The ϕ / ψ / ω angle distributions for Man 5.

GlcNAc β(1-4) GlcNAc	ϕ	ψ	ω	Pop (%)
Cluster 1	-78.1 (11.2)	-130.2 (18.3)	-	94.6
Cluster 2	-81.9 (11.4)	65.5 (11.2)	-	5.4
Man β(1-4) GlcNAc	ϕ	ψ	ω	Pop (%)
Cluster 1	-76.0 (12.9)	-125.9 (15.6)	-	93.0
Cluster 2	-70.5 (10.7)	72.9 (10.9)	-	3.5
Cluster 3	-170.7 (12.3)	-146.8 (7.83)	-	2.5
Cluster 4	48.6 (12.9)	-116.3 (5.5)	-	1.0
Man α(1-3) Man (1-3)	ϕ	ψ	ω	Pop (%)
Cluster 1	71.7 (8.1)	141.6 (14.5)	-	73.9
Cluster 2	68.8 (9.4)	99.7 (10.5)	-	26.1
Man α(1-3) Man (1-6)	ϕ	ψ	ω	Pop (%)
Cluster 1	72.1 (9.3)	138.6 (15.1)	-	62.5
Cluster 2	67.6 (9.9)	99.9 (10.4)	-	37.5
Man α(1-6) Man	ϕ	ψ	ω	Pop (%)
Cluster 1	71.1 (10.7)	-172.5 (17.0)	56.1 (10.8)	48.6
Cluster 2	67.8 (10.0)	-175.4 (14.0)	-175.4 (12.7)	33.3
Cluster 3	79.2 (16.4)	86.7 (13.4)	50.9 (10.0)	12.4
Cluster 4	82.6 (8.6)	-76.5 (10.8)	-150.2 (10.5)	5.7
Man α(1-6) Man (1-6)	ϕ	ψ	ω	Pop (%)
Cluster 1	70.3 (10.3)	-171.2 (15.9)	54.8 (10.5)	80.4
Cluster 2	69.6 (8.5)	-173.3 (13.4)	-80.7 (12.6)	5.7
Cluster 3	69.6 (7.6)	-120.2 (13.6)	-64.3 (10.5)	5.6
Cluster 4	69.7 (7.6)	-168.7 (17.5)	-164.4(9.4)	4.4
Cluster 5	70.9 (10.5)	82.9 (11.6)	47.9 (8.8)	3.9

Man 6 I
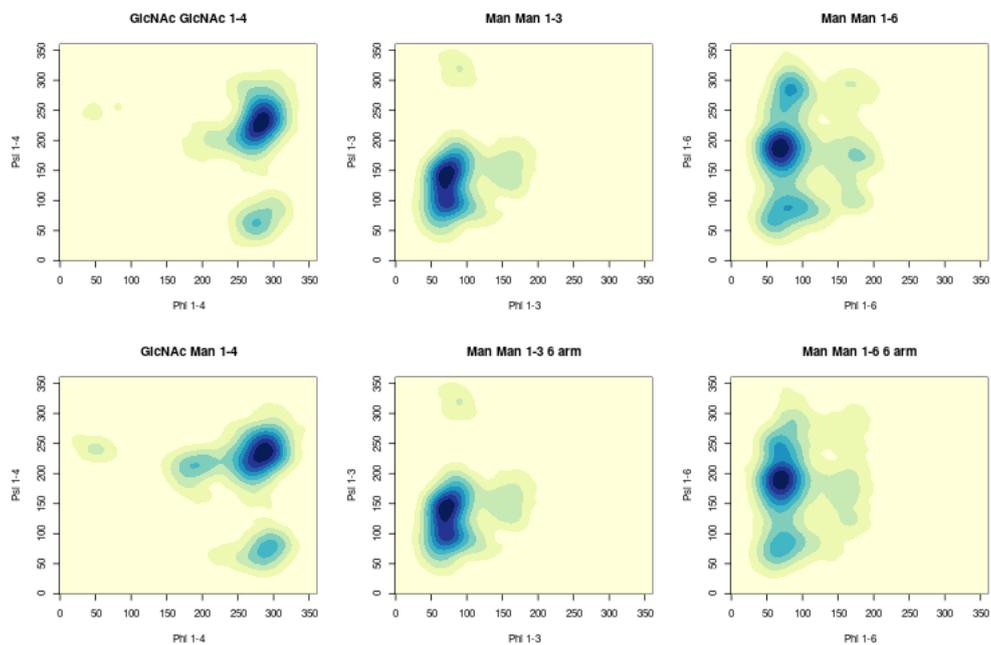
Figure B.3: 2D Kernel density estimates for the ϕ / ψ angle distributions for Man 6 I.

Table B.2: The ϕ / ψ / ω angle distributions for Man 6 I.

GlcNAc β(1-4) GlcNAc	ϕ	ψ	ω	Pop (%)
Cluster 1	-78.0 (10.9)	-130.3 (17.4)	-	96.7
Cluster 2	-82.5 (10.7)	64.7 (11.1)	-	3.3
Man β(1-4) GlcNAc	ϕ	ψ	ω	Pop (%)
Cluster 1	-76.2 (12.9)	-126.4 (15.4)	-	92.4
Cluster 2	-68.5 (10.5)	74.4 (10.94)	-	4.8
Cluster 3	-169.1 (12.1)	-146.2 (8.3)	-	2.8
Man α(1-3) Man (1-3)	ϕ	ψ	ω	Pop (%)
Cluster 1	72.1 (9.1)	142.1 (14.4)	-	67.8
Cluster 2	71.1 (9.9)	98.5 (10.4)	-	32.2
Man α(1-3) Man (1-6)	ϕ	ψ	ω	Pop (%)
Cluster 1	72.3 (9.2)	138.6 (15.1)	-	62.2
Cluster 2	67.7 (9.7)	99.7 (10.3)	-	37.8
Man α(1-6) Man	ϕ	ψ	ω	Pop (%)
Cluster 1	71.0 (10.7)	-172.2 (15.8)	55.9 (10.8)	41.1
Cluster 2	67.7 (9.9)	-175.1 (14.1)	-175.8 (12.3)	37.5
Cluster 3	76.6 (16.5)	85.1 (14.6)	50.3 (9.9)	10.9
Cluster 4	83.3 (8.5)	-75.3 (11.7)	-149.0 (10.8)	7.1
Cluster 5	70.7 (8.7)	-177.6 (13.6)	-71.9 (10.1)	3.4
Man α(1-6) Man (1-6)	ϕ	ψ	ω	Pop (%)
Cluster 1	70.2 (10.4)	-171.3 (15.7)	54.7 (10.3)	76.0
Cluster 2	69.2 (8.7)	-173.9 (13.6)	-82.4 (13.3)	7.8
Cluster 3	68.7 (7.8)	-119.6 (14.1)	-65.1 (10.8)	8.5
Cluster 4	70.2 (7.8)	-175.5 (17.5)	-163.6 (8.9)	4.4
Cluster 5	71.3 (9.3)	82.0 (10.1)	47.5 (8.0)	3.2
Man α(1-2) Man (1-3)	ϕ	ψ	ω	Pop (%)
Cluster 1	74.3 (8.7)	150.7 (15.1)	-	72.7
Cluster 2	70.0 (9.4)	107.2 (11.6)	-	27.3

Man 6 II
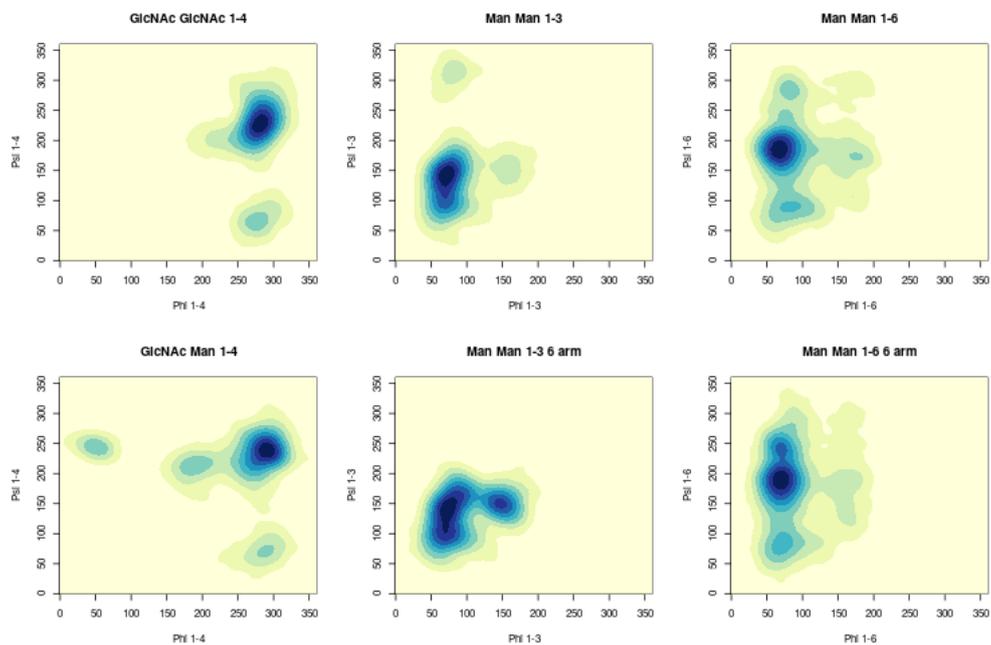
Figure B.4: 2D Kernel density estimates for the ϕ / ψ angle distributions for Man 6 II.

Table B.3: The ϕ / ψ / ω angle distributions for Man 6 II.

GlcNAc β(1-4) GlcNAc	ϕ	ψ	ω	Pop (%)
Cluster 1	-78.9 (10.3)	-131.7 (16.7)	-	98.0
Cluster 2	-83.0 (9.4)	65.8 (9.0)	-	2.0
Man β(1-4) GlcNAc	ϕ	ψ	ω	Pop (%)
Cluster 1	-71.5 (11.8)	-122.8 (13.8)	-	94.2
Cluster 2	-70.3 (7.8)	70.7 (6.5)	-	1.0
Cluster 3	-170.2 (12.0)	-146.2 (8.0)	-	2.8
Cluster 4	50.7 (8.2)	-116.0 (5.7)	-	2.0
Man α(1-3) Man (1-3)	ϕ	ψ	ω	Pop (%)
Cluster 1	72.3 (9.2)	138.6 (15.1)	-	62.2
Cluster 2	67.7 (9.7)	99.7 (10.3)	-	37.8
Man α(1-3) Man (1-6)	ϕ	ψ	ω	Pop (%)
Cluster 1	76.1 (11.3)	144.6 (16.6)	-	57.0
Cluster 2	70.3 (10.5)	100.3 (9.5)	-	24.8
Cluster 3	147.4 (10.5)	150.2 (9.6)	-	18.3
Man α(1-6) Man	ϕ	ψ	ω	Pop (%)
Cluster 1	71.0 (10.7)	-172.2 (15.8)	56.2 (11.4)	58.2
Cluster 2	65.4 (16.3)	-178.9 (11.8)	-177.4 (11.2)	33.9
Cluster 3	82.1 (16.4)	89.2 (13.7)	51.8 (9.9)	6.3
Cluster 4	81.7 (7.4)	-75.6 (9.1)	-149.2 (9.2)	1.6
Man α(1-6) Man (1-6)	ϕ	ψ	ω	Pop (%)
Cluster 1	70.3 (10.5)	-171.4 (16.2)	54.7 (10.6)	70.28
Cluster 2	71.3(7.1)	179.1 (10.22)	-72.0 (8.3)	3.9
Cluster 3	69.5 (7.2)	-119.1 (14.0)	-67.1 (9.5)	14.5
Cluster 4	70.7 (9.7)	-173.3 (23.8)	-164.6 (13.2)	6.9
Cluster 5	71.5 (12.0)	84.2 (13.6)	49.0 (10.2)	4.3
Man α(1-2) Man (1-3)	ϕ	ψ	ω	Pop (%)
Cluster 1	72.3 (8.0)	149.3 (13.4)	-	74.8
Cluster 2	69.6 (7.9)	112.5 (10.1)	-	25.2

Man 6 III
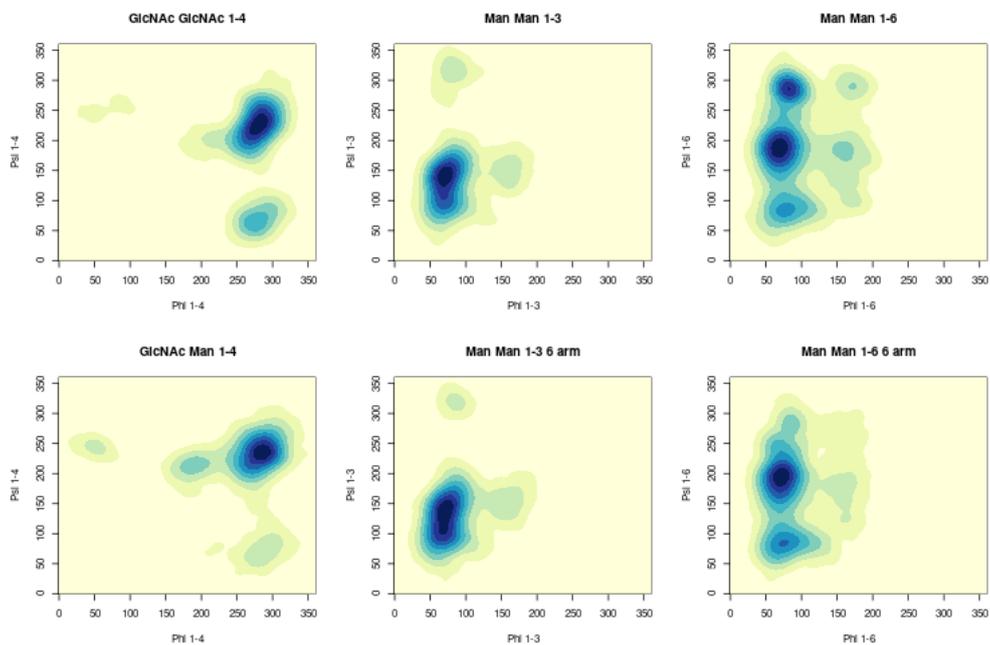
Figure B.5: 2D Kernel density estimates for the ϕ / ψ angle distributions for Man 6 III.

Table B.4: The ϕ / ψ / ω angle distributions for Man 6 III.

