Exploring the role of the “glycan-shield” of human immunodeficiency virus in susceptibility to, and escape from, broadly neutralising antibodies.

Roux-Cil Ferreira

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor Philosophiae in the South African National Bioinformatics Institute, University of the Western Cape.

Supervisors: Prof. Simon A. Travers

Dr. Natasha Wood

October 2018

Exploring the role of the “glycan-shield” of human immunodeficiency virus in susceptibility to, and escape from, broadly neutralising antibodies.

Keywords

HIV-1

Envelope glycoprotein

Glycans

Molecular dynamics

Glycan N301

N301A mutation

Broadly neutralising antibodies

HIV-1 vaccine

Neutralisation resistance

HIV-1 selection pressure

Declaration

I, Roux-Cil Ferreira, hereby declare that all the work presented is my own work, that it has not been submitted for any degree or examination in any other university, and all the sources I have used or quoted have been indicated and acknowledged by complete references.

Full name: Roux-Cil Ferreira Date: 5 October 2018

Signed: 

Acknowledgements

I am deeply grateful to a number of people who have helped and supported me during this project. I would like to thank Prof. Simon Travers for welcoming me into his group and providing me with the equipment required for this project as well his continuous support and encouragement. I would also like to thank my co-supervisor Dr. Natasha Wood for reading all those overly complicated first drafts. I greatly appreciate her continuous support and her willingness to have daily progress report chats that were instrumental in the write-up process. Her confidence in me has helped me to become the independent scientist I am today.

Special thanks to Dr. Oliver Grant, who wrote the glycosylation tool that helped create the models that were used during this project. He also provided numerous scripts that helped to initialise the systems and helped to run the molecular dynamics simulations. I would also like to thank Dr. Robert Woods for allowing us to run the molecular dynamics simulations on their super computers.

There are also many people outside of the project who have supported and encouraged me. To my mother, thanks for always listening to all my complaints and always encouraging me to keep at it. To my partner, Phillip Labuschagne who's always dragging me to the gym, stating that deadlifts are a cure all even for incomplete, uncooperative analyses that never seem to lend themselves to easy explanation. I can't put into words how much you have helped me during this project and how thankful I am for your support. Finally, to my late father, you were always in my corner, always encouraging me to aim for that 100%. I sincerely believe that the sheer stubbornness I have inherited from you has allowed me to complete this project. Thank you to all of you for sticking with me. I am looking forward to the next chapter in my career and life with all of you as a part of it.

Abstract

The HIV-1 envelope (Env) glycoprotein is the primary target of the humoral immune response and a critical vaccine candidate. However, Env is densely glycosylated and thereby substantially protected from neutralisation. Despite the importance of the HIV-1 Env glycans, limited computational analyses have been employed to analyse these glycans.

Here, the Env glycans of two HIV-1 wild-type subtype C isolates are examined, in detail, using computational approaches. These particular strains were used since *in vitro* data showed that the removal of a single glycan had a substantially different impact on the neutralisation sensitivity of the two strains. Molecular dynamics simulations, and the subsequent analyses, were carried out on the computationally determined, fully glycosylated, Env structures of these two wild-type strains and their N301A mutant counterparts.

Detailed comparison of the molecular dynamics simulations demonstrated that unique glycan dynamics and conformations emerged and that, despite shared HXB2 reference sequence positions, the glycans adopted distinct conformations specific to each wild-type model. Furthermore, different changes in conformations were observed for each wild-type model compared to its N301A mutant counterpart and, interestingly, these N301A mutant model-specific glycan conformations were directly associated with the protein residues ultimately found to be exposed, which may explain the varied resistance to neutralising antibodies observed, *in vitro*, for the two N301A mutant strains.

A further detailed analysis was carried out focussing on the strain that remained resistant to frequently elicited neutralising antibodies, and displayed increased resistance to a CD4-binding site bNAb, despite the loss of glycan N301. The results complemented the laboratory study and revealed that, *in silico*, the glycan shield of this strain retained its ability to shield the protein residues even after the removal of glycan N301 from the wild-type model. Moreover, the change in the glycan landscape, and cascade of events, that contributed towards the maintenance of this glycan shield

and the increased resistance to a bNAb epitope located at the edge of the cascade, were discernible.

This study has demonstrated that the landscape of the glycan shield contains immense diversity, and that glycans that share an HXB2 reference sequence position will not necessarily conform in shape, dynamics and function between different strains. These conformational differences ultimately determine whether or not the glycan shield can compensate for the loss of a glycan, which directly relates to the fundamental function of the glycans in their ability to shield and protect the virus from immune surveillance. Additionally, this study highlights the remarkable potential of molecular dynamics simulations, which can be used as a powerful predictive technique to facilitate, and direct, laboratory studies focussing on vaccine research.

Table of content

List of Abbreviations

μ	Chemical potential
3D	Three-dimensional
$\boldsymbol{a}$	Acceleration
Å	Angstrom
AASA	Antibody accessible surface area
AIDS	Acquired immunodeficiency syndrome
Asn or N	Asparagine
BLAST	Basic local alignment search tool
bNAbs	Broadly neutralising antibodies
C	Carbon
C1-C5	Constant regions 1 to 5
CCR5	C-C chemokine receptor type 5
CD4bs	CD4-binding site
cDNA	complementary DNA
CPU	Central processing unit
CRFs	Circulating recombinant forms
cryo-EM	Cryo-electron microscopy
CXCR4	C-X-C chemokine receptor type 4
DNA	Deoxyribonucleic acid
DOPE	Discrete optimised protein energy
E	Energy
ER	Endoplasmic reticulum
Env	Envelope
$\boldsymbol{F}$	Force
Glc	Glycose
GlcNAc	N-acetylglucosamine
GPU	Graphics processing unit
H	Hydrogen
HIV	Human immunodeficiency virus
K	Kelvin
kcal/mol	kilocalorie per mole

M	Mutant
m	mass
mAb	Monoclonal antibodies
Man	Mannose
MD	Molecular dynamics
MPER	Membrane-proximal external region
N	Nitrogen
n	Number
nm	Nanometre
ns	Nanosecond
nPT	Isobaric-isothermal ensemble
O	Oxygen
OH	Hydroxyl
P	Pressure
PBC	Periodic boundary conditions
PCA	Principal component analysis
PDB	Protein data bank
PNGS	Potential N-linked glycosylation site
r	Position
RNA	Ribonucleic acid
SASA	Solvent accessible surface area
Ser	Serine
SIV	Simian immunodeficiency virus
T	Temperature
t	Time
Thr	Threonine
V1-V5	Variable loop regions 1 to 5
V	Volume
WT	Wild-type

List of Figures

List of tables

Introduction

The human immunodeficiency virus-1 (HIV-1) envelope (Env) glycoprotein trimer is extensively glycosylated with host-derived glycans. These glycans impede the engagement of frequently elicited neutralising antibodies, effectively shielding the virus from immune surveillance. However, the dense glycosylation on the Env trimer surface also limits the actions of the glycan-processing enzymes, which leads to an abundance of under-processed oligomannose glycans. Since this feature of the glycan shield is conserved, antibodies that target these oligomannose glycans achieve remarkable breadth. While glycan-targeting broadly neutralising antibodies (bNAbs) provide hope for the development of a vaccine, the isolation of viral isolates harbouring escape mutations from infected individuals who develop bNAbs, is of concern. These escape mutations have been related to shifted or deleted N-glycosylation sites and, therefore, the glycan shield is of particular interest both in its susceptibility to, and escape from, bNAbs. However, despite the importance of these glycans, to date, the computational analyses employed to analyse HIV-1 Env glycans are limited. Thus, this thesis explores the characteristics of the HIV-1 Env glycan shield using molecular dynamics simulations, with particular focus on the effect of removing a glycan from a wild-type model, as well as the application of various analysis techniques to this context.

Chapter 1 is aimed at understanding both the features of HIV-1 Env glycans, as well as the requirements necessary to perform, and analyse, molecular dynamics simulations of the HIV-1 Env protein and its glycans. Therefore, a broad background on HIV, glycans and molecular dynamics simulations is provided. The first section includes an overview of the Env glycoprotein focusing on its biosynthesis and the experimental and computational approaches that have been employed to understand the HIV-1 Env glycans. The second section describes structural and conformational considerations relevant to the study of HIV-1 Env glycans, and the chapter concludes with a description of the underlying theory of molecular dynamics simulations as well as the requirements for applying it to molecular models.

In Chapter 2, the two wild-type HIV-1 subtype C isolates used throughout this thesis are introduced. These particular strains were of interest since *in vitro* data showed that the absence of glycan N301 had a substantially different effect on the neutralisation sensitivity of the two strains. Therefore, in order to understand the difference in the glycan shields of these two strains, and the differential impact of removing glycan N301, the fully glycosylated Env structures of these strains were computationally determined. Thus, Chapter 2 presents the computational approaches used to both predict the three-dimensional (3D) structures of experimentally undetermined HIV-1 Env trimers as well as to glycosylate these trimers. Since numerous modelling options and parameters exist, for each of the software packages used, the prerequisites are described and the various features investigated. Finally, with the established framework for determining the HIV-1 Env glycoprotein trimer models, the preliminary analysis aimed at understanding the impact of removing glycan N301 by using these static state representations, is described.

The static state representations determined in Chapter 2, however, offered relatively limited information since the molecules are dynamic, changing conformation over time. Therefore, Chapter 3 extended the analysis by comparing the molecular dynamics simulations of the modelled structures (two wild-type models and their N301A mutant counterparts). Several methods were applied to individual, or clusters of, glycan/s in order to describe the surroundings, conformations and potential functions of the glycans. These results were then compared between the wild-type models, between a wild-type model and its N301A mutant counterpart, and between the N301A mutant models. The differences between the N301A mutant model simulations are discussed in relation to their wild-type model counterparts, with specific focus on the integrity of the modelled glycan shields of the two strains. Therefore, Chapter 3 built upon Chapter 2 and included information on the changes in glycan conformations over time for each model.

During the analysis performed in Chapter 3, glycans that were conformational heterogeneous between each wild-type model and its N301A mutant counterpart were identified. Interestingly, some of these glycans were not in the immediate vicinity of the N301A mutation. The conformational heterogeneity of these glycans that are not located near the N301A mutation may explain why one of the strains remained

resistant to frequently elicited neutralising antibodies and displayed increased resistance to a CD4-binding site bNAb despite the loss of glycan N301. Therefore, in order to further investigate the resistant strain and directly associate the conformational heterogeneity of its glycans with the N301A mutation, a more detailed comparison between this wild-type model and its N301A mutant counterpart was carried out in Chapter 4. A method that determines the glycan nearest to each amino acid residue over time is described and used to illustrate how changes in glycan conformations impact their ability to protect surrounding residues of the underlying protein. Finally, the cascade of conformational changes that occurred after glycan N301 was removed is described. This analysis offers one explanation of how glycans further from the N301A mutation site could be impacted by the N301A mutation. Furthermore, results are presented that link the cascade of glycan conformational changes with the increased resistance to the CD4-binding site bNAb (observed *in vitro*), thereby providing a definitive hypothesis for, and connecting the computational analysis with, the laboratory findings.

Finally, Chapter 5 provides a summary of the main findings of this thesis and presents clear avenues for future work.

Chapter 1 Background

1.1 The Human Immunodeficiency Virus

Human immunodeficiency virus (HIV) is a lentivirus of the retrovirus family that can cause acquired immunodeficiency syndrome (AIDS) in humans. Despite the development of antiretroviral therapies, HIV remains one of the greatest global health burdens. Approximately 1.8 million people were newly infected with HIV in 2017 and 36.9 million people were already living with HIV at the time[1]. An effective prophylactic vaccine is therefore essential, but remains elusive despite tremendous progress in our understanding of HIV infection and the substantial investments in vaccine research and clinical trials[2–10]. A key scientific challenge in the field of HIV vaccine development is the design of immunogens that elicit antibodies capable of neutralising HIV despite its immense global diversity[2–10].

Two types of HIV have been characterised, HIV-1 and HIV-2, which originated from simian immunodeficiency virus (SIV) through multiple cross-species transmission events from non-human primates to humans[11–17]. Compared to HIV-1 infection, those infected with HIV-2 generally have longer asymptomatic stages, lower virus titers, and lower transmission rates; these particular differences contribute to the confinement of HIV-2 to West Africa[18]. HIV-1, on the other hand, is the cause of the majority of HIV infections globally and is currently the main focus for HIV researchers.

HIV-1 is an assortment of genetically related, but phylogenetically diverse, viruses, which are classified into four groups (M, N, O, and P)[19–21]. Group M is responsible for the current worldwide pandemic and is further divided into nine major subtypes (A, B, C, D, F, G, H, J, and K)[19–21]. Superinfection or co-infection with one or more subtypes within an individual has led to more than 100 circulating recombinant forms (CRFs)[19–21]. Subtypes A-D, and the recombinant forms CRF01_AE and CRF02_AG, cause the majority of HIV-1 infections; however, subtype C alone accounts for half of all HIV-1 infections worldwide[20].

HIV-1 is transmitted by the exchange of bodily fluids, which includes sexual contact, needle sharing by injecting drug users, and transmission from mother to child[22]. The main transmission route of HIV-1 infection is via sexual contact and requires that the virus cross a mucosal barrier[22]. Therefore, the first cells to encounter HIV-1 are the dendritic cells and macrophages located in vaginal mucosa, oral mucosa and male foreskin[22]. Both dendritic cells and macrophages express low levels of the surface protein, CD4, required by HIV-1 for binding and entry[23–25]. Following transmission, the progression from initial to chronic infection includes the spread of HIV-1 throughout the body's $CD4^+$ T cells, i.e. T cells that express a CD4 receptor[26]. The depletion of these cells is a hallmark of HIV-1 infection[26].

Like many other viruses, HIV-1 uses the cellular biosynthesis pathways of host cells to replicate[27]. This requires the virus to enter a host cell, a process initiated by the HIV-1 surface glycoprotein, envelope (Env). Briefly, the Env glycoprotein comprises two subunits, gp120 and gp41, and of these, the gp120 subunit binds the CD4 receptor of the targeted host cell and initiates a series of conformational changes that enables binding to a chemokine co-receptor, either C-C chemokine receptor type 5 (CCR5) or CXC-chemokine receptor type 4 (CXCR4)[28–30]. After co-receptor binding, structural changes in the gp41 subunit of the HIV-1 envelope glycoprotein mediate fusion between the virion and host cell membranes[27].

Upon fusion, the HIV-1 capsid, which contains the viral RNA and enzymes, enters the host cell cytoplasm[31]; uncoating and degradation of the viral core takes place and the RNA is liberated from attached viral proteins and coded into complementary DNA (cDNA) by reverse transcriptase[31,32]. The cDNA forms a double-stranded viral DNA molecule and is transported into the cell nucleus by another viral enzyme, integrase[33]. The final step, assembly of a new HIV-1 virion, begins at the plasma membrane of the host cell. The polyproteins associate with the inner surface of the plasma membrane along with the HIV-1 RNA, after which the immature virion buds from the host cell and is then assembled to produce a mature HIV-1 virion capable of infecting another cell[33].

This HIV-1 replication cycle, where viral RNA is reverse transcribed into DNA[33], is relatively short, lasting only approximately 2.5 days[34]. This process is performed by the viral enzyme reverse transcriptase, which is highly error-prone[35] compared to other

transcription processes performed by mammalian cells where various proofreading and error correction mechanisms exist[36]. The short replication cycle and error-prone reverse transcription leads to diverse populations of viral particles, which are commonly referred to as the quasispecies[37]. While certain mutations may result in non-functioning viral particles[38], others may result in increased fitness, particularly in the presence of a immune reaction[38]. As will be noted in section 1.1.1, a subset of individuals develop particularly broad and potent neutralising antibodies capable of neutralising a wide range of HIV-1 viruses[39–43]. However, this does not provide a clinical benefit for these individuals since escape mutations, present on the Env glycoprotein, flourish under the selection pressure leading to a "new" quasispecies, rendering the neutralising antibodies ineffective against the viruses expressing the mutated, and thus resistant, Env glycoproteins[42,44].

1.1.1 The HIV-1 Env glycoprotein

HIV-1 Env, which mediates cell entry (as described in section 1.1), is the only surface-exposed viral protein (Figure 1.1 A). It is, therefore, targeted by the humoral immune response and a good candidate for vaccine research. This has consequently led to extensive analyses of the Env glycoprotein (Figure 1.1 B), which, as the name suggests, comprises protein residues and glycans (carbohydrates; described in detail in section 1.2).

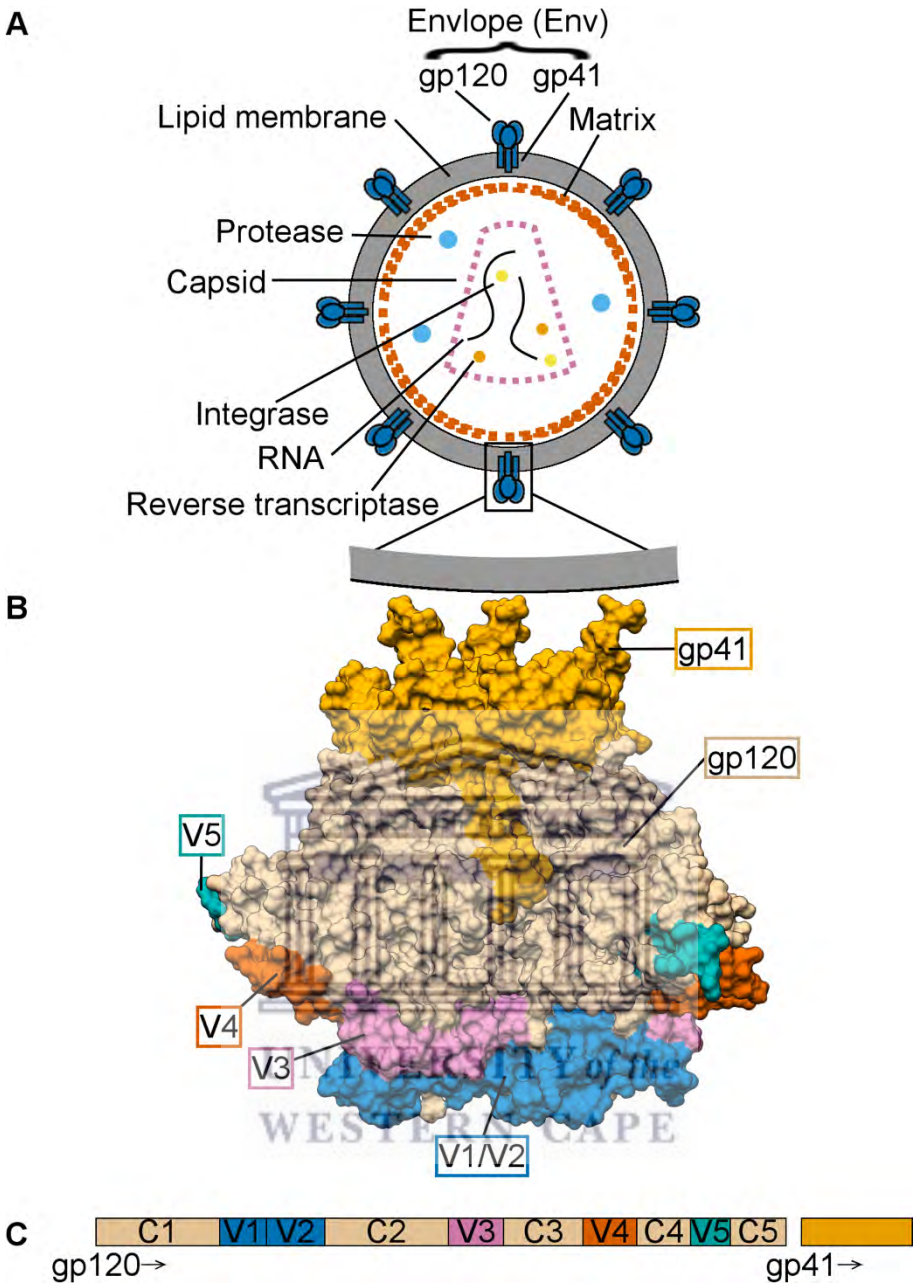

Figure 1.1: Schematic diagram of the HIV-1 virion and its surface glycoprotein

(A) The assembled HIV-1 virion; (B) enlarged view of the surface protein, Env[a]; and (C) linear schematic of the *env* gene. For the Env protein structure (B) the gp41, gp120 and variable loop (V1-V5) regions of gp120 are labelled and coloured, for the *env* gene schematic (C), these regions as well as the conserved regions (C1-C5) of gp120 are labelled and coloured to illustrate the order of gene sections.

The Env glycoprotein is initially synthesised as a precursor gp160 molecule, which is cleaved by the host protease, furin, during its transportation through the trans-Golgi

[a] *The depicted orientation, where the lipid membrane is located at the top and the V1/V2 loop regions are at the bottom of the figure, is used throughout this thesis.*

network, into gp41 and gp120[28]. The surface of the gp120 subunit, in turn, consists of 5 variable loop (V1-V5) and 5 conserved (C1-C5) regions (Figure 1.1 C). Prior to protein folding, cleaving and trimerisation, precursor Glc3Man9GlcNAc2 (Glc: glycose; Man: mannose; GlcNAc: N-acetylglucosamine; Figure 1.2 A) glycans are added, where possible, to the amide groups of asparagine (Asn) residues at potential N-linked glycosylation sites (PNGSs), determined by an Asn-X-Threonine (Thr) or Asn-X-Serine (Ser) motif, where X is any amino acid except proline[45]. The gp120 subunit has, depending on the viral strain, 20-30 PNGSs, while the gp41 subunit has 4-5 PNGSs, and HIV-1 Env is one of the most heavily glycosylated molecules known to date[46,47].

After the precursor glycans are attached to the PNGSs, they are trimmed by the endoplasmic reticulum (ER) α-glucosidase I and II to yield monoglucosylated N-linked glycans, GlcMan9GlcNAc2, that interact with the folding chaperones calnexin and calreticulin[45,48] (Figure 1.2 A). Once gp160 has correctly folded, the ER α-glucosidase II further trims the GlcMan9GlcNAc2 glycans into Man9GlcNAc2 glycans[45] (Figure 1.2 A). These are then further processed by ER and Golgi α-mannosidases, which yields unprocessed, immature, oligomannose-type glycans (Man5GlcNAc2) (Figure 1.2 B).

Finally, N-acetylglucosaminyltransferase I transfers an N-acetylglucosamine residue, which initiates complex-type glycosylation and diversification[45] (Figure 1.2). It is important to note that the exact glycoforms of hybrid- and complex-type glycans are cell type and tissue-specific, which can lead to a range of different glycoforms at different glycan sites[49,50], as well as micro-heterogeneity at a particular site[49]. Ultimately, the matured, and functional, Env glycoprotein is a metastable trimer formed by non-covalently associated heterodimers of gp120 and gp41, with both subunits extensively modified by numerous glycans comprising different glycoforms.

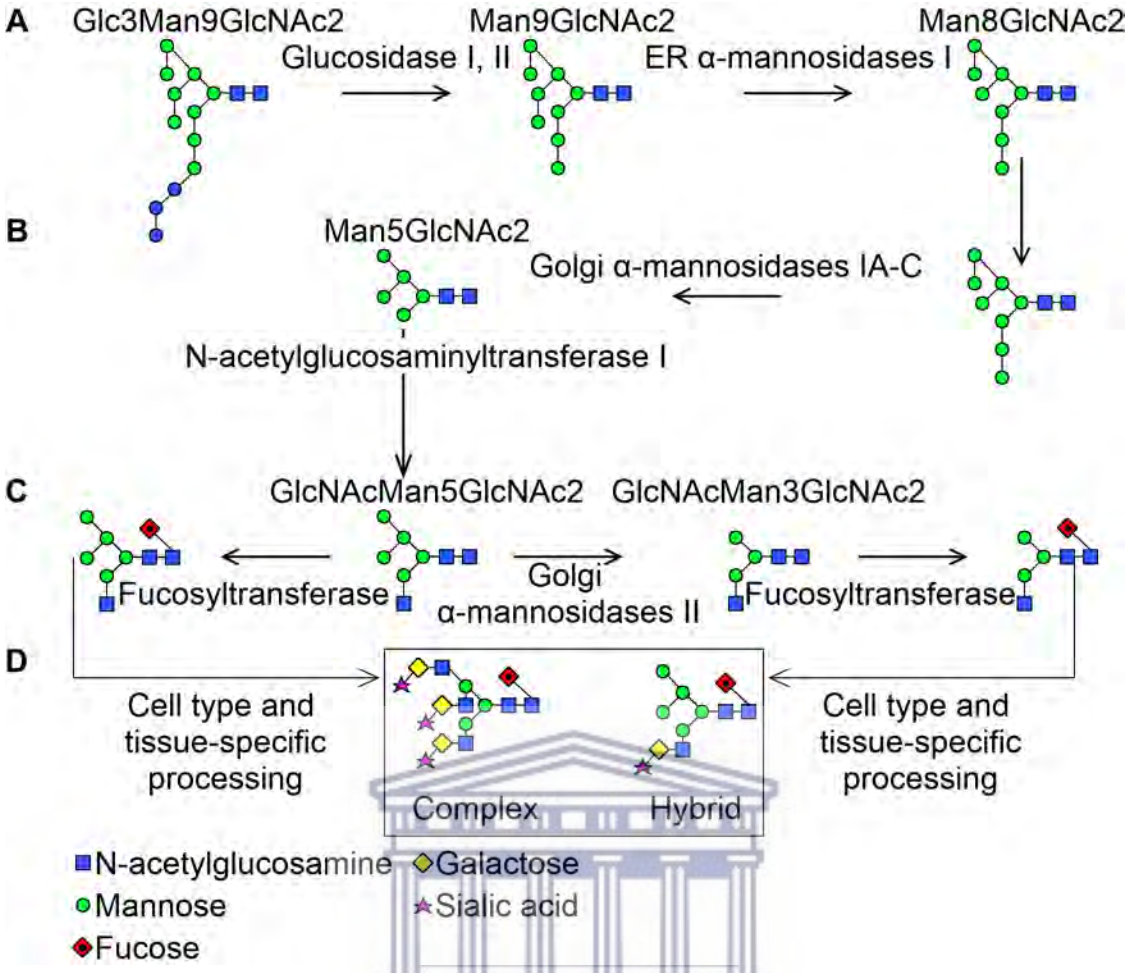

Figure 1.2: Glycosylation pathway

(A) In the ER, Glc3Man9GlcNAc2 glycans are attached to Asn residues of PNGSs on the Env glycoprotein. The three terminal glucose molecules are removed by α-glucosidases I and II. This is followed by the ER α-mannosidases I, which further trim the glycans to form Man8GlcNAc2 glycans. (B) These glycans are further processed by Golgi α-mannosidases IA-C to yield Man5GlcNAc2 glycans. Thereafter, an N-acetylglucosamine residue is transferred, by N-acetylglucosaminyltransferase I, to the Man5GlcNAc2 glycans. (C) The GlcNAcMan5GlcNAc2 glycans are the substrate for Golgi-resident fucosyltransferase. Alternatively, further mannose trimming by the Golgi α-mannosidase II can occur, which leads to GlcNAcMan3GlcNAc2 glycans. (D) Finally, the glycan can be further possessed into hybrid- or complex-type glycans. This figure was adapted from a book by Pantophlet *et al.*[51] and glycan representations are shown according to the nomenclature of Harvey *et al*.[52].

This extensive glycan coating is an important feature of the HIV-1 Env glycoprotein, as it is used as a defence mechanism to evade recognition by neutralising antibodies. Since these glycans are synthesised by the host-cell glycosylation machinery, it limits the range of B-cell responses due to antigen self-tolerance. The shielding capacity of these glycans is demonstrated by observations that the removal or addition of specific N-linked glycans modulate the susceptibility of HIV-1 to neutralising antibodies[42,44,53–

[55]. Therefore, the HIV-1 Env glycans have often been referred to as the “glycan shield” [56,57].

It was, therefore, surprising when a breakthrough publication by Walker *et al.*[58] revealed not only the presence of a distinct family of antibodies with greater neutralisation potency and breadth, but also that the epitopes of these antibodies contained glycans[58]. Further studies of such broadly neutralising antibodies (bNAbs) showed that the epitopes tend to cluster around “sites of vulnerability” (Figure 1.3), which include the CD4-binding site[59], the membrane-proximal external region of gp41[60], the glycan outer domain[61], the V1V2 apex region around glycan N160[58], the V3 base around glycans N301 and N332[62], and the gp120/gp41 interface[63]. However, despite the presence of bNAbs in the serum of HIV-1 infected individuals, circulating plasma viruses generally escape, resulting in continued infection[64,65]. In these infected individuals, the virus is under constant pressure from both antibodies that target exposed protein epitopes and antibodies that target N-linked glycans[42,44]. Thus, the glycan shield is constantly evolving and shifting in response to the host immune system[42,44]. Taken together, these observations have led to the emerging view that the glycosylation profiles of HIV-1 strains are homogeneous overall, but within-host virus populations have diminished or increased numbers of glycans, or shifted glycan positions, depending on the specific pressure exerted by the host immune system.

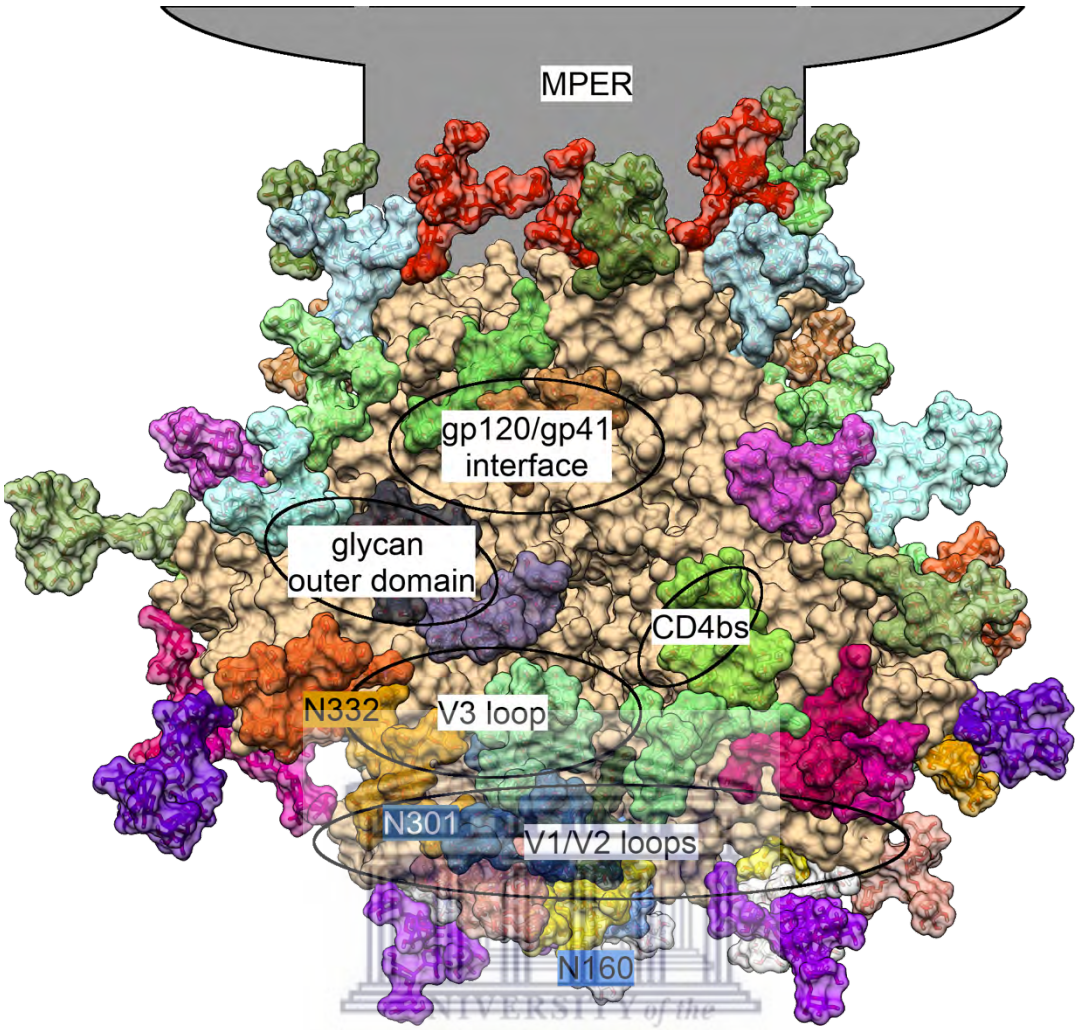

Figure 1.3: HIV-1 Env sites of vulnerability

Anti-HIV-1 bNAbs target six conserved regions on the Env glycoprotein: The CD4-binding site (CD4bs); the membrane-proximal external region (MPER); the glycan outer domain; the V1V2 apex region around glycan N160; the V3 base around glycan N301 and N332; and the gp120/gp41 interface.