GlcNAc β(1-4) GlcNAc	ϕ	ψ	ω	Pop (%)
Cluster 1	-78.8 (11.4)	-133.8 (17.5)	-	93.5
Cluster 2	-82.0 (11.8)	66.0 (11.7)	-	6.5
Man β(1-4) GlcNAc	ϕ	ψ	ω	**Pop (%)**
Cluster 1	- 76.3 (12.6)	-125.5 (14.4)	-	97.9
Cluster 2	-170.4 (10.9)	-146.2 (7.5)	-	2.1
Man α(1-3) Man (1-3)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	76.1 (11.3)	144.6 (16.6)	-	57.0
Cluster 2	70.3 (10.5)	100.3 (9.5)	-	24.8
Man α(1-3) Man (1-6)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	72.3 (9.2)	138.6 (15.1)	-	62.2
Cluster 2	67.7 (9.7)	99.7 (10.3)	-	37.8
Man α(1-6) Man	ϕ	ψ	ω	**Pop (%)**
Cluster 1	71.2 (10.7)	-172.9 (17.2)	56.0 (10.7)	36.1
Cluster 2	67.5 (10.7)	-174.3 (15.2)	-175.9 (12.4)	28.6
Cluster 3	79.1 (14.8)	86.4 (12.9)	50.1 (10.7)	9.7
Cluster 4	82.3 (7.7)	-74.5 (9.8)	-151.5 (9.7)	24.2
Cluster 5	71.6 (6.9)	-173.2 (11.5)	-69.2 (8.3)	1.4
Man α(1-6) Man (1-6)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	71.9 (10.2)	-166.8 (16.7)	55.1 (10.1)	68.2
Cluster 2	69.2 (8.2)	-173.8 (12.6)	- 83.8 (11.2)	4.0
Cluster 3	68.9 (7.6)	-120.6 (11.9)	- 65.4 (10.0)	4.0
Cluster 4	69.0 (8.9)	-175.7 (23.2)	-165.9 (10.6)	4.5
Cluster 5	78.2 (16.7)	86.2 (10.1)	49.8 (10.1)	12.3
Cluster 6	71.9 (10.4)	82.6 (12.5)	-166.3 (17.7)	6.1
Cluster 7	84.9 (5.7)	- 78.2 (7.5)	- 79.57 (6.4)	1.0
Man α(1-2) Man (1-6)(1-6)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	74.7 (8.6)	152.1 (15.1)	-	79.1
Cluster 2	71.3 (9.6)	106.4 (12.2)	-	20.9

Man 7 I
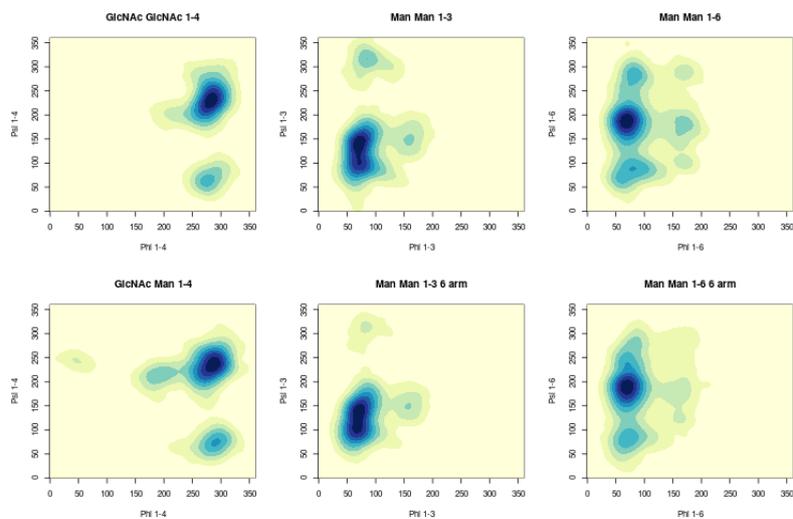
Figure B.6: 2D Kernel density estimates for the ϕ / ψ angle distributions for Man 7 I.

Table B.5: The ϕ / ψ / ω angle distributions for Man 7 I.

GlcNAc β(1-4) GlcNAc	ϕ	ψ	ω	Pop (%)
Cluster 1	-77.9 (10.9)	-130.2 (17.5)	-	95.2
Cluster 2	-82.1 (10.8)	65.5 (11.4)	-	4.8
Man β(1-4) GlcNAc	ϕ	ψ	ω	**Pop (%)**
Cluster 1	-76.6 (12.8)	-125.5 (15.4)	-	89.8
Cluster 2	-69.3 (10.9)	73.5 (11.3)	-	7.7
Cluster 3	-167.7 (12.2)	-145.2 (8.3)	-	2.5
Man α(1-3) Man (1-3)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	72.1 (9.0)	141.5 (14.1)	-	63.3
Cluster 2	71.3 (10.3)	97.6 (10.7)	-	36.7
Man α(1-3) Man (1-6)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	71.9 (9.2)	138.6 (14.8)	-	61.1
Cluster 2	67.5 (9.9)	99.5 (10.4)	-	38.9
Man α(1-6) Man	ϕ	ψ	ω	**Pop (%)**
Cluster 1	70.7 (10.8)	-172.3 (15.8)	55.4 (11.2)	46.2
Cluster 2	68.0 (10.0)	-175.6 (14.1)	-175.6 (12.3)	33.1
Cluster 3	76.6 (16.5)	84.9 (14.7)	49.8 (9.8)	11.7
Cluster 4	82.0 (8.3)	-76.0 (10.9)	-148.8 (10.6)	4.6
Cluster 5	71.5 (9.1)	-178.9 (14.6)	-71.8 (11.5)	4.4
Man α(1-6) Man (1-6)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	70.2 (10.4)	-171.5 (15.4)	54.4 (10.4)	78.1
Cluster 2	69.1 (8.3)	-175.9 (12.8)	-79.3 (12.2)	5.7
Cluster 3	69.2 (7.8)	-118.2 (13.2)	-65.1 (10.1)	5.4
Cluster 4	69.8 (8.5)	-172.0 (18.5)	-165.6 (10.0)	7.4
Cluster 5	71.3 (9.5)	84.0 (10.3)	47.7 (8.0)	3.4
Man α(1-2) Man (1-3)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	74.6 (9.0)	152.5 (15.4)	-	72.0
Cluster 2	70.0 (9.7)	105.3 (11.6)	-	28.0
Man α(1-2) Man (1-3)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	74.1 (8.8)	151.4 (15.0)	-	76.0
Cluster 2	70.0 (9.3)	106.5 (11.6)	-	24.0

Man 7 II
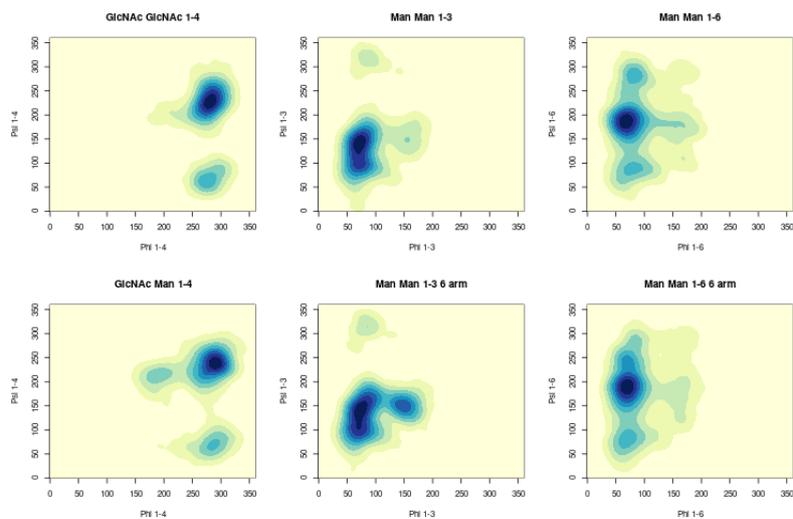
Figure B.7: 2D Kernel density estimates for the ϕ / ψ angle distributions for Man 7 II.

Table B.6: The ϕ / ψ / ω angle distributions for Man 7 II.

GlcNAc β(1-4) GlcNAc	ϕ	ψ	ω	Pop (%)
Cluster 1	-78.8 (10.2)	-131.6 (16.0)	-	92.8
Cluster 2	-81.8 (11.8)	65.5 (11.4)	-	7.2
Man β(1-4) GlcNAc	ϕ	ψ	ω	Pop (%)
Cluster 1	-71.7 (12.8)	-125.5 (15.4)	-	92.3
Cluster 2	-72.5 (11.4)	69.3 (11.6)	-	5.0
Cluster 3	-170.2 (12.3)	-146.7 (8.0)	-	2.7
Man α(1-3) Man (1-3)	ϕ	ψ	ω	Pop (%)
Cluster 1	72.5 (9.2)	142.0 (14.1)	-	67.6
Cluster 2	70.8 (10.2)	98.2 (10.5)	-	32.4
Man α(1-3) Man (1-6)	ϕ	ψ	ω	Pop (%)
Cluster 1	75.9 (11.1)	138.6 (14.8)	-	59.0
Cluster 2	70.4 (10.7)	100.2 (9.6)	-	26.5
Cluster 3	148.9 (10.2)	150.6 (9.7)	-	14.5
Man α(1-6) Man	ϕ	ψ	ω	Pop (%)
Cluster 1	71.1 (12.5)	-171.9 (15.6)	56.1 (11.9)	55.5
Cluster 2	66.7 (10.2)	-177.4 (13.0)	-175.7 (12.4)	31.7
Cluster 3	79.6 (15.6)	87.9 (13.3)	50.1 (9.7)	6.5
Cluster 4	82.4 (8.8)	- 77.6 (10.8)	-150.0 (10.3)	5.2
Cluster 5	71.6 (7.2)	-177.4 (9.8)	- 71.7 (7.9)	1.2
Man α(1-6) Man (1-6)	ϕ	ψ	ω	Pop (%)
Cluster 1	70.3 (10.5)	-170.8 (16.1)	54.8 (10.5)	71.9
Cluster 2	69.8 (7.8)	-179.8 (14.0)	- 74.4 (9.6)	5.0
Cluster 3	70.0 (7.5)	-117.8 (13.4)	- 67.3 (10.1)	8.5
Cluster 4	70.7 (9.8)	-172.4 (24.5)	-163.9 (12.9)	8.3
Cluster 5	71.0 (12.2)	84.4 (13.6)	48.8 (9.9)	4.4
Cluster 6	67.8 (10.7)	73.6 (12.5)	-177.1 (10.1)	1.9
Man α(1-2) Man (1-3)	ϕ	ψ	ω	Pop (%)
Cluster 1	74.2 (8.7)	150.6 (14.7)	-	72.5
Cluster 2	70.0 (9.2)	107.4 (11.6)	-	27.5
Man α(1-2) Man (1-3)(1-3)	ϕ	ψ	ω	Pop (%)
Cluster 1	72.4 (8.1)	149.5 (13.6)	-	75.9
Cluster 2	69.7 (8.2)	111.7 (10.6)	-	24.1

Man 7 III
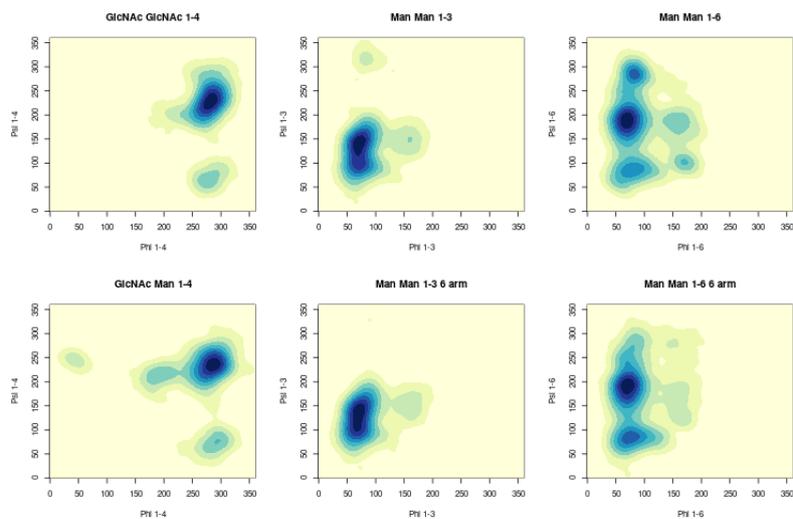
Figure B.8: 2D Kernel density estimates for the ϕ / ψ angle distributions for Man 7 III.

Table B.7: The ϕ / ψ / ω angle distributions for Man 7 III.