Due to their importance, the HIV-1 Env glycans have been extensively investigated using a variety of techniques. The outcomes of these techniques include, but are not limited to, remodelling of the glycan shield[53–55,66–71], determining the glycoforms at each glycosylated site[49,72–82], and determining the structural characteristics of glycans[83–92], which are each described in more detail below.

Glycan shield remodelling refers to a process by which one, or more, glycans are edited or removed from the Env surface. This involves the use of glycosidase inhibitors[93,94] to manipulate the glycoforms on Env, and/or the addition or removal of glycosylation sites through site-directed mutagenesis[53–55,66–71], which mutates the

Asn, and/or Thr or Ser, residues. These methods have allowed for the identification of potential escape mutations, as well as N-linked glycans, and their glycoforms, that form part of epitopes of particular HIV-1 Env bNAbs[53–55,66–71].

Although remodelling of the glycan shield presents a crude determination of the glycoforms necessary for antibody binding, it does not provide information regarding the native glycoforms. However, the native glycoforms of each site can be, and are commonly, determined by carrying out mass spectrometry using HIV-1 Env glycoproteins[49,72–82]. Broadly speaking, these studies have demonstrated that complex-type glycans are more often found on the C1, V1/V2 and V3 regions, whereas oligomannose-type glycans mainly reside on the C2 and C3 regions, i.e. the outer domain of the Env glycoprotein (Figure 1.1 C). The presence of a dense cluster of glycans on the gp120 outer domain is thought to decrease the accessibility for glycosidases, which subsequently leads to stalled, or reduced, trimming of glycans forming part of these clusters[49,72–82]. This results in an N-linked glycosylation pattern dominated by minimally trimmed glycans, which differs from that which is generally observed on human glycoproteins[49,72–82]. Additionally, recent studies using mass spectrometry have provided information on how often each PNGS is glycosylated, as well as quantification of the individual glycoform micro-heterogeneity at each glycosylated site[49,95]. The presence or absence of a glycan/s, as well as the specific glycoforms, have been shown to have a substantial impact on the binding of bNAbs[95].

However, in order to determine exactly how different glycoforms impact the binding of bNAbs, the Env glycoprotein, together with the relevant bNAbs, need to be characterised in detail. This requires alternative methods over and above glycan shield remodelling and mass spectrometry. X-ray crystallography is one such method, however, the structural characterisation of a stable trimer is particularly difficult. The reason for this is that the assembly of functional Env trimers is inefficient and leads to dimers, malformed trimers, and higher-order aggregates, which do not represent the native Env trimer[96,97]. Additionally, Env often sheds the gp120 component[98]; together, these features of the HIV-1 Env glycoprotein contribute to the huge challenge of obtaining intact crystal structures of Env. Despite these challenges, the crystal structures of peptide fragments, gp120 core monomers, and scaffolded gp120 and gp41 loops have been determined for a variety of viral strains by introducing various

stabilising mutations, and by using antibodies and/or CD4 for additional stabilisation (some examples are[83–92]).

The crystal structures of the Env glycans are even more difficult to characterise using X-ray crystallography than the protein component[99]. Crystallising any glycan-part beyond the protein-proximal GlcNAc residues is extremely difficult due to the extensive conformational heterogeneity[99]. Nevertheless, analysis of these protein-proximal GlcNAc residues bound to proteins has shown a strong preference for glycosidic linkages with specific *phi* and *psi* dihedral angles (bond between four linked atoms; see section 1.2.2) with minimal energy conformations[99–101]. Recently, the atomic-level details of clades A, B and G Env glycoproteins were determined by Stewart *et al.*[88], with 15, 17 and 16 glycans, respectively, resolved (to different degrees/lengths from the base) beyond the protein-proximal GlcNAc residues. The authors demonstrated that the HIV-1 Env glycan shield comprises prominent ridges of hydrogen-bonded oligosaccharides, which are ordered by glycan clustering and protrude 20Å above the Env glycoprotein surface[88].

Due to the difficulties surrounding the determination of the structural characteristics of glycans with regards to their impact on various Env functions, studies commonly include computational approaches to complement experimental-based studies[53,95,102–113]. For these Env-focussed studies, the structures of the respective Env glycoproteins were either undetermined, or additional structural characteristics were sought. The computational approaches that were used include molecular dynamics simulations, which are particularly well suited for glycan and glycoprotein systems, i.e. systems that are either experimentally undetermined or poorly resolved. While molecular dynamics simulation studies have provided valuable insights on Env and its glycans, few analytical approaches exist in the context of the Env glycans, with the majority of studies focussing on the protein only[103–113], or on the overlap, or binding, of glycans to antibodies[88,113]. Only two studies[86,89] to date investigated glycans in more detail using, as measure, a solvent accessible surface area (SASA) calculation (with a probe of 0-10Å[86] or 10Å[89]), network analysis (based on distances[89] or hydrogen bonds[86]), glycan-antibody overlap[86,89], glycan volumes[86,89], and SASA when a glycan is removed[89]. Given the variation between the glycan shields of different viral isolates, as well as the glycoform micro-heterogeneity at different glycosylated sites, the application of

analytical techniques to investigate the HIV-1 Env glycan shield in a novel way is essential.

1.2 Glycans and glycoform heterogeneity

As introduced in section 1.1.1, the glycans of HIV-1 Env are diverse in their glycoforms and distribution, and characterising this diversity remains a challenge for HIV-1 Env research. A detailed investigation and comparison between different HIV-1 glycan shields is presented in Chapter 2-4 of this thesis, and background on the composition and conformational features of glycans is provided below.

Glycans, or carbohydrates, are biomolecules consisting of chains of monosaccharides and follow the general formula $(C \cdot H_2O)_n$ from which their name, hydrates of carbon, is derived[114]. They are also referred to as sugars or, specifically, glycans when they are attached to other biomolecules such as lipids or proteins.

1.2.1 Monosaccharides

Monosaccharides, the most basic unit of carbohydrates, have a linear and unbranched carbon skeleton with one carbonyl (C=O) group on the first carbon atom and hydroxyl (OH) groups on all the remaining carbon atoms[114] (Figure 1.4 A). The carbon atoms are numbered from 1 to *n* along the backbone, starting from the end that is closest to the carbonyl (C=O) group[114] (Figure 1.4 A). Two monosaccharides, with the same molecular formula and location of the carbonyl group, can still be distinct in their 3D arrangements of the bonds of certain atoms[114]. For example, monosaccharides are classified as a member of either the D- or L-series, depending on the orientation of the hydroxyl furthest from the carbonyl group (Figure 1.4 B). Most naturally occurring monosaccharides belong to the D-series[114].

Monosaccharides can exist in a linear or cyclic form; the latter results from an intra-molecular interaction between the oxygen atom of a hydroxyl group and the carbonyl carbon[114] (Figure 1.4 C). Several ring-forms are possible depending on which hydroxyl interacts with the carbonyl carbon. For example, monosaccharides can form five-

membered (furanoses) or six membered (pyranoses) rings, provided that there are the appropriate number of carbons[114]. The pyranose form is the most common for the majority of monosaccharides[114]. Furthermore, monosaccharides are classified as either α or β, based on two alternative hydroxyl configurations at the carbonyl carbon[114] (Figure 1.4 D). Monosaccharides combine to form more complex molecules such as disaccharides, polysaccharides and glycans[114].

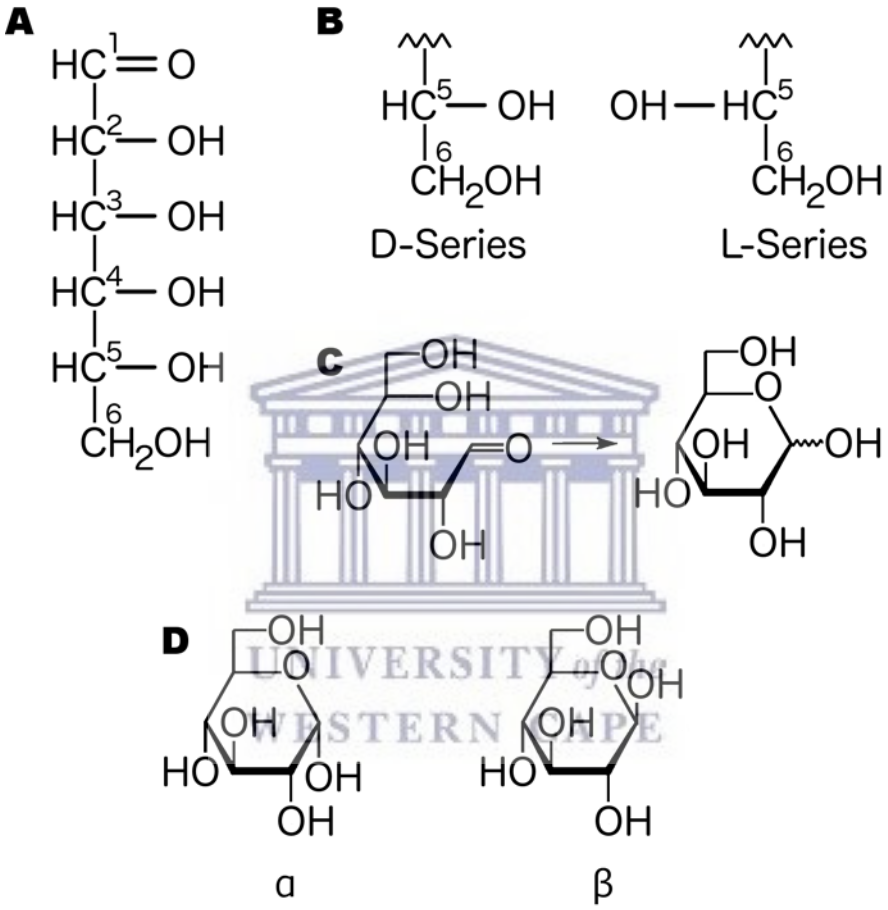

Figure 1.4: Monosaccharide configurations

Representations of the (A) linear carbon skeleton, (B) distinction between D- and L-series, (C) cyclic form, and (D) α and β classifications of monosaccharides. This figure was adapted from a book by Collins *et al.*[114]

1.2.2 Glycan conformations

Disaccharides, and eventually glycans, are formed when the hydroxyl groups of two monosaccharides form a C-O-C bond, which releases a water molecule[114]. The C-O-C bond is called the glycosidic linkage and occurs between carbon 1 (the first carbon) of one monosaccharide and either carbon 2, 3, 4, or 6 of another monosaccharide, forming a 1-2, 1-3, 1-4 or 1-6 linkage[114]. The dihedral angles, formed by four linked atoms, of these linkages are called ϕ (H1-C1-O-Cx) and ψ (C1-O-Cx-Hx)[114] (Figure 1.5 A). The 1-6 linkage has an additional dihedral angle called ω (O-C6'-C5'-H5')[114] (Figure 1.5 B). Glycosidic linkages and dihedral angles determine the overall glycan conformations.

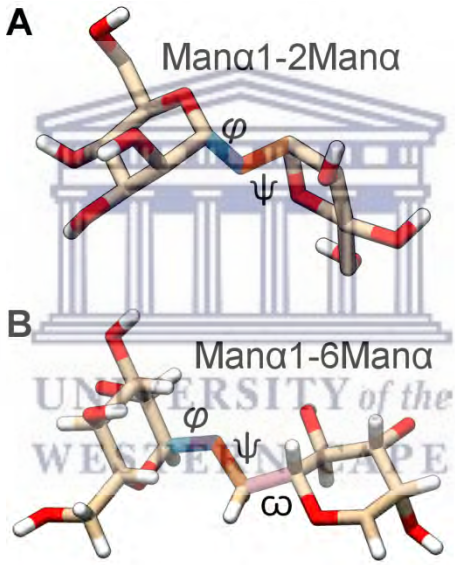

Figure 1.5: The dihedral angles of glycosidic linkages

3D representations of sections of a Man-9 glycan, demonstrating the (A) ϕ, ψ and (B) ϕ, ψ and ω dihedral angles.

For example, and in this thesis, the majority of analyses were performed using Man-9 glycans with the specific form: Manα1-2Manα1-6[Manα1-2Manα1-3]Manα1-6[Manα1-2Manα1-2Manα1-3]Manβ1-4GlcNAcβ1-4GlcNAcβ1. During attachment of the glycans to the protein structures, these dihedral angles were adjusted, within reason, to facilitate glycan attachment, which is discussed in section 2.2.2.

1.2.3 The HIV-1 Env glycan shield

The HIV-1 Env glycoprotein trimer is characterised by an unusually high abundance of glycans[46,47]. As detailed in section 1.1.1, these glycans shield the underlying Env surface from the immune system and continue to do so over the course of infection[42,44,53–56]. The HIV-1 Env glycans are, therefore, key features during investigations on structural and antigenic properties of Env, and while recent experimentally determined structures have greatly extended our knowledge of the interlocking glycan networks of the Env glycans, these cryo-electron microscopy[115–117] and X-ray crystallography[87,88,91,92] studies can only capture a single static state of any individual glycan (as described in section 1.1.1). However, the potential glycan conformational heterogeneity and movement can be illustrated and described using a computational method of generating molecular dynamics simulations. As example, Lemmin *et al.*[86] used a molecular dynamics approach and observed that neutralising antibodies generally bind at the interface between stable patches of glycans.

1.3 Molecular Dynamics Simulations

Molecular dynamics is a computational approach that simulates the physical movement of atoms in a given system. Atoms are typically represented as single point masses inside van der Waals potentials, i.e. they mostly act as hard spheres, and bond- and angle restraints are represented as harmonic oscillations[118–120]. Atoms can transfer both energy and momentum to one another via electrostatics and van der Waals interactions[118–120]. Atomic movements are determined by numerically solving Newton's equations of motion for the system of N interacting atoms[118–120]. By Newton's second law, the force ($\boldsymbol{F}$) on an object, in this case an atom, is equal to the mass (m) of that object multiplied by the acceleration ($\boldsymbol{a}$) of the object[118–120]:

$$\boldsymbol{F} = m\boldsymbol{a} \tag{1}$$

The force ($\boldsymbol{F}$) is further given by the negative gradient of the potential energy $V(r_1, r_2, \dots, r_N)$[118–120]:

$$\boldsymbol{F} = -\nabla V(\boldsymbol{r}) \tag{2}$$

Equations *(1)* and *(2)* are solved simultaneously in small time steps (Δt), and the atom is accelerated in the direction of the force using a leap-frog integrator to determine the new velocity (v) and position (r) of the atom[118–120]:

$$v\left(t+\frac{1}{2}\Delta t\right) = v\left(t-\frac{1}{2}\Delta t\right) + \frac{\Delta t}{m}\boldsymbol{F}(t) \tag{3}$$

$$r(t+\Delta t) = r(t) + \Delta t\, v\left(t+\tfrac{1}{2}\Delta t\right) \tag{4}$$

Thus, in order to perform molecular dynamics simulations, several properties should be understood and specified. The potential energy $V(\boldsymbol{r})$, calculated by solving a mathematical expression referred to as a force field, the thermodynamic ensemble that governs the simulation, the solvent model, and the boundary around the simulated system are important features that can have a large impact on the resultant molecular dynamics simulation[118–120]. Each of these features are described in further detail below.

1.3.1 Force Fields

An important consideration during the preparation of a system prior to the simulation, is the choice of a force field. Force fields are the mathematical equations used to model the interactions between atoms, and have two components, a mathematical expression and the parameters required by that expression[118–120]. The mathematical expression has a general form, which can be considered as comprising of the sum of two terms representing the bonded and non-bonded interactions[118–120]:

$$V(\boldsymbol{r}) = V_{bonded} + V_{Non-bonded} \tag{5}$$

The bonded and non-bonded interactions are, in turn, the sum of their own respective terms. Bonded interactions are calculated for the interactions formed between two (bonds), three (angles) and four (dihedrals/torsion) atoms[118–120]:

$$V_{bonded} = \sum_{i<j}^{Bonds} \frac{1}{2}k_{ij}^{b}\left(r_{ij}-r_{ij}^{0}\right)^2 + \sum_{i<j<k}^{Angles} \frac{1}{2}k_{ijk}^{\theta}\left(\theta_{ijk}-\theta_{ijk}^{0}\right)^2 + \sum_{i<j<k<l}^{Dihedrals} \frac{1}{2}k_{\phi}(1+\cos(n\phi+\gamma)) \tag{6}$$

Non-bonded interactions are calculated for the van der Waals and electrostatic interactions using a Lennard-Jones potential and Coulomb's law, respectively[118–120]:

$$V_{Non-bonded} = \sum_{i<j}^{Van\ der\ Waals} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_{i<j}^{Electrostatics} \frac{q_i q_j}{4\pi\varepsilon_0\epsilon_r r_{ij}} \quad (7)$$

The exact form of each equation (e.g. well depth and location) is determined by the parameters specified by the particular force field, which are derived using experimental data[118–120]. Several different force fields exist and the choice depends on the composition of the system and the ability of the force field to produce accurate results while maintaining reasonable computational efficiency[118–120]. The ff14SB[121] and GLYCAM06j-1[122] force fields were used in this thesis for the protein and glycan atoms, respectively. During molecular dynamics simulations, the equation of motion is governed by both the force field as well as the constraints enforced by the particular ensemble.

1.3.2 Ensemble

A further consideration during the preparation of a system is the choice of statistical ensemble. An ensemble represents all the microstates accessible by a system during a simulation. For each microstate, the ensemble also provides the probability of that microstate[118–120]. For example, if the system under study is a bottle of beer, then the ensemble can be thought of as a huge warehouse filled with individual, non-interacting, bottles of beer. A typical system has three intrinsic properties, number of atoms, volume and energy (Table 1.1) that would remain constant under an ideal molecular dynamics simulation[118–120]. Alternatively, each intrinsic property can be modulated by an external source, maintaining the chemical potential, pressure, and temperature instead (Table 1.1)[118–120]. For example, during a molecular dynamics simulation, energy is continuously lost due to rounding errors in the motion calculation[118–120], and thus, the decision to keep the energy of a system constant requires that a thermostat be specified that governs when and how much energy is added to the system. Therefore, the selection of a statistical ensemble involves the decisions around which intrinsic properties remain constant and which are modulated by an external source.

Table 1.1: The intrinsic properties of a molecular dynamics system and their related properties. An ensemble is defined by selecting a property in each row that will be kept constant.

Intrinsic property	Related property
Number of atoms (n)	Chemical potential (μ)
Volume (V)	Pressure (P)
Energy (E)	Temperature (T)

The isobaric-isothermal ensemble (nPT) used in this thesis, is the ensemble where the number of atoms, pressure and temperature are held constant[118–120]. A thermostat and barostat are used to maintain the desired temperature and pressure, respectively[118–120]. Numerous thermostat and borostat algorithms exist to facilitate temperature and pressure control. Those available in the molecular dynamics package AMBER[123], include the Berendsen, Andersen and Langevin thermostats, and the Berendsen and Monte Carlo borostats[119]. Of these, the Langevin thermostat and the Berendsen borostat were used here.

1.3.3 Solvent Models

While it is possible to proceed with molecular dynamics simulations after the specification of the force field and ensemble, it is common to include solvent molecules in order to better ascertain the behaviour of the biomolecule in a biologically relevant environment. One of two methods can be used to represent the solvent[118–120]. The first method uses a mathematical formula to approximate the influence of the solvent on the surface of the biomolecule, referred to as implicit solvent[118–120]. While an implicit solvent model is less computational expensive than the alternative method, the inability of the biomolecule to form hydrogen bonds with solvent molecules is a severe limitation. The alternative is to use an explicit solvent model that includes the atomic representation of the solvent similar to those of the biomolecule. Despite the increased computational burden, systems with explicit solvent offer a more accurate representation, and more detail, of the solvation environments of biomolecules.

There are various representations of the solvent molecules[118–120]. Each representation varies in the number of atoms used to represent the H_2O molecules as well as the distances and angels between the atoms. The TIP3P solvent model (Figure 1.6 A) is most commonly used during molecular dynamics simulations since it is computationally efficient[118–120]. TIP3P has three interaction points corresponding to the three atoms of a water molecule[118–120]. Each atom has a point charge representing the electrostatic distribution around water. More complex water models add "dummy" atoms (also called extra points) to the TIP3P solvent model to refine the electrostatic distribution around the water molecules[118–120]. The TIP5P solvent model[124] (Figure 1.6 B), for example, augments the TIP3P solvent model with two extra points that represent the electron lone pairs on the oxygen[124]. While the TIP5P solvent model is more computationally expensive[124], it has been shown to give the best quantitative agreement with experimental free energy data for small saccharides when used in conjunction with the GLYCAM force field[125]. For this reason, the TIP5P solvent model was used in this thesis.

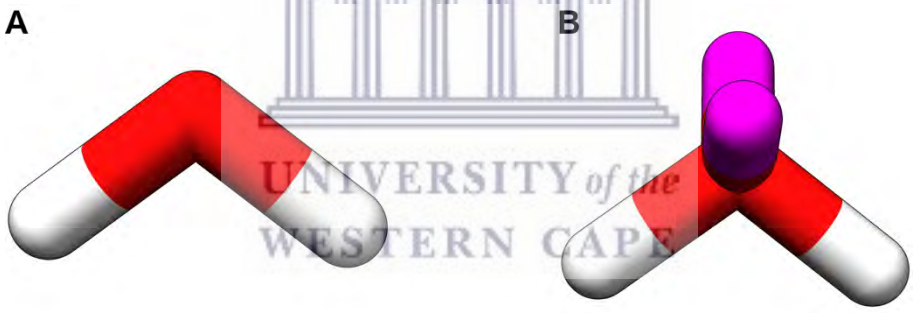

Figure 1.6: Explicit solvent models

Diagrams of the explicit solvent models (A) TIP3P and (B) TIP5P. Each model is coloured according to its atom composition, with hydrogen (white), oxygen (red), and the dummy atom (purple) shown.

1.3.4 Periodic Boundary Conditions

Once the solute (biomolecule) has been placed within the solvent, a boundary must be defined that will prevent the solvent molecules from dispersing too far from the biomolecule[118–120]. However, when such a boundary is introduced, the solvent molecules that are in contact with void space along the edges give rise to calculation

artefacts[118–120]. This “edge effect” can be addressed by using periodic boundary conditions (PBC, Figure 1.7)[118–120].

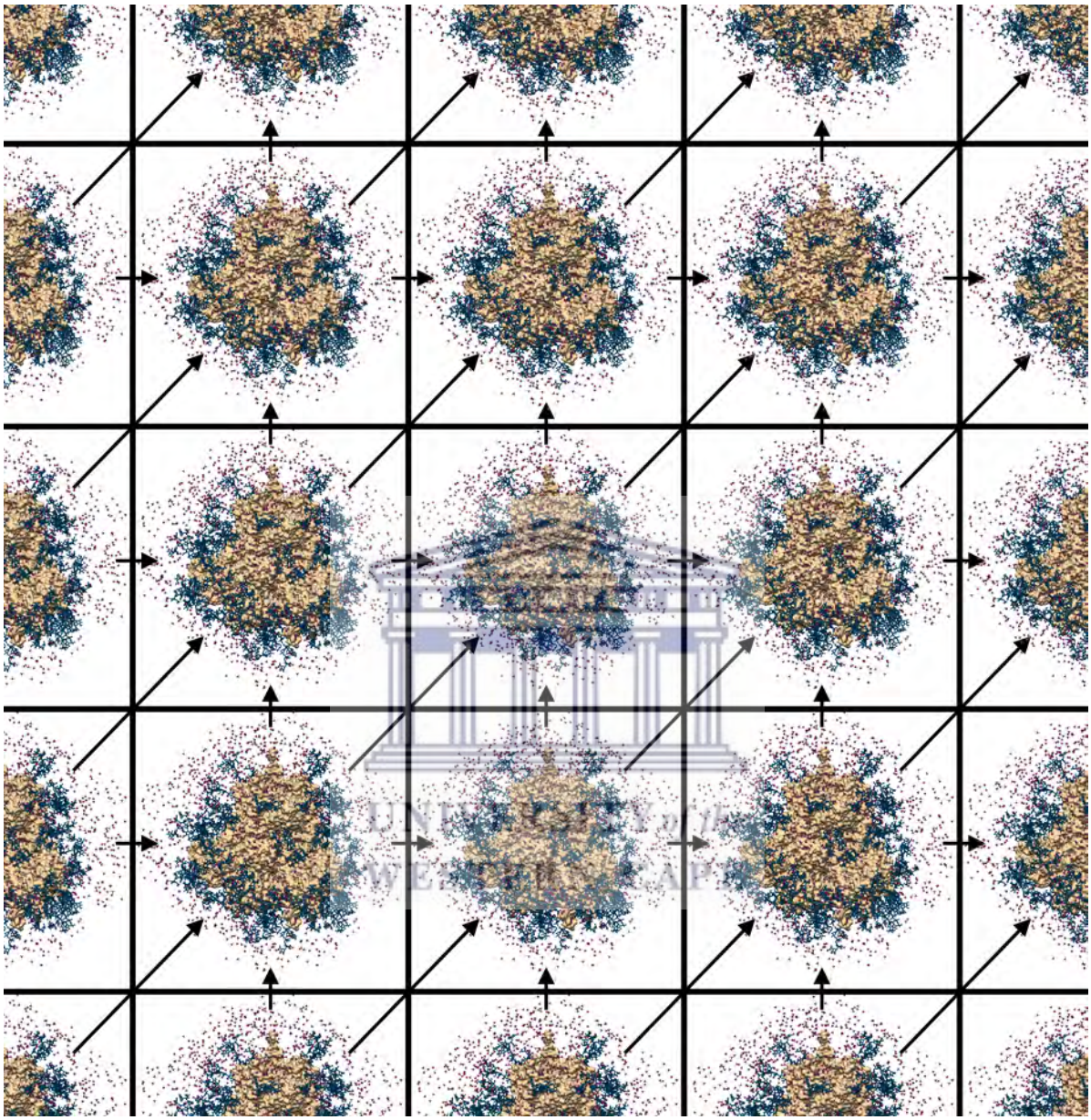

Figure 1.7: Periodic boundary conditions

Illustration of an HIV-1 Env system under periodic boundary conditions. As a molecule crosses the boundary on one side of the primary simulation, it re-enters on the opposite side, unaware of the boundaries.

PBC essentially results in a system that is infinitely copied in all directions such that the edge of every system borders its opposing edge in the adjacent image[118–120]. Therefore, if a solvent molecule happens to cross the edge of the solvent box, i.e. the periodic boundary, it exits the system on one side and immediately re-enters on the

opposite side (Figure 1.7). Consequently, this kind of system requires a finite number of atoms to represent a fully solvated biomolecule and the problems associated with edge effects are eliminated[118–120].

There is, however, a caveat to using periodic boundary conditions, which should be addressed to avoid the introduction of calculation artefacts. The minimum image convention states that the shortest edge of the simulated system must be at least twice the value of the non-bonded interaction cut-off (usually 8Å)[118–120]. Therefore, during molecular dynamics simulations, the size of the water box should be large enough such that the biomolecule cannot interact with itself across the boundary, which would lead to erroneous results[118–120].

1.3.5 Molecular dynamics simulations of HIV-1 Env proteins

Molecular dynamics simulations have been carried out on several HIV-1 Env proteins and recently, more studies have started including glycans. For example, Yang *et al.*[89] used molecular dynamics to calculate how often the glycans that are targeted by bNAbs, sample the bNAb-bound conformation (as determined in crystal structures). This "pre-organisation" is important since glycans that more frequently sample the conformation/s required for bNAb-binding potentially enhance the number of opportunities for bNAbs to do so, and are, thus, comparatively, better targets[89]. Therefore, despite the use of homogeneous, oligomannose, glycoforms in the Yang *et al.*[89] study (Man5NAc2 or Man9NAc2, depending on the site), the ability to rigorously define the composition of the system and subsequently study its molecular details resulted in valuable insights that can be used in a predictive fashion to facilitate vaccine design.

Chapter 2 Variability in the HIV-1 Env glycan shield influences its vulnerability after the loss of a glycan

As described in Chapter 1, the glycans on HIV-1 Env serve multiple functions. One important function is that these glycans shield epitopes from antibodies; for example, glycan N301 is implicated in the shielding of V3-loop and CD4-binding site epitopes. In this chapter, two subtype C strains are investigated, where for each of these isolates, the removal of glycan N301 had substantially different effects during an *in vitro* neutralisation investigation. Specifically, glycan N301 appeared to shield the epitopes of frequently elicited antibodies on one of the strains from recognition, but did not do so on the second strain. The molecular modelling of the Env glycoproteins, described here, suggests that the removal of glycan N301 likely exposed a greater surface area of specific protein residues in the V3-loop and C4 regions on one model compared to the other.

2.1 Introduction

The surface of HIV-1 Env, despite the protection afforded by the glycan shield, is susceptible to bNAbs. As described in section 1.1.1, there are six known sites of vulnerability on the Env glycoprotein that are targeted by bNAbs[58–63] (Figure 1.3). Of these, the class of bNAbs targeting the V3 glycans may be the easiest to elicit via vaccination, since these are among the most widespread and potent bNAb responses in infected individuals[126–129]. There is, therefore, considerable research focussed on understanding the virus-antibody dynamics that enables the development of these bNAbs[40,42,130–135]. In addition, the mutations that facilitate escape from bNAbs are also extensively studied[42,44,64,65,102,134].

One potential escape pathway is the loss of a glycan targeted by a specific bNAb. As example, glycans N301, N332 and, to a lesser extent, N334 are targeted by the V3-loop glycan-binding bNAbs. However, the loss of a glycan may afford an opportunity for strain-specific neutralising antibodies, that target the residues previously shielded by this lost glycan, to bind and neutralise. Since these antibodies that target the V3-loop glycans are candidates for both passive infusion, as therapy, and induction by a

future vaccine, it is important to evaluate the tolerance of HIV-1 Env glycoproteins for the loss of the targeted glycan/s. The comparison between the resistance of mutated and wild-type HIV-1 Env glycan shields to frequently elicited neutralising antibodies, facilitates this assessment.

For this reason, collaborators at the University of Cape Town (working in the CPGR[b] laboratories at the time) investigated the key glycans around the base of the V3-loop, at positions 301, 332 and 334, during an *in vitro* study of two subtype C strains, CAP45.2.00.G3 (further referred to as CAP45.G3) and Du156.12, which share approximately 90% sequence identity. During this study, the glycan shield was remodelled and tested against a panel of sera (n=64) from chronically HIV-1 infected individuals. This was done to assess whether the epitopes of frequently elicited neutralising antibodies were exposed on the remodelled shield. It was shown that the Du156.12 strain required glycan N301 for maintenance of its glycan shield, whereas the CAP45.G3 strain was, comparatively, less reliant on glycan N301, since it remained resistant to the panel of sera despite the loss of this glycan. However, the differences between the glycan shields that bring about the distinct tolerances of these strains, for the loss of a glycan, remained unknown.

As described in section 1.1.1, computational analysis complements laboratory studies where structures are required but experimentally undetermined. For the described study, our research group carried out the computational analyses to predict the CAP45.G3 and Du156.12 Env glycoprotein trimers. In this chapter, the portion of our published manuscript, Moyo *et al.*[53], focussing on the computational analyses is included[c] and described in further detail (section 2.3). The primary focus was to describe the differences in the glycans shields of the CAP45.G3 and Du156.12 models, however, several steps preceded the final analyses and are also included in this chapter (section 2.2.1 and 2.2.2).

[b] *Center for Proteomic & Genomic Research (CPGR), University of Cape Town*
[c] *Text adapted from peer reviewed publication Moyo* et al.[53]

2.2 Methods

2.2.1 Protein Structural Modelling

In order to determine the unknown 3D protein structures of the CAP45.G3 and Du156.12 protein sequences, homology modelling was performed with the computer program MODELLER[136,137]. The MODELLER algorithm predicts 3D structures of proteins by satisfying spatial restraints. These restraints are derived, by MODELLER, from known, related, structures (i.e. templates), and from the alignment between the template and target sequences, in this case either the CAP45.G3 or Du156.12 sequences. These spatial restraints can be placed on distances, angles, dihedral angles, pairs of dihedral angles and other spatial features defined by atoms, or pseudo-atoms. Subsequently, a 3D structure is obtained by optimising a molecular probability density function. Thus, during homology modelling, MODELLER predicts the 3D structure by performing the following operations:

1. Determine the spatial restraints on the target sequence from its alignment with the templates.
    a. Generate the molecular topology for the target sequence.
    b. Calculate the coordinates for atoms with equivalent atoms in the templates by using the average coordinates over all the templates.
    c. Build the remaining coordinates using internal coordinates from the CHARMM topology library.
    d. Determine the restraints (stereo-chemical, homology-derived and spatial).
2. Predict a model that satisfies, to the largest extent, the restraints determined in step 1.
    a. Randomise the initial structure determined in 1.b. and 1.c. by adding a random number to the atomic coordinates.
    b. Optimise the model.
3. Determine the remaining restraint violations.
4. Write the violations and final model to separate files.

Therefore, in order to use MODELLER to predict the unknown 3D structures, a known related structure and its alignment with the target sequence is required.

2.2.1.1 Structural templates for HIV-1 sequences

Since MODELLER requires a template structure, an appropriate template or combination of templates for the HIV-1 Env sequences, was sought. At the time of the analysis (2015/05/22), there were 108 protein data bank (PDB) structures identified by BLAST (www.rcsb.org) that matched the HIV-1 HXB2 gp120 protein sequence. Most of the structures were modified Env proteins with shortened, or absent, variable loop sequences (see section 1.1.1). Therefore, in order to evaluate each of these structures, their protein sequences were aligned to HIV-1 HXB2 gp120 and the regions where the amino acids structures were unresolved, were highlighted (Figure 2.1). Furthermore, the structures were ordered based on their structural resolution, from highest to lowest (Figure 2.1 A).

The PDB entries 2QAD[138], 2B4C[139], 4TVP[87], 4NCO[91] and 3J5M[115] were the only structures with resolved V3-loop regions (Figure 2.1 A). Additionally, the PDB entries 4TVP[87], 4NCO[91] and 3J5M[115] were the only structures with resolved V1V2-loop regions and, of these, 4TVP[87] had the highest resolution (Figure 2.1 A). Therefore, the 4TVP[87] crystal structure was the primary candidate as a homology modelling template. However, since there were some unresolved amino acids, a combination of 4TVP[87] with 4NCO[91] and 2B4C[139], that adds 3 and 11 resolved amino acids, respectively, in the V4-loop region of the 4TVP[87] structure, was also considered (Figure 2.1 B). The HIV-1 Env sequences used for crystallisation, and any structures that were co-crystallised with these proteins, are shown in Table 2.1. The co-crystallised structures were removed prior to modelling.

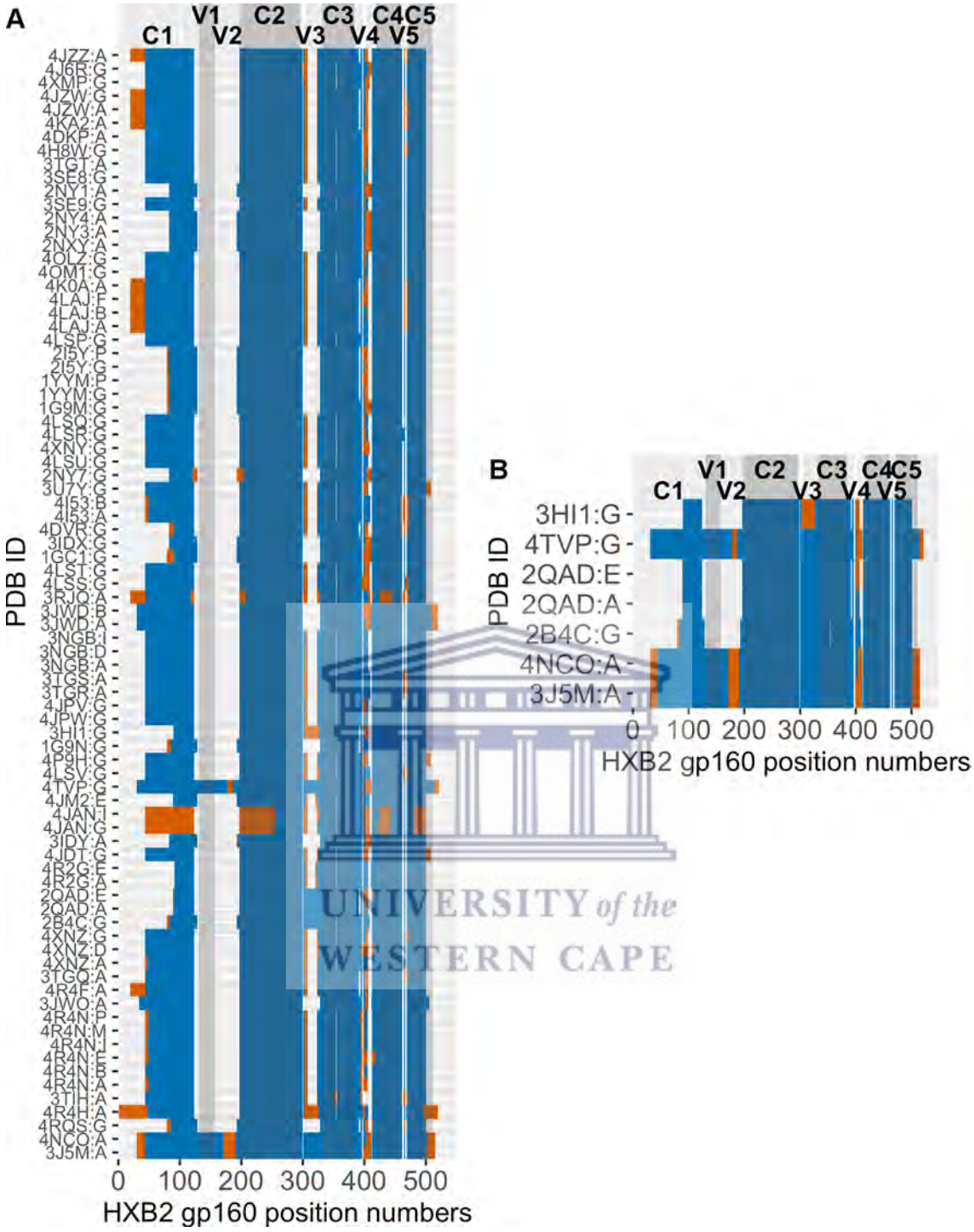

Figure 2.1: Overview of the HIV-1 Env crystallised structures

Alignment of PDB protein sequences to the HIV-1 HXB2 gp120 protein sequence for (A) all and (B) a subset of the PDB structures (PDB ID:CHAIN) identified by BLAST. Regions where the amino acids structures were unresolved are highlighted (orange). PDB structures are ordered from highest to lowest resolution. The gp120 conserved (C1-C4) and variable (V1-V5) regions are labelled and shaded to indicate the borders of each region.

Table 2.1: Crystal structures considered for modelling.

PDB ID	Sequence name	Co-crystallised structures
4TVP[87]	BG505 SOSIP.644	PGT122 and 35O22
4NCO[91]	BG505 SOSIP.644	PGT122
2B4C[139]	JR-FL	CD4 and X5

2.2.1.2 Comparing template combinations

In order to decide whether to use only 4TVP[87] as the template, or in combination with 4NCO[91] and 2B4C[139], the different template combinations were used to generate 3D models and the results compared. The comparison relied on the assignment of numerical scores to the generated structures, which are indicative of how similar the model is to the native structure. Generally, the native structure has the lowest free energy (under the native conditions, e.g. temperature and pressure), however, the evaluation of an accurate free energy function is computationally expensive[140]. An alternative approach, where a scoring function is constructed with a global minimum that corresponds to a sample of native structures deposited in the Protein Data Bank, is commonly used[140–144]. The reliance of this scoring function on known protein structures led to the term knowledge-based, or statistical, potentials[140]. The discrete optimised protein energy (DOPE) is such a knowledge-based scoring function[140] and is implemented in MODELLER. Lower DOPE scores represent models that are more native-like and a representation of the DOPE score by residue (DOPE score profile) further facilitates a detailed analysis of different template combinations[140].

Hence, ten homology models of both wild-type protomers (CAP45.G3 and Du156.12) were generated for each of the three different template combinations: 4TVP[87] alone; 4TVP[87], 4NCO[91] and 2B4C[139]; and 4NCO[91] and 2B4C[139]. The last combination, 4NCO[91] and 2B4C[139], was included to evaluate a 4TVP-independent template combination. For the comparison between the template combinations, only the best (lowest DOPE score) models generated for the wild-type CAP45.G3 and Du156.12 protomer sequences were considered. For the best models, the per residue DOPE scores were calculated and normalised for the number of restraints, since the model

generated using 4TVP[87], 4NCO and 2B4C[139] as templates has more restraints than those generated by using only 4TVP[87] or 4NCO[91] and 2B4C[139] (Figure 2.2).

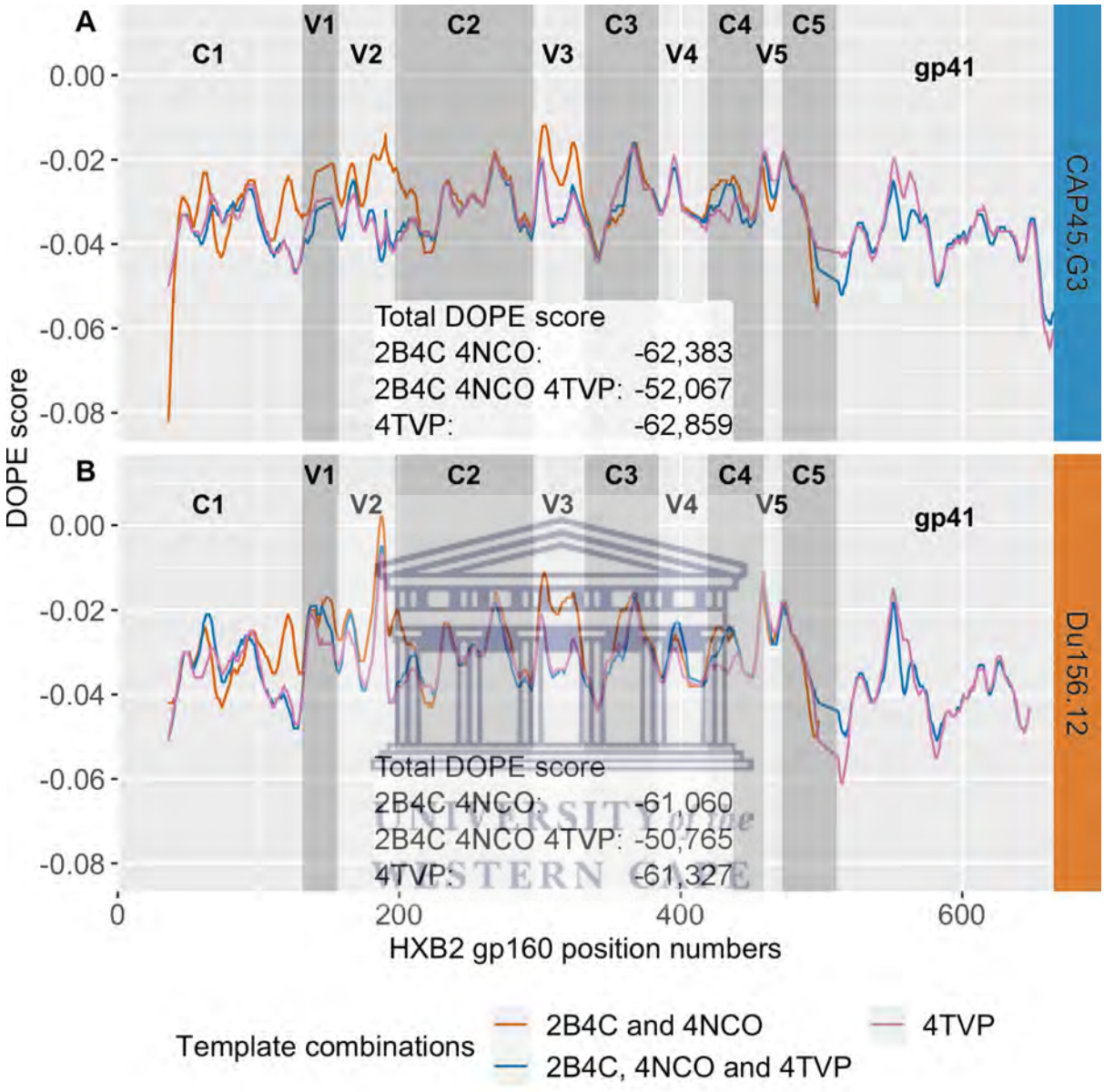

Figure 2.2: Comparison of protein-template combinations

The DOPE score profiles of the best (A) CAP45.G3 and (B) Du156.12 protomer models, using different template combinations. The total DOPE score for the best model is shown for each template combination. The gp160 conserved (C1-C4), variable (V1-V5) and gp41 sequence regions are labelled and shaded to indicate the borders of each region.

Despite the additional structural information in the V4-loop region, provided by the inclusion of 4NCO[91] and 2B4C[139], the DOPE score profile values were similar to those of the models generated with 4TVP[87] as the only template (Figure 2.2). Conversely,

including 4TVP[87] as one of the templates in the combined example markedly improved the DOPE sore profile values for the V1V2-loop (Figure 2.2 A) and V3-loop residues (Figure 2.2 A and B). Furthermore, an interesting difference in the DOPE score profiles was observed between the V1V2-loop regions of the CAP45.G3 and Du156.12 protomer models (Figure 2.2). The Du156.12 models had higher DOPE values for these regions, which was possibly due to the longer variable loop regions of the Du156.12 sequence and, hence, higher uncertainty in the modelling.

While the models generated with only 4TVP[87] as the template were the best models based on the DOPE scores, the addition of 4NCO[91] and 2B4C[139] was a more cautious approach as it included further information for modelling and did not substantially increase the DOPE scores or the DOPE score profiles (Figure 2.2). Hence, 4TVP[87], 4NCO[91] and 2B4C[139] were used as templates for generating the 10 models of the wild-type CAP45.G3 and Du156.12 protomer sequences. Each protomer model was then individually triplicated, and the three copies were aligned to the protomers of the 4NCO trimer structure to generate the trimeric models.

Thereafter, each of these trimeric models (20 in total) were duplicated and a computational mutation was introduced to the duplicated models. Specifically, the duplicated models of the CAP45.G3 strain were computational mutated such that the PNGS at position 334 was shifted to position 332. For the Du156.12 duplicated models, a computational mutation was introduced that shifted the PNGS at position 332 to 334. Therefore, the final sets included 10 Env trimer models each for the CAP45.G3 wild-type, Du156.12 wild-type, CAP45.G3^{332} variant (PNGS shifted from position 334 to 332), and Du156.12^{334} variant (PNGS shifted from position 332 to 334) strains.

2.2.2 Glycosylating the protein models

In order to computationally glycosylate the 40 trimeric models, PNGSs were determined by identifying the Asn-X-Thr/Ser motifs, where X is any amino acid except a proline[45]. A glycan was attached, where possible, to each PNGS of the trimer protein models, with the exception of the second PNGS that occurred within an NNTT, or

similar, motif. At the start of the research project, the question was raised as to whether two glycans can be present within this kind of grouped N-linked glycosylation site, since glycans are attached in a linear fashion in the Golgi and the attachment of the first may exclude the enzymes from attaching the second[d]. This decision was pertinent for the CAP45.G3 strain, since glycans N334 and N335 occurred within a NNST motif. However, since glycan N334 is important in the context of bNAbs[42] and is relatively more abundant than glycan N335[47], glycan N334 was glycosylated instead of glycan N335. Moreover, for consistency, this handling of neighbouring glycosylation sites within a partially shared PNGS motif was applied to all similar motifs, i.e. a glycan was only attached at the first Asn residue.

During the computational glycosylation, glycans were attached to the protein structures using a prototype tool under development for the Glycam-Web suite of web tools[145]. This tool adjusts the conformation of the input glycan such that there are no overlaps between the glycan and any previously added glycans or protein residues. The tool achieves this by executing the following sequence of operations:

1. Superimpose an input glycan onto the target glycoprotein structure.
2. Determine if there are overlapping atoms between the superimposed glycan (step 1.) and any other atoms of the target glycoprotein:
    a. If an overlap is present, calculate the glycan residue with the greatest overlap.
    b. For the glycan residue with the greatest overlap, determine and trace the sequential glycosidic linkages back to the Asn-GlcNac linkage located at the base of the glycan.
3. Find a glycan conformation, for the superimposed glycan, that has no overlap with any atoms of the target glycoprotein.
    a. For each glycosidic linkage identified during the trace-back (step 2.b.), starting with the Asn-GlcNac linkage:
        i. Determine the number of dihedral angles present in the glycosidic linkage.

[d] *Recent unpublished mass spectrometry data shared by a collaborator, demonstrated that the attachment of a second glycan in these motifs is possible.*

ii. Increment each dihedral angle by plus or minus 5°, within a 20° range.
iii. Generate “new” conformations by considering all the permutations of these angles, for example 2 dihedral angles would each have 8 new angles (step 3.a.(ii)), resulting in 64 (8x8) new conformations.
iv. For those conformations determined in step 3.a.(iii), proceed with the conformation with the least number of overlaps.
v. Repeat steps 3.a.(i)-(iv) for the next glycosidic linkage until a conformation with no overlaps is found.

4. If no solution is found proceed to the next input glycan and repeat.

The input glycans required by the attachment tool were generated using the carbohydrate builder available on the GLYCAM web server[145]. The carbohydrate builder generates the different rotamers of the specified glycan and then creates four copies of each in order to, in turn, link each glycan rotamer to one of the four Asn rotamers. For example, the four different rotamers of a Man5GlcNAc2 glycan (Figure 2.3) are each attached to one of the four possible Asn residue conformations, resulting in sixteen conformations, which are then used as input for the attachment tool.

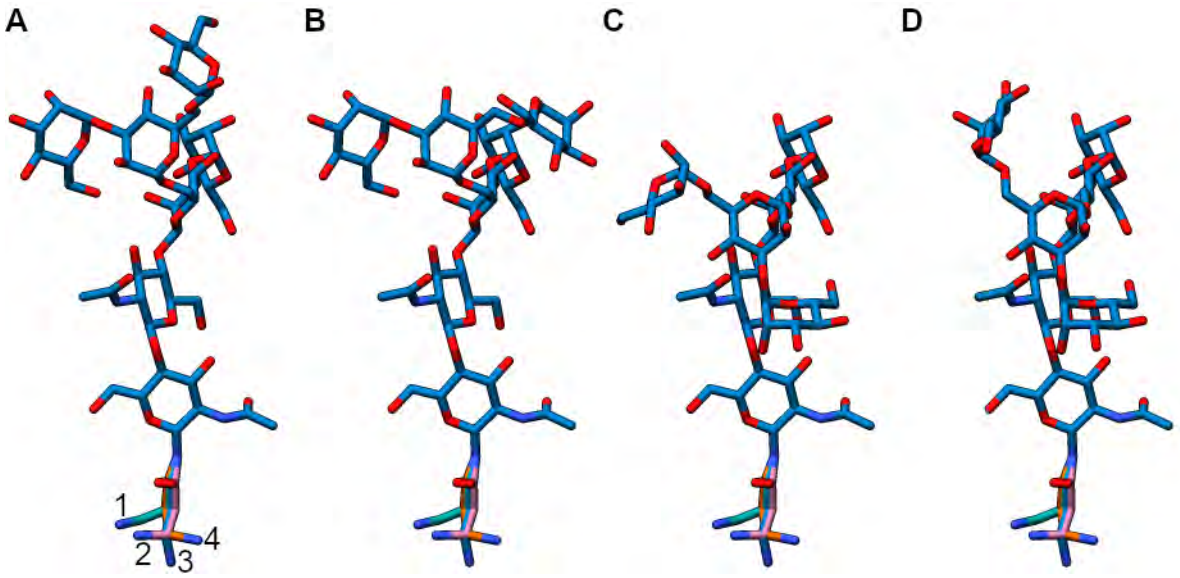

Figure 2.3: The Man5GlcNac2 rotamers

Each conformation (A-D) represents one of the four Man5GlcNac2 rotamers as well as its attachment to one of the four Asn rotamers (A1-4). The oxygen (red) and nitrogen (blue) atoms of each conformation are coloured.

While the current implementation of the attachment tool is remarkably efficient at attaching glycans, it is still limited to rotating and calculating overlaps for a single glycan at a time. Therefore, to ensure that the maximum number of glycans were attached to the CAP45.G3 and Du156.12 trimer models, the order of attachment as well as different glycoforms were taken into consideration.

2.2.2.1 Glycan attachment order

Since only a single glycan is considered at a time, previous, successfully added glycans are obstacles for the addition of subsequent glycans. This is of concern when previously added glycans block the current glycan by occupying the available space around the targeted PNGS. Therefore, the order in which glycans are attached can impact the total number of glycans successfully attached to a protein structure. To computationally glycosylate the majority of sites, those PNGSs with the least space available to accommodate a glycan should be considered first. Here, the SASA values for all the PNGSs of the two models (CAP45.G3 wild-type and Du156.12 wild-type) with the lowest DOPE scores were determined using a 3Å probe, which approximates the size of a monosaccharide[99] (Figure 2.4). Subsequently, the order of the

computational glycosylation was set by these SASA values, glycosylating those sites with the lowest SASA values first (Figure 2.4).

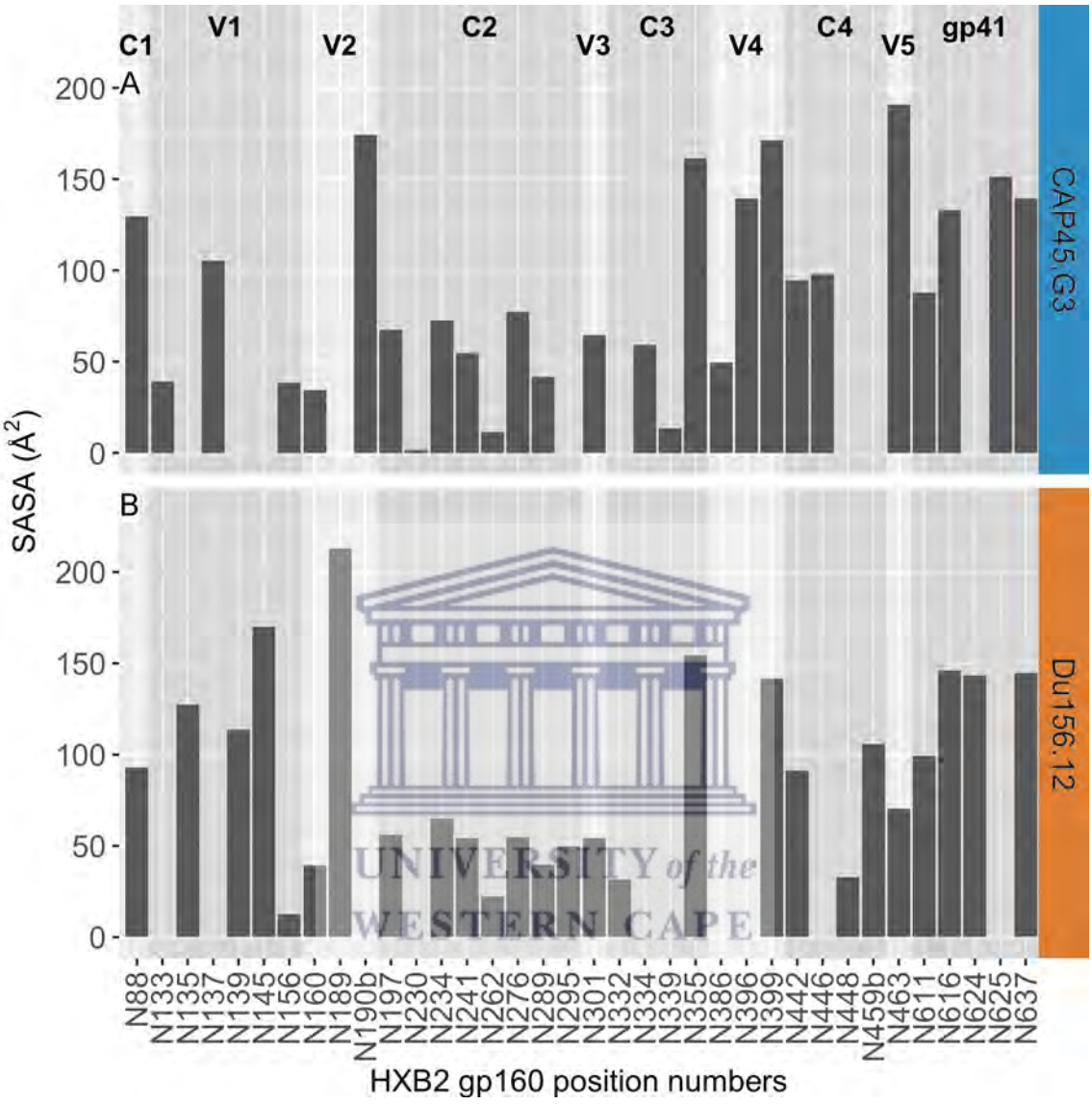

Figure 2.4: SASA of each PNGS of the CAP45.G3 and Du156.12 models

The SASA (Å^2) for each PNGS of the (A) CAP45.G3 and (B) Du156.12 wild-type models. The gp160 conserved (C1-C4), variable (V1-V5) and gp41 sequence regions are labelled and shaded to indicate the borders of each region.

2.2.2.2 Assessing the attachment efficiency of different glycoforms

Apart from the attachment order, the total number of glycans attached can also be affected by the glycoform, i.e. the type of glycan. This is due to the size difference between glycoforms. Since larger glycoforms (e.g. Man9GlcNAc2) occupy more space

than smaller ones (e.g. Man5GlcNAc2), the proximity of HIV-1 Env glycans may result in more unsuccessful attachments for the larger glycoforms than for the smaller glycoforms. Here, three glycoforms (Man5GlcNAc2, Man8GlcNAc2 and Man9GlcNAc2; Figure 2.5) were attached to the CAP45.G3 and Du156.12 trimer models (generated using the lowest DOPE score protomer models) according to the attachment order detailed in section 2.2.2.1.

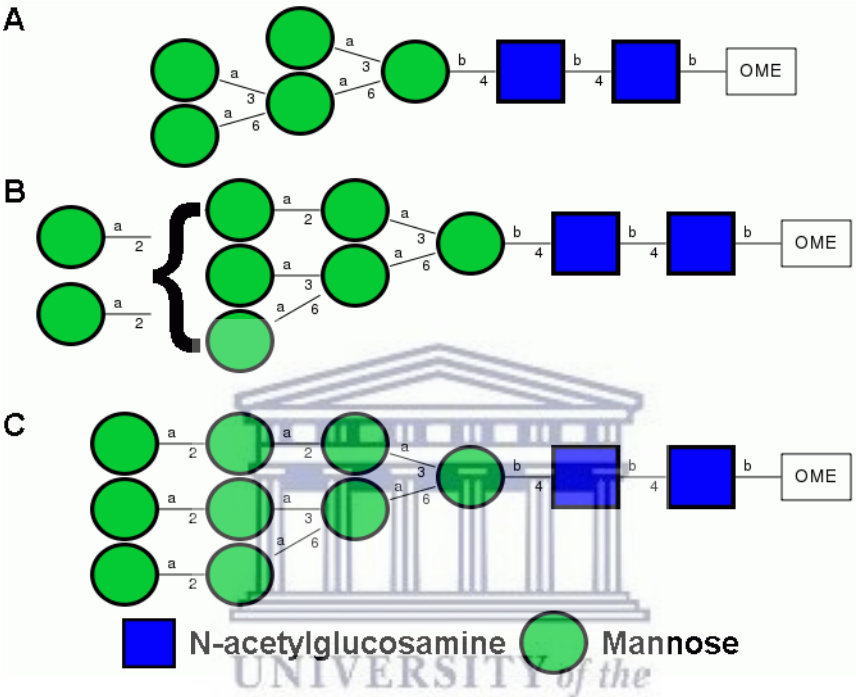

Figure 2.5: Glycoforms consdered during computational glycosylation

Representations of the (A) Man5GlcNAc2, (B) Man8GlcNAc2 and (C) Man9GlcNAc2 glycans considered during the computational glycosylation step. Each diagram is annotated to show the carbon atom involved in the glycosidic linkage (2 – 6; section 1.2.2) as well as the α (a) and β (b) classification of each monosaccharide (section 1.2.1).

The total number of glycans attached were similar across the glycoforms for the Du156.12 model (Table 2.2), however, for the CAP45.G3 model, three less Man9GlcNAc2 glycans were attached compared to either Man5GlcNAc2 or Man8GlcNAc2 glycans (Table 2.2). However, Man8GlcNAc2 glycans were unsuitable for the primary analysis due to the ambiguity of this glycoform (Figure 2.5 B). This is also the case for hybrid and complex glycans. On the other hand, Man5GlcNAc2 and Man9GlcNAc2 are commonly used during molecular studies[86,89,103–113] as there is no ambiguity in their glycoforms (Figure 2.5 A and C), and these glycoforms can easily

be imitated during *in vitro* studies using inhibitors of the processing enzymes that govern glycosylation[93,94] (section 1.1.1). Here, the Man9GlcNAc2 glycan was preferred due its abundance as one of the predominant glycoforms identified during mass spectrometry studies on HIV-1 Env glycans[49,95].

Table 2.2: Total number of attached glycans for each glycoform.

Sequence	Total PNGSs	Man5GlcNAc2	Man8GlcNAc2	Man9GlcNAc2
CAP45.G3	81	80	80	77
Du156.12	84	76	75	75

Since multiple rotamers (Figure 2.3) of each glycoform were provided to the attachment tool, a further investigation to determine the exact conformations of the glycans that were attached to each glycoprotein model was carried out (Figure 2.6). This revealed that the vast majority of the attached glycans were the first glycan conformation (rotamer 1 with Asn rotamer 1; Figure 2.3) supplied to the attachment tool (Figure 2.6). This implies that the glycan conformations on these glycoproteins were biased towards a specific glycan conformation, which is later addressed by molecular dynamics simulations (Chapter 3 and Chapter 4).

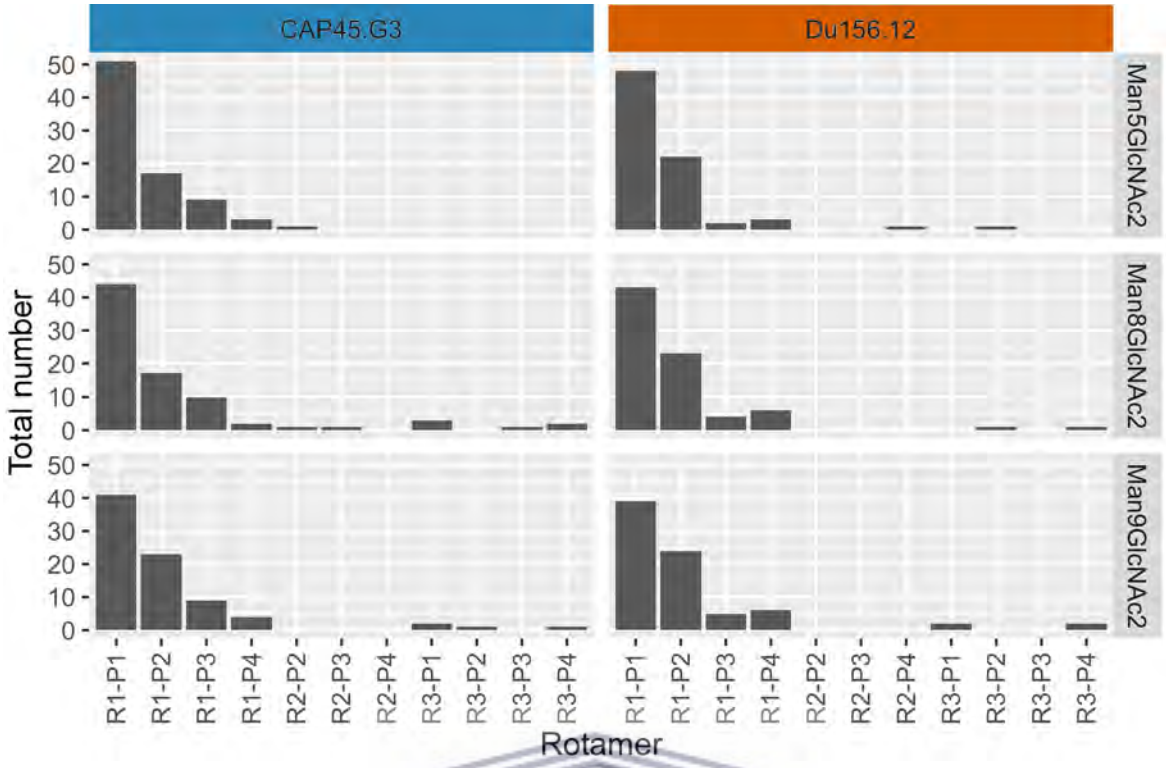

Figure 2.6: Conformations of the attached glycans

Histograms of the number of times a glycan with a specific conformation was attached to the CAP45.G3 or Du156.12 trimer models for each glycoform. The conformations are defined by the glycan rotamer (R1 to R4) as well as the Asn rotamer (P1 to P4). For example, R1-P1 indicates the first glycan rotamer (R1) combined with the first Asn rotamer (P1).