GlcNAc β(1-4) GlcNAc	ϕ	ψ	ω	Pop (%)
Cluster 1	-78.9 (11.5)	-131.9 (18.3)	-	97.7
Cluster 2	-82.7 (9.8)	65.3 (9.0)	-	2.3
Man β(1-4) GlcNAc	ϕ	ψ	ω	Pop (%)
Cluster 1	-75.4 (12.7)	-125.3 (14.8)	-	94.0
Cluster 2	-68.4 (10.4)	73.5 (10.4)	-	3.1
Cluster 3	-167.5 (12.4)	-145.6 (8.0)	-	2.9
Man α(1-3) Man (1-3)	ϕ	ψ	ω	Pop (%)
Cluster 1	72.1 (9.0)	141.5 (14.5)	-	66.5
Cluster 2	70.4 (9.9)	98.3 (10.4)	-	33.5
Man α(1-3) Man (1-6)	ϕ	ψ	ω	Pop (%)
Cluster 1	71.8 (9.1)	139.1 (14.6)	-	65.1
Cluster 2	67.8 (9.5)	100.4 (9.8)	-	34.9
Man α(1-6) Man	ϕ	ψ	ω	Pop (%)
Cluster 1	71.1 (10.7)	-172.4 (16.1)	56.0 (10.7)	42.6
Cluster 2	67.2 (9.4)	-175.5 (15.4)	-176.2 (11.7)	20.2
Cluster 3	78.0 (16.2)	85.3 (13.5)	49.5 (10.3)	14.5
Cluster 4	82.4 (7.7)	- 74.3 (9.9)	-151.0 (9.4)	11.3
Cluster 5	73.3 (9.5)	-160.8 (20.5)	- 66.1 (12.0)	7.8
Cluster 6	72.8 (8.6)	- 98.6 (10.0)	- 73.0 (9.1)	1.9
Cluster 7	169.8 (7.8)	102.0 (7.2)	171.2 (7.6)	1.7
Man α(1-6) Man (1-6)	ϕ	ψ	ω	Pop (%)
Cluster 1	71.3 (10.3)	-169.3 (16.5)	55.0 (10.3)	56.7
Cluster 2	69.2 (9.1)	-174.3 (15.0)	- 79.7 (13.6)	6.2
Cluster 3	71.2 (8.9)	-113.5 (17.1)	- 67.5 (11.4)	5.0
Cluster 4	69.4 (9.6)	-178.5 (21.7)	-166.2 (12.1)	9.2
Cluster 5	79.6 (17.5)	85.8 (11.4)	49.8 (9.6)	14.5
Cluster 6	71.9 (11.1)	80.9 (11.2)	-168.0 (14.6)	8.4
Man α(1-2) Man (1-3)	ϕ	ψ	ω	Pop (%)
Cluster 1	74.2 (8.7)	150.4 (14.8)	-	72.1
Cluster 2	69.8 (9.1)	107.5 (11.3)	-	27.9
Man α(1-2) Man (1-6)(1-6)	ϕ	ψ	ω	Pop (%)
Cluster 1	74.5 (8.5)	151.5 (14.5)	-	78.0
Cluster 2	70.9 (9.1)	106.7 (11.8)	-	22.0

Man 7 IV
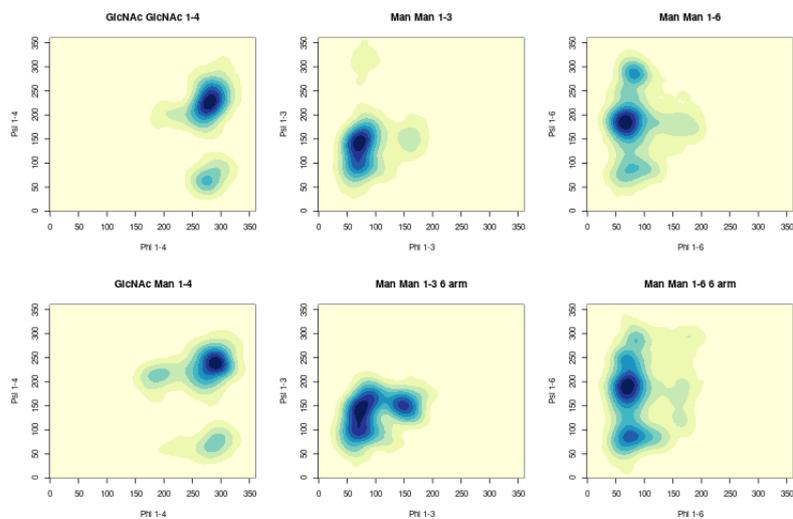
Figure B.9: 2D Kernel density estimates for the ϕ / ψ angle distributions for Man 7 IV.

Table B.8: The ϕ / ψ / ω angle distributions for Man 7 IV.

GlcNAc β(1-4) GlcNAc	ϕ	ψ	ω	Pop (%)
Cluster 1	-79.8 (10.2)	-133.1 (16.5)	-	95.8
Cluster 2	-82.3 (10.6)	65.8 (11.8)	-	4.2
Man β(1-4) GlcNAc	ϕ	ψ	ω	**Pop (%)**
Cluster 1	-71.8 (12.0)	-122.9 (14.1)	-	95.3
Cluster 2	-70.0 (10.7)	72.3 (10.9)	-	2.6
Cluster 3	-169.2 (10.7)	-145.3 (7.7	-	2.1
Man α(1-3) Man (1-3)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	72.9 (9.1)	142.4 (14.1)	-	77.3
Cluster 2	69.3 (9.1)	100.1 (10.3)	-	22.7
Man α(1-3) Man (1-6)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	76.6 (11.8)	145.7 (16.9)	-	56.9
Cluster 2	70.6 (10.7)	100.0 (9.7)	-	23.7
Cluster 3	148.4 (10.7)	151.5 (10.3)	-	19.4
Man α(1-6) Man	ϕ	ψ	ω	**Pop (%)**
Cluster 1	70.5 (12.4)	-172.4 (15.5)	55.7 (11.5)	46.0
Cluster 2	66.3 (9.9)	-178.1 (13.4)	-176.7 (11.7)	40.8
Cluster 3	78.4 (14.7)	87.6 (12.9)	50.3 (9.7)	5.1
Cluster 4	82.0 (7.5)	- 74.7 (10.1)	-151.1 (9.8)	8.1
Man α(1-6) Man (1-6)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	71.1 (10.3)	-170.2 (16.6)	54.9 (10.4)	61.8
Cluster 2	70.2 (7.3)	177.6 (10.8)	- 72.6 (8.8)	3.0
Cluster 3	70.1 (7.5)	-118.6 (12.0)	- 66.9 (9.8)	7.5
Cluster 4	69.9 (9.2)	-179.5 (22.1)	-165.8 (11.2)	5.4
Cluster 5	79.8 (16.5)	86.1 (11.9)	50.5 (9.5)	12.5
Cluster 6	71.6 (10.2)	85.4 (12.6)	-158.7 (21.5)	8.8
Cluster 7	84.9 (5.7)	- 76.2 (7.5)	- 80.6 (6.1)	1.2
Man α(1-2) Man (1-3)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	74.5 (8.8)	151.5 (14.9)	-	77.2
Cluster 2	70.6 (9.3)	106.2 (12.2)	-	22.8
Man α(1-2) Man (1-6)(1-6)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	72.2 (8.0)	149.4 (13.6)	-	74.1
Cluster 2	69.7 (8.1)	112.1 (10.4)	-	25.9

Man 8 I
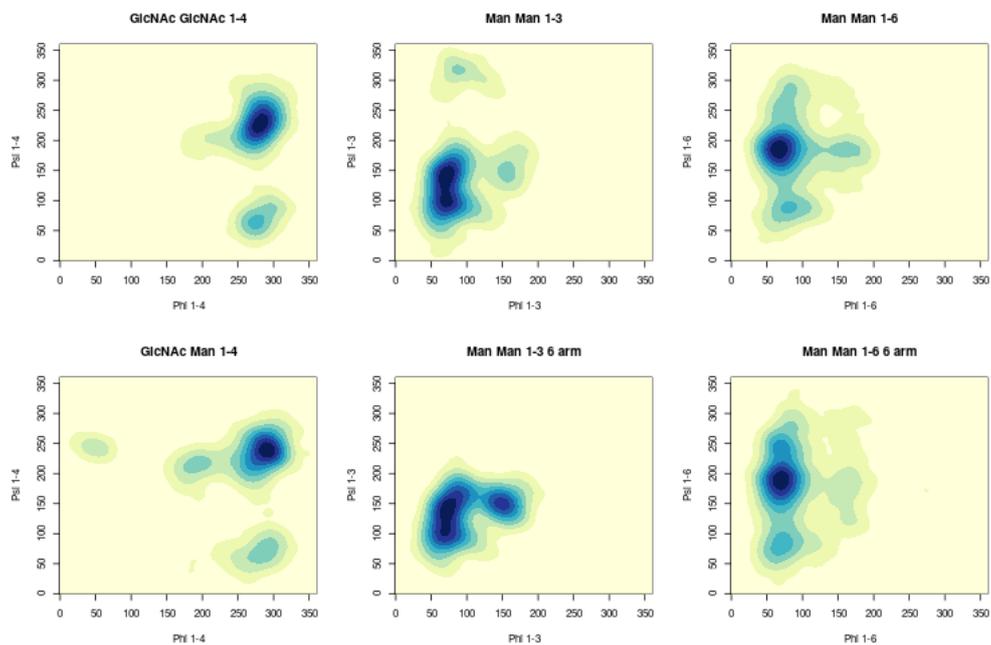
Figure B.10: 2D Kernel density estimates for the ϕ / ψ angle distributions for Man 8 I.

Table B.9: The ϕ / ψ / ω angle distributions for Man 8 I

GlcNAc β(1-4) GlcNAc	ϕ	ψ	ω	Pop (%)
Cluster 1	-78.8 (10.1)	-131.7 (16.2)	-	94.9
Cluster 2	-82.2 (11.5)	66.7 (12.6)	-	5.1
Man β(1-4) GlcNAc	ϕ	ψ	ω	**Pop (%)**
Cluster 1	-71.3 (11.7)	-122.2 (14.1)	-	94.8
Cluster 2	-72.4 (10.9)	70.3 (11.3)	-	3.2
Cluster 3	-167.7 (11.9)	-145.0 (8.1)	-	2.2
Man α(1-3) Man (1-3)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	72.2 (9.0)	141.3 (14.6)	-	58.5
Cluster 2	71.2 (10.4)	96.9 (10.9)	-	41.5
Man α(1-3) Man (1-6)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	75.8 (11.6)	143.9 (16.7)	-	53.5
Cluster 2	70.2 (10.7)	99.9 (9.5)	-	26.5
Cluster 2	70.2 (10.7)	150.2 (10.0)	-	19.9
Man α(1-6) Man	ϕ	ψ	ω	**Pop (%)**
Cluster 1	69.5 (12.3)	-173.0 (15.5)	54.1 (12.6)	55.3
Cluster 2	66.3 (9.8)	-178.0 (12.6)	-176.1 (12.1)	38.2
Cluster 3	81.4 (13.1)	88.3 (11.3)	49.3 (10.3)	4.8
Cluster 4	82.0 (8.3)	-76.0 (10.9)	-148.8 (10.6)	0.0
Cluster 5	75.0 (7.31)	-171.1 (9.5)	-64.8 (8.2)	1.7
Man α(1-6) Man (1-6)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	70.3 (10.4)	-171.5 (15.8)	54.7 (10.7)	74.9
Cluster 2	71.2 (7.4)	-178.9 (14.1)	-73.2 (8.8)	5.4
Cluster 3	69.4 (7.2)	-117.9 (12.3)	-67.0 (9.7)	9.4
Cluster 4	69.8 (9.2)	-175.0 (19.6)	-164.4 (11.3)	6.2
Cluster 5	70.3 (10.6)	83.9 (12.5)	48.0 (9.3)	4.1
Man α(1-2) Man (1-3)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	75.4 (9.0)	151.7 (15.6)	-	70.8
Cluster 2	71.1 (9.7)	105.9 (11.6)	-	29.2
Man α(1-2) Man (1-3)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	73.9 (8.8)	151.2 (15.2)	-	74.8
Cluster 2	70.1 (9.2)	106.6 (11.7)	-	25.2
Man α(1-2) Man (1-6)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	73.9 (8.8)	151.2 (15.2)	-	74.8
Cluster 2	70.1 (9.2)	106.6 (11.7)	-	25.2

Man 8 II

Figure B.11: 2D Kernel density estimates for the ϕ / ψ angle distributions for Man 8 II.

Table B.10: The ϕ / ψ / ω angle distributions for Man 8 II.

GlcNAc β(1-4) GlcNAc	ϕ	ψ	ω	Pop (%)
Cluster 1	-79.4 (11.4)	-132.7 (17.5)	-	95.4
Cluster 2	-82.4 (11.2)	65.6 (11.0)	-	4.6
Man β(1-4) GlcNAc	ϕ	ψ	ω	Pop (%)
Cluster 1	-75.1 (12.4)	-125.0 (14.9)	-	93.9
Cluster 2	-72.8 (8.2)	69.6 (8.0)	-	1.1
Cluster 3	-167.9 (12.5)	-145.3 (8.2)	-	3.3
Cluster 4	47.9 (8.5)	-117.23 (5.7)	-	1.7
Man α(1-3) Man (1-3)	ϕ	ψ	ω	Pop (%)
Cluster 1	72.2 (9.1)	141.7 (14.6)	-	61.0
Cluster 2	71.0 (10.4)	97.0 (10.8)	-	39.0
Man α(1-3) Man (1-6)	ϕ	ψ	ω	Pop (%)
Cluster 1	72.1 (9.2)	139.7 (14.8)	-	64.7
Cluster 2	67.4 (9.6)	100.2 (10.3)	-	35.3
Man α(1-6) Man	ϕ	ψ	ω	Pop (%)
Cluster 1	71.1 (10.8)	-171.6 (16.3)	55.4 (11.2)	47.5
Cluster 2	67.4 (9.9)	-173.9 (14.9)	-175.6 (12.0)	18.0
Cluster 3	79.6 (15.1)	85.7 (13.3)	48.9 (10.4)	11.8
Cluster 4	81.9 (7.6)	- 74.7 (9.6)	-150.7 (9.4)	16.6
Cluster 5	72.6 (8.9)	-164.8 (19.3)	- 67.6 (11.7)	6.0
Man α(1-6) Man (1-6)	ϕ	ψ	ω	Pop (%)
Cluster 1	71.4 (10.2)	-168.5 (16.4)	54.8 (10.3)	67.5
Cluster 2	69.4 (8.8)	-175.8 (13.9)	- 80.6 (13.5)	5.0
Cluster 3	69.1 (6.3)	-119.2 (8.8)	- 63.6 (7.6)	2.4
Cluster 4	69.5 (9.2)	-175.5 (21.3)	-166.7 (10.9)	6.7
Cluster 5	79.0 (16.6)	86.2 (11.7)	50.1 (9.8)	13.1
Cluster 6	71.6 (10.6)	81.9 (11.4)	-166.1 (13.5)	5.4
Man α(1-2) Man (1-3)	ϕ	ψ	ω	Pop (%)
Cluster 1	74.7 (9.0)	152.3 (15.5)	-	62.9
Cluster 2	72.2 (10.0)	97.8 (17.8)	-	37.1
Man α(1-2) Man (1-3)	ϕ	ψ	ω	Pop (%)
Cluster 1	74.3 (8.4)	152.1 (15.1)	-	75.6
Cluster 2	70.0 (9.2)	106.6 (11.7)	-	24.4
Man α(1-2) Man (1-6)(1-6)	ϕ	ψ	ω	Pop (%)
Cluster 1	74.3 (8.8)	152.1 (15.2)	-	75.6
Cluster 2	70.0 (9.2)	106.7 (11.7)	-	24.4

Man 8 III
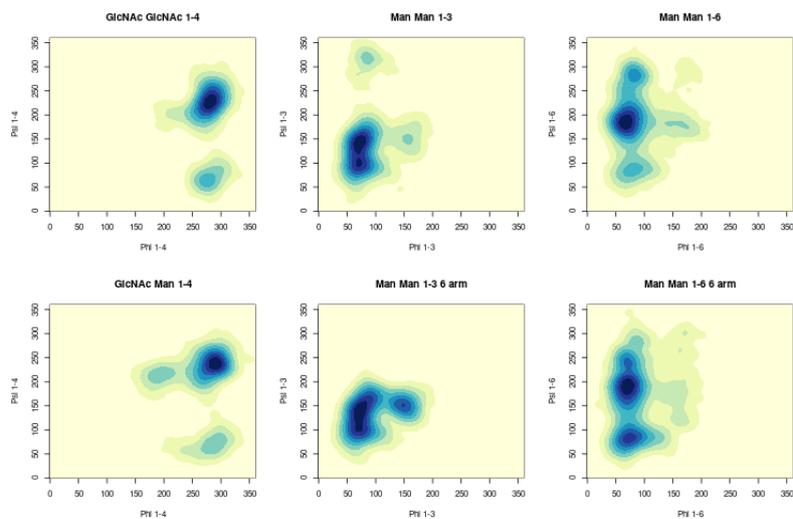
Figure B.12: 2D Kernel density estimates for the ϕ / ψ angle distributions for Man 8 III.