2.2.3 Predicting the exposed surface area

In order to understand the potential impact of removing glycan N301 and why the impact varied between the CAP45.G3 and Du156.12 strains, it was necessary to identify the protein residues that became exposed once glycan N301 was removed. The computationally glycosylated Env models (ten for each of CAP45.G3, Du156.12, CAP45.G3[332], and Du156.12[334]) predicted here (sections 2.2.1 and 2.2.2), facilitated this analysis. By using these models, the change in the SASA of the protein residues were determined and, in turn, compared between Env models.

Thus, for each homology model the relative SASA was calculated in the presence of glycan N301, and then glycan N301 was removed and the relative solvent accessible surface was calculated. These relative SASA calculations were carried out using NACCESS[146] and a 1.4Å probe, which approximates the size of water. The VDW radii

used for glycan atoms during the solvent accessibility calculation, were based on the force fields published by the Complex Carbohydrate Research Center version j1[122] available online (www.glycam.com). The change in relative SASA was then calculated using R[147]. Amino acid residues that showed an increase in relative SASA of 10% or greater when glycan N301 was removed, were identified. The number of occurrences of an increase of ≥10% for each amino acid was tabulated. It should be noted that each amino acid is limited to a maximum of 30 occurrences, since only ten minimised state trimer models were determined, each of which consists of three protomers and, therefore, only three duplicates of each amino acid exists. All amino acid numbers correspond to the HXB2 gp160 position numbers (Genbank accession number K03455).

To further visualise the potential differences in glycan clustering, each homology model was glycosylated a second time with glycan N301 excluded from the models. Here, therefore, the change in clustering when glycan N301 is absent and other glycans are permitted to reshuffle, was modelled.

2.3 Results

To gain further insight into the surface likely shielded by glycan N301 in each virus, the fully glycosylated Env trimer structures were modelled for the CAP45.G3 and Du156.12 wild-type sequences. Additionally, since each strain had a PNGS at either position 332 or 334 (sites known to play a role in neutralisation sensitivity), a variant of each was also considered during the study. These variants were denoted by CAP45.G3^{332} and Du156.12^{334} to indicate the variants where the PNGS, and consequently the glycan, was shifted from its wild-type position to either 332 or 334.

For each of the four target sequences, ten energy state-minimisation conformations were generated, glycosylated, and analysed. For each replicate the solvent accessible surface area (SASA) was calculated, using a 1.4Å probe, and subsequently, compared to those calculated after glycan N301 was removed (section 2.2.3). The number of occurrences of a specific protein residue, among those with a 10% or

greater increase in its SASA when glycan N301 was removed, across all the protomers of the replicates, was determined (Table 2.3).

Table 2.3: Comparison of the number of times an increase of 10% or greater was observed in the SASA measure after the removal of glycan N301.

		V3 region					C4 region		
		N301[a]	T303	I323	G324	D325	E440[c]	G441[c]	N442
332 equitable comparators	**CAP45.G3^{332} variant**	30	6	6	0	0	0	8	14
	Du156.12 wild-type	29	6	14	2	2	8	14	7
334 equitable comparators	**CAP45.G3 wild-type**	30[b]	6	6	0	0	0	8	11
	Du156.12^{334} variant	30	5	10	0	0	5	13	12

[a] HXB2 gp160 reference numbering
[b] Number of occurrences where an increase of ≥10% was observed in the SASA measure. The occurrences were counted across ten replicates each with three protomers, i.e. max=30
[c] Suspected co-receptor binding site residue

This analysis revealed that several residues were more frequently shielded by glycan N301 on the Du156.12 wild-type models compared to the CAP45.G3 wild-type models (Table 2.3). Between the 332 equitable comparators[e] (Du156.12 wild-type and CAP45.G3^{332} variant), access to three V3-loop residues (323, 324 and 325) and two C4 residues (440 and 441) were blocked by glycan N301 in a higher proportion of the Du156.12 wild-type models than in CAP45.12^{332} variant models (Table 2.3). For the 334 equitable comparators (CAP45.G3 wild-type and Du156.12^{334} variant), similar results were observed, with the exception of residues 324 and 325, which were not differentially blocked (Table 2.3). On the other hand, only residue 442 (C4 region) was more frequently shielded by glycan N301 on the CAP45.12^{332} variant models compared to the Du156.12 wild-type models, which was not the case when comparing the 334 equitable comparators (CAP45.G3 wild-type model and Du156.12^{334} variant model; Table 2.3).

[e] *Comparison between models where the glycan of the mutually exclusive PNGS pair, 332 or 334, was in the same location.*

In order to visualise the areas that were likely shielded by glycan N301, the wild-type and variant homology models were superimposed and the models of the equitable comparators, 332 (Figure 2.7) and 334 (Figure 2.8), compared. From these figures it is clear that the exposed protein residues are clustered around the N301A mutation site. Although previous reports indicated that glycan N301 often shields the CD4-binding site from antibody binding[54,66,67,69,71,148], here, these amino acids (Figure 2.7 and Figure 2.8) did not overlap with residues known to participate in CD4 binding[69,85]. Despite this, the differences in increased SASA were consistent with the *in vitro* neutralisation results[53] that demonstrated that the loss of glycan N301 resulted in varied resistance to neutralising antibodies for the two HIV-1 subtype C viruses. Specifically, the removal of glycan N301 resulted in substantial increases in neutralisation sensitivity of the Du156.12 viruses to sera from chronically HIV-1 infected individuals (18/64), while comparatively few sera (3/64) exhibited a similar increase for the CAP45.G3 viruses.

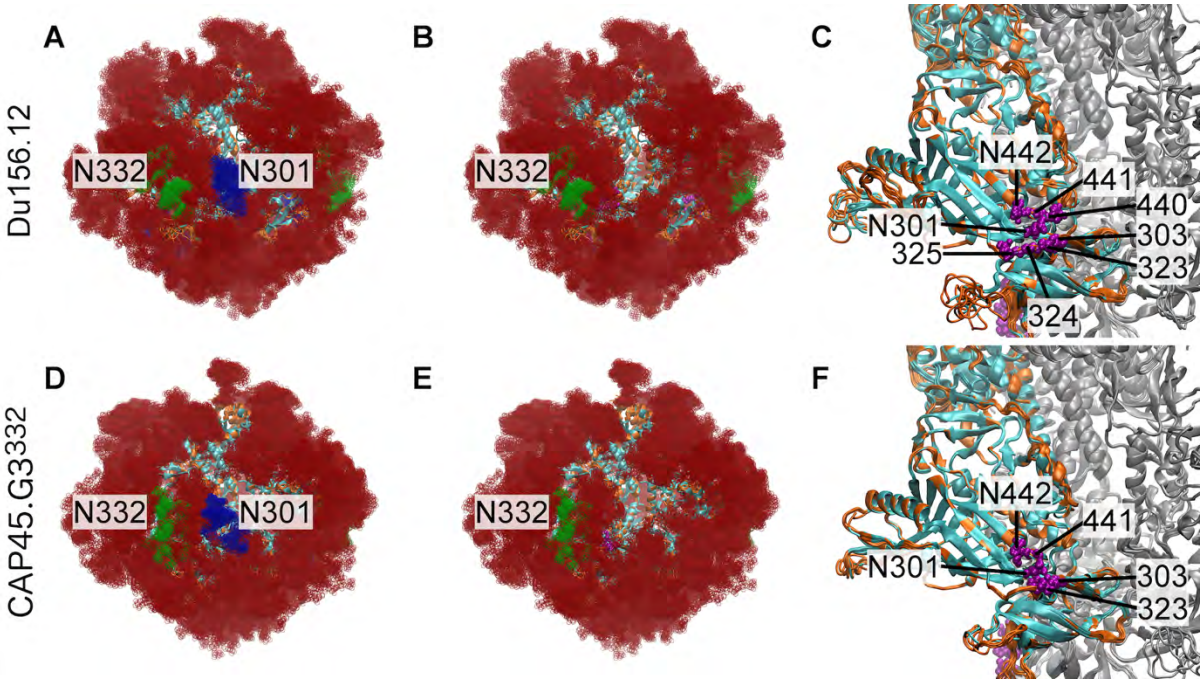

Figure 2.7: Glycosylated trimer models of the glycan-N332 variants

The predicted models of the (A-C) Du156.12 wild-type and (D-F) CAP45.G3^{332} variant sequences are superimposed to illustrate the effect of removing glycan N301 (dark blue) in the presence of glycan N332 (dark green) and other glycans (red). C and F provide further detail around glycan N301. Amino acid residues with a relative SASA greater than 30% regardless of the presence or absence of glycan N301 are shown in orange. The backbone of the gp120 subunit and the associated partial gp41 subunit are shown in cyan. The amino acid residues with a change in relative SASA greater than 10%, when glycan N301 is removed, are shown in a purple ball representation (B, C, E and F).

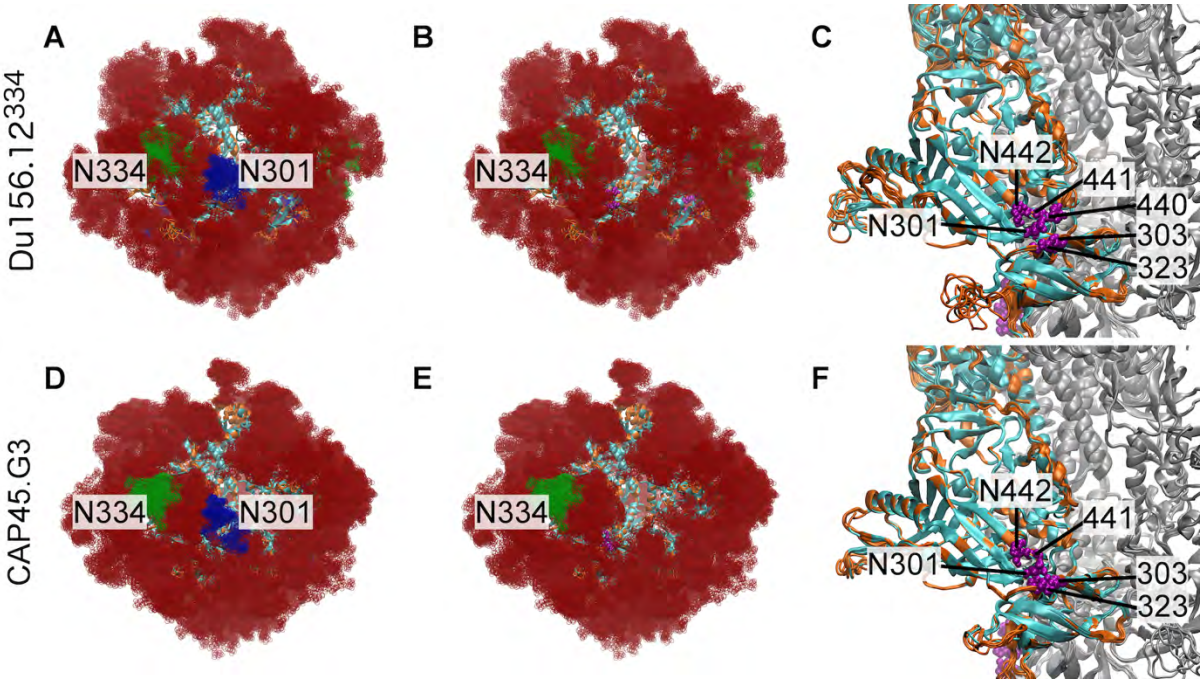

Figure 2.8: Glycosylated trimer models of the glycan-N334 variants

The predicted models of the (A-C) Du156.12^{334} variant and (D-F) CAP45.G3 wild-type sequences are superimposed to illustrate the effect of removing glycan N301 (dark blue) in the presence of the glycan N334 (dark green) and other glycans (red). C and F provide further detail around glycan N301.Amino acid residues with a relative SASA greater than 30% regardless of the presence or absence of glycan N301 are shown in orange. The backbone of the gp120 subunit and the associated partial gp41 subunit are shown in cyan. The amino acid residues with a change in relative SASA greater than 10%, when glycan N301 is removed, are shown in a purple ball representation (B, C, E and F).

Since the above SASA results did not account for the reshuffling of glycans that surround position 301 after the glycan was removed, an additional analysis was performed where glycan N301 was excluded during the computational glycosylation process in order to allow the remaining glycans to “reshuffle” on the models. Visual inspection of these glycosylated models (without glycan N301) illustrated the influence that glycan clustering may have. Apart from the space created by removing glycan N301, no clear overall differences were observed between the glycan shields (Figure 2.9). Additionally, there was also no clear difference in the volumes of space occupied by glycan N334 on the models where this glycan is present (Figure 2.9 E, F, G and H). However, glycan N332 on the CAP45.G3^{332} variant models (Figure 2.9 C) clearly occupied a broader spatial range than those on the CAP45.G3^{332} variant models without glycan N301 (Figure 2.9 D). This comparison supports the hypothesis that, after the removal of one glycan, the clustering, conformational heterogeneity, and

reshuffling of glycans are likely to play a role in the accessibility and formation of particular antibody epitopes.

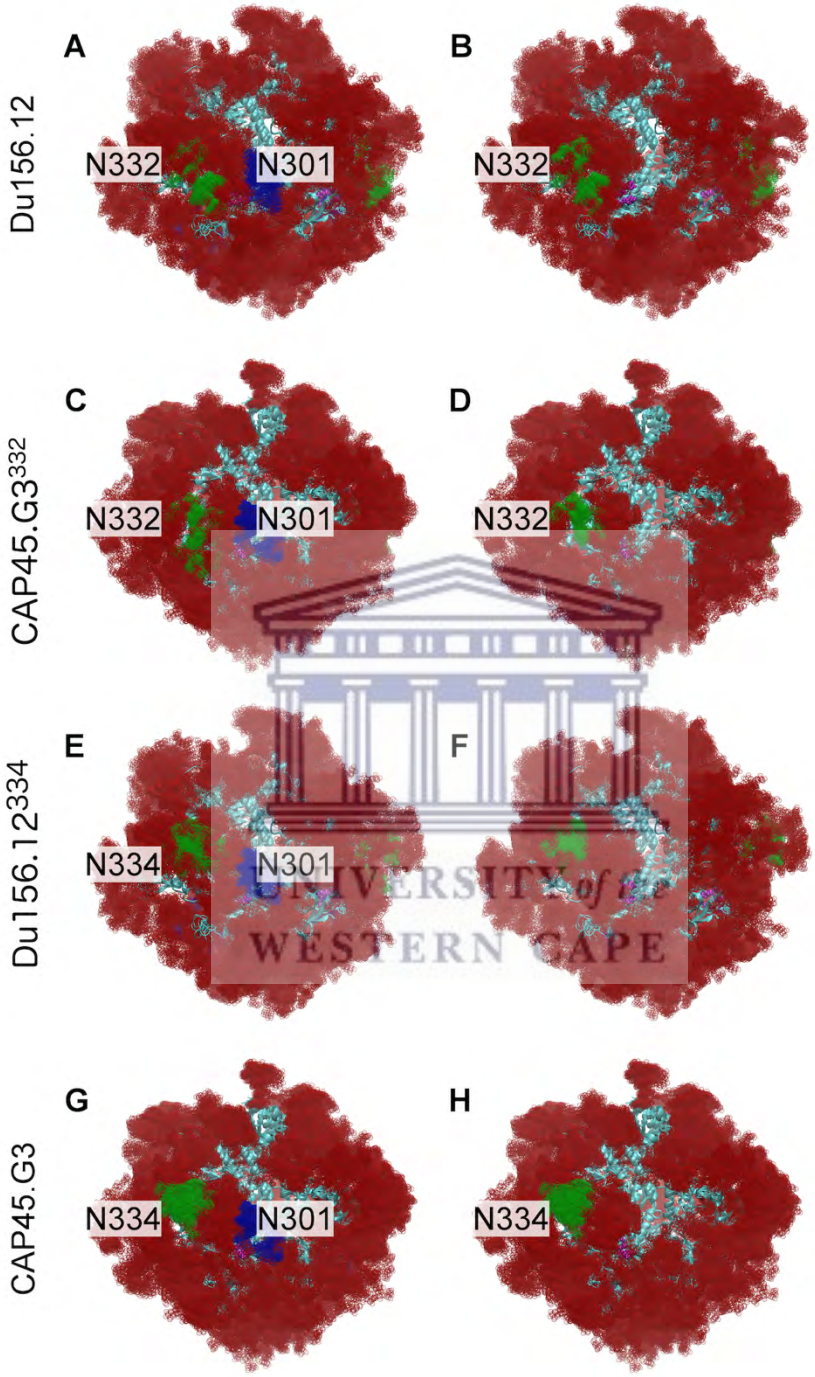

Figure 2.9: Taking glycan reshuffling into account

Superimposed structures to illustrate the effect of including (A, C, E and G) or excluding (B, D, F and H) glycan N301 prior to glycan-attachment for the two pairs, 332 and 334, of equitable comparators. Glycan N301 is shown in dark blue, glycans N332/N334 are shown in dark green, and other glycans are in red.

2.4 Discussion

Glycans comprise approximately half of the molecular weight of the gp120 subunit of HIV-1 Env and substantially shield the Env trimer from recognition by antibodies[56,57]. Conversely, glycans also often form important parts of epitopes of broadly neutralising antibodies. While the effect of these opposing selective pressures, for and against a glycan, from the different categories of antibodies has been observed in HIV-1 infected individuals[42], the phenomenon remains incompletely understood. Studying the *in vitro* changes in neutralisation sensitivity upon the removal of PNGSs (and therefore the glycan/s), along with structural modelling, can broaden our understanding of this observation.

Here, the state minimisation molecular models were modelled for the strains and their variants considered during the *in vitro* analysis, i.e. the CAP45.G3 wild-type, CAP45.G3^{332} variant, Du156.12 wild-type and Du156.12^{334} variant strains. The analysis of these state minimisation molecular models revealed a notable increase in the SASA upon removal of glycan N301 for the Du156.12 models as compared to the CAP45.G3 models, irrespective of whether the models had a glycan present at position 332 or 334 (Table 2.3). The regions more often blocked by glycan N301 in the Du156.12 glycan models were in the V3-loop and C4 regions. Although no further laboratory experiments were carried out to determine to what extent any particular residue is part of frequently recognised neutralising epitopes, the models clearly suggest that the surface area blocked by glycan N301 is larger for the Du156.12 models than for the CAP45.G3 models.

The particular residues identified that are likely to be more exposed on the Du156.12 models than the CAP45.G3 models upon removal of glycan N301, were residues 323–325 (V3) and residues 440 and 441 (C4). Amino acids 440 and 441 are thought to be part of the co-receptor binding site[149,150]. None of the identified residues formed part of the CD4-binding site, which was partially expected based upon the *in vitro* results of the presented work, and those of other studies[54,66,69,71,148]. It is possible, however, that the exposure of the CD4-binding site once glycan N301 is removed is related to changes in dynamic constraints on the conformation of the V3-loop[111], which may be

incompletely modelled here using only a state minimisation approach. Nevertheless, several key observations regarding the dynamic properties of HIV-1 Env glycans were described. Specifically, glycans adopted distinct conformations on different HIV-1 Env structures and reshuffling is likely to occur upon the removal of glycan N301, and both of these observations may affect the accessibility and formation of antibody epitopes. However, since the state minimisation models provided only limited information on the conformational heterogeneity of the modelled HIV-1 Env glycan shields, the next step was to carry out a more rigorous investigation, including molecular dynamics simulations, to analyse the spatial and temporal characteristics of these HIV-1 Env glycans.

Chapter 3 Glycans adopt distinct conformations on different HIV-1 Envelope structures

Although it is well established that the HIV-1 Env glycans form a “glycan shield” that protects the underlying protein from recognition by neutralising antibodies, Chapter 2 introduced the concept of a “glycan hole”, which refers to the notion that when glycans are removed, an opening is created in the glycan shield that can be exploited by neutralising antibodies. This chapter further investigates the concept of glycan holes, and the work is carried out on the two HIV-1 wild-type subtype C strains, which present different glycan landscapes, introduced in Chapter 2. The analysis considered two pairs of molecular dynamics simulations, where each pair comprised a simulation of a wild-type model and its N301A mutant counterpart. The analysis of these simulations revealed distinct conformations for equivalent glycans, i.e. glycans that shared an aligned HXB2 sequence position, between wild-type models. These distinct, and potentially favoured, conformations were specifically influenced by the surrounding glycan landscape, and these glycan landscapes were also important in the context of the glycan holes, since it shaped how the removal of glycan N301 impacted each glycan shield uniquely on the N301A mutant models. Due to the fact that glycan conformation were dependent on the overall landscape, glycans adopted distinct conformations on different HIV-1 Env structures, even in the case where the overall landscape deviated by only a single glycan.

3.1 Introduction

As the only surface-exposed molecule of the HIV-1 virion, the Env glycoprotein trimer is targeted by the immune system during infection. Due to the selection pressure this creates, HIV-1 *env* is the most variable gene[151–153] in an already diverse genome (section 1.1). This variability is demonstrated by the disparity in the lengths of gp120 subunits between viral strains, which range from approximately 484 to 543 amino acids[154,155]. For this reason, the HIV-1 Env amino acid positions are normalised to the HXB2 reference sequence to enable comparison across viral strains. These comparisons between viral strains have demonstrated that, despite the high levels of sequence and length diversity, many PNGSs are conserved[47]. For example, the

PNGSs at positions 88, 156, 160, 197, 230, 234, 241, 262, 276, 289, 301, 332, 356, 386, 392, 442 and 448 are abundant among subtype C strains[47]. These conserved PNGSs and their glycans compose wholly, or partially, the epitopes recognised by many of the bNAbs[58,62,63,83,127,156–165] discussed previously (section 1.1.1). While the HIV-1 Env glycans have long been considered as a shield, it is now evident that the glycan shield itself can be targeted by antibodies. Therefore, understanding the conformational properties of the Env glycans will substantially increase our knowledge of bNAb epitopes, which is vitally important in guiding vaccine design.

The structural investigation of HIV-1 Env and its glycans has traditionally fallen in the domain of X-ray crystallography[87,88,91,92] and cryo-electron microscopy[115–117]. While these studies have provided important insights into the structural features of Env and its glycans, the single static state image these studies produce lacks information regarding the dynamic nature of glycans and the glycan shield they constitute. Computational methods that provide atomic level molecular dynamics simulations offer an alternative approach with great potential for further structural investigations as already demonstrated, in particular, by two studies[86,89]. Yang *et al.*[89] showed that despite conformational heterogeneity of individual glycans over the course of a simulation, glycans can form a network of glycan-glycan interactions. In addition, the authors showed that the partial pre-organisation of the glycans potentially favours binding by bNAbs[89]. Similarly, Lemmin *et al.*[86] showed that glycans and protomer scissoring work in unison to restrict access to the underlying protein. Furthermore, the authors distinguish patches of glycans that form stable microdomains, with neutralising antibodies generally binding at the interface between these glycan microdomains[86]. While these studies provide valuable insight into key features of the glycan shield, they do not compare features between simulations of different Env glycoprotein trimers. This important comparison is the focus of the current chapter.

With the goal of defining, understanding and comparing conformational properties of the Env glycan shield, four molecular dynamics simulations of fully glycosylated Env trimers were analysed. These models included the computationally-modelled glycosylated Env structures of the two subtype C strains introduced in Chapter 2. The remaining two simulations were of the N301A mutant forms of these models. The N301A mutant simulations were carried out to add insight on “glycan holes”, a concept

that implicates the removal of a glycan in the exposure of residues of the underlying protein. Thus, while any glycan was comparable between a wild-type and its N301A mutant counterpart, only glycans that shared an aligned HXB2 sequence position (i.e. equivalent glycans) were comparable within type, i.e. between wild-type or between N301A mutant models.

The comparison between a subset of wild-type equivalent glycans revealed that both the conformations and interaction networks of these glycan were distinct between the wild-type models. This was more so for some glycans (e.g. glycans N446/N448 and N442) than others (e.g. glycans N156 and N262). For the comparison between each wild-type and its N301A mutant model counterpart (paired models), at least one glycan neighbouring the mutation site displayed an altered conformation. However, an additional, expanded comparison between the paired models demonstrated that glycans further from the mutation site also displayed relatively large conformational variations. Finally, collectively, these results, and a comparison between the N301A mutant models, directly associated the N301A mutant glycan conformation with the protein residues ultimately found to be exposed, which may explain the varied resistance to neutralising antibodies observed, *in vitro*, for the two N301A mutant strains.

3.2 Methods

3.2.1 Structural modelling and molecular dynamics simulations

The protomer structures were generated using MODELLER[136,137] (section 2.2.1). The CAP45.G3 (GenBank accession number DQ435682) and Du156.12 (GenBank accession number DQ411852) sequences were used as the targets, and three reference structures (PDB IDs 4NCO[91], 4TVP[87] and 2B4C[139]) were used as starting templates (sections 2.2.1.1 and 2.2.1.2). The modelling was repeated ten times for each strain, and models were ranked according to their DOPE scores[140] (section 2.2.1.2). The models of the CAP45.G3 and Du156.12 sequences with the lowest DOPE scores were then selected, triplicated, and the three copies were aligned to the protomers of the 4NCO[91] trimer structure to generate the trimeric model.

Thereafter, for each PNGS, where possible, a Man9GlcNAc2 glycan (Manα1-2Manα1-6[Manα1-2Manα1-3]Manα1-6[Manα1-2Manα1-2Manα1-3]Manβ1-4GlcNAcβ1-4GlcNAcβ1-) was attached using a prototype tool under development for the Glycam-Web suite of web tools[145] (section 2.2.2). However, for the CAP45.G3 model, no attempt was made to glycosylate two of the PNGSs, N335 and N678; site 335 occurs in an NNST motif and due to the importance of glycan N334 in the context of bNAbs[42] and since glycan N334 is relatively more abundant than glycan N335[47], glycan N334 was glycosylated instead of glycan N335 (section 2.2.2), and N678 falls outside the modelled domain. For the Du156.12 model, one of the PNGSs, N625, was not glycosylated, for consistency, since it occurs in an NNTT motif (section 2.2.2).

Of the remaining 81 and 84 PNGSs for the CAP45.G3 and Du156.12 models, 79 and 77 were computationally glycosylated, respectively. During the first round, glycosylation of sites N160 (CAP45.G3, protomer C as well as Du156.12, protomer B), N295 (Du156.12, protomer B), N339 (Du156.12 protomer A, B and C), N393 (Du156.12 protomer A, B and C), N399 (CAP45.G3, protomer B), N386 (CAP45.G3 protomer C) and N442 (Du156.12, protomer B) failed. However, after 30 ns simulations (according to the steps described below), a second glycosylation attempt was carried out and sites N386 (CAP45.G3, protomer C), N295 (Du156.12, protomer B) and N442 (Du156.12, protomer B) were successfully glycosylated, whereas sites N160 (CAP45.G3, protomer C and Du156.12, protomer B), N339 (Du156.12 protomer A, B and C), N393 (Du156.12 protomer A, B and C) and N399 (CAP45.G3, protomer B) remained unglycosylated. These ‘fully’ glycosylated models were then used to create the N301A mutant models by replacing the Asn residue at position 301 (HXB2 numbering) with an alanine residue and removing the glycan for both the CAP45.G3 and Du156.12 wild-type models.

The systems for simulation were created using the tLEaP package contained in AmberTools 14[123]. The ff14SB[121] force field was used for the protein and the GLYCAM06j-1[122] force field for the glycans. The wild-type and N301A mutant systems were immersed in a truncated octahedron water box containing TIP5P[124] water molecules (section 1.3.3), since TIP5P was found to produce the best quantitative agreement with experimental data[125]. The box size was set such that all protein and

glycan atoms were 15Å from the edge of the box (section 1.3.4). Chloride ions were added to neutralise the system.

The molecular dynamics simulations were produced using AMBER 14[123]. The systems were minimised by running 10,000 steps of steepest descent and 10,000 steps of conjugate gradient minimisation. During minimisation, restraints were placed on all non-hydrogen protein and glycan atoms. The systems were equilibrated by running 0.4 ns simulations under nPT (1 bar, 300 K) on a CPU cluster. During the first equilibration stage, Cartesian restraints (5.0 kcal/mol) were placed on all non-hydrogen protein and glycan atoms. Thereafter, the restraints were removed, and the equilibration was extended by another 1 ns on a GPU cluster (to ensure stability across clusters) before the production run was initiated on the GPU cluster. The 520 ns production runs were generated on a GPU cluster using AMBER GPU acceleration pmemd[123] and 0.002 ps time steps, with coordinates written to the trajectory file every 10,000 steps.

3.2.2 Analyses

3.2.2.1 Root mean square deviation (RMSD) to investigate equilibrium states

The conformational stability of each protein model was assessed by calculating the RMSD between the protein backbone atoms (C, C-alpha, N, O) of the starting structure and the structure at each time point during the production run, i.e. for each of the saved 26,000 frames during the simulation. This analysis was carried out using the whole protein as well as, separately, using only the conserved and only the variable loop regions (section 3.3.2). Equilibrium was reached by 20 ns for each system; the first 20 ns of each trajectory was therefore discarded, and further analyses carried out on the remaining 500 ns, including 25,000 frames, of each simulation (section 3.3.2).

3.2.2.2 Hydrogen-bond analyses to determine and visualise the N301-glycan clusters

Since glycans form dense clusters on the HIV-1 Env surface[88,89], the first objective was to investigate the clusters surrounding position 301 of the CAP45.G3 and Du156.12 wild-type models. Hydrogen-bond analyses were carried out to determine

which glycans formed part of the N301-glycan clusters. The hbond function of cpptraj[166] (AmberTools 14), and the default cut-off values for the distance between the heavy atoms (3Å) and the angle between the acceptor and donor atom (135 degrees), were used. For each frame in each trajectory (25,000 frames), it was noted whether hydrogen bonds were formed between any of the hydrogen atoms of two different glycans. The total number of times a hydrogen bond was formed between any two glycans was calculated, and thus, for each glycan pair (including inter-protomer pairs), the percentage time a hydrogen bond existed during the entire trajectory was determined. These percentages were further grouped ([0-25%), [25-50%), [50%-75%), and [75%-100%][f]) and plotted on a circular graph, using edgebundleR[167] and R[147].

3.2.2.3 Glycan heterogeneity of the wild-type models

The hydrogen-bond analyses revealed that the N301-glycan clusters for the CAP45.G3 and Du156.12 wild-type models varied in size and conformation. Since these differences were likely to affect the subsequent changes once glycan N301 was removed, it was important to first determine the heterogeneity of the wild-type model glycans. In order to describe the conformational differences between glycans on the CAP45.G3 and Du156.12 wild-type models, an "average" conformation of each glycan was projected, separately, onto its first principal component.

The first principal component, and subsequently the projections, were determined using the matrix, diagmatrix and projection functions of cpptraj[166]. The visualisations of the movement along the first principal component required an average conformation. This average conformation was determined by first calculating the RMSD between all the conformations of the glycan under consideration using the 2drms function of cpptraj[166]. Then, by using the upper triangular matrix this generates, the average RMSD of each conformation was calculated by averaging across the rows. The conformation with the smallest average RMSD, i.e. the conformation that was the most similar to all other conformations, was defined as the average

[f] *Standard interval notation where square, '[' and ']', and round, '(' and ')', brackets indicate whether the endpoints are included or not, respectively.*

conformation. This method of determining the average conformation was used instead of the standard method (that averages coordinates across the trajectory), since the standard method results in a distorted glycan conformation. This is due to the relatively larger variability in glycan conformations over the course of the simulation compared to those of protein residues. Once the average conformation was determined for each glycan, it could be projected onto the first principal component by adding, to its coordinates, the projection matrix multiplied by an index value. Subsequent projections were, therefore, determined by altering the index value. The number, as well as the exact numerical values, of the indices were determined by using the lower and upper quartiles of the first principal component as the interval endpoints, with the numerical values spread within these endpoints (each incremented by one). For example, if the first principal component values had upper and lower quartile values of -1 and 5, there would be seven indices with values of -1, 0, 1, 2, 3, 4 and 5. This implies that a glycan with more variable conformations and, therefore, greater variability in its first principal component was represented by more projections of its average glycan conformation than a glycan with less variability. Visual representations of these projections were generated using UCSF Chimera[168] and by setting these projection conformations to opaque, with the average conformation represented as a "solid" form.