Table B.11: The ϕ / ψ / ω angle distributions for Man 8 III.

GlcNAc β(1-4) GlcNAc	ϕ	ψ	ω	Pop (%)
Cluster 1	-79.3 (10.5)	-132.7 (16.4)	-	94.0
Cluster 2	-82.3 (10.6)	65.9 (11.1)	-	6.0
Man β(1-4) GlcNAc	ϕ	ψ	ω	Pop (%)
Cluster 1	-71.8 (11.9)	-123.4 (13.9)	-	94.0
Cluster 2	-71.6 (10.3)	69.8 (10.5)	-	3.2
Cluster 3	-167.7 (11.2)	-145.4 (7.7)	-	2.7
Man α(1-3) Man (1-3)	ϕ	ψ	ω	Pop (%)
Cluster 1	72.4 (9.0)	142.1 (14.1)	-	61.2
Cluster 2	71.8 (9.8)	97.3 (10.2)	-	37.8
Man α(1-3) Man (1-6)	ϕ	ψ	ω	Pop (%)
Cluster 1	75.4 (11.1)	143.5 (16.2)	-	58.3
Cluster 2	71.1 (10.2)	100.6 (9.1)	-	27.7
Cluster 3	148.3 (9.3)	151.5 (8.7)	-	14.1
Man α(1-6) Man	ϕ	ψ	ω	Pop (%)
Cluster 1	70.6 (11.7)	-172.7 (14.8)	56.2 (11.7)	36.7
Cluster 2	66.3 (9.8)	-178.4 (12.7)	-176.7 (11.7)	35.9
Cluster 3	80.2 (11.7)	86.2 (9.7)	48.0 (9.3)	4.3
Cluster 4	82.9 (8.0)	- 76.2 (9.39)	-150.8 (9.31)	8.1
Cluster 5	74.5 (9.7)	-152.8 (16.7)	- 64.6 (11.8)	13.0
Cluster 6	72.4 (7.9)	-100.0 (8.2)	- 72.1 (8.0)	2.2
Man α(1-6) Man (1-6)	ϕ	ψ	ω	Pop (%)
Cluster 1	71.2 (10.1)	-170.4 (16.0)	55.1 (10.3)	50.0
Cluster 2	70.2 (8.5)	-178.3 (14.4)	- 76.2 (10.7)	5.9
Cluster 3	69.9 (6.8)	-117.7 (10.2)	- 66.0 (8.5)	8.2
Cluster 4	69.5 (8.4)	179.8 (16.4)	-166.4 (9.6)	5.9
Cluster 5	79.4 (15.8)	85.4 (11.0)	49.5 (9.4)	12.0
Cluster 6	71.7 (11.3)	81.4 (11.8)	-166.5 (14.2)	18.0
Man α(1-2) Man (1-3)	ϕ	ψ	ω	Pop (%)
Cluster 1	73.7 (8.6)	149.5 (14.3)	-	70.4
Cluster 2	69.7(8.9)	107.5 (11.3)	-	29.5
Man α(1-2) Man (1-6)(1-3)	ϕ	ψ	ω	Pop (%)
Cluster 1	72.3 (7.9)	149.5 (13.3)	-	72.2
Cluster 2	69.9 (8.1)	111.1 (10.3)	-	27.8
Man α(1-2) Man (1-6)(1-6)	ϕ	ψ	ω	Pop (%)
Cluster 1	72.3 (7.9)	149.5 (13.3)	-	72.2
Cluster 2	69.1 (8.1)	111.0 (10.4)	-	27.8

Man 9
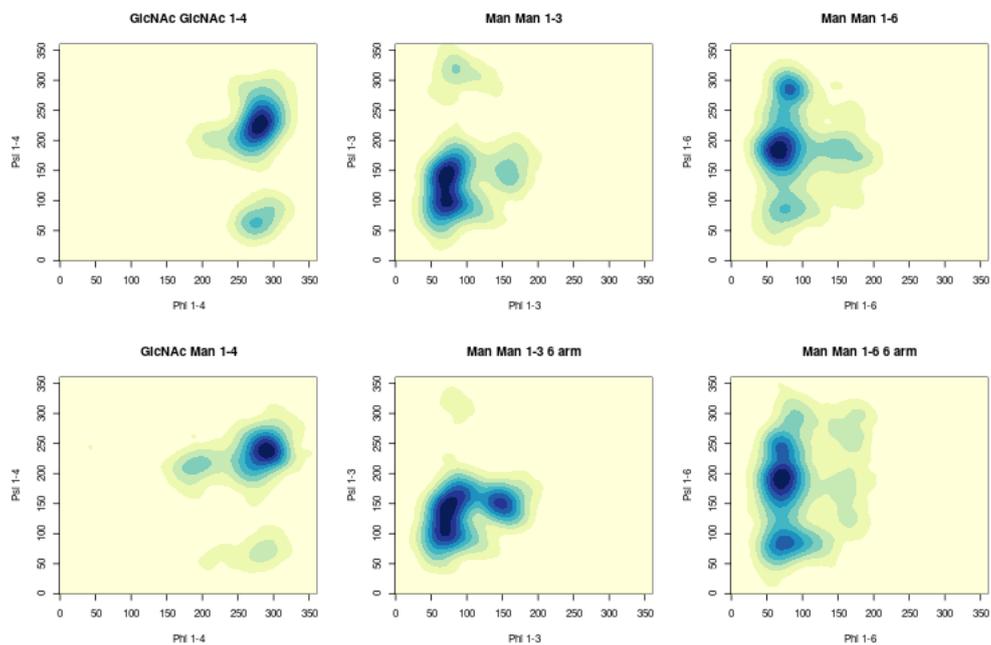
Figure B.13: 2D Kernel density estimates for the ϕ / ψ angle distributions for Man 9.

Table B.12: The ϕ / ψ / ω angle distributions for Man 9

GlcNAc β(1-4) GlcNAc	ϕ	ψ	ω	Pop (%)
Cluster 1	-78.8 (10.5)	-131.7 (17.8)	-	95.8
Cluster 2	-82.3 (11.8)	65.9 (11.5)	-	4.2
Man β(1-4) GlcNAc	ϕ	ψ	ω	**Pop (%)**
Cluster 1	-71.8 (11.7)	-123.1 (13.5)	-	97.5
Cluster 3	-168.4 (12.5)	-146.1 (8.7)	-	2.5
Man α(1-3) Man (1-3)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	72.3 (9.2)	141.5 (14.7)	-	58.5
Cluster 2	71.4 (10.8)	96.7 (11.1)	-	41.5
Man α(1-3) Man (1-6)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	76.1 (12.1)	144.8 (16.9)	-	55.4
Cluster 2	69.8 (10.9)	100.2 (9.7)	-	26.4
Cluster 2	147.8 (10.9)	150.8 (10.4)	-	18.2
Man α(1-6) Man	ϕ	ψ	ω	**Pop (%)**
Cluster 1	70.3 (12.6)	-173.3 (15.5)	55.1 (12.4)	38.0
Cluster 2	66.0 (9.8)	-178.6 (13.0)	-176.9 (11.6)	37.2
Cluster 3	77.9 (11.2)	86.9 (10.5)	46.2 (9.3)	2.8
Cluster 4	82.0 (7.5)	-74.4 (9.6)	-151.1 (9.6)	11.9
Cluster 5	73.2 (9.4)	-157.3 (20.6)	-65.4 (11.3)	7.0
Cluster 6	75.9 (8.6)	-95.6 (9.7)	-74.8 (8.9)	1.8
Cluster 7	147.3 (9.3)	-171.6 (7.8)	-151.1 (7.4)	1.3
Man α(1-6) Man (1-6)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	71.4 (10.2)	-168.8 (16.6)	57.6 (10.3)	56.9
Cluster 2	71.2 (6.7)	-180.0 (14.6)	-73.2 (7.8)	3.8
Cluster 3	70.2 (7.4)	-118.1 (11.7)	-67.4 (10.0)	11.6
Cluster 4	69.2 (9.5)	-179.3 (22.9)	-165.7 (11.4)	7.5
Cluster 5	78.8 (15.92)	86.3 (11.8)	50.8 (9.65)	4.1
Cluster 6	72.6 (11.8)	82.9 (11.8)	-163.9 (17.6)	10.0
Man α(1-2) Man (1-3)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	75.4 (9.2)	151.8 (15.7)	-	69.6
Cluster 2	71.1 (9.1)	105.9 (11.5)	-	30.4
Man α(1-2) Man (1-3)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	74.1 (8.8)	151.7 (15.4)	-	74.7
Cluster 2	70.1 (9.6)	106.3 (11.8)	-	25.3
Man α(1-2) Man (1-6)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	74.1 (8.8)	151.8 (15.4)	-	74.7
Cluster 2	70.1 (9.6)	106.3 (11.7)	-	25.3
Man α(1-2) Man (1-6)(1-6)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	74.4 (8.8)	151.2 (15.1)	-	76.5
Cluster 2	70.5 (9.5)	106.6 (12.2)	-	23.5

B.3   FcγRC Tables and Figures.

FcγRC: Man5 N45
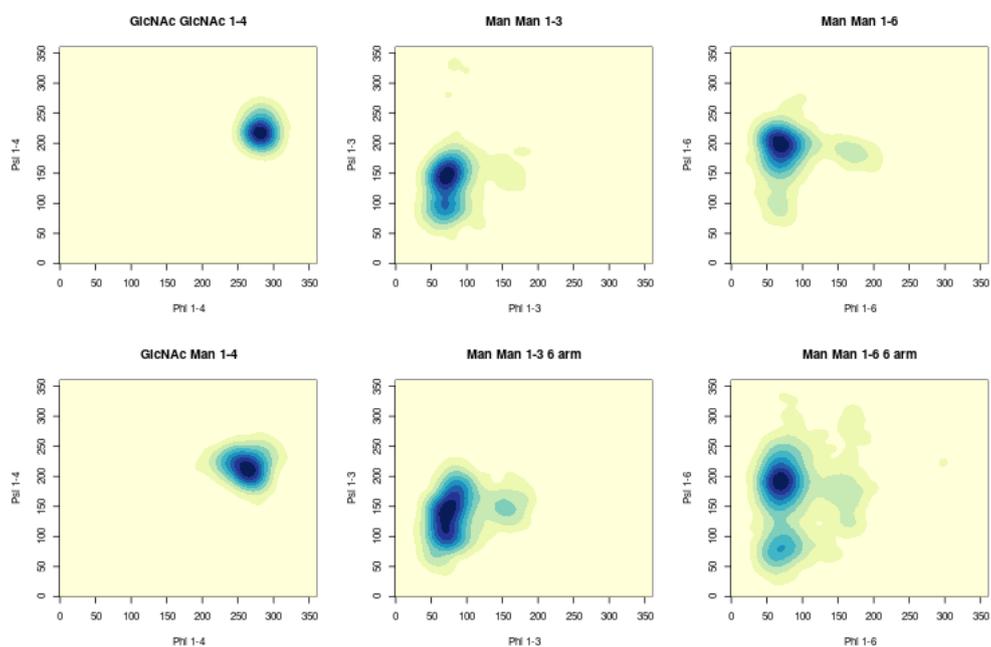
Figure B.14: 2D Kernel density estimates for the ϕ / ψ angle distributions for Man 5 (N45).

Table B.13: The ϕ / ψ / ω angle distributions for Man 5 on Fcγ-RC N45

GlcNAc β(1-4) GlcNAc	ϕ	ψ	ω	Pop (%)
Cluster 1	-78.62 (6.6)	-142.4 (9.5)	-	100
Man β(1-4) GlcNAc	ϕ	ψ	ω	**Pop (%)**
Cluster 1	-96.9 (10.7)	-146.3 (11.3)	-	93.0
Man α(1-3) Man (1-3)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	72.1 (8.1)	145.9 (12.7)	-	82.9
Cluster 2	68.2 (8.4)	97.2 (10.1)	-	17.1
Man α(1-3) Man (1-6)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	75.4 (9.5)	147.4 (15.0)	-	65.9
Cluster 2	70.6 (8.3)	108.72 (10.6)	-	34.1
Man α(1-6) Man	ϕ	ψ	ω	**Pop (%)**
Cluster 1	69.1 (9.5)	-162.1 (13.9)	-172.2 (10.6)	100
Man α(1-6) Man (1-6)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	69.7 (10.2)	-172.2 (15.1)	53.8 (10.7)	44.2
Cluster 2	69.2 (9.4)	-166.3 (15.3)	-87.0 (15.8)	36.9
Cluster 3	70.6 (10.3)	-155.6 (18.9)	-161.1 (10.6)	12.3
Cluster 4	69.2 (10.6)	79.9 (11.5)	-166.9 (12.5)	6.7

FcγRC: Man5 N162
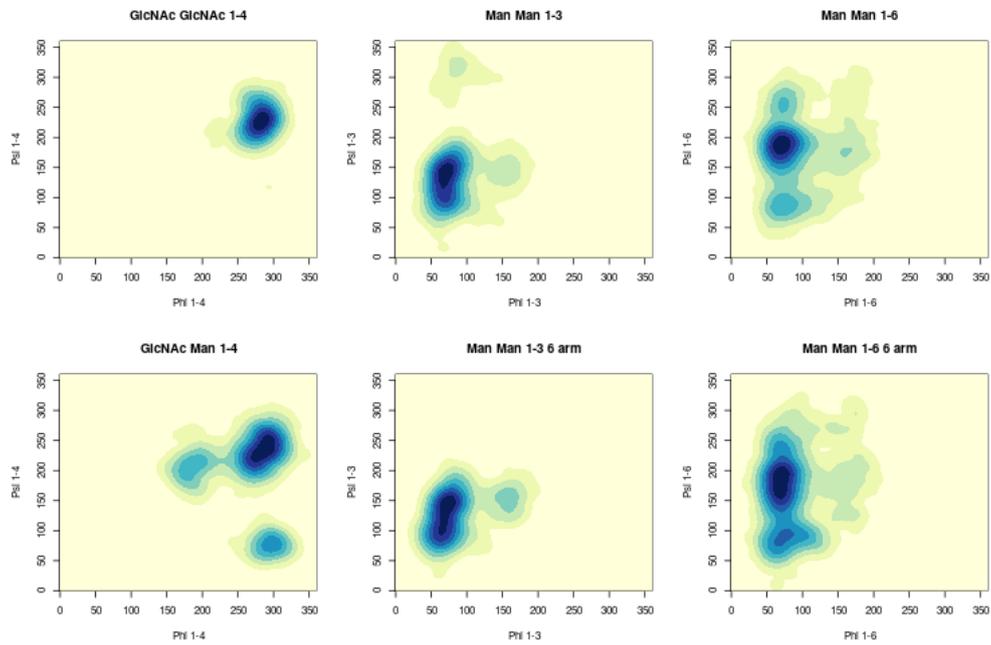
Figure B.15: 2D Kernel density estimates for the ϕ / ψ angle distributions for Man 5 on Fcγ-RC N162.

Table B.14: The ϕ / ψ / ω angle distributions for Man 5 on Fcγ-RC N162

Man α(1-6) Man (1-6)	ϕ	ψ	ω	Pop (%)
Cluster 1	-78.2 (9.3)	-132.8 (14.9)	-	100
Man β(1-4) GlcNAc	ϕ	ψ	ω	**Pop (%)**
Cluster 1	-74.12 (12.9)	-125.7 (17.2)	-	86.7
Cluster 2	-62.9 (9.6)	76.7 (8.9)	-	8.3
Cluster 3	-175.3 (10.8)	-152.2 (11.8)	-	5.0
Man α(1-3) Man (1-3)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	70.9 (8.8)	141.5 (13.9)	-	72.0
Cluster 2	68.7 (8.9)	100.9 (10.0)	-	28.0
Man α(1-3) Man (1-6)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	73.5 (8.9)	144.5 (11.1)	-	52.1
Cluster 2	65.9 (9.4)	101.9 (13.6)	-	47.9
Man α(1-6) Man	ϕ	ψ	ω	**Pop (%)**
Cluster 1	69.1 (11.2)	-175.0 (16.1)	52.3 (11.9)	9.6
Cluster 2	71.6 (11.9)	-172.8 (16.7)	-169.3 (14.9)	82.5
Cluster 3	74.0 (13.3)	87.6,(12.8)	52.6 (11.7)	7.8
Man α(1-6) Man (1-6)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	70.3 (10.1)	179.1 (22.7)	57.4 (11.0)	81.1
Cluster 2	75.1 (15.8)	89.4 (12.6)	53.1 (9.7)	18.9

FcγRC: Man9 N45
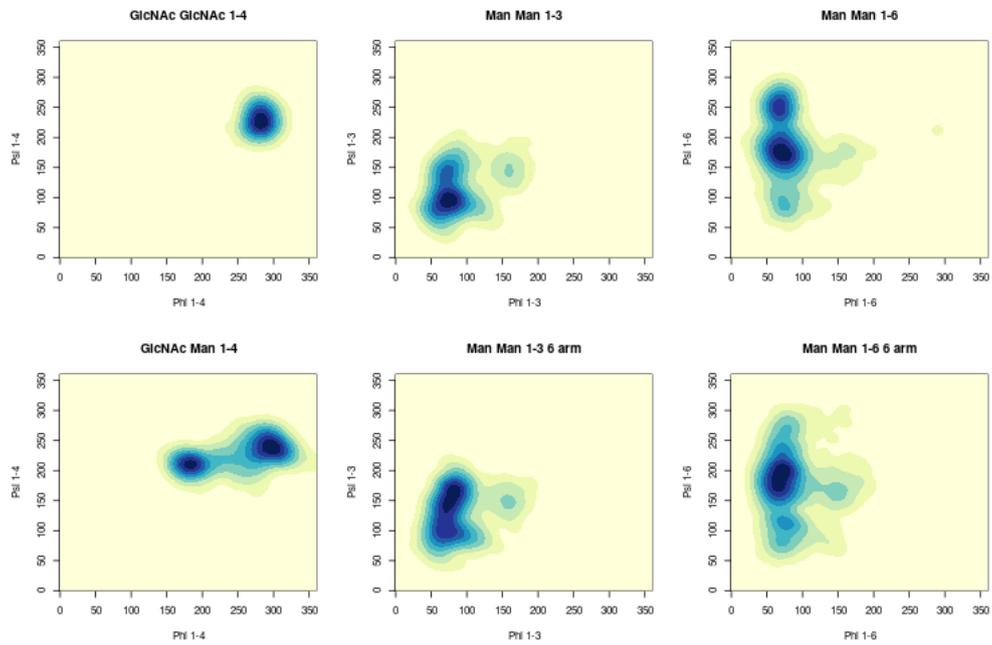
Figure B.16: 2D Kernel density estimates for the ϕ / ψ angle distributions for N45.

Table B.15: The ϕ / ψ / ω angle distributions for Man 9 on Fcγ-RC N45

GlcNAc β(1-4) GlcNAc	ϕ	ψ	ω	Pop (%)
Cluster 1	-77.9 (7.5)	-132.6 (11.8)	-	100
Man β(1-4) GlcNAc	ϕ	ψ	ω	**Pop (%)**
Cluster 1	-65.9 (14.1)	-122.3 (11.6)	-	64.6
Cluster 2	-176.2 (12.7)	-149.6 (7.1)	-	35.4
Man α(1-3) Man (1-3)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	72.2 (7.9)	138.1 (12.7)	-	23.1
Cluster 2	73.7 (11.4)	93.4 (11.1)	-	76.9
Man α(1-3) Man (1-6)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	78.3 (9.1)	155.1 (15.3)	-	60.2
Cluster 2	70.8 (12.0)	101.0 (12.4)	-	39.8
Man α(1-6) Man	ϕ	ψ	ω	**Pop (%)**
Cluster 1	71.2 (11.5)	173.8 (13.1)	59.4 (15.2)	57.1
Cluster 2	71.8 (8.7)	-173.6 (14.9)	-75.0 (11.3)	15.1
Cluster 3	67.1 (7.2)	-111.6,(12.1)	-69.1 (12.4)	27.9
Man α(1-6) Man (1-6)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	71.0 (9.8)	174.8 (14.8)	- 76.7 (14.1)	17.2
Cluster 2	65.7 (9.5)	-178.4 (22.2)	-170.2 (10.8)	33.9
Cluster 3	70.5 (10.2)	-164.1 (14.2)	56.4 (11.7)	42.9
Cluster 4	76.1 (9.4)	113.3 (8.3)	- 87.5 (7.9)	5.9
Man α(1-2) Man (1-3)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	74.3 (8.4)	148.2 (13.2)	-	61.1
Cluster 2	72.7 (10.1)	105.3 (11.1)	-	38.9
Man α(1-2) Man (1-3)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	73.2 (8.3)	149.3 (14.3)	-	69.1
Cluster 2	70.1 (8.5)	106.0 (10.9)	-	30.9
Man α(1-2) Man (1-6)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	72.2 (7.6)	146.9 (12.4)	-	77.4
Cluster 2	69.9 (7.9)	107.1 (10.9)	-	22.6
Man α(1-2) Man (1-6)(1-6)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	74.5 (9.0)	151.9 (13.6)	-	78.0
Cluster 2	72.2 (8.1)	104.2 (11.3)	-	22.0

FcγRC: Man9 N162
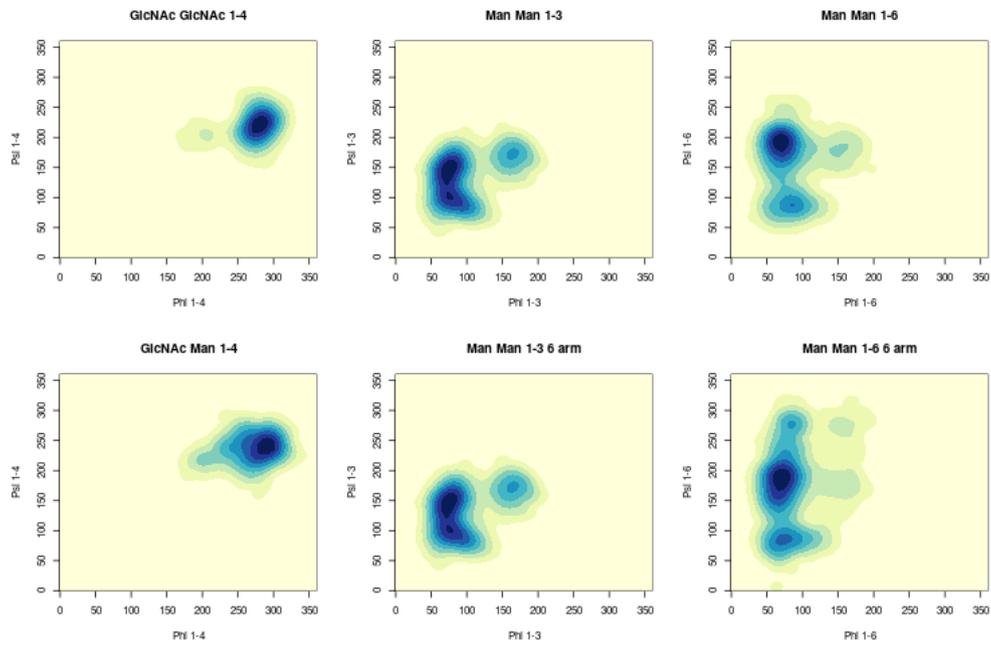
Figure B.17: 2D Kernel density estimates for the ϕ / ψ angle distributions for Man9 on the Fcγ-RC N162

Table B.16: The ϕ / ψ / ω angle distributions for Man 9 on Fcγ-RC N162

GlcNAc β(1-4) GlcNAc	ϕ	ψ	ω	Pop (%)
Cluster 1	-79.8 (9.6)	-138.9 (13.2)	-	100
Man β(1-4) GlcNAc	ϕ	ψ	ω	**Pop (%)**
Cluster 1	-75.1 (17.0)	-120.5 (13.8)	-	100
Man α(1-3) Man (1-3)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	72.0 (9.2)	145.4(15.2)	-	58.7
Cluster 2	79.0 (12.6)	96.5 (10.4)	-	34.6
Cluster 3	163.7 (11.2)	170.9(12.0)	-	6.7
Man α(1-3) Man (1-6)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	77.9 (9.4)	156.2 (17.2)	-	70.8
Cluster 2	70.3 (9.4)	108.8 (10.7)	-	29.2
Man α(1-6) Man	ϕ	ψ	ω	**Pop (%)**
Cluster 1	70.1 (10.6)	-170.6 (16.0)	54.6 (11.1)	82.6
Cluster 2	83.4 (14.3)	88.2 (12.2)	47.8 (9.5)	17.4
Man α(1-6) Man (1-6)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	70.2 (10.9)	-172.8 (18.8)	54.8 (11.2)	52.8
Cluster 2	69.9 (9.7)	176.3 (19.8)	-76.6 (12.8)	9.5
Cluster 3	83.1 (8.0)	-85.2 (10.9)	-80.0 (10.1)	4.7
Cluster 4	74.6 (15.6	84.7 (11.1)	48.8 (10.8)	16.8
Cluster 5	69.9 (10.1)	175.5 (21.1)	-168.1 (14.0)	16.2
Man α(1-2) Man (1-3)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	76.6 (8.3)	156.5 (13.8)	-	77.8
Cluster 2	68.5 (8.1)	100.1 (9.6)	-	22.2
Man α(1-2) Man (1-3)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	74.4 (9.0)	154.8 (16.1)	-	93.9
Cluster 2	66.6 (4.8)	107.4 (5.2)	-	6.1
Man α(1-2) Man (1-6)	ϕ	ψ	ω	**Pop (%)**
Cluster 1	71.8 (7.6)	146.4 (12.9)	-	79.8
Cluster 2	70.0 (7.4)	111.1 (10.1)	-	20.2
Man α(1-2) Man (1-6)(1-6)	ϕ	ω	ψ	**Pop (%)**
Cluster 1	74.7 (8.7)	150.9 (14.9)	-	73.8
Cluster 2	70.4 (8.3)	105.3 (11.6)	-	26.2

B.4   DBSCAN Parameters

Dihedral Angle's	eps (ϵ)	Minimum points
GlcNAc β(1-4) GlcNAc	7.5	75
Man β(1-4) GlcNAc	7.5	100
Man α(1-3) Man (1-3)	10	1000
Man α(1-3) Man (1-6)	10	1000
Man α(1-6) Man	16	400
Man α(1-6) Man (1-6)	16	400
Man α(1-2) Man	10	1000
Man α(1-2) Man (2)	10	1000
Man α(1-2) Man (3)	10	1000
Man α(1-2) Man (4)	10	1000

Table B.17: List of parameters used for the clustering analysis of Man 9. Note: some of these parameters may need to be tweaked as clusters can bleed into each other. This algorithm will take all RAM it's given, best practice to test on smaller data sets before scaling up to production datasets in order to prevent crashes.