3.2.2.4 Conformational heterogeneity between wild-type and N301A mutant model glycans

The conformational analyses of the wild-type glycans revealed several differences, which provided a proxy for the detailed analysis between the glycans of each wild-type and N301A mutant model pair. Multiple comparisons of glycan conformations, from various sources, i.e. different time points, protomers and models, were carried out in order to obtain a detailed account of the glycan conformational heterogeneity. The 2drms function in cpptraj[166], which computes the RMSD between every pair of frames of the supplied trajectory, was used, and pseudo-trajectories were generated specifically for the heterogeneity analysis. These trajectories were constructed such that the first frames were from the relevant wild-type simulation, starting with those from protomer A, followed by the frames from protomers B and C, and then followed

by the N301A mutant model simulation frames, protomer A through to C. All the frames of these pseudo-trajectories were aligned to the first frame of the applicable trajectory using the protein backbone atoms. Subsequently, to restrict the analysis to only the glycan under consideration (N156, N197, N262, N332/N334, N442 or N446/N448), all other atoms were removed from the trajectory. Therefore, running the 2drms routine on one of these trajectories resulted in an upper triangular matrix containing the RMSD calculated between all the atoms in the frames indicated by the row and column. For example, the cell in row 5 and column 10 constrained the RMSD calculated between all the atoms in frames 5 and 10. Therefore, by selecting certain combinations of rows and columns from this RMSD triangular matrix, it was possible to evaluate conformational differences across time, between protomers or between models. For example, Table 3.1 shows the RMSD matrix and the regions selected for each analyses, i.e across time (Table 3.1, blue), between protomers (Table 3.1, orange) and between the protomers of the wild-type and N301A mutant model counterparts (Table 3.1, pink). The box plots of these RMSD distributions were subsequently plotted using R[147].

Table 3.1: The 2drms upper triangular matrix and the sections used during the comparisons over time (blue), between promoters (orange), and between the protomer of the wild-type and N301A mutant model counterparts (pink).

		Wild-type protomers			N301A mutant protomers		
		A [1,25k][a]	B (25k,50k]	C (50k,75k]	A (75k,100k]	B (100k,125k]	C (125k,150k]
Wild-type protomers	A						
	B						
	C						

[a] Columns of the represented matrix.

3.2.2.5 Extended investigation using principal component analysis (PCA)

Since the initial investigation focused only on glycans that neighbour the N301A mutation, further analysis that would extend this initial work was required. Therefore, PCA was performed to investigate the conformational heterogeneity of all the glycans, as well as the protein, between the wild-type and N301A mutant model counterparts. While PCA is relatively capable of dealing with “wide” data, it is still preferable to

reduce the data to the minimal working set to circumvent the curse of dimensionality that can decay the power of the PCA. For this reason, the comparison between the wild-type and N301A mutant model counterparts were limited to the protomers, and for each protomer the protein residues and glycans were considered separately. The pseudo-trajectories for this analysis were assembled such that the first frames originated form the wild-type model simulations, followed by the frames from the relevant N301A mutant model simulation. These pseudo-trajectories were created, separately, for each protomer (A, B and C) and wild-type model (CAP45.G3 and Du156.12). Additionally, as before, the frames were aligned to the first frame of each pseudo-trajectory, using the protein backbone atoms, before the analysis was carried out. Thus, for the protein PCA, the glycan residues were removed and, similarly, for the glycan PCA, the protein residues were removed. The PCA was then performed on these pseudo-trajectories using the matrix, diagmatrix and projection functions of cpptraj[166] and graphed using R[147].

After determining that the first principal components reasonably differentiated between the wild-type and N301A mutant model frames, the factor loadings of the first principal components were determined by squaring each value of the first eigen vector. Since these values represented the factor loadings of individual atoms, they were summed to obtain the factor loadings of the total protein residues or glycans. Given that small changes in the factor loadings could potentially impact which glycans, or protein residues, were identified during the analysis, a median factor-loading value was determined instead. These medians were determined using 100 moving-blocks bootstrap replicates. The 50 ns blocks were defined such that each block could only start on whole nanoseconds to ensure adequate sampling of the whole trajectory. There were, therefore, 450 possible blocks for each of the wild-type and N301A mutant simulations, with ten random blocks required for each bootstrap replicate. The pseudo-trajectories required for each bootstrap replicate were generated by first randomly selecting 10 blocks from the relevant wild-type protomer simulation, followed by 10 blocks randomly selected for the corresponding N301A mutant protomer simulation. Moving-blocks bootstrap was used instead of a normal bootstrap approach to conserve the correlation between sequential observations.

To address the possibility that random variation caused the observed differences between the wild-type models and their N301A mutant counterparts, it was necessary to quantify each glycans' random variation. For this purpose, the conformational heterogeneity between the protomers of the same wild-type model were used to disqualify any protein residues, or glycans, with relatively large random variation. This assessment of the variation was determined in a similar way to that between the wild-type and N301A mutant models, i.e. with 100 bootstrap replicates used to determine the medians. The key difference was the pseudo-trajectory used. In this instance, the pseudo-trajectories were assembled such that the first frames originated from protomer A, followed by the frames of protomers B and C. These pseudo-trajectories were created, separately, for each wild-type simulation.

3.2.2.6 Antibody accessible surface area for protein residues

Taken together, the results from the previous analyses showed that not only did each glycan shield respond differently to the N301A mutation, but the conformational shifts were distinct for each of the N301A mutant models. Therefore, to describe these unique responses to the N301A mutation in a meaningful way, the antibody accessible surface area (AASA) was calculated. This calculation was performed using a 10Å probe (as an approximation of the size of an antibody[87]) with Naccess[146] for both the CAP45.G3, and Du156.12, wild-type and N301A mutant model simulations. The van der Waals radii of the glycan atoms were defined as described for the GLYCAM06j-1[122] force field. The AASA was calculated for 2,500 evenly spaced frames across the 500 ns trajectories.

Differences between the AASA values of the wild-type and N301A mutant models could be attributed to either the protein and/or glycan movements and, therefore, the AASA values were normalised to remove any changes in the AASA due to protein movements. This was done by determining the "base/maximum" AASA by removing the glycans and re-calculating the AASA for these non-glycosylated frames. The final AASA ratio, per frame, was the ratio of the glycosylated and non-glycosylated AASA values, calculated by dividing the AASA of the glycosylated protein by the AASA of

the non-glycosylated protein. Averages were calculated using these AASA ratios over time.

The mean AASA ratio, per residue, of the N301A mutant model simulations were compared to those of their wild-type model counterparts to determine whether there were statistically significant increases. The AASA ratio distribution of each residue, under the null-hypothesis (wild-type and N301A mutant model AASA ratio means are equal), was assessed by using 100 moving-blocks bootstrap replicates. The AASA ratio datasets for each model were mean normalised to satisfy the null-hypothesis before bootstrap replicates were drawn. These bootstrap replicates were constructed similar to those of the PCA. Variance normalisation was not performed since numerous residues had average AASA ratios (over time) equal to zero. The significant residues were further filtered to include only those where the difference between the average AASA ratio for the wild-type and N301A mutant model over time was 10% or greater.

3.3 Results

As described in section 3.1, the main goal of this chapter was to define and compare conformational features of glycans of two glycosylated Env-modelled molecular dynamics simulations. The first step towards this goal was to establish the differences between the PNGS profiles of the wild-type models (section 3.3.1), which was followed by verifying the stability of the molecular dynamics simulations (section 3.3.2) and the selection of a subset of glycans for further analyses (section 3.3.3). Thereafter, the conformations of the glycans in this subset were compared between the wild-type models (section 3.3.4) as well as between each wild-type model and its N301A mutant counterpart (i.e. paired models, section 3.3.5). The comparison between the paired models was further extended to include protein residues and all the remaining glycans (section 3.3.6). Finally, the integrity of each N301A mutant model was evaluated and compared between the strains (section 3.3.7). These results offer a detailed account of the conformational heterogeneity, *in silico*, of HIV-1 Env glycans.

3.3.1 Overview of the PNGS profiles of the wild-type models

To examine which PNGSs were contained within the modelled domains of the CAP45.G3 and Du156.12 strains, the locations of Asn-X-Thr/Ser motifs were determined (Figure 3.1). The CAP45.G3 wild-type strain contained 27 PNGSs, which included PNGSs at positions 334 and 446, whereas the Du156.12 wild-type strain had 28 PNGSs, with PNGSs at positions 332 and 448. These positions (332/334 and 446/448) represent mutually exclusive PNGSs that were inverted in the two wild-type strains. Both strains contained a PNGS at position 301, which was removed to create the N301A mutants. Although each PNGS could not be glycosylated during the computational glycosylation, it is known that there is variation in glycan occupancy, i.e. not all PNGSs are glycosylated 100% of the time[49,72–82,95,169]. Therefore, there are potentially several different glycosylated forms of Env glycoproteins, and the models used here (Figure 3.2) are representations of one of the possible forms of the Env glycoprotein of the CAP45.G3 and Du156.12 strains.

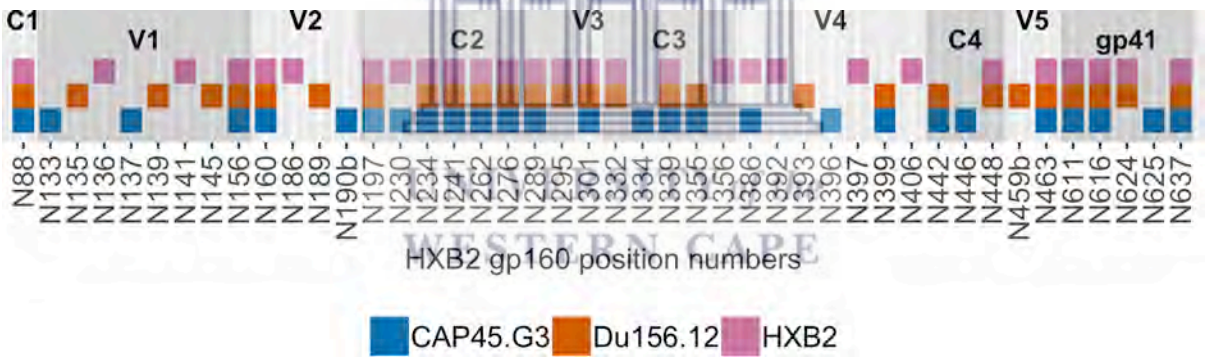

Figure 3.1: Distribution of PNGSs

PNGSs contained within the CAP45.G3 (blue) and Du156.12 (orange) strains in relation to the HIV-1 reference strain, HXB2 (pink) are shown. The gp160 conserved (C1-C4), variable (V1-V5) and gp41 sequence regions are labelled and shaded to indicate the borders of each region.

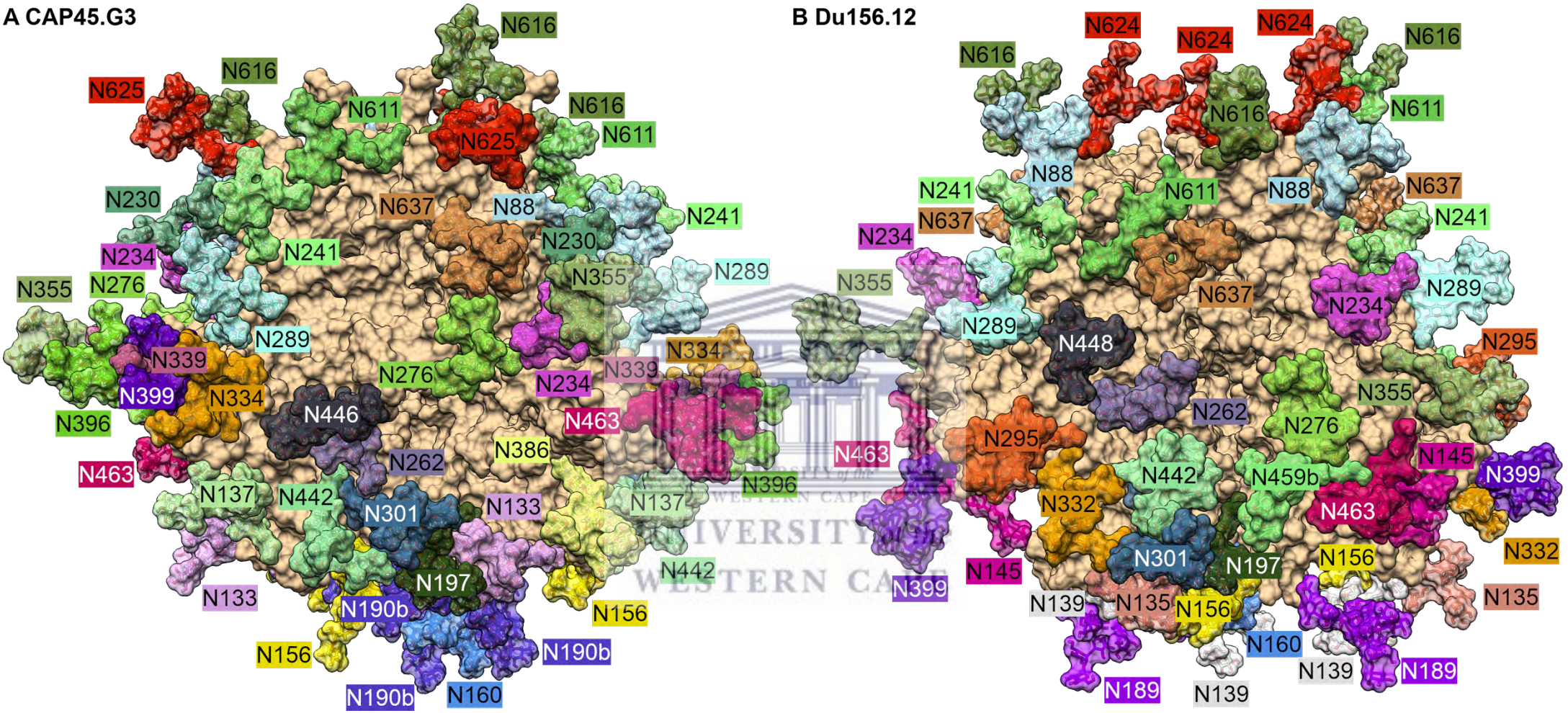

Figure 3.2: Glycan landscapes of the wild-type models

3D representation of the N-linked glycosylation sites of the (A) CAP45.G3 and (B) Du156.12 Env models. The protein residues and glycans are shown as surfaces and the glycans are labelled according to HXB2 numbering.

3.3.2 Assessing the stability of the molecular dynamics simulations

Molecular dynamics simulations evolve the conformation of a molecule forward in time. If the initial conformation is unrepresentative of an equilibrium conformation, the initial samples (i.e. frames) of the trajectory will correspond to the conformation in the tails of the equilibrium distribution. In short simulations, as analysed here, one would not expect to observe the initial conformations and this portion of the trajectory is commonly discarded. To determine which portion to exclude from the analysis, the RMSD was calculated between each frame and the starting frame for all backbone atoms (Figure 3.3; A and B). Since the HIV-1 Env variable regions (V1-V5) are generally more flexible than the conserved regions (C1-C5), the RMSD for the conserved (Figure 3.3; C and D) and variable (Figure 3.4) regions were calculated separately.

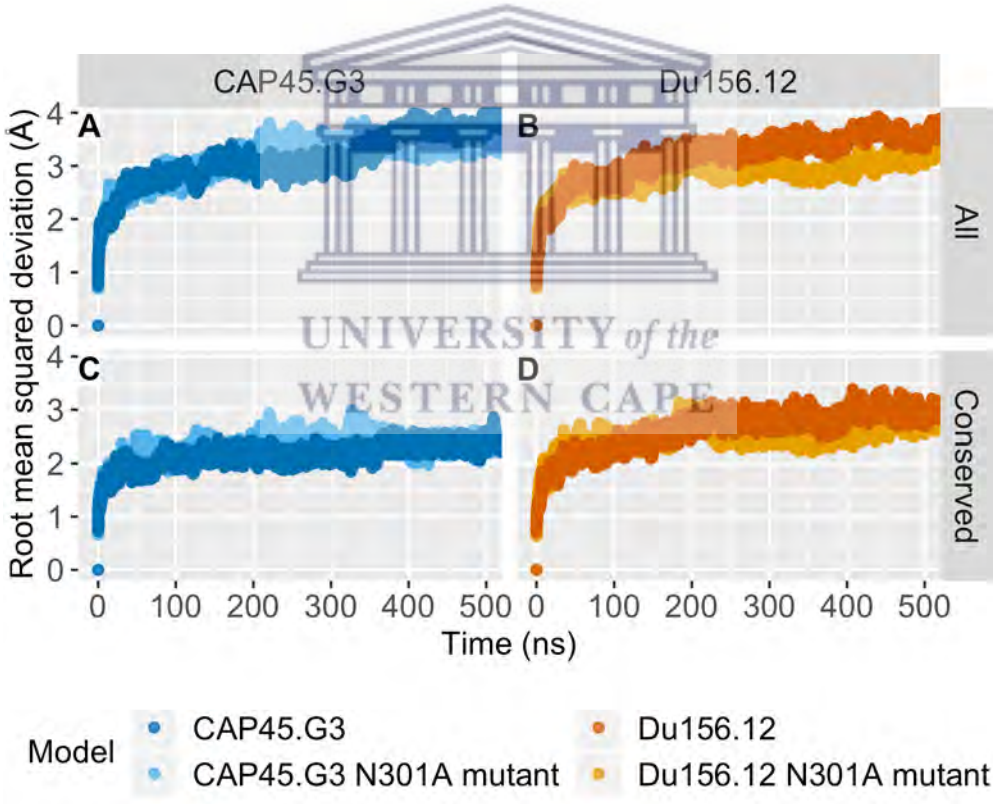

Figure 3.3: RMSD calculated for all the protein residues or only those in the conserved regions

The RMSD (Å), over time (ns), for each wild-type and N301A mutant model simulation, calculated for all backbone atoms (A and B), and only the atoms of the conserved regions (C and D). For each RMSD calculation, the first frame was used as reference.

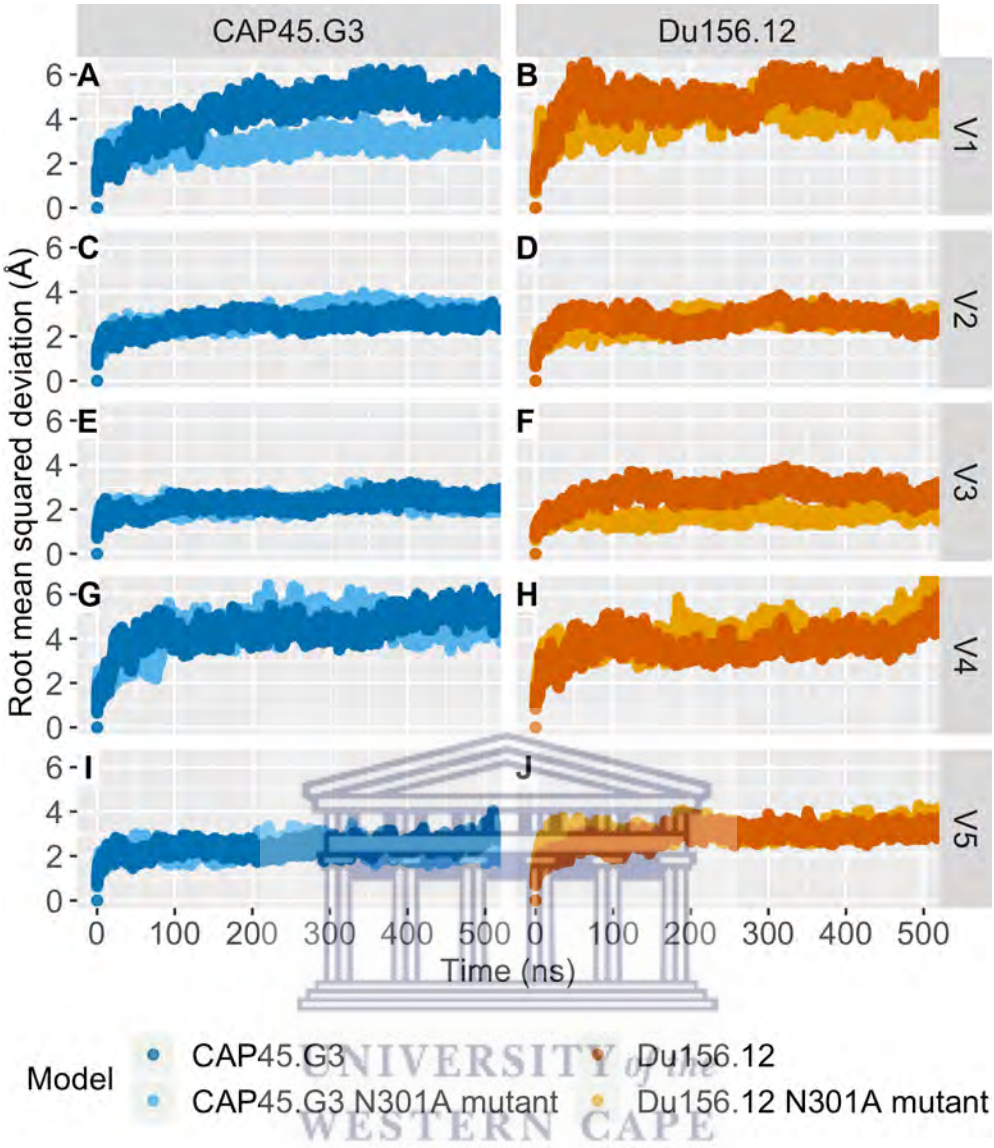

Figure 3.4: RMSD calculated for the variable regions

The RMSD (Å) over time (ns) for each wild-type and N301A mutant simulation, calculated for backbone atoms of the variable regions (V1-V5).

An upward trend over time was observed for the RMSD calculated using the entire protein (Figure 3.3; A and B); however, this trend was far less pronounced when the variable regions were removed from the analysis, i.e. calculating the RMSD using only the residues found in the conserved regions (Figure 3.3; C and D). After the first 20 ns, the RMSD of the conserved regions fluctuated between 1.5 and 3.5Å, suggesting that the conformations of the protein backbones were more stable during the latter portion of the simulation. In comparison, the variable regions fluctuated between 0.8 and 6.8Å after the first 20 ns (Figure 3.4), with substantial variation between the individual variable regions, which is expected given the large degree of variability in

the sequence lengths and composition of these regions. The relative stability of the conserved regions after 20 ns provided an indication that an equilibrium state was reached; the first 20 ns of each simulation was thus discarded, and the remaining 500 ns was used for all subsequent analyses.

There was no clear difference between the RMSD results of the wild-type and N301A mutant models for either the conserved or variable regions (Figure 3.3 and Figure 3.4), which suggests that the N301A point mutation did not cause any apparent disruptions to the conformations of the protein backbone of either the CAP45.G3 or Du156.12 models.

In summary, the RMSD analyses demonstrated that the models had relatively stable conformations after the first 20 ns, and that the N301A mutation, and loss of this glycan, did not have a discernible influence on the conformations of the protein backbones. Given these results, most of the analysis was focused on comparing the glycan (and not protein) conformations.

3.3.3 The organisation of the N301 glycan-glycan interaction networks on the wild-type models

Glycans form dense clusters on the HIV-1 Env surface[88,89] and the glycan arrangements are therefore likely to be affected by the loss of a glycan. Here, the aim was to determine the composition of the N301-glycan clusters of the CAP45.G3 and Du156.12 wild-type models. Describing these clusters allows for the subsequent investigation of how each cluster changes after the removal of one of the glycans forming part of that cluster – in this case glycan N301. Steward *et al.*[88] and Yang *et al.*[89] defined glycan clusters using their proximity to one another. However, instead of using distance-based calculations that are intractable due to time constraints, glycan clusters were determined using the glycan-glycan interaction networks (hydrogen bonds). For each glycan-glycan interaction, the proportion of frames in which the interaction is present during the simulation was calculated. These results are represented on circular network graphs that illustrate both within and between protomer (cross-protomer) interactions (Figure 3.5 and Figure 3.6).

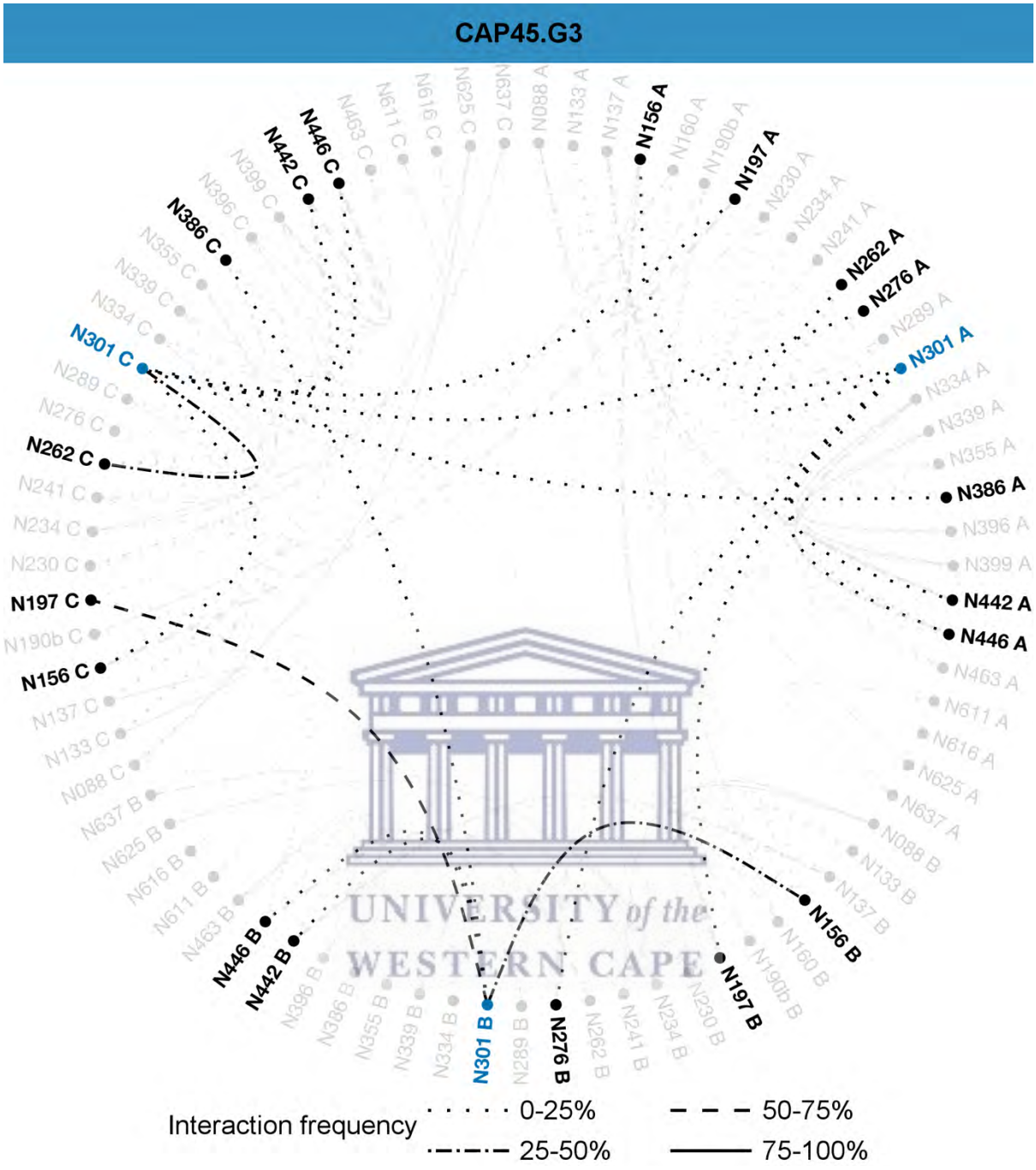

Figure 3.5: CAP45.G3 wild-type model N301 glycan-glycan interaction network

Interaction networks between glycan N301 and any other glycan on each protomer (A, B and C) are depicted. Specific line types represent different interaction frequencies (0-100%).

The interaction partners of glycan N301 on the CAP45.G3 wild-type model varied slightly between the protomers; in total seven glycans (N156, N197, N262, N276, N386, N442 and N446) were identified (Figure 3.5). Four glycans (N156, N197, N442

and N446) formed part of the N301-glycan cluster on all the CAP45.G3 wild-type protomers.

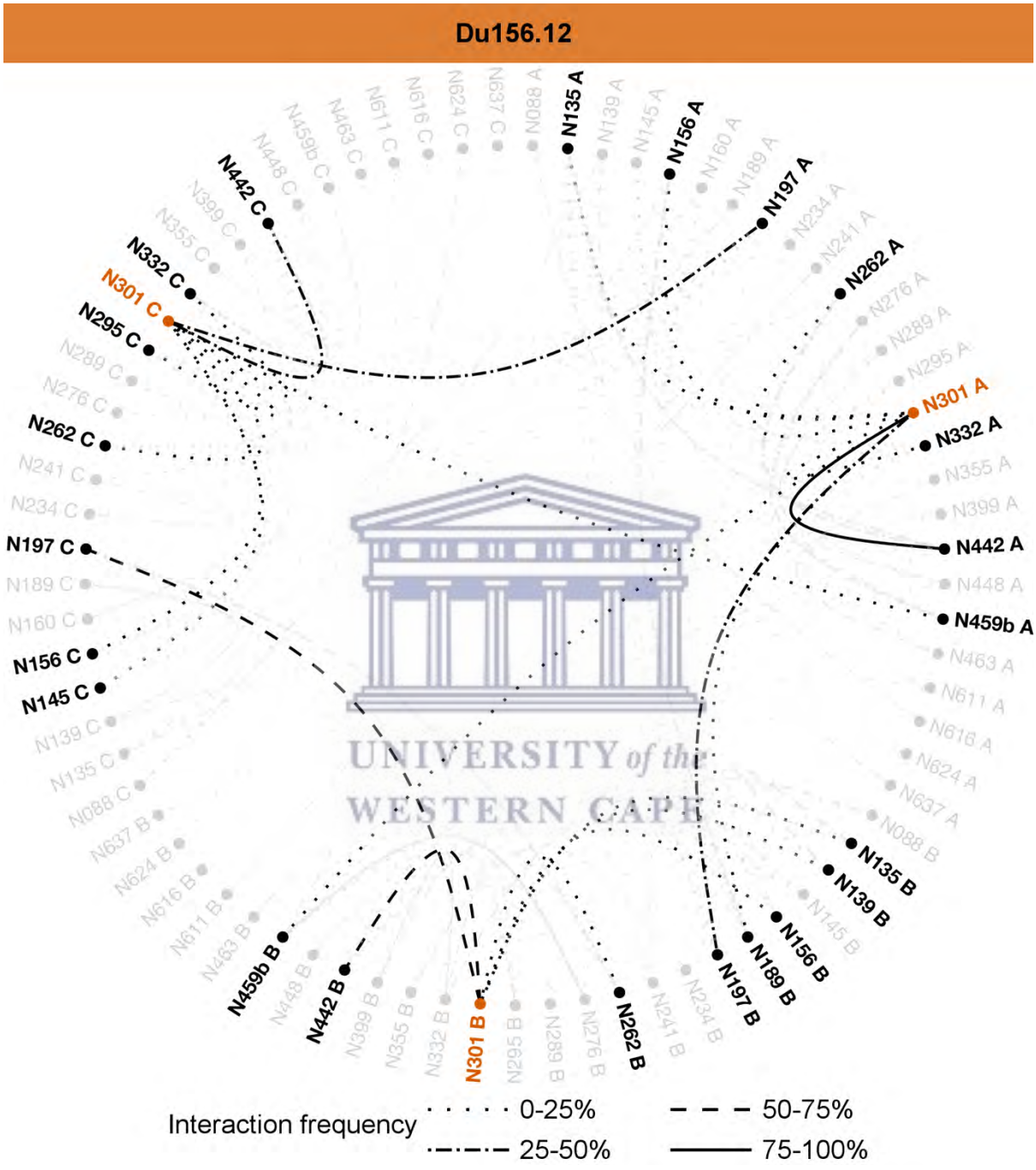

Figure 3.6: Du156.12 wild-type model N301 glycan-glycan interaction network

Interaction networks between glycan N301 and any other glycan on each protomer (A, B and C) are depicted. Specific line types represent different interaction frequencies (0-100%).

For the Du156.12 wild-type model, the N301-glycan clusters varied to a larger degree between protomers and in total eleven glycans (N135, N139, N145, N156, N189, N197, N262, N295, N332, N442 and N459b) were identified (Figure 3.6). However, despite the variation, glycans N156, N197, N262 and N442 were found on all the protomers.

Taken together, the glycan-glycan interaction network analysis showed that glycans N156, N197 and N442 were consistently found to interact with glycan N301 on both the CAP45.G3 and Du156.12 wild-type models. Additionally, glycan N446 formed part of the N301-glycan cluster on the CAP45.G3 model, and N262 on the Du156.12 wild-type model. Therefore, the loss of glycan N301 from the cluster is likely to have a pronounced effect on the conformations of each of these glycans. However, the glycan-glycan interaction network analysis only presents information on the interaction frequencies and, therefore, does not provide insight on the directionality, or shape, of each cluster. If the conformations of the N301-glycan clusters are irregular across protomers and models, the loss of glycan N301 would likely impact each cluster in a different way. Thus, to gain a better understanding of the conformational characteristics of each cluster, a preliminary analysis was carried out on the three glycans (N156, N197, and N442) that formed part of the N301-glycan clusters on both wild-type models.

3.3.4 Evaluation of the conformational variance between equivalent glycans present on the wild-type models

In section 3.3.3, the glycan-glycan interaction networks of glycan N301 were investigated and compared between the CAP45.G3 and Du156.12 Env wild-type models. Although several differences were observed, three glycans (N156, N197 and N442) were present in both wild-type N301-glycan interaction networks. The conformational differences for each of these glycans common to each distinct model are important considerations, since these differences are likely to impact the way the glycan conformations change upon removal of glycan N301.

Therefore, to investigate the extent of conformational heterogeneity between the N301-glycan clusters on each model, the extended hydrogen-bond networks of glycans N156, N197 and N442, were determined. The conformations of each of these glycans and their interaction partners were then visualised and compared between the wild-type models.

The visual representations were obtained by determining the predominant movement of each glycan using principal component analysis. Each shared glycan, and its interaction partners, were projected onto their individual, separate, first principal components. These interaction partners were restricted to include only those glycans where interaction frequencies of 10% or greater were observed during the relevant simulation. This criterion meant that glycan N301 itself was not always included in the comparison, as its interaction frequency was lower than 10% for several protomers of the wild-type models during the simulations. Detailed analysis of each of these networks highlighted several interesting conformational differences as well as important contextual characteristics, i.e. how the glycan distribution and "landscape" varied around glycan N301.

The visual representation of glycan N156 and its interaction partners illustrated that glycan conformations can be heterogeneous across models (Figure 3.7). For the N156 glycans on the CAP45.G3 wild-type model, three interaction partners were present across protomers. Glycan N190b, which is located on the Env protein apex, and either glycan N197 or N301, which appear to compete as interaction partners (Figure 3.7 A, C and E). On the Du156.12 wild-type model, the interaction partners of the N156 glycans, glycans N139, N160 and N189, were all on the protein apex (Figure 3.7 B, D and F). This distinction, where a greater number of interaction partners were available on the protein apex of the Du156.12 wild-type model, may be one reason for the differences in glycan conformations observed between the CAP45.G3 and Du156.12 wild-type models. As can be seen in Figure 3.7, the N156 glycans on the Du156.12 wild-type model displayed conformations that were all orientated more toward the protein apex than those on the CAP45.G3 wild-type model. However, the lack of apex interaction partners for glycan N156 on protomer B (Figure 3.7 D), that also displayed the distinct conformation orientated towards the protein apex, is perplexing, as it

implies that frequent interactions with the apex partners were not the sole cause for the orientation of the N156 glycans on the Du156.12 model.

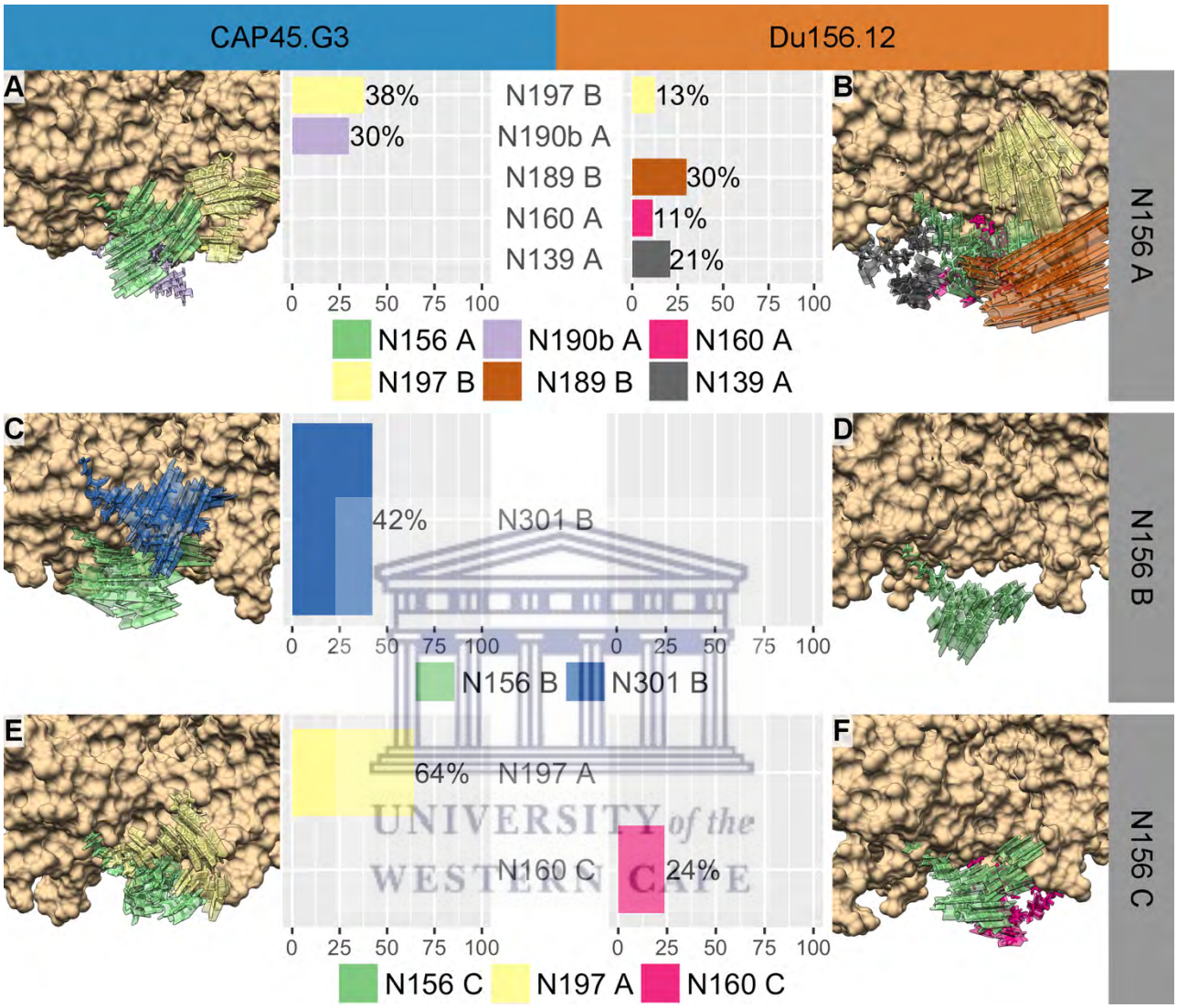

Figure 3.7: Movement and interaction networks of glycan N156

Interaction frequencies, shown as horizontal bar graphs, and 3D movement representations for the glycans forming part of the N156-glycan interaction network of each protomer during the CAP45.G3 and Du156.12 wild-type model simulations.

The next comparison, focussing on the N197-glycan interaction networks, revealed interesting differences in the intra- and cross-protomer interactions. For the CAP45.G3 wild-type model, each N197 glycan had two interaction partners. One of these partners formed a cross-protomer interaction, either glycan N156 or N301, and the other an intra-protomer interaction, either glycan N133 or N386 (Figure 3.8 A, C and E). These interaction partners effectively confine the N197 glycans, limiting their movements.

However, the N197 glycans on the Du156.12 wild-type model exclusively formed cross-protomer interactions with glycans N156, N262 or N301, depending on the protomer. Consequently, the conformations of the N197 glycans on the Du156.12 wild-type model appeared less restricted, which resulted in larger observed conformational differences between the protomers of this model (Figure 3.8 B, D and F).

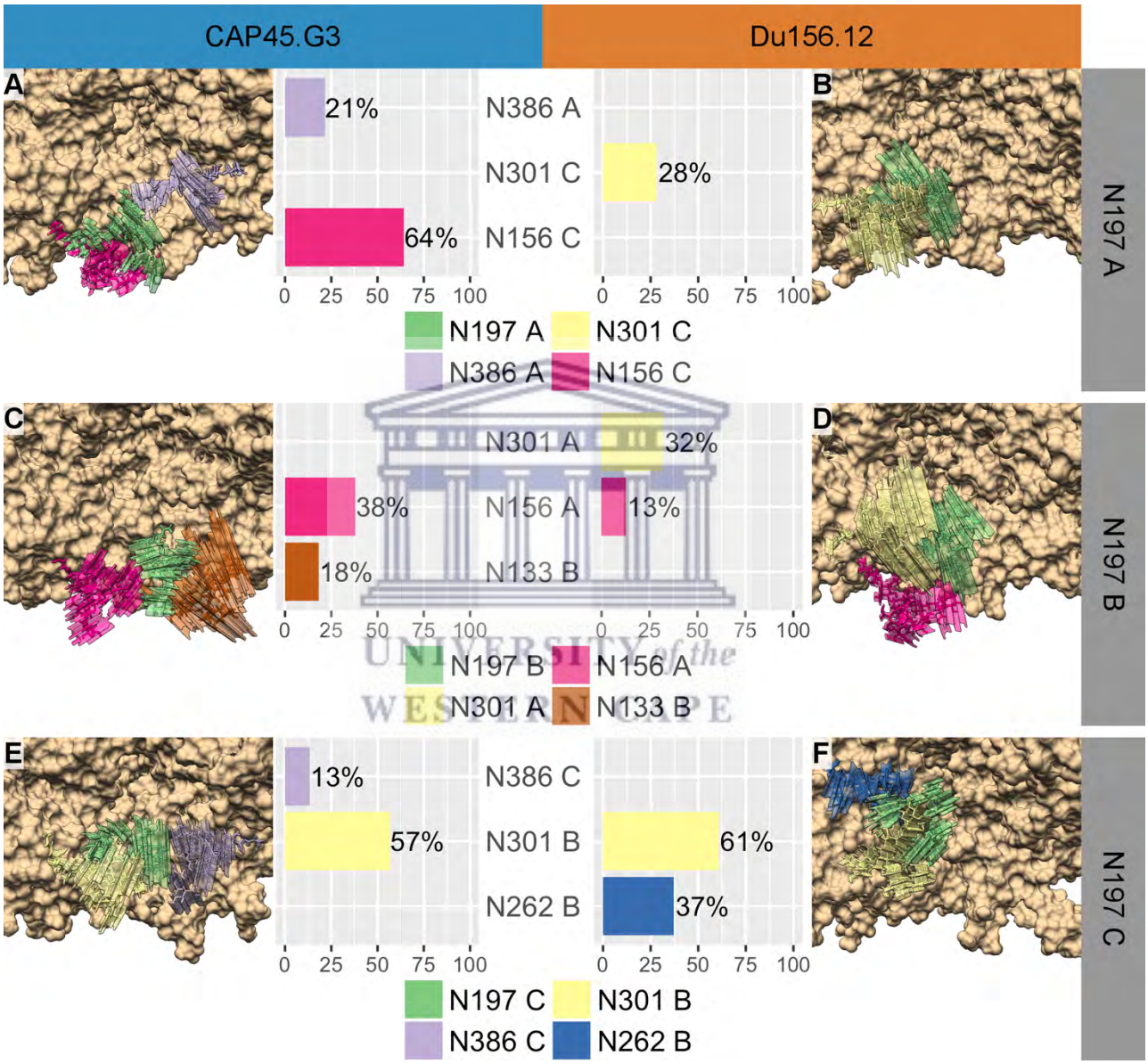

Figure 3.8: Movement and interaction networks of glycan N197

Interaction frequencies, shown as horizontal bar graphs, and 3D movement representations for the glycans forming part of the N197-glycan interaction network of each protomer during the CAP45.G3 and Du156.12 wild-type model simulations.

The final wild-type glycan interaction network comparison was for glycan N442 (Figure 3.9), which appeared to move substantially on both models compared to the N156 and

N197 glycans. The N442 glycans on the CAP45.G3 wild-type model consistently displayed conformations orientated towards the protein apex for all three protomers (Figure 3.9 A, C and E). Furthermore, glycan N137 was consistently an interaction partner of the N442 glycans, which was also observed on all three protomers (Figure 3.9 A, C and E). In contrast, the N442 glycans on the Du156.12 wild-type were orientated towards the gp41 region and displayed substantial interactions with the N301 glycans (Figure 3.9 B, D and F). These conformations may, in part, be due to the N301-N442 glycan interactions and in part due to structural hindrance by V1V2 glycans (N135 or N145).

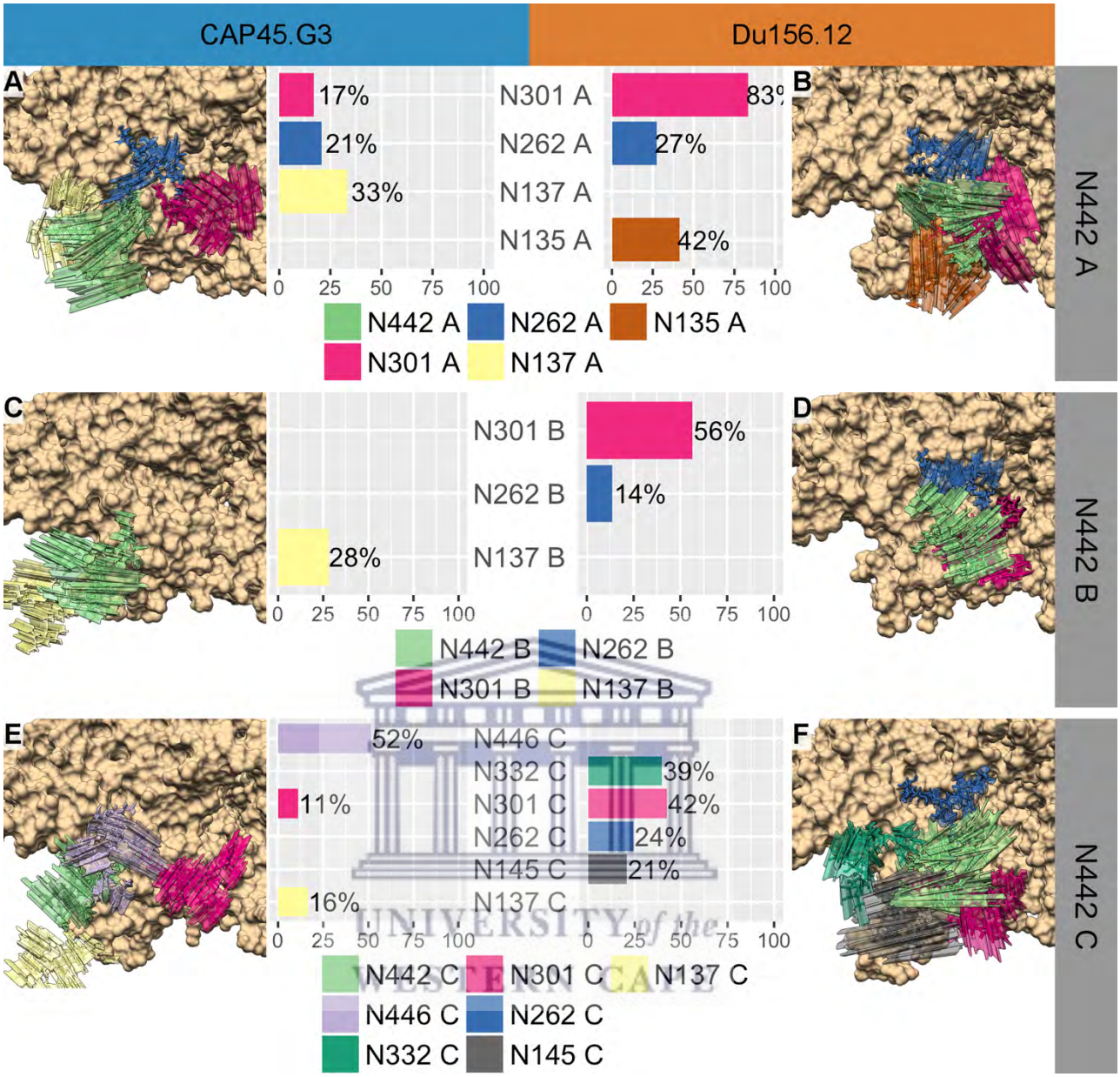

Figure 3.9: Movement and interaction networks of glycan N442

Interaction frequencies, shown as horizontal bar graphs, and 3D movement representations for the glycans forming part of the N442-glycan interaction network of each protomer during the CAP45.G3 and Du156.12 wild-type simulations.

Taken together, the glycan interaction network comparison between the wild-type models clearly illustrated how different the N301-glycan cluster formations are, despite including the same, shared, interaction partners. Given these conformational differences of the N301-glycan clusters, it would be naïve to expect that the loss of glycan N301 from these wild-type models would have the same impact on the conformations of the remaining glycans. In the following section (section 3.3.5), the conformational heterogeneity comparison is expanded to include further neighbouring

glycans and describes, in detail, the extent of glycan conformational heterogeneity within, and between, protomers and paired models.

3.3.5 Removing glycan N301: model-specific differences between the conformations of neighbouring glycans

Apart from glycans N156, N197, and N442, which were shared glycans for each wild-type N301-glycan cluster, glycans N446 and N262 formed consistent interactions across protomers with glycan N301 on the CAP45.G3 and Du156.12 wild-type models, respectively (section 3.3.3). Therefore, these two glycans were included in the new glycan subset for the detailed conformational heterogeneity analysis. In addition, glycans N448, N332 and N334 were also included in the analysis; glycan N448 is the mutually exclusive counterpart of glycan N446, and glycans N332 and N334 (also a mutually exclusive pair) were key consideration during the *in vitro* analysis[53].

To adequately describe and compare the conformational heterogeneity of the glycans within the new glycan subset, three glycan-group terms were defined: unique glycans, glycan siblings, and glycan families. The unique glycan group defines each individual glycan and is specified by the glycan position (e.g. “N156”), a protomer (e.g. “protomer A”) and a model (e.g. “CAP45.G3 wild-type”). Each model includes a set of three unique glycans, one from each protomer, that can then further be defined as glycan siblings, i.e. glycans N156 of the CAP45.G3 wild-type model are “N156 glycan siblings”. Finally, a glycan family includes all glycans at a specific position, regardless of the protomer or model. The 12 glycans (3 protomers A, B and C, of 4 models) attached to position 156 constitute the “N156 glycan family”. As described in the methods section 3.2.2.4, the conformational heterogeneity of glycans was investigated using the 2dms routine, which calculates the RMSD between all atoms in all frames of the supplied trajectory. By sub-setting and concatenating specific sections of the simulated trajectories, it is possible to examine conformational heterogeneity, over time, for unique glycans (section 3.3.5.1), between siblings (section 3.3.5.2), and between glycans from paired models (section 3.3.5.3).

To further illustrate the framework for these analyses, an example of each is presented in Figure 3.10. For each analysis, a conformation in group 1 was compared, using RMSD, to a conformation in group 2.

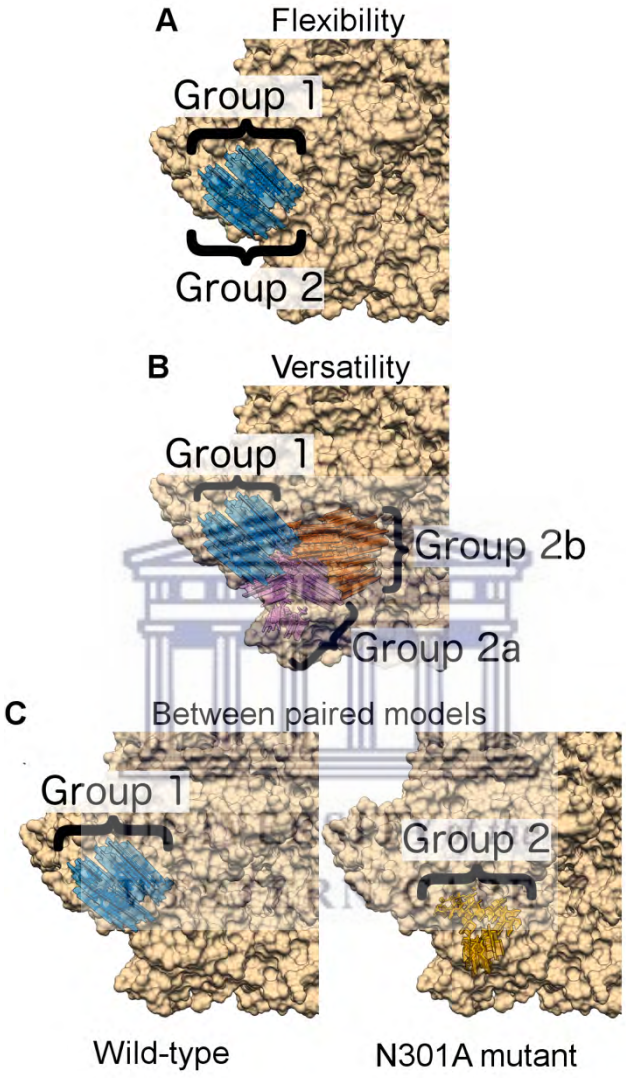

Figure 3.10: Framework of the conformational heterogeneity analyses

Each schematic (A, B and C) represents one of the three conformational heterogeneity analyses. (A) The comparison over time for unique glycans, termed flexibility; (B) the comparison between glycan siblings (versatility); and (C) the comparison between unique glycans from the wild-type and N301A mutant paired models.

For the first analysis focusing on unique glycans, group 1 and 2 are the same conformations (e.g. all conformations of glycan N156 on a specific protomer and model). Thus, the resulting RMSD distribution represents the flexibility of a unique glycan (Figure 3.10 A).

The second analysis follows on from the first, where each unique glycan is compared to each of its siblings, and these siblings to each other. Group 1 represents the conformations of a unique glycan and group 2a and 2b the confirmations of its siblings (Figure 3.10 B). The distributions of these RMSD values represent the versatility of a glycan.

For the final analysis (wild-type vs N301A mutant model), group 1 contained the conformations of a unique glycan (without its siblings) and group 2 those of its N301A mutant model counterpart (Figure 3.10 C). This analysis was extended to include, in group 1, all the conformations of a unique glycan and its siblings from the wild-type model, and in group 2 the glycan conformations of the glycan siblings on the N301A mutant model. However, these combined results did not provide meaningful information due to the large interquartile ranges and low overall differences between the glycan families and were not included for discussion (Supplementary Figure S3.1).

These three analyses are inherently related, since each expands the preceding analysis by including, successively, more distant (siblings) and parallel (paired models) glycan conformations in the analysis set. The results for each are presented in sub-sections 3.3.5.1-3.3.5.3.

3.3.5.1 The conformational heterogeneity of glycans over time

When a glycan is removed from a densely glycosylated area, one possible consequence is that the glycan shield's established equilibrium is maintained. This implies that the observed flexibility of the glycans surrounding the space created by the missing glycan (vacated space), increases, without substantially altering their average conformations. Under this hypothesis, the glycans of the CAP45.G3 and Du156.12 models surrounding the N301A mutation that displayed model specific flexibilities were distorted, i.e. had increased/decreased flexibility, by the removal of glycan N301. As described above (Figure 3.10 A), the flexibilities are defined as the distribution of RMSD values calculated between the collective conformations of a unique glycan over time. Despite the observation of large discrepancies between the

flexibilities of several glycans at an individual protomer level, which makes conclusive assertions regarding model specific flexibilities difficult, there were particular glycans that displayed model specific flexibilities. In these cases, glycans displayed different flexibilities in the paired model comparison, i.e. between the wild-type and N301A mutant models.

3.3.5.1.1 Glycan flexibilities differentiating the CAP45.G3 paired models

The comparison of glycan flexibilities between the wild-type and N301A mutant model revealed three glycans that showed evidence of model specific flexibilities. The glycans at positions 446 (Figure 3.11 K and L), 334 (Figure 3.11 I and J), and 442 (Figure 3.11 G and H), had altered flexibilities in the N301A mutant model, suggesting that these glycans moved to a greater, or lesser, extent as result of the vacated space. Only for glycan N446 was the overall flexibility reduced for all protomers on the N301A mutant model, whereas for glycans N334 and N442, the change in flexibility differed across protomers.

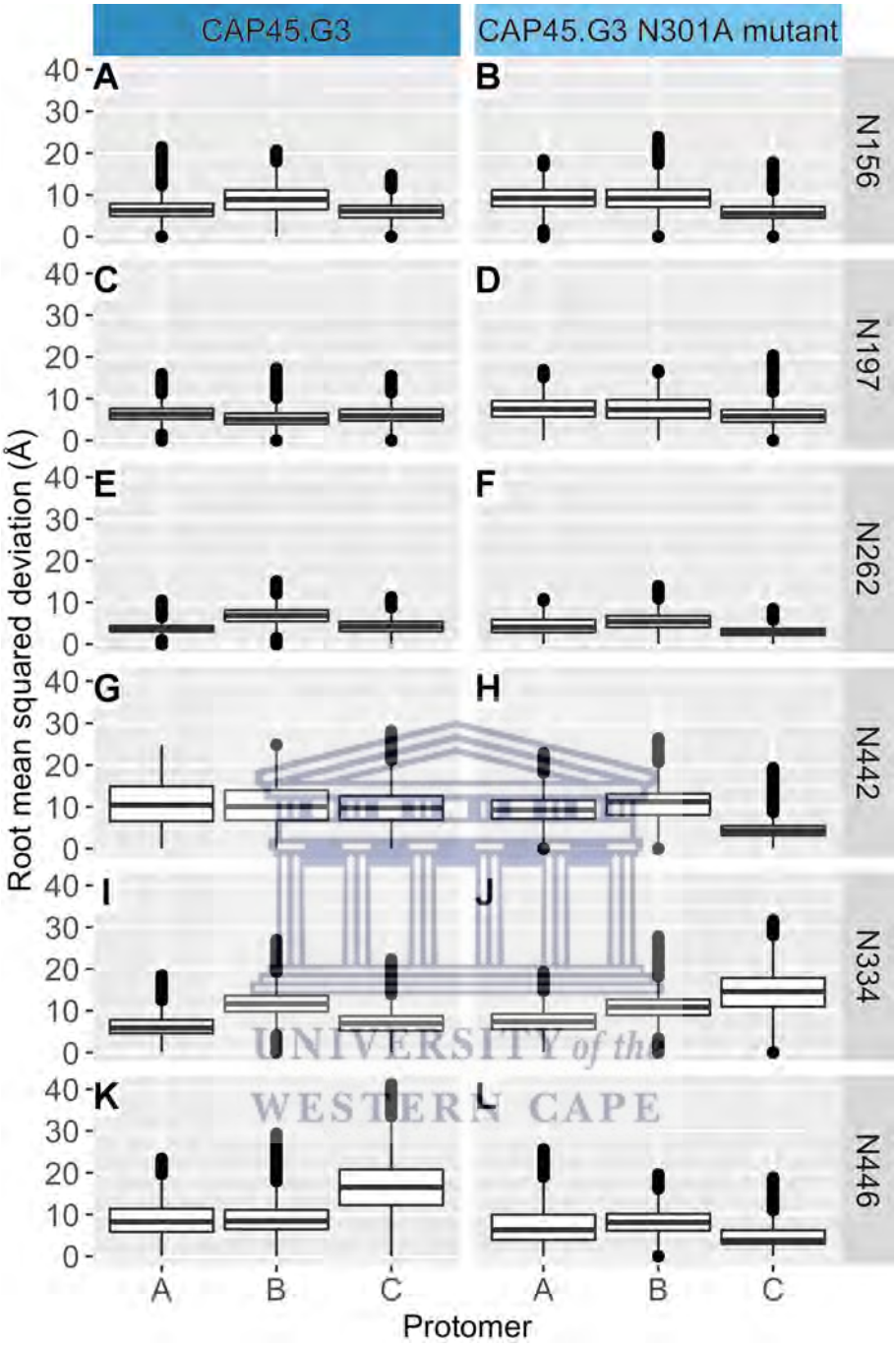

Figure 3.11: Glycan flexibilities of the CAP45.G3 paired models

Box and whisker[g] plots of the RMSD (Å) of specific glycans (N156, N197, N262, N442, N334 and N446) calculated for each protomer (A, B and C). The RMSD, calculated between all adopted conformations of a single glycan over time, represents the flexibility of that glycan.

[g] For all Box and whisker plots, the horizontal line within the box indicates the median, the boundaries of the box indicate the 25th- and 75th-percentile, whiskers indicate the lowest/highest datum within 1.5 of the inter-quartile range of the lower/upper quartile, and circles indicate data outliers.

3.3.5.1.2 Glycan flexibilities differentiating the Du156.12 paired models

Similar to the CAP45.G3 paired model comparison, there were three Du156.12 glycans that displayed different flexibilities between the wild-type and N301A mutant model. Glycans N332 (Figure 3.12 I and J), N442 (Figure 3.12 G and H), and N448 (Figure 3.12 K and L) all presented evidence of model specific flexibility, even though the altered flexibilities were inconsistently lower or higher for different protomers. Interestingly, the N301A mutant model N442 glycans were all more flexible than their wild-type model counterparts when comparing the median values. However, one of the wild-type N442 glycans also displayed relatively large flexibility (Figure 3.12 G, protomer C), suggesting that the range of movement of glycan N442 is not exclusively affected by the removal of glycan N301.

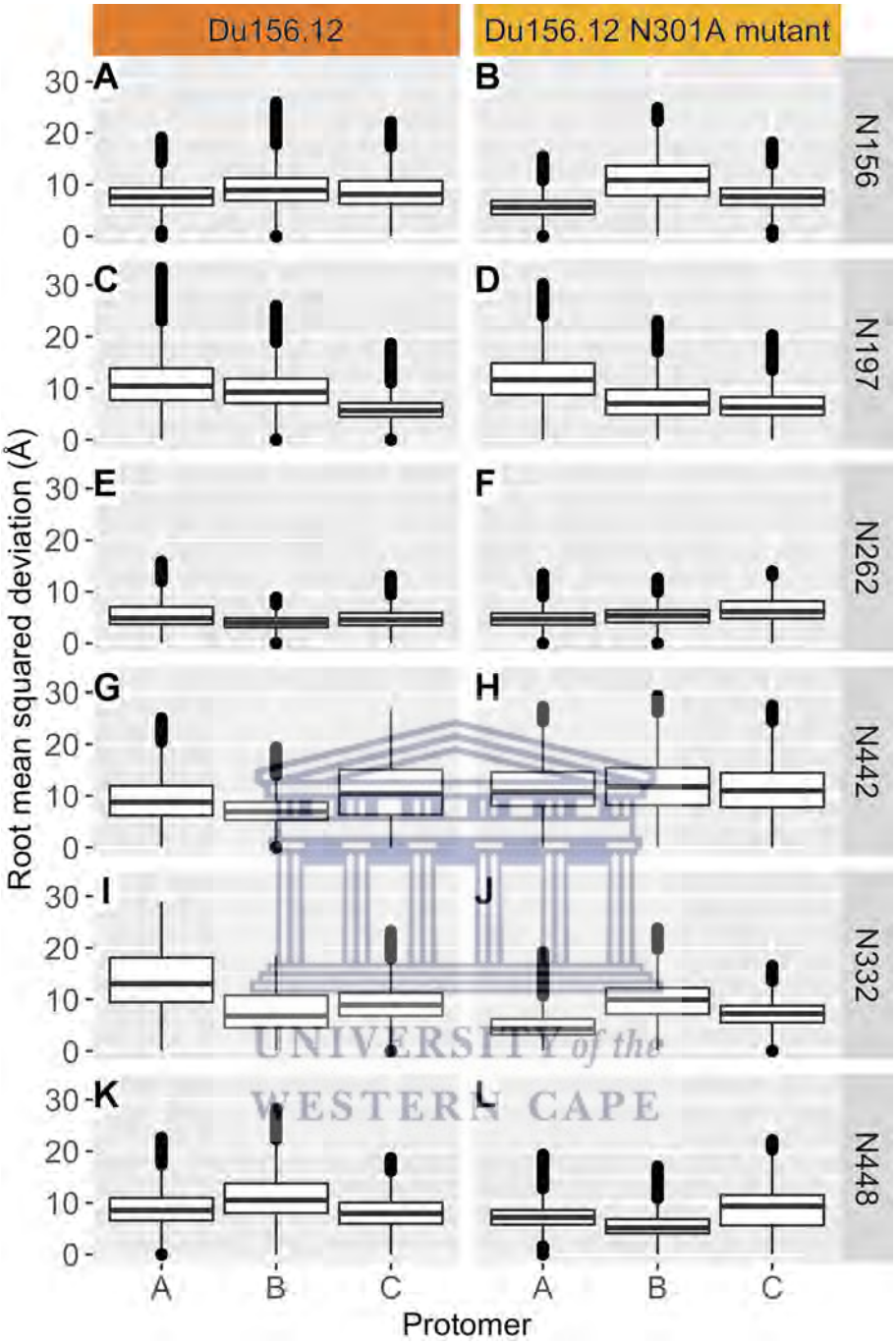

Figure 3.12: Glycan flexibilities of the Du156.12 paired models

Box and whisker plots of the RMSD (Å), representing the flexibilities, of specific glycans N156, N197, N262, N442, N332 and N448) calculated for each protomer (A, B and C)

In summary, the glycan flexibilities revealed insufficient evidence to differentiate the wild-type and N301A mutant models. However, the RMSD results nonetheless presented several interesting observations. Glycans N156 and N262 had the lowest flexibilities on all models, which suggests that the movements of these glycans were

restricted regardless of the presence of glycan N301. It has been reported that glycan N262 is buried[170] and, with the models used here, glycan N156 was densely packed within a cluster of glycans, which provides one explanation for the low flexibilities observed for these two glycans. Furthermore, although there were no definitive directional (less/greater) differences between glycan flexibilities when comparing the wild-type and N301A mutant paired models, glycans N332 (Du156.12 model), N334 (CAP45.G3 model), N442 (CAP45.G3 and Du156.12 models), N446 (CAP45.G3 model) and N448 (Du156.12 model) had altered flexibilities. These observations provide a first overview of how the movements of individual glycans were altered by the removal of glycan N301, but also indicates that further analyses, beyond individual glycans, are required.