B.5 Man 9 / 8(II) Distance Measurements
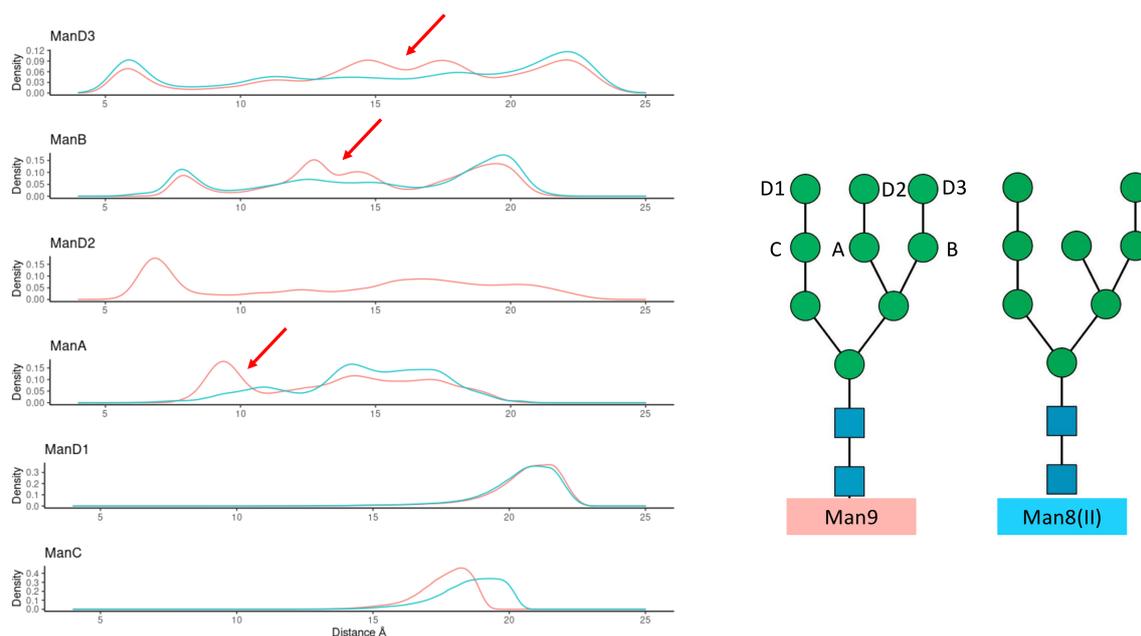
Figure B.18: KDE analysis of the distances between the anomeric protons of the reducing GlcNAc and that of the specified mannose residues in the legend for Man9 (red) and Man8(II) (blue) obtained from our simulations. The red arrows highlight shorter distances only observed in Man9, which indicate the higher occurrence of folded structures, in agreement with a progressively higher occurrence of arm-arm interactions with arm elongation, as described in the main text. KDE analysis made with r and diagrams with RStudio (www.rstudio.com).

C   Supporting Information for Chapter 5

C.1   Computational Method

MD simulations: System set-up and protocol. The SARS-CoV2 S models were generated by homology modelling using SWISS MODEL [10] from the cryo-EM structure 6VYB (3.2 Åresolution) of the SARS-CoV-2 S 2P recombinant glycoprotein, with the reference sequence NCBI YP_009724390.1. The loops not resolved in the cryo-EM structure were reconstructed using the SWISS MODEL structural libraries of backbone fragments, based on similar sequence structure. This resulted in a model with 54 N-glycosylation sites, 18 per protomer. The SARS-CoV2 S model, bearing the N370 sequon, was generated by introducing the A372T mutation. Complete glycoforms were reconstructed at the N-glycosylation sites, by aligning conformationally equilibrated N-glycan structures from our in-house GlycoShape library [11, 12] (under construction) to the GlcNAc residues resolved in the cryo-EM structure, with slight adjustments of the torsion angles to resolve steric clashes with the surrounding protein, when necessary. Selection of the glycan at each N-glycosylation site was determined based primarily on work by Wantanabe *et al* [13] with consideration to other studies available5, 6. O-glycosylation at T323 was chosen based on results from Shajahan *et al* [14] The N-glycosylation at N370 in the SARS-CoV2 S mutant was chosen to be the same as reported for the SARS-CoV-1 S [15] Summary of the specific glycosylation at each site is shown in Table C.1 for all three protomers.

In all MD simulations the protein and counterions (200 mM) were represented by the AMBER ff14SB [16] parameter set, whereas the glycans were represented by the GLYCAM06j-1 version of the GLYCAM06 force field [17]. Water molecules were represented by the TIP3P model. All simulations were run with v18 of the AMBER software package [18] The following running protocol was used for all MD simulations. The energy of the S ectodomains models was minimized in two steps of 50,000 cycles of the steepest descent algorithm each. During the first minimization all the heavy atoms were kept harmonically restrained using a potential weight of 5 $kcal\ mol^{-1}\text{Å}^{-2}$, while the solvent, counterions and hydrogen atoms were left unrestrained. The minimization step was repeated with only the protein heavy atoms were kept restrained, while the glycans, solvent, counterions and hydrogens were left unrestrained. After energy minimization the system was equilibrated in the NVT ensemble with the same restraints scheme, where heating was performed in two stages over a total time of 1 ns, from 0 to 100 K (stage 1) and then 100 to 300 K (stage 2). During equilibration the SHAKE algorithm was used to constrain all bonds to hydrogen atoms. The Van der Waals interactions were truncated at 11 Åand Particle Mesh Ewald (PME) was used to treat long range electrostatics with B-spline interpolation of order 4. Langevin dynamics with collision frequency of 1.0 ps^{-1} was used to control temperature, which a pseudo-random variable seed to ensure there are no synchronization artefacts. Once the system was brought to 300 K an equilibration phase in the NPT ensemble of 1 ns was used to set the pressure to 1 atm. The pressure was held constant with isotropic pressure scaling and a pressure relaxation time of 2.0 ps. At this point all restraints on the protein heavy atoms were removed, allowing the system to evolve for 15 ns of conformational equilibration before production. At this stage different replicas for each model were generated starting from different velocities. The conformational equilibration phase for each replica was further extended to include the first 300 ns of production to allow the glycans shield to adapt to the protein structure and vice-versa. In the analysis this 300 ns initial phase was discarded. The MD simulations were performed on PRACE (www.prace-ri.eu) resources allocated on CINECA Marconi100, using 4 V100 GPUs per replica simulation, with a benchmark standard of approximately 25 ns/day. The total simulation times for each replica, including equilibration time, are shown in Table

Table C.1: . Site specific glycosylation chosen for the SARS-CoV-2 S glycoprotein models, with 18 N-glycan sites and 1 O-glycan site per protomer. Residue numbering (Resid) corresponds to the PDB 6VYB. The glycosylation sites where the glycosylation was changed between models are highlighted in red. N-glycans are represented by the Oxford nomenclature.

Resid / Glycan Type	N234-Man5	N234-Man3	N234-Man9	N234-Man9*	N234-Man9** +N370	N234-Man5*
N61	Man 5	Man 5	Man 5	Man 5	Man 5	Man 5
N74	FA2G2	FA2G2	FA2G2	FA2G2	FA2G2	FA2G2
N122	Man 5	Man 5	Man 5	Man 5	Man 5	Man 5
N149	FA2G2	FA2G2	FA2G2	FA2G2	FA2G2	FA2G2
N165	FA2G2	FA2G2	FA2G2	A2B	A2B	Man 5
N234	Man 5	Man 3	Man 9	Man 9	Man 9	Man 9
N282	FA2G2	FA2G2	FA2G2	FA2G2	FA2G2	FA2G2
T323	Sial T	Sial T	Sial T	Sial T	Sial T	Sial T
N331	FA2G2	FA2G2	FA2G2	FA2G2	FA2G2	FA2G2
N343	FA2G2	FA2G2	FA2G2	FA2B	FA2B	Man 5
N370	*Null*	*Null*	*Null*	*Null*	FA2G2	*Null*
N603	Man 5	Man 5	Man 5	Man 5	Man 5	Man 5
N616	A2G2	A2G2	A2G2	A2G2	A2G2	A2G2
N657	FA2G2	FA2G2	FA2G2	FA2G2	FA2G2	FA2G2
N709	Man 5	Man 5	Man 5	Man 5	Man 5	Man 5
N717	Man 5	Man 5	Man 5	Man 5	Man 5	Man 5
N801	Man 5	Man 5	Man 5	Man 5	Man 5	Man 5
N1074	Man 5	Man 5	Man 5	Man 5	Man 5	Man 5
N1098	A2G2	A2G2	A2G2	A2G2	A2G2	A2G2
N1134	FA2G2	FA2G2	FA2G2	FA2G2	FA2G2	FA2G2

* In this model the N-glycans at N165 and N234 are both Man5
** In this model the N-glycans at N165 and N234 are FA2B and A2B complex, with a bisecting GlcNAc, respectively.

Table C.2: RMSD values (Å) calculated for the RBD residues (330 to 530) in chain A, B and C, defined according to the original PDB used as a template to build the all SARS-CoV2 S models studied in this work. The RBDs in chains A and C are closed, while the RBD in chain B is open. Standard deviation values are shown in parenthesis.

Model System	RBD (A) (Closed)	RBD (B) (Open)	RBD (C) (Closed)
N234-Man5 (R1)	5.8 (0.9)	9.9 (1.1)	4.6 (0.6)
N234-Man5 (R2)	4.9 (0.6)	9.8 (1.3)	6.2 (0.8)
N234-Man5 (R3)	6.5 (0.4)	9.1 (0.8)	4.1 (0.3)
N234-Man3 (R1)	6.6 (0.9)	12.2 (1.6)	4.0 (0.7)
N234-Man3 (R2)	6.2 (0.9)	5.9 (1.3)	4.0 (0.6)
N234-Man3 (R3)	3.2 (0.8)	10.7 (0.4)	5.0 (0.2)
N234-Man9 (R1)	4.6 (0.6)	7.9 (0.7)	4.1 (0.3)
N234-Man9 (R2)	3.4 (0.4)	6.9 (0.5)	4.3 (0.2)
N234-Man9** N370 (R1)	3.1 (0.4)	12.8 (1.2)	3.9 (0.7)
N234-Man9** N370 (R2)	3.1 (0.6)	4.0 (0.8)	2.2 (0.4)
N234-Man9**	2.6 (0.4)	5.3 (0.6)	3.3 (0.5)
N234-Man9**	4.9 (0.4)	8.8 (0.6)	4.3 (0.4)
All Man5 (R1)	5.4 (0.8)	12.0 (1.0)	3.6 (0.6)
All Man5 (R2)	4.9 (0.4)	11.5 (1.6)	3.5 (0.3)
All Man5 (R3)	3.9 (0.9)	9.0 (1.7)	4.2 (0.4)
N234-Man9* (R1)	5.1 (0.6)	9.5 (1.2)	6.1 (0.7)

* In this model the N-glycans at N165 and N234 are both Man5
** In this model the N-glycans at N165 and N234 are FA2B and A2B complex, with a bisecting GlcNAc, respectively.

Figure C.1: Time evolution of the three RBDs backbone root mean square deviation (RMSD) values for six simulations chosen as examples to illustrate our rationale for discarding the first 300 ns (dashed line) of the MD from the analysis, considering it as part of the conformational equilibration stage. Not in every case the 300 ns threshold is probably as necessary, see R1 and R2 N234 Man9** (see Table C.2 footnotes for nomenclature) and N370 (R2), which in these cases is further indication of the glycosylation supporting a "wide open" RBD corresponding to the cryo-EM PDB 6VYB used as a starting structure. Nevertheless, the same threshold was used for all simulations for consistency.
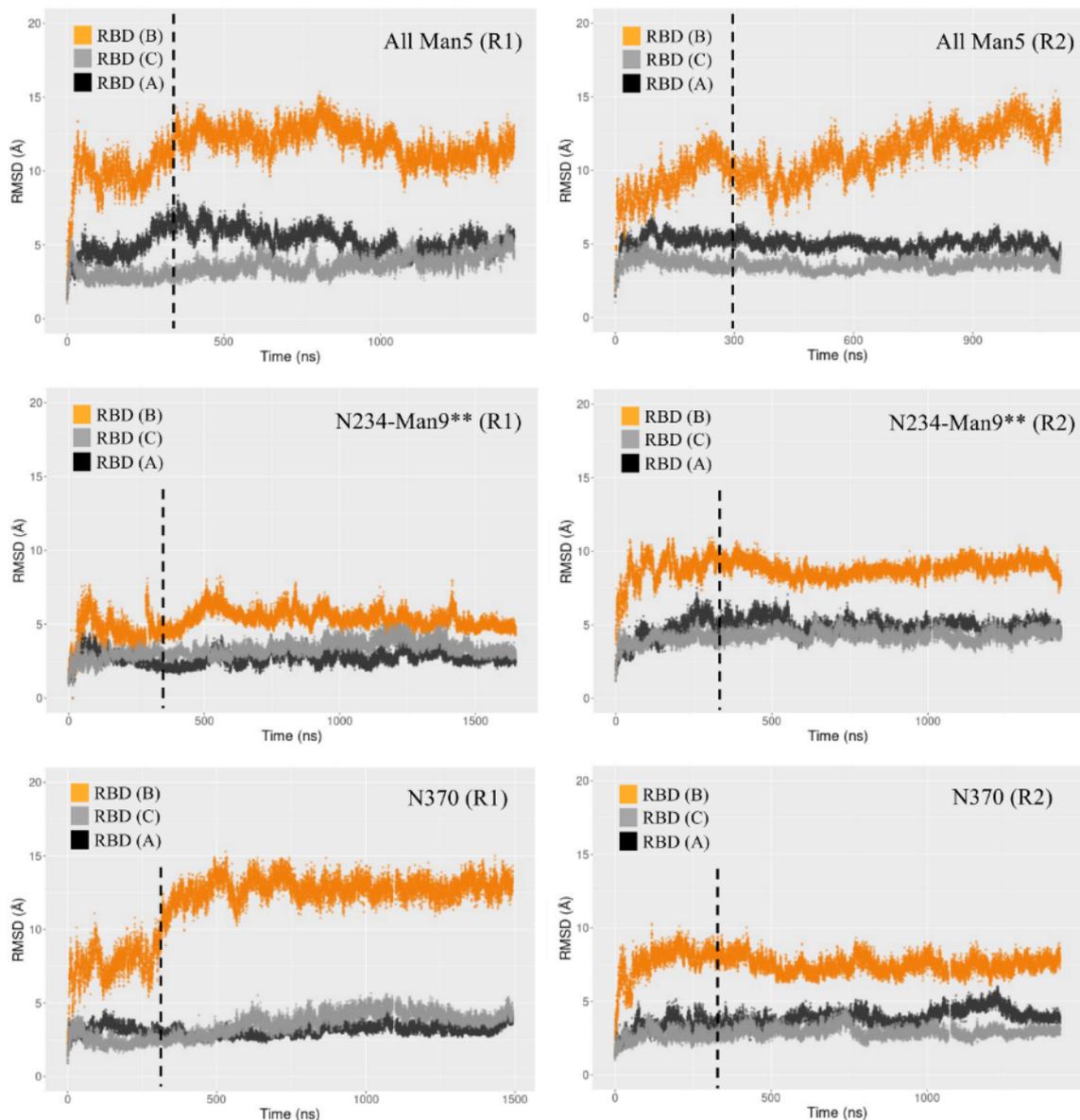
Bioinformatics. We obtained a list of SARS-CoV-2 S protein homologs from UniProt [19] using a blastp search of the SARS-CoV-2 S protein sequence (UniProt Accession P0DTC2) against "Virus" proteins with an E-value threshold of 0.01, auto-selection of matrix, allowing gaps and with a maximum of 1000 hits. We further filtered this list of proteins to remove duplicates and only retain those with "spike" in the name. We aligned these proteins with Clustal Omega v1.2.2 [20, 21] . A phylogenetic tree was created from

the multiple sequence alignment using FastTree v2.1.10 without SSE3, and 1000 as the
bootstrap parameter [22] Ancestral sequence reconstruction was performed based on this
multiple sequence alignment and the associated phylogenetic tree, using CodeML from
the PAML v4.9e with WAG amino acid substitution matrix and molecular clock turning
on [22]. A multiple sequence alignment of reconstructed sequences was parsed from the
CodeML output. The locations of glycosylation sequons (Asn-Xaa-Ser/Thr, Xaa≠Pro)
from the sequences before the alignment were mapped onto the aligned sequence positions
including gaps in the aligned sequences. Using the script *spike_protein.tree_traversal.py*, we
traversed the tree from the root to a specific extant sequence. This script utilized a list of
extant proteins of interest, the multiple sequence alignment before ancestral sequence re-
construction, the multiple sequence alignment after ancestral sequence reconstruction, the
phylogenetic tree including annotation of reconstructed nodes, and the generated dataset
of N-linked sequons with their associated positions in the multiple sequence alignment af-
ter ancestral sequence reconstruction (including ancestral nodes). We recorded the status
of each sequon from the ancestral sequence to each of its direct descendants, including
whether the sequon had been lost, gained, or retained.