3.3.5.2 The conformational heterogeneity between glycan siblings

The preceding section explored the hypothesis that the flexibility of glycans allowed the established equilibrium of the glycan shield to be maintained. An alternative hypothesis concerns the formation of a "new" equilibrium, where glycans surrounding the N301A mutation adopt different conformations. The observation that certain glycans were, unexpectedly, less flexible in the N301A mutant models (where intuitively there was more space for movement) compared to the wild-type models, fits the alternative hypothesis where a new equilibrium was established. These alternate conformations were inconsistent among protomers; therefore, the next investigation focussed on the glycan siblings, i.e. the same glycan on all protomers of the same model. As previously described (section 3.3.5 and Figure 3.10 B), the glycan conformational diversity between protomers was defined as the versatility. The versatility represents the distribution of RMSD values calculated between the conformations of a unique glycan and each of its siblings, and between these siblings. The versatilities were then compared between paired models, and glycans (surrounding the N301A mutation) with model specific versatilities, that were the direct result of the removal of glycan N301, were identified.

3.3.5.2.1 Glycan versatilities of the CAP45.G3 paired models

Glycans N197, N332, N442 and N446 all showed evidence of different versatilities on the CAP45.G3 N301A mutant model when compared to the wild-type model (Figure 3.13). Glycans N197 (Figure 3.13 C and D) and N442 (Figure 3.13 G and H), which are in the immediate vicinity of glycan N301, had higher versatilities on the N301A mutant model, whereas glycans N334 (Figure 3.13 I and J) and N446 (Figure 3.13 K and L), had lower versatilities.

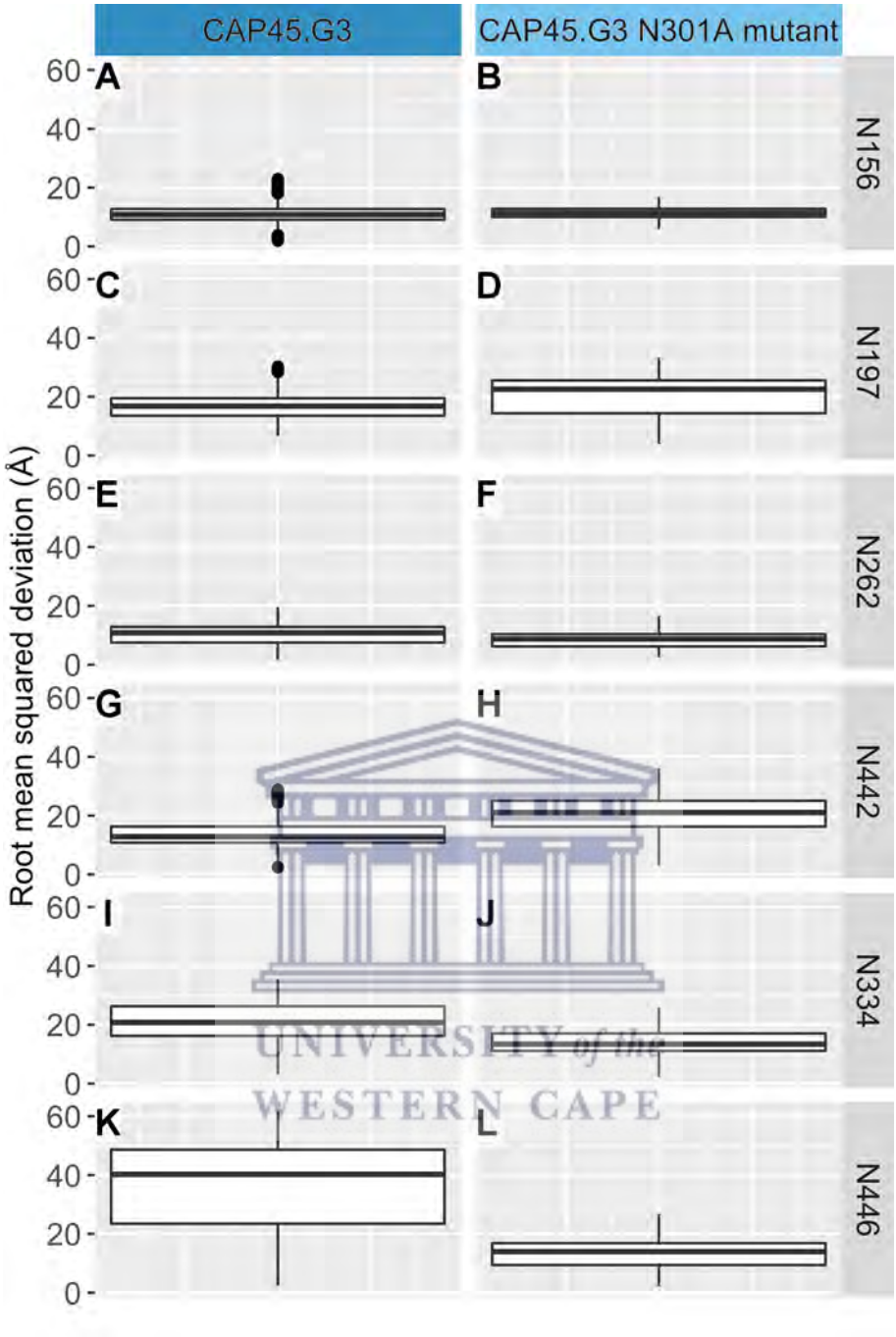

Figure 3.13: Glycan versatilities of the CAP45.G3 paired models

Box and whisker plots of the RMSD (Å) of specific glycans (N156, N197, N262, N442, N334 and N448). The RMSD was calculated between the conformations of a unique glycan and each of its siblings, and between these siblings.

3.3.5.2.2 Glycan versatilities of the Du156.12 paired models

Apart from glycan N197, all the remaining glycans (N156, N262, N442, N332 and N448) had increased versatilities on the Du156.12 N301A mutant model (Figure 3.14).

This suggests that the glycans surrounding the N301A mutation were able to explore a larger number of equilibrium conformations than those on the wild-type model. However, since these RMSD results were uninformative regarding the type of movement, further analysis was required to determine whether these additional conformations were distinct from those of the wild-type model.

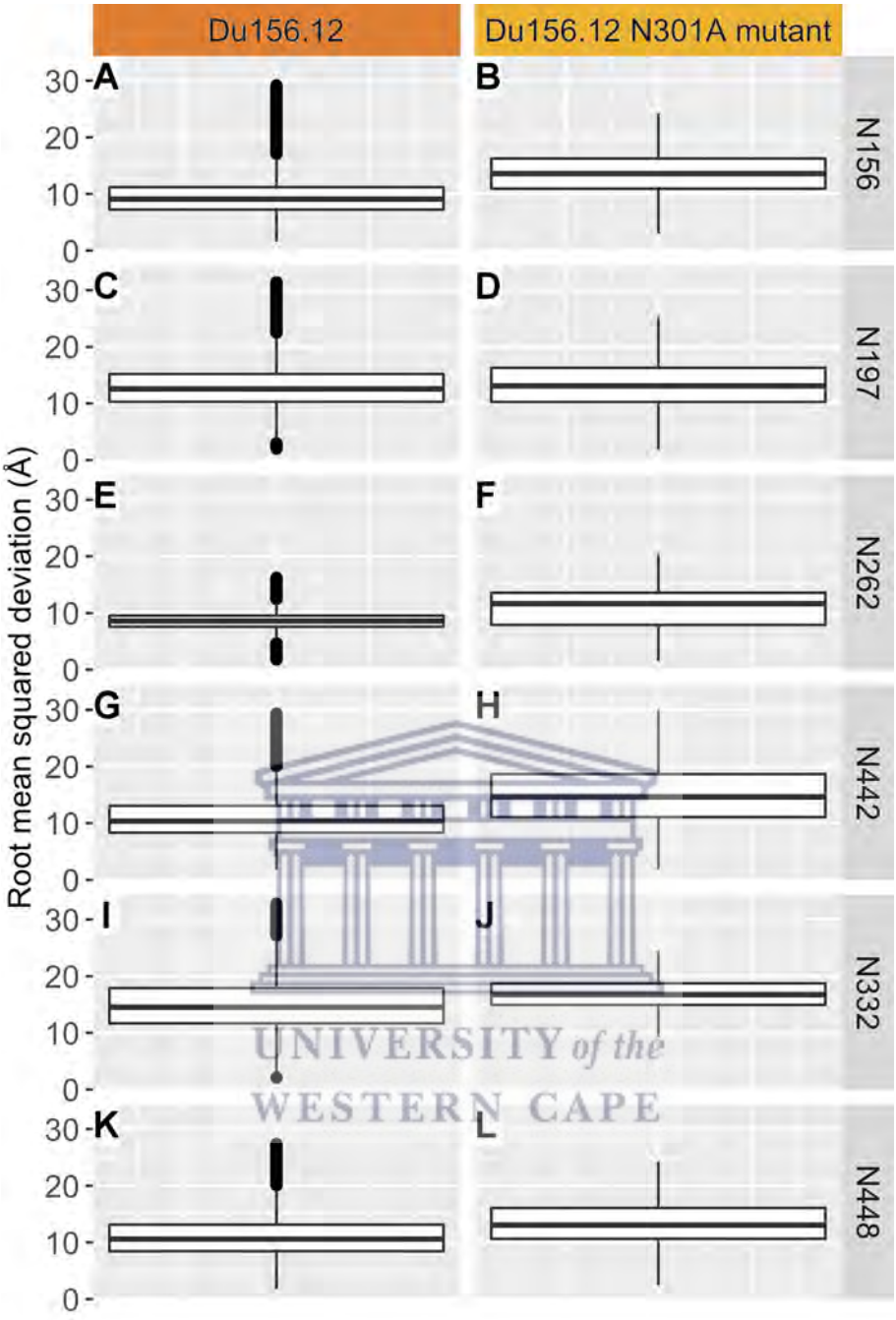

Figure 3.14: Glycan versatilities of the Du156.12 paired models

Box and whisker plots of the RMSD (Å) of specific glycans (N156, N197, N262, N442, N332 and N446), which was calculated between the conformations of a unique glycan and each of its siblings, and between these siblings.

3.3.5.3 The difference between the conformational heterogeneity of glycans on each wild-type and its N301A mutant model counterpart

The final conformational heterogeneity analysis focussed specifically on the differences between the conformations of N301A mutant glycans compared to their wild-type model counterparts (sections 3.3.5 and Figure 3.10 C). Hence, the RMSD distribution was calculated between identical glycans on the paired models (Figure 3.10 C). When evaluating these RMSD distributions to determine which glycans have substantially different conformation between the wild-type and N301A mutant model, the cause of the difference becomes important. Conformation differences can either be caused by the N301A mutation or by random variation, which was observed during the conformational heterogeneity analysis between glycan siblings (section 3.3.5.2). For this reason, RMSD distributions calculated here were compared to the relevant versatility results (Figure 3.13 and Figure 3.14). Only glycans with conformational differences greater than their versatilities were considered as heterogeneous between the wild-type and N301A mutant models. The possibility of identifying glycans with conformational heterogeneity due to random variation is also reduced by the fact that the starting conformation of the paired glycans, and all the surrounding glycans, were identical, and that this analysis was limited to the glycans close to the N301A mutation. Despite these precautions, it is still possible that some of the glycans identified during this analysis were heterogeneous due to random variation instead. Therefore, the results should be interpreted taking this into account.

3.3.5.3.1 Glycan conformations differentiating the CAP45.G3 models

For three glycans (N197, N442 and N446) particular protomer/s showed noticeably higher conformational heterogeneity when the CAP45.G3 N301A mutant glycan conformations were compared to those of the wild-type model. For glycans N197 and N446, the larger conformational heterogeneity was observed for protomer C, while for glycan N442 it was observed for protomers A and B. The heterogeneity between paired glycans was the largest for glycan N446, followed by glycan N442 (protomer A) and then glycan N197. The magnitude of these conformational differences (ranging from 12 to 43Å) compared to those observed for random variation (between the

CAP45.G3 wild-type protomers; section 3.3.5.2) for each particular glycan, increases the confidence that these conformational differences were likely due to the N301A mutation as opposed to random variation.

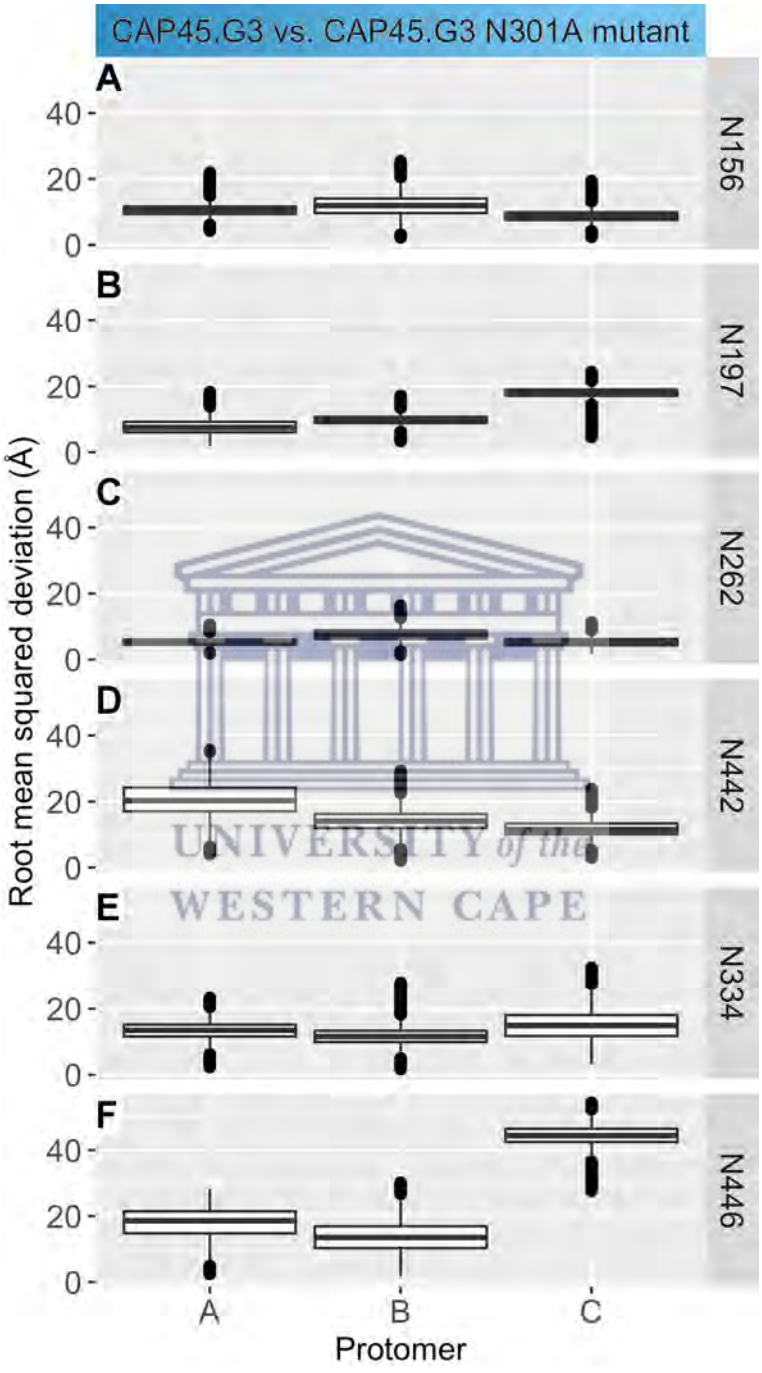

Figure 3.15: RMSD distribution of identical glycans on the CAP45.G3 paired models

Box and whisker plots of the RMSD (Å) of specific glycans (N156, N197, N262, N442, N334 and N446) calculated, for each glycan, between the conformations of the wild-type glycan and the conformation of its N301A mutant glycan counterpart.

3.3.5.3.2 Glycan conformations differentiating the Du156.12 models

With the exception of glycan N197, all the remaining glycans (N156, N262, N442, N332 and N448) had at least one protomer with considerable conformational heterogeneity when the Du156.12 N301A mutant glycan conformations were compared to the wild-type glycan conformations. Unlike the CAP45.G3 comparison, the differences were also more consistent across the protomers for the Du156.12 paired models, where the same glycans from multiple protomers had considerable conformational heterogeneity, e.g. for glycan N442 on protomers A, B and C, for glycan N156 on protomers A and B, and for glycan N448 on protomers B and C. However, the magnitude of the conformational differences still varied between the protomers. Despite these differences in magnitudes, even the lowest variation was still larger than that observed for the random variation (i.e., between the Du156.12 wild-type protomers; section 3.3.5.2) of the particular glycan. This implies, similar to the CAP45.G3 paired model comparison, that these conformational differences were likely due to the N301A mutation rather than random variation.

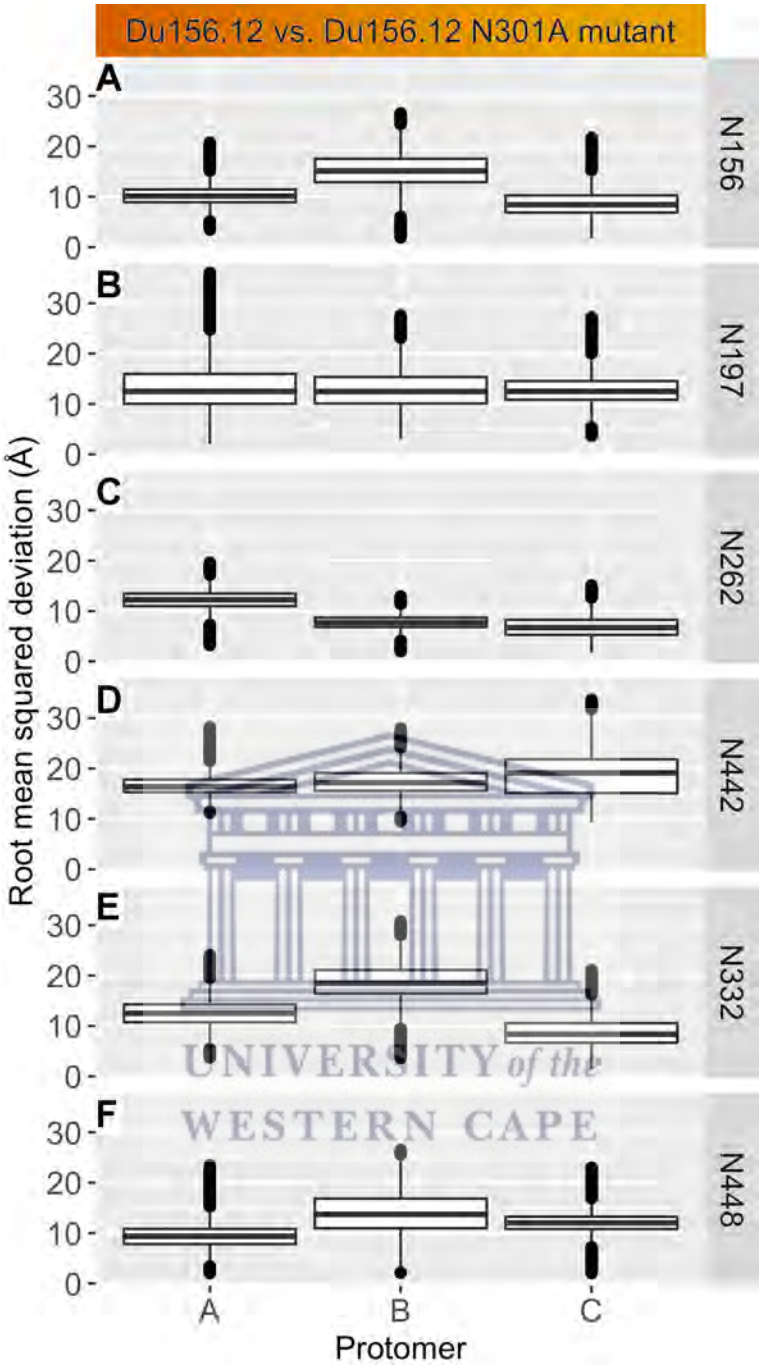

Figure 3.16: RMSD distribution of identical glycans on the Du156.12 paired models

Box and whisker plots of the RMSD (Å) of specific glycans (N156, N197, N262, N442, N332 and N448) calculated, for each glycan, between the conformations of the wild-type glycan and the conformation of its N301A mutant glycan counterpart.

3.3.6 Altered glycan conformations extended beyond those neighbouring glycan N301

In the previous section (section 3.3.5), the extent of the conformational heterogeneity of glycans N156, N197, N262, N332/N334, N442 and N446/N448 within, and between, protomers and models was described in detail. Subsequently, the next focus was to extend this analysis beyond only these glycans that neighbour the N301A mutation, to all the glycans as well as the protein residues. To achieve this goal, each glycan, or protein residue, was ranked based on its capability to distinguish between frames of the wild-type and the corresponding N301A mutant model. For example, suppose the conformations of glycan N442 were perfectly distinct between paired models; if presented with a randomly chosen frame, one merely needs to consider the conformation of glycan N442 to determine if the frame emerged during the N301A mutant or wild-type model simulation. Therefore, glycans, or protein residues, exceedingly capable of correctly classifying frames as wild-type or N301A mutant model frames, were also the most conformational heterogeneous between paired models.

In order to determine which glycans or protein residues were the best classifiers, an approach capable of dealing with "wide" data, i.e. where there are more variables/columns than observations/rows (many more atoms than frames), is advantageous. Principal component analysis (PCA), already implemented to accommodate simulation data, was suitable for this analysis. Even though PCA is relatively capable of dealing with "wide" data, it is still preferable to reduce the data to the minimal working set. The reason for this is that the number of variables is directly related to the volume of the resulting space, which increases exponentially with each additional variable, resulting in the available data becoming sparse. This sparsity is detrimental for any method that requires statistical significance. Therefore, the comparisons between the N301A mutant models and their wild-type counterparts were limited to the protomers. Additionally, for each protomer-level comparison, the protein residues and glycans were considered separately.

Once these "reduced" simulation trajectories were assembled (see Methods section 3.2.2.5), PCA was performed on pseudo-trajectories comprising frames originating from both the wild-type and N301A mutant simulations of paired models. For the initial

analysis, focussing only on the glycans, the first principal component differentiated the frames of the wild-type and N301A mutant models perfectly, since a vertical line completely separated each group of frames on the standard PCA plots (e.g. Figure 3.17 A and C, and Supplementary figures Figure S3.2 and Figure S3.3). The separation was not as clear for the protein analysis (Supplementary figures Figure S3.4 and Figure S3.5), but still provided an indication of the frame separation for the paired models and was thus sufficient for further analysis.

The next step was to isolate which glycans, or protein residues, were contributing most to the first principal component. In this analysis, the principal components are linear combinations of the atom coordinates weighted by their contributions towards explaining the variance in a particular orthogonal dimension. Therefore, since the first principal component differentiates between frames of the wild-type and N301A mutant, the squared weights (factor loadings) rank the glycans, or protein residues, by the extent of their variation between the wild-type and N301A mutant models. However, due to small changes in the factor loadings potentially altering the rankings substantially, especially if glycans or protein residues were equally distinct between paired models, each glycan or protein residue was ultimately ranked based on its median factor loading. These medians were determined for the factor loadings by repeating each PCA on 100 moving-block bootstrap replicates of the appropriate pseudo-trajectories.

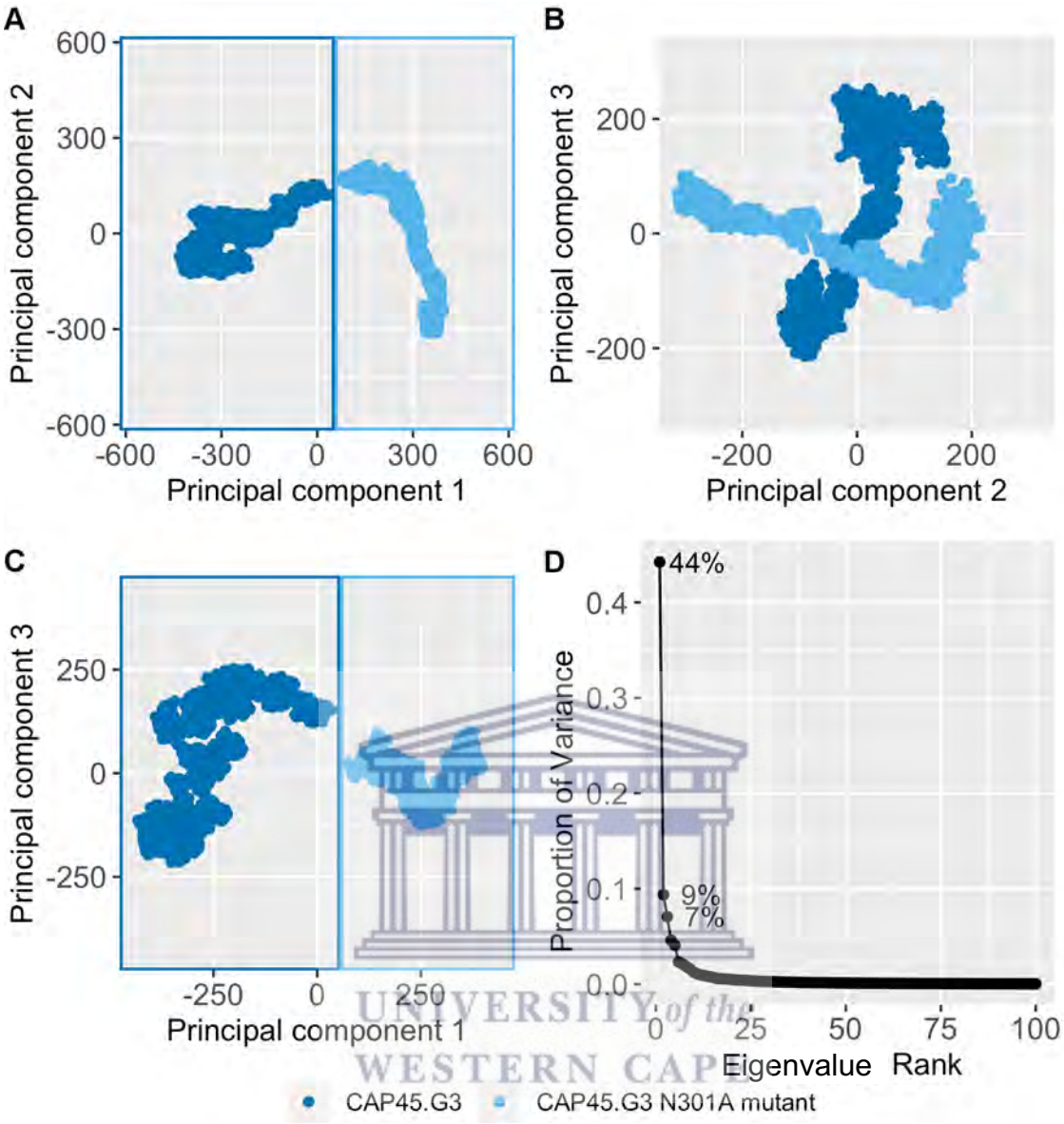

Figure 3.17: Example of the standard PCA plots for the CAP45.G3 model

The standard PCA plots for the CAP45.G3 paired-model comparison performed on the glycans of protomer A. The figures demonstrate the projection of the frames onto the (A) first and second, (B) second and third and (C) first and third principal components as well as (D) the proportion of the variance explained by each principal component.

In studying the factor loadings of the first principal components of each protomer and paired models, the true range and scope of conformational heterogeneity between the N301A mutant models and the wild-type counterparts can be appreciated (Figure 3.18, Figure 3.19, Figure 3.20 and Figure 3.21). Due to the factor loadings only narrowly distinguishing successively ranked glycans, or protein residues, it was necessary to specify a threshold above which a practical number of glycans, or protein residues,

were identified. Since the same glycans were rarely identified for each protomer, the two top ranking glycans within the gp120 and gp41 region of each protomer were considered. For the protein analysis, the top five protein residues for each protomer were identified. For consistency and reference, the rankings of glycans N156, N197, N262, N332/N334, N442 and N446/N448, which have been discussed in detail in the previous sections (sections 3.3.4 and 3.3.5) are also shown (Figure 3.18 and Figure 3.19).

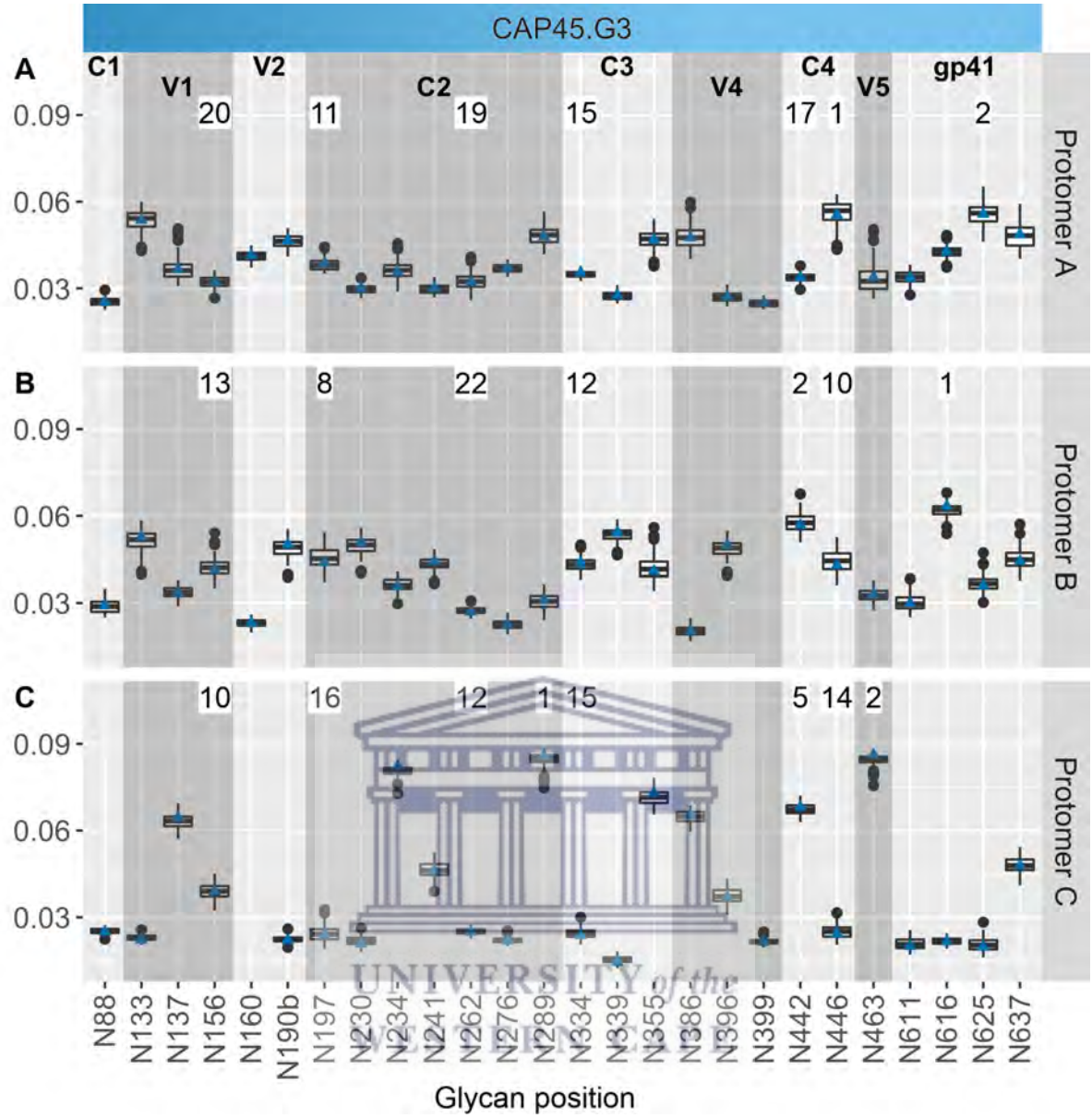

Figure 3.18: Factor loadings of the CAP45.G3 paired-model comparison focussing on glycans

Box and whisker plots of the factor loadings of all glycans for (A) protomer A, (B) protomer B and (C) protomer C. Blue triangles represent the factor loadings calculated using the simulated trajectories. The rankings based on median factor loadings are shown for the top two glycans as well as glycans N156, N197, N262, N334, N442 and N446.