C.2   Open RBD range of motions
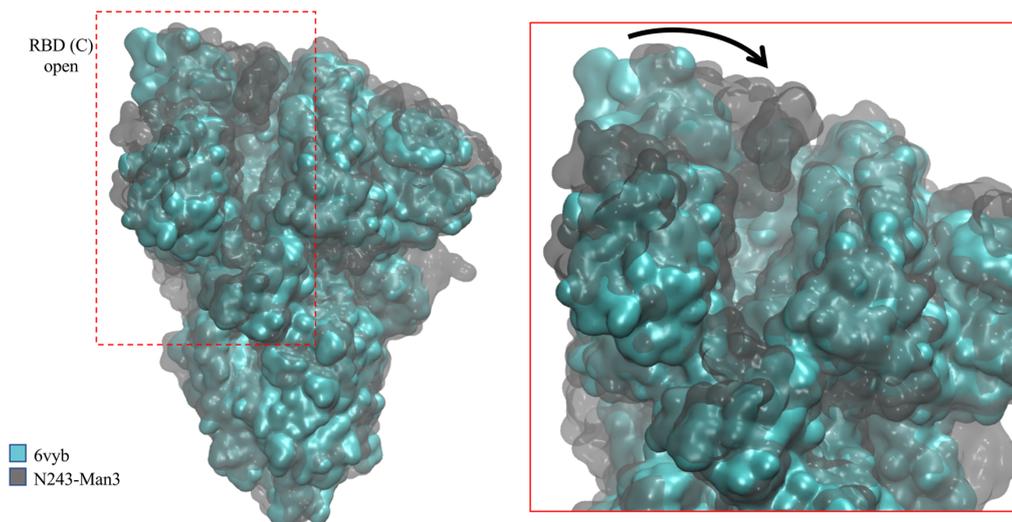
Figure C.2: . Structural alignment (CA backbone) of the SARS-COV2 S cryo-EM
structure PDBid 6vyb and a representative snapshot from the N234-Man3 trajectory
(R1) to illustrate the wide range of motion of the open RBD (chain C in the PDB
numbering) we observed in function of glycosylation. The proteins are represented
by QuickSurf mode with VMD [9], where 6vyb is rendered in cyan (opaque surface)
and N234-Man3 in grey (glass3 surface). The insert on the right-hand side shows a
close up of the section highlighted within the dashed rectangle on the figure, with
an arrow as a visual guide connecting the two orientations. All glycans are omitted
here for clarity.

C.3   N234-Man5
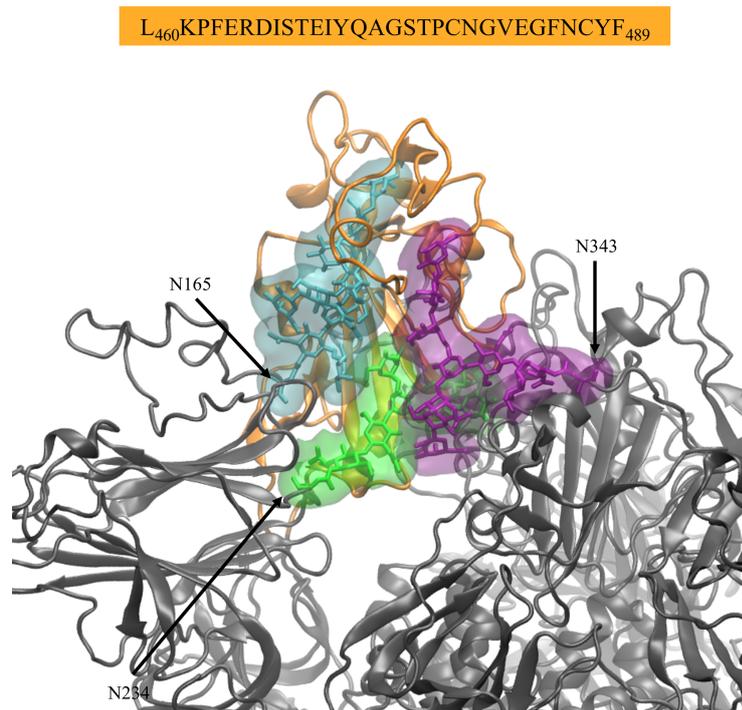
Figure C.3: Close-ups on two representative snapshots of the N234-Man5 simulation R2 illustrating the orientation of the Man5 at N234 and corresponding interactions of the RBD (sequence indicated above) with the N-glycans at N165 and. N343. The Man5 at N234 is shown in green, while the FA2G2 at N165 and N343 are shown in cyan and purple, respectively. The open RBD (B) is shown in orange with a cartoon representation, while the rest of the protein is in grey. The glycans at all other sites are omitted for clarity. Molecular rendering with VMD [9]

C.4   All - Man5
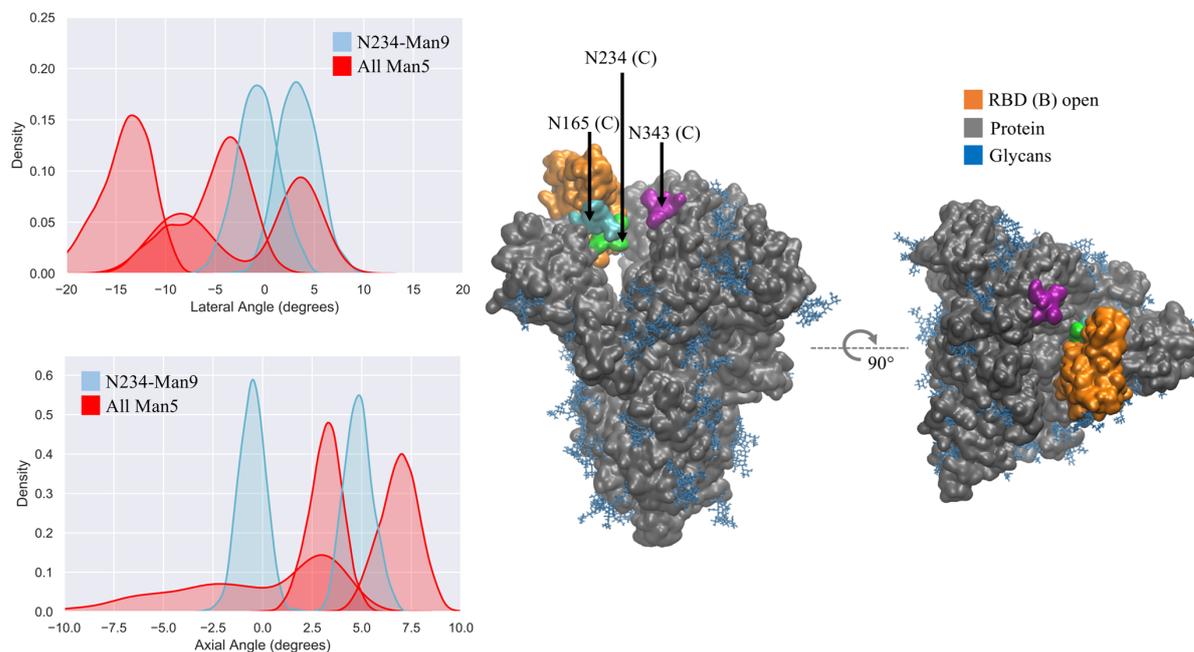
Figure C.4: . Left KDE analysis of the lateral and axial angles distributions calculated through the uncorrelated MD trajectories obtained for All Man5 (red), replicas R1-3, and for N234-Man9 (cyan), replicas R1,2 for comparison. Right Close-ups on a representative snapshot of the All Man5 (R3) simulation from the side and top, respectively. Man5 at N234 is shown in green, while the Man5 at N165 and N343 are shown in cyan and purple, respectively. The solvent accessible surface of the open RBD is shown in orange, while the rest of the protein is shown in grey. All other glycans are shown in blue. Data analysis and graphs were done with seaborn (www.seaborn.pydata.org) and molecular rendering with VMD [9]

C.5   N234-Man9* (N165/N343 Man5)
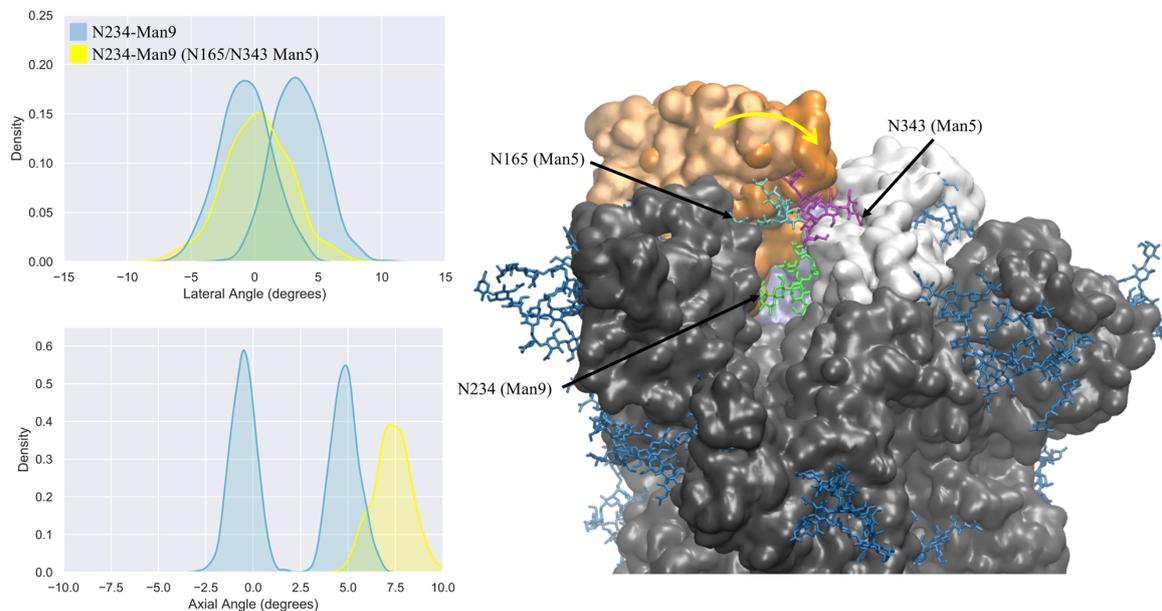
Figure C.5: . Left KDE analysis of the lateral and axial angles distributions calculated through the uncorrelated MD trajectories obtained for N234-Man9 with Man5 at N165 and N343 (yellow), and for N234-Man9 (cyan), replicas R1,2 for comparison. Right Close-ups on a representative snapshot of the N234-Man9 with Man5 at N165 and N343 simulation from the side. Man9 at N234 is shown in green, while the Man5 at N165 and N343 are shown in cyan and purple, respectively. The solvent accessible surfaces of the open RBD (B) and closed RBD (C) are shown in orange and white, respectively. The displacement of the open RBD observed through the 2.1 μs trajectory is highlighted with a yellow arrow, where the initial RBD conformation is rendered in pale (brushed metal) orange. The rest of the protein is shown in grey and all other glycans are shown in blue. Data analysis and graphs were done with seaborn (www.seaborn.pydata.org) and molecular rendering with VMD [9]

C.6   N234-Man3
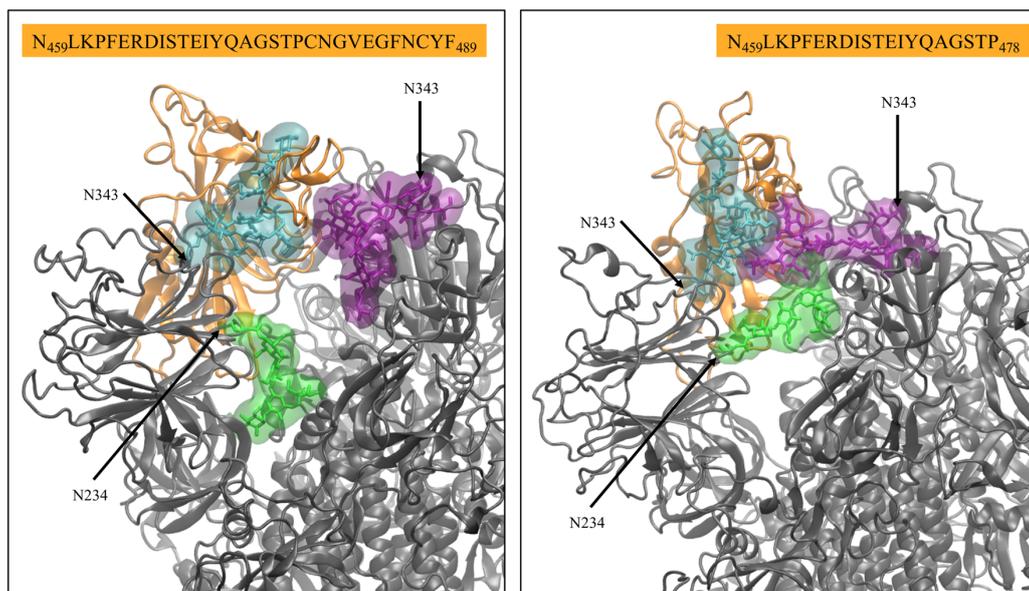
Figure C.6:   Close-ups on two representative snapshots of the N234-Man3 simulations R1 (left) and R2 (right) illustrating different orientations of the Man3 at N234 and corresponding interactions of the RBD (sequence indicated above each panel) with the N-glycans at N165 and. N343. The Man3 at N234 is shown in green, while the FA2G2 at N165 and N343 are shown in cyan and purple, respectively. The open RBD (B) is shown in orange with a cartoon representation, while the rest of the protein is in grey. The glycans at all other sites are omitted for clarity. Molecular rendering with VMD [9]
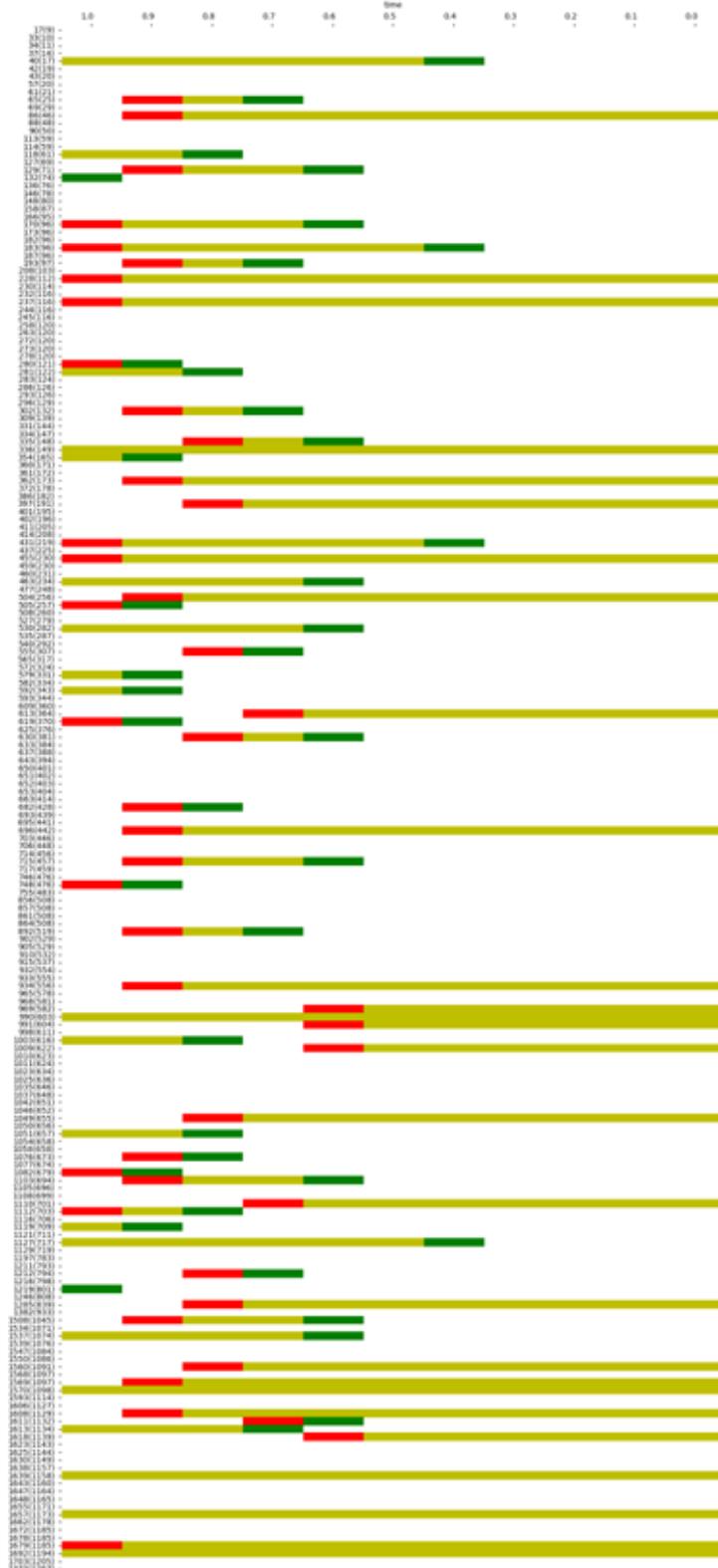
Figure C.7: 6 Conservation (in olive), gain (in green), or loss (in red) of sequons in SARS-CoV2 S. The x-axis starts with the original node in the phylogeny on the right (0.0), and the extant sequence on the left. Residue numbers are indicated in the y-axis, where the first number indicates the residues position in the alignment of all (∼30) extant sequences included in the analysis, and the residue number corresponding to the PDB 6VYB is indicated in parenthesis.
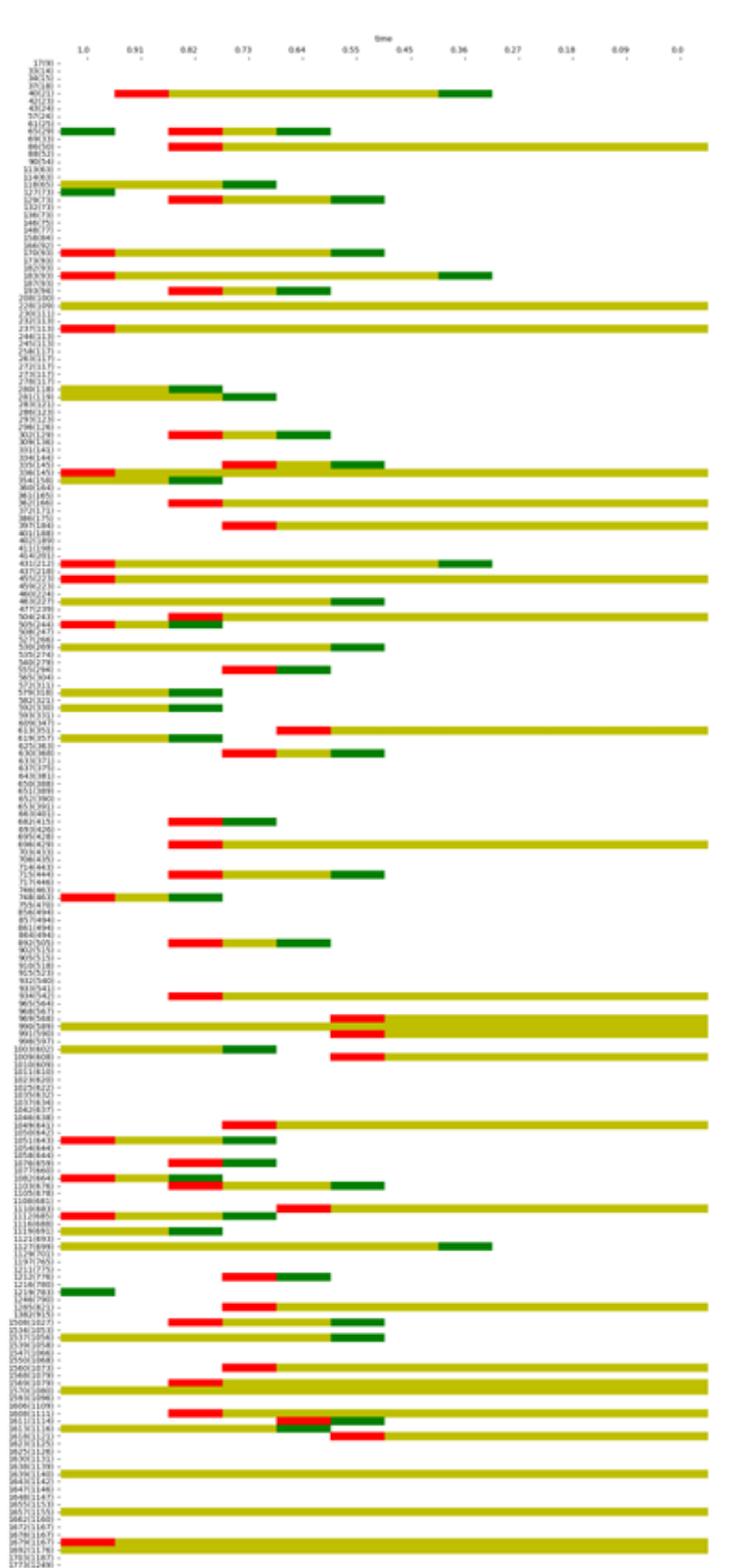
Figure C.8: Conservation (in olive), gain (in green), or loss (in red) of sequons in SARS-CoV (SARS) S. The x-axis starts with the original node in the phylogeny on the right (0.0), and the extant sequence (1.0) on the left. Residue numbers are indicated in the y-axis, where the first number indicates the residues position in the alignment of all (∼30) extant sequences included in the analysis, and the residue number corresponding to the SARS numbering, where N370 in SARS-CoV2 corresponds to position 357, is indicated in parenthesis.

```
             CoV1 sp|P59594|SPIKE_SARS   GVIADYNYKLPDDFMGCVLAWNTRNIDATSTGNYNYKYRYLRHGKLRPFERDISNVPFSP   462
      Pangolin CoV QLR06869.1             GRIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQA   471
             CoV2 YP_009724390.1         GKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQA   475
           RaTG13 QHR63300.2              GKIADYNYKLPDDFTGCVIAWNSKHIDAKEGGNFNYLYRLFRKANLKPFERDISTEIYQA   475
                                          * ************ ***:***:.::*:.  **:** ** :*:.:*:*******.  :.


             CoV1 sp|P59594|SPIKE_SARS   DGKPCTP-PALNCYWPLNDYGFYTTTGIGYQPYRVVVLSFELLNAPATVCGPKLSTDLIK   521
      Pangolin CoV QLR06869.1             GSTPCNGVEGFNCYFPLQSYGFHPTNGVGYQPYRVVVLSFELLNAPATVCGPKQSTNLVK   531
             CoV2 YP_009724390.1         GSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVK   535
           RaTG13 QHR63300.2              GSKPCNGQTGLNCYYPLYRYGFYPTDGVGHQPYRVVVLSFELLNAPATVCGPKKSTNLVK   535
                                          ...**.   .:***:** **** * *:*:.*************:********* **:*:*
```

Figure C.9: Close up of a snapshot of the MD simulation of the N234-Man9 glycosylated at N370 (R2). The FA2G2 N-glycan at N370 of chain A (yellow) is bound to the surface of the closed RBD (C) shown in as surface (left) in white. The N-glycan FA3G2 (bisecting GlcNAc) at N165 (C) is shown in cyan. On the top right image, the residues in contact with the N-glycan are shown in light green. The positions of Y449 and F490 are highlighted for ease of orientation. The sequence alignment of a section of the RBD in the CoV1, CoV2, pangolin CoV and of the bat RaTG13 S is shown at the bottom, where the residues lining the cleft, where the F2A2G2 (1-6) arm lies, are highlighted in light green. Rendering was done with VMD [9] and sequence alignment with Clustal Omega[20].

Bibliography

[1] Case, D. et al. AMBER 2018. University of California: San Francisco, CA, USA **2018**,

[2] Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; González-Outeiriño, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. GLYCAM06: A generalizable biomolecular force field. Carbohydrates. Journal of Computational Chemistry **2007**, 29, 622–655.

[3] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. The Journal of Chemical Physics **1983**, 79, 926–935.

[4] Harbison, A. M.; Brosnan, L. P.; Fenlon, K.; Fadda, E. Sequence-to-structure dependence of isolated IgG Fc complex biantennary N-glycans: a molecular dynamics study. Glycobiology **2018**, 29, 94–103.

[5] Goga, N.; Rzepiela, A. J.; de Vries, A. H.; Marrink, S. J.; Berendsen, H. J. C. Efficient Algorithms for Langevin and DPD Dynamics. Journal of Chemical Theory and Computation **2012**, 8, 3637–3649.

[6] Scott, R. Computer Simulation of Liquids. 1991.

[7] Berendsen, H. J.; Postma, J. v.; Van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. The Journal of chemical physics **1984**, 81, 3684–3690.

[8] Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. Journal of Computational Physics **1977**, 23, 327–341.

[9] Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. Journal of Molecular Graphics **1996**, 14, 33–38.

[10] Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F. T.; de Beer, T. A. P.; Rempfer, C.; Bordoli, L., et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic acids research **2018**, 46, W296–W303.

[11] Fogarty, C. A.; Harbison, A. M.; Dugdale, A. R.; Fadda, E. How and why plants and human N-glycans are different: Insight from molecular dynamics into the "glycoblocks" architecture of complex carbohydrates. Beilstein journal of organic chemistry **2020**, 16, 2046–2056.

[12] Harbison, A. M.; Brosnan, L. P.; Fenlon, K.; Fadda, E. Sequence-to-structure dependence of isolated IgG Fc complex biantennary N-glycans: a molecular dynamics study. Glycobiology **2019**, 29, 94–103.

[13] Watanabe, Y.; Allen, J. D.; Wrapp, D.; McLellan, J. S.; Crispin, M. Site-specific glycan analysis of the SARS-CoV-2 spike. Science **2020**, 369, 330–333.

[14] Shajahan, A.; Supekar, N. T.; Gleinich, A. S.; Azadi, P. Deducing the N- and O-glycosylation profile of the spike protein of novel coronavirus SARS-CoV-2. Glycobiology **2020**, 30, 981–988.

[15] Watanabe, Y.; Berndsen, Z. T.; Raghwani, J.; Seabright, G. E.; Allen, J. D.; Pybus, O. G.; McLellan, J. S.; Wilson, I. A.; Bowden, T. A.; Ward, A. B., et al. Vulnerabilities in coronavirus glycan shields despite extensive glycosylation. Nature communications **2020**, 11, 2688.

[16] Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. Journal of chemical theory and computation **2015**, 11, 3696–3713.

[17] Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; González-Outeiriño, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. GLYCAM06: a generalizable biomolecular force field. Carbohydrates. Journal of computational chemistry **2008**, 29, 622–655.

[18] Case, D.; Ben-Shalom, I.; Brozell, S.; Cerutti, D.; Cheatham III, T.; Cruzeiro, V.; Darden, T.; Duke, R.; Ghoreishi, D.; Gilson, M., et al. AMBER 2018; 2018. University of California, San Francisco **2018**,

[19] Apweiler, R.; Bairoch, A.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M., et al. UniProt: the universal protein knowledgebase. Nucleic acids research **2004**, 32, D115–D119.

[20] Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T. J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J., et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology **2011**, 7, 539.

[21] Price, M. N.; Dehal, P. S.; Arkin, A. P. FastTree 2–approximately maximum-likelihood trees for large alignments. PloS one **2010**, 5, e9490.

[22] Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. Molecular biology and evolution **2007**, 24, 1586–1591.