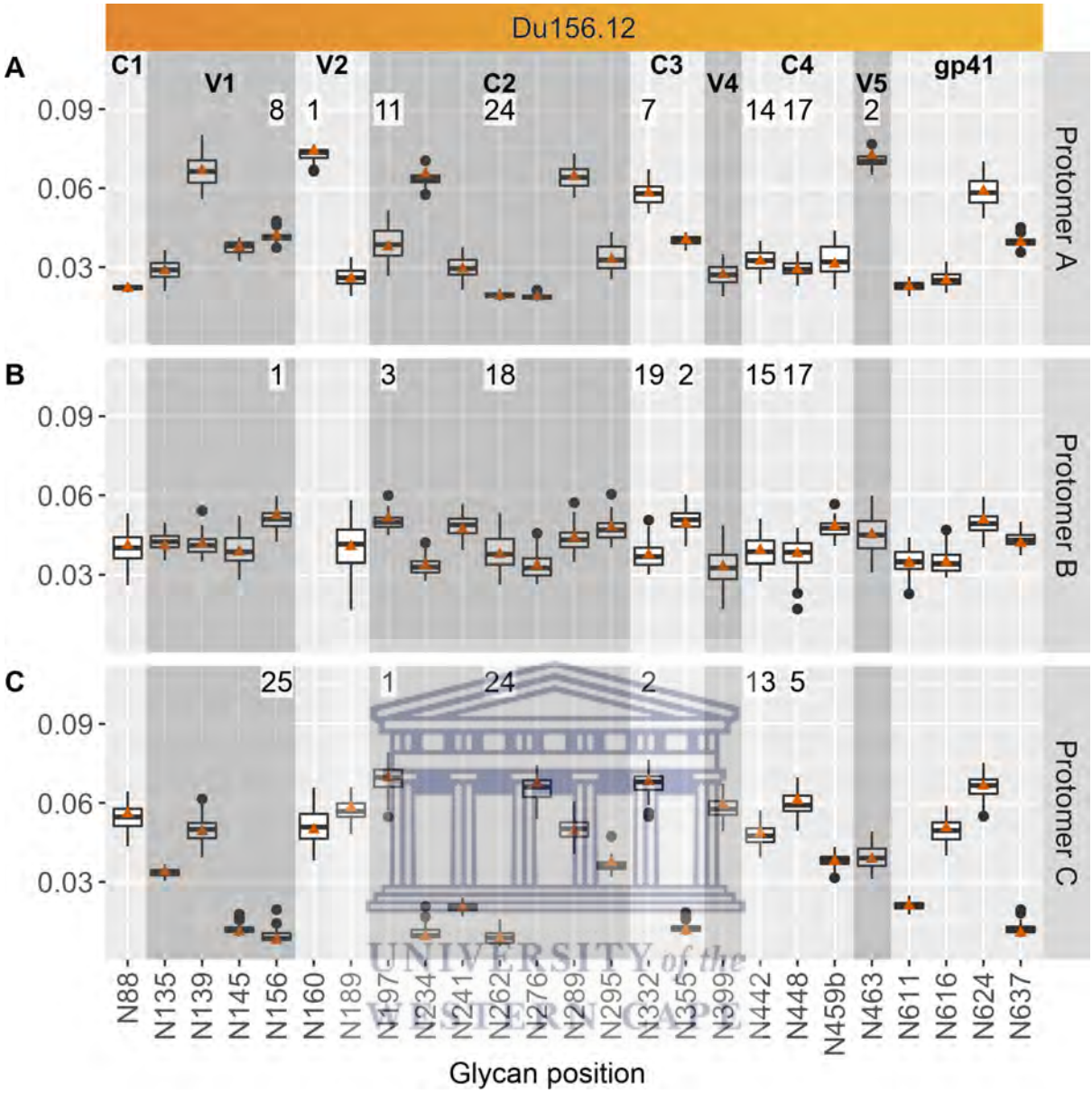

Figure 3.19: Factor loadings of the Du156.12 paired-model comparison focussing on glycans

Box and whisker plots of the factor loadings of all glycans for (A) protomer A, (B) protomer B and (C) protomer C. Orange triangles represent the factor loadings calculated using the simulated trajectories. The rankings based on median factor loadings are show for the top two glycans as well as glycans N156, N197, N262, N332, N442 and N448.

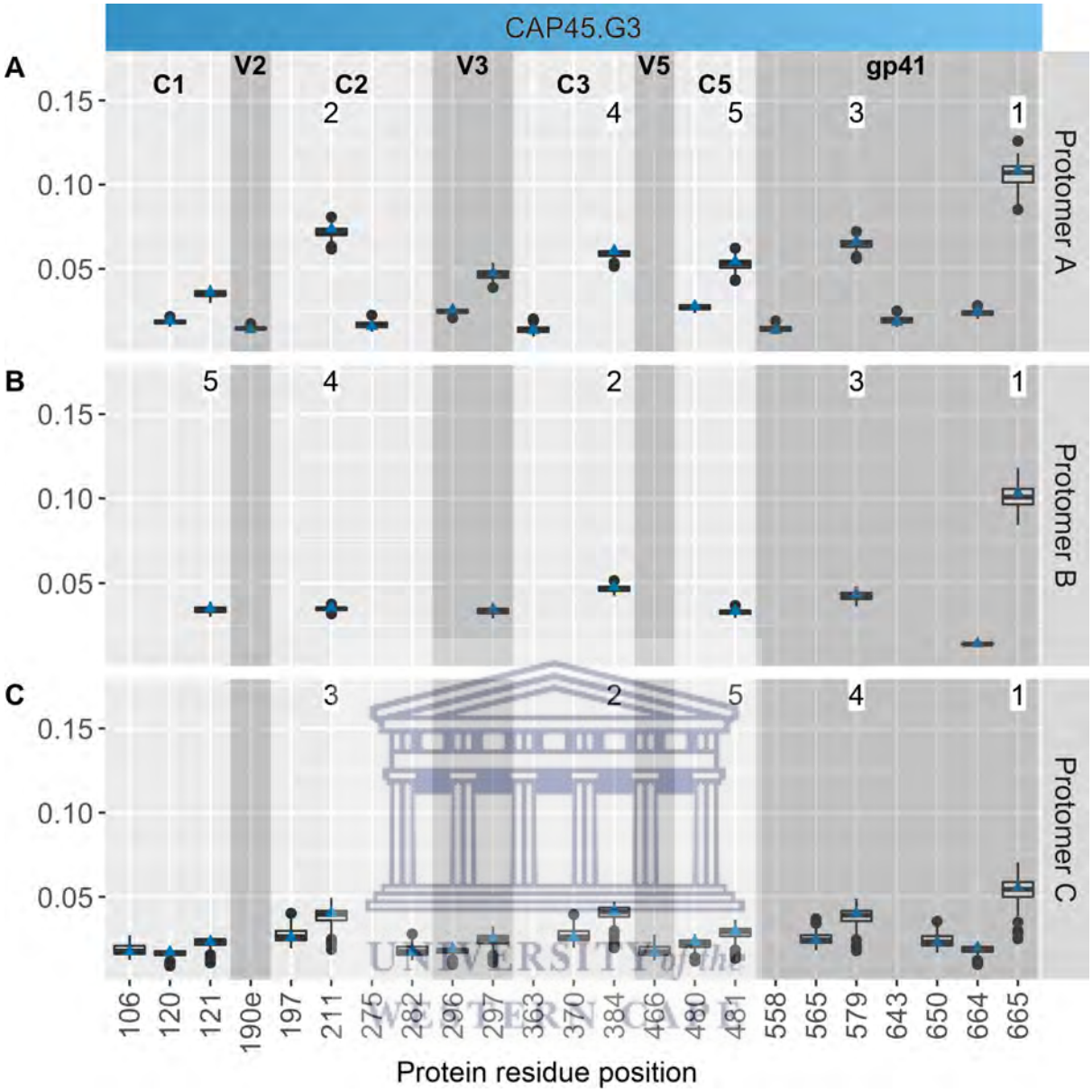

Figure 3.20: Factor loadings of the CAP45.G3 paired-model comparison focussing on protein residues

Box and whisker plots of the factor loadings of protein residues for (A) protomer A, (B) protomer B and (C) protomer C with a median value above a threshold (0.013) such that only the top protein residues are displayed. Blue triangles represent the factor loadings calculated using the simulated trajectories. The rankings based on median factor loadings are show for the top five protein residues.

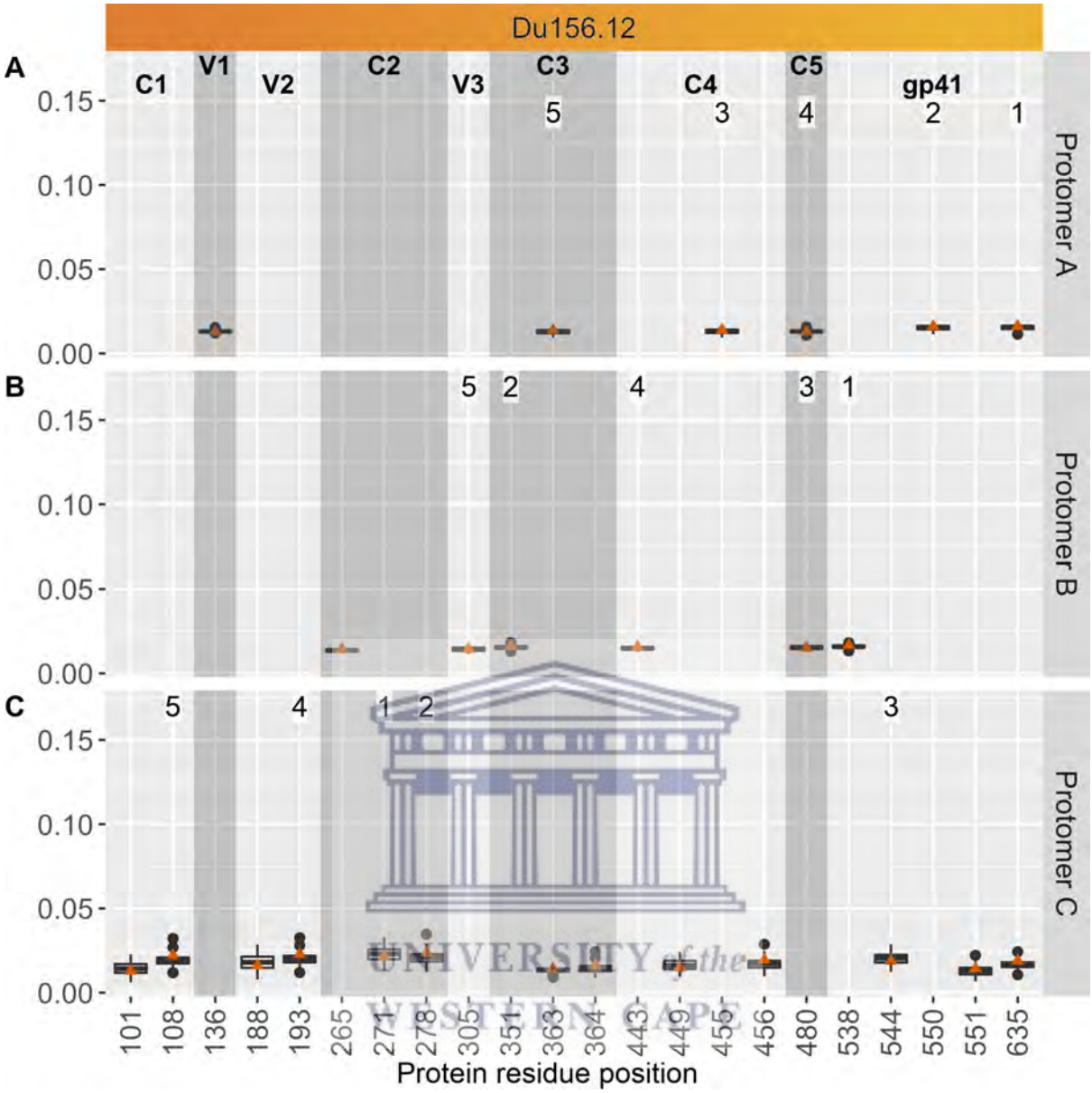

Figure 3.21: Factor loadings of the Du156.12 paired-model comparison focussing on protein residues

Box and whisker plots of the factor loadings of protein residues for (A) protomer A, (B) protomer B and (C) protomer C with a median value above a threshold (0.013) such that only the top protein residues are displayed. Orange triangles represent the factor loadings calculated using the simulated trajectories. The rankings based on median factor loadings are show for the top five protein residues.

Of these glycans, the N442 and N446 glycans were identified by the PCA as among the glycans with the most distinct conformations between the CAP45.G3 wild-type and N301A mutant models for the gp120 region (Figure 3.18). Additionally, and similarly located in the gp120 region, glycans N289 and N463 also had considerable conformational heterogeneities between the CAP45.G3 paired models (Figure 3.18

C). In contrast, for the Du156.12 gp120 region, the paired-model comparison showed that the N197 and N332 glycans (that were also considered in the previous sections 3.3.4 and 3.3.5), were among the most distinct glycans, along with glycans N160, N289 and N463 (Figure 3.19).

For the gp41 region, both paired-model comparisons had at least one gp41 glycan that was identified as distinct between the wild-type and N301A mutant models (Figure 3.18 and Figure 3.19). Specifically, glycans N616 and N625 for the CAP45.G3, and glycan N624 for the Du156.12, paired-model comparisons. Given that these glycans are located in the gp41 region, where there are relatively fewer glycans compared to gp120, and consequently more space, it is plausible that these distinct conformations could be attributed to random variation and not the N301A mutation. Thus, these glycans were not of interest in the context of the N301A mutation.

In addition, the analysis of the protein revealed that residues 211, 384, 579 and 665 consistently had distinct conformations between the CAP45.G3 N301A mutant model and its wild-type counterpart (Figure 3.20). However, for the Du156.12 paired-model comparison, no protein residues were identified that consistently had distinct conformations between the models (Figure 3.21). Furthermore, since the difference in the factor loadings of successively ranked protein residues were larger for the CAP45.G3 model protein residues than those for Du156.12, the factor loadings of the CAP45.G3 paired-models readily attributed the conformational heterogeneity to specific residues, whereas for the Du156.12 paired-model comparison, several, diverse residues were identified.

Taken together the PCA analysis demonstrated that glycans (and protein residues) were conformationally heterogeneous between the wild-type models and their N301A mutant counterparts, and some more so than others. Of these top ranked glycans, only glycans N197, N332, N442 and N446 were previously identified using a crude method (hydrogen-bond interaction networks, section 3.3.3). This suggests that only a portion of the glycans potentially impacted by the N301A mutation were in the immediate vicinity of the mutation. Thus, it is important to investigate glycans further from the N301A mutation.

However, as previously described (section 3.3.5.3), a reasonable level of caution needs to be employed when considering glycans progressively further from the N301A mutation, since these glycans are more likely to display conformational heterogeneity due to random variation rather than due to the N301A mutation. Hence, the next step was to address this issue by quantifying the random variation of the glycans identified during the PCA and considering them relative to other glycans.

As described in the methods (section 3.2.2.5), to quantify the random variation of each glycan, the individual glycan variation between the protomers of the wild-type models was used as a proxy. For example, a glycan that was conformationally heterogeneous between the protomers of the wild-type models, i.e. displayed higher random variation, was also more likely to exhibit distinct conformations between the wild-type and N301A mutant paired models due to this random variation, and not due to the N301A mutation.

As before, the "reduced" simulation trajectories were used, however, the pseudo-trajectories on which the PCA was performed consisted of frames from only the wild-type simulation. Here, the first two principal components differentiated the frames of the wild-type protomer models perfectly, since only a vertical and horizontal line were required to separate each group of frames on the standard PCA plots for the glycans and protein residues (e.g. Figure 3.22 and Supplementary figures Figure S3.6-Figure S3.8).

Similar to the paired-model comparisons, the first two principal components can be deconstructed into their factor loadings to determine which glycans, or protein residues, were contributing the most to each principal component. These glycans and protein residues were subsequently ranked according to their contributions based on their median factor loadings. As before, the medians were determined by repeating the PCA on 100 moving-block bootstrap replicates of the appropriate pseudo-trajectories (Methods section 3.2.2.5).

Quantifying and comparing the random variation of glycans N289, N442, N446, N463, N616 and N625 (CAP45.G3 model), and glycans N160, N197, N289, N332 and N468 (Du156.12 model) was of particular interest, since these glycans displayed

exceptionally distinct conformations during the PCA between the appropriate paired models (Figure 3.18 and Figure 3.19). Additionally, the rankings of the glycans considered during the RMSD analysis are shown for reference and consistency (Figure 3.23 and Figure 3.24 glycans N156, N197, N262, N332/N334, N442 and N446/N448).

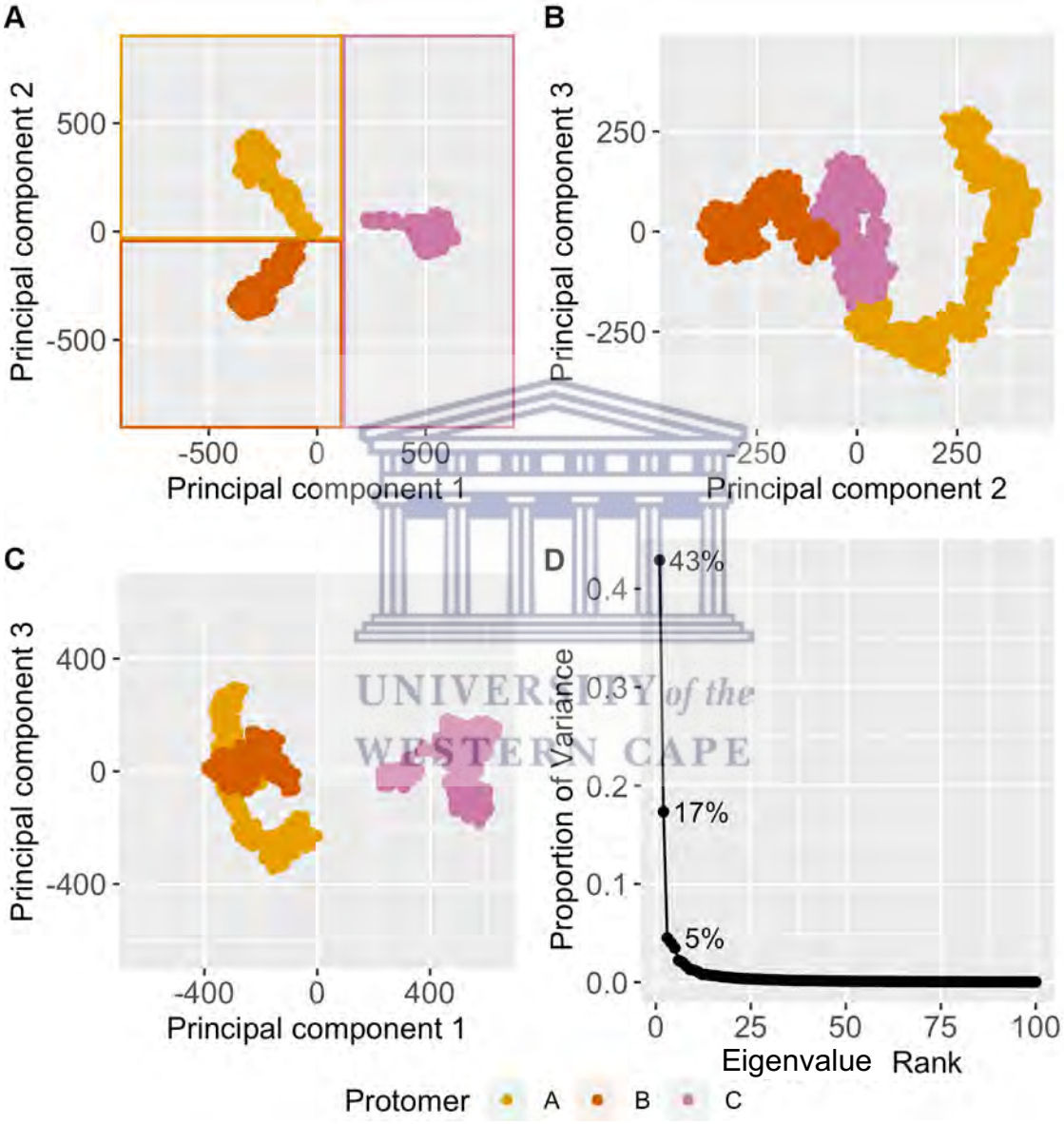

Figure 3.22: Example of the standard PCA plots for the Du156.12 model

The standard PCA plots for the Du156.12 protomer comparison performed on the glycans. The figures demonstrate the projection of the frames onto the (A) first and second, (B) second and third and (C) first and third principal components as well as (D) the proportion of the variance explained by each principal component.

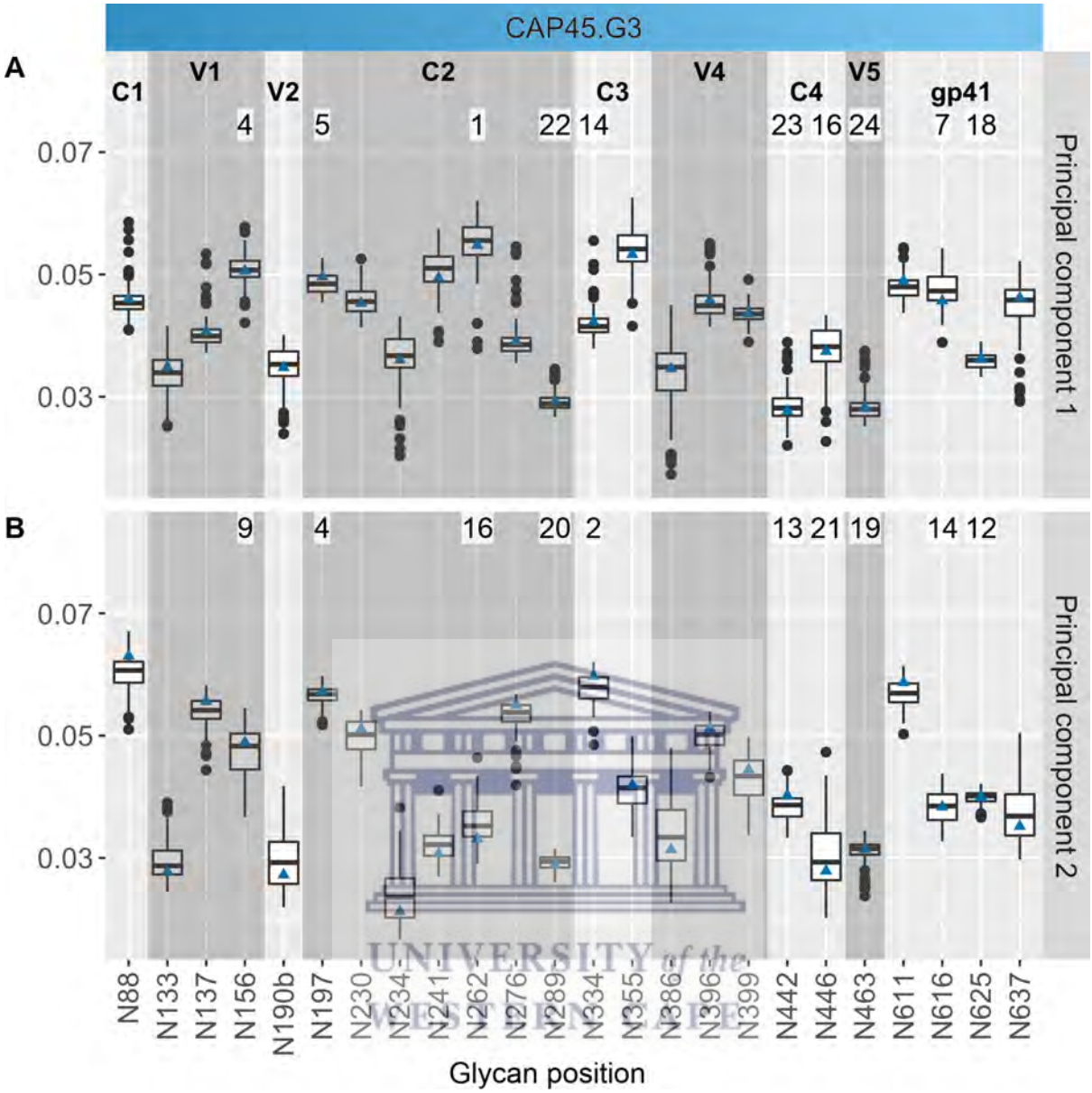

Figure 3.23: Factor loadings of the CAP45.G3 wild-type protomer comparison focussing on glycans

Box and whisker plots of the factor loadings of all glycans for the (A) first and (B) second principal components. Blue triangles represent the value of the factor loadings calculated using the simulated trajectories. The rankings based on median factor loadings are shown for glycans N156, N197, N262, N289, N334, N442, N446, N463, N616 and N625.

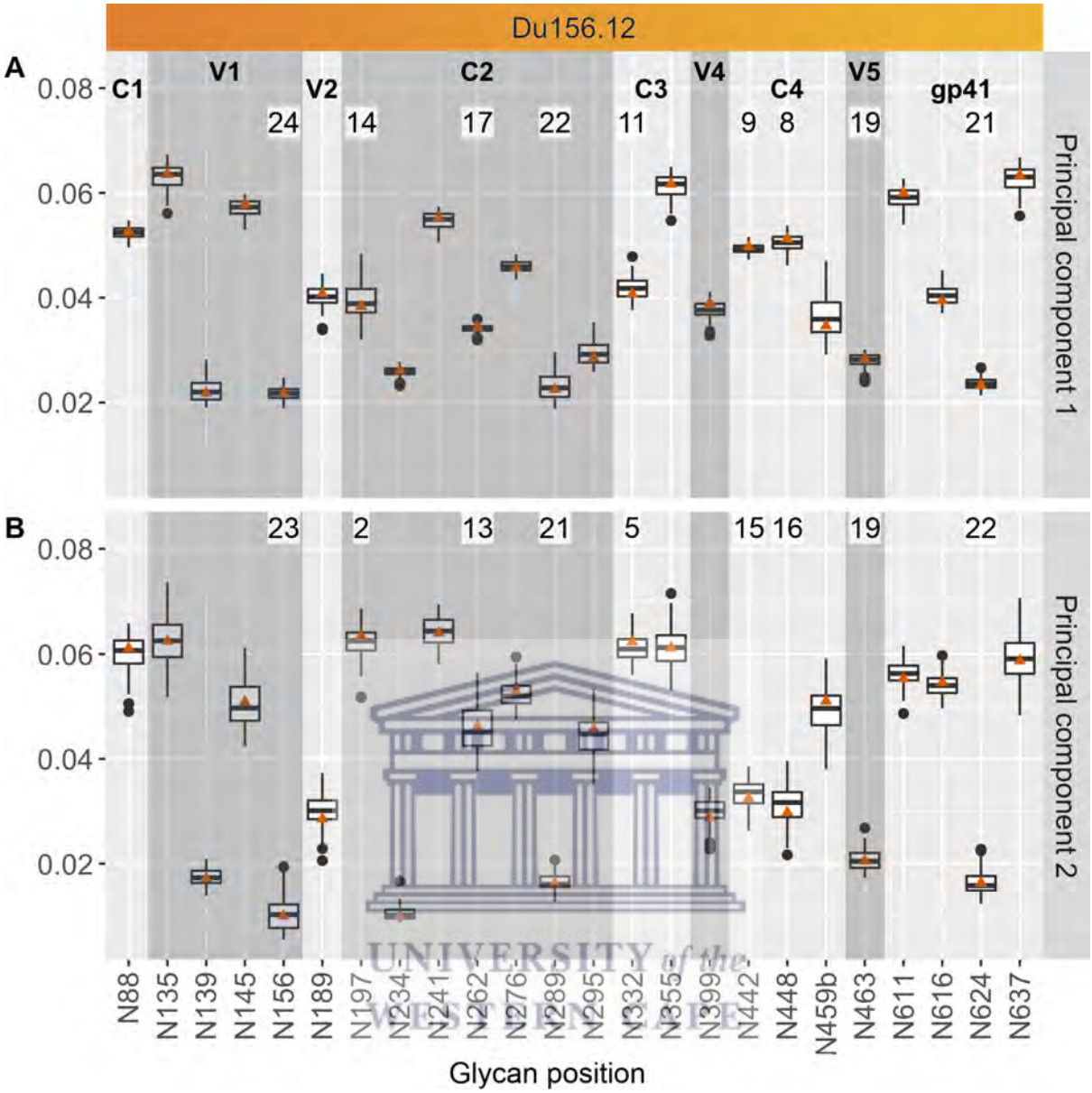

Figure 3.24: Factor loadings of the Du156.12 wild-type protomer comparison focussing on glycans

Box and whisker plots of the factor loadings of all glycans for the (A) first and (B) second principal components. Orange triangles represent the factor loadings calculated using the simulated trajectories. The rankings based on median factor loadings are shown for glycans N156, N197, N262, N289, N332, N442, N448, N463 and N624.

None of the glycans (N289, N442, N446 and N463) identified in the gp120 region during the PCA for the CAP45.G3 paired-model comparison were ranked within the top twelve (Figure 3.23). For the gp41 region, where two glycans were identified, glycan N616 was ranked 7th and 14th, while glycan N625 was ranked 18th and 12th by the first two principal components, respectively (Figure 3.23). Interestingly, glycans

N156 and N262 were ranked within the top ten by one or both principal components (Figure 3.23). This suggests that these glycans had distinct conformations on each wild-type protomer but were minimally affected by the N301A mutation, since the wild-type and N301A mutant model conformations were comparable (Figure 3.18).

During the Du156.12 paired-model comparison, five glycans (N160, N197, N289, N332 and N463) were identified. Glycan N160 was not glycosylated on protomer B (section 3.2.1) and was therefore not included during the wild-type protomer model comparison. Of the remaining glycans, N197 and N332 were ranked 2nd and 5th by the second principal component (Figure 3.24 B), while glycans N289 and N463 had relatively low ranks (Figure 3.24 rank 22 and 21, and 19 and 19). This suggests that while these glycans (N197 and N332) were potentially affected by the N301A mutation, they also varied between the wild-type protomer models. Therefore, further analysis would be required to unequivocally determine whether the conformations of these glycans were directly affected by the N301A mutation.

During the protein analysis, only the CAP45.G3 paired-model comparison revealed protein residues (211, 384, 579 and 665) that were identified for all protomers (A, B and C). However, these were also the top four ranked protein residues during the wild-type protomer analysis (Figure 3.25), which indicates that these protein residues were likely distinct due to random variation and not due to the N301A mutation.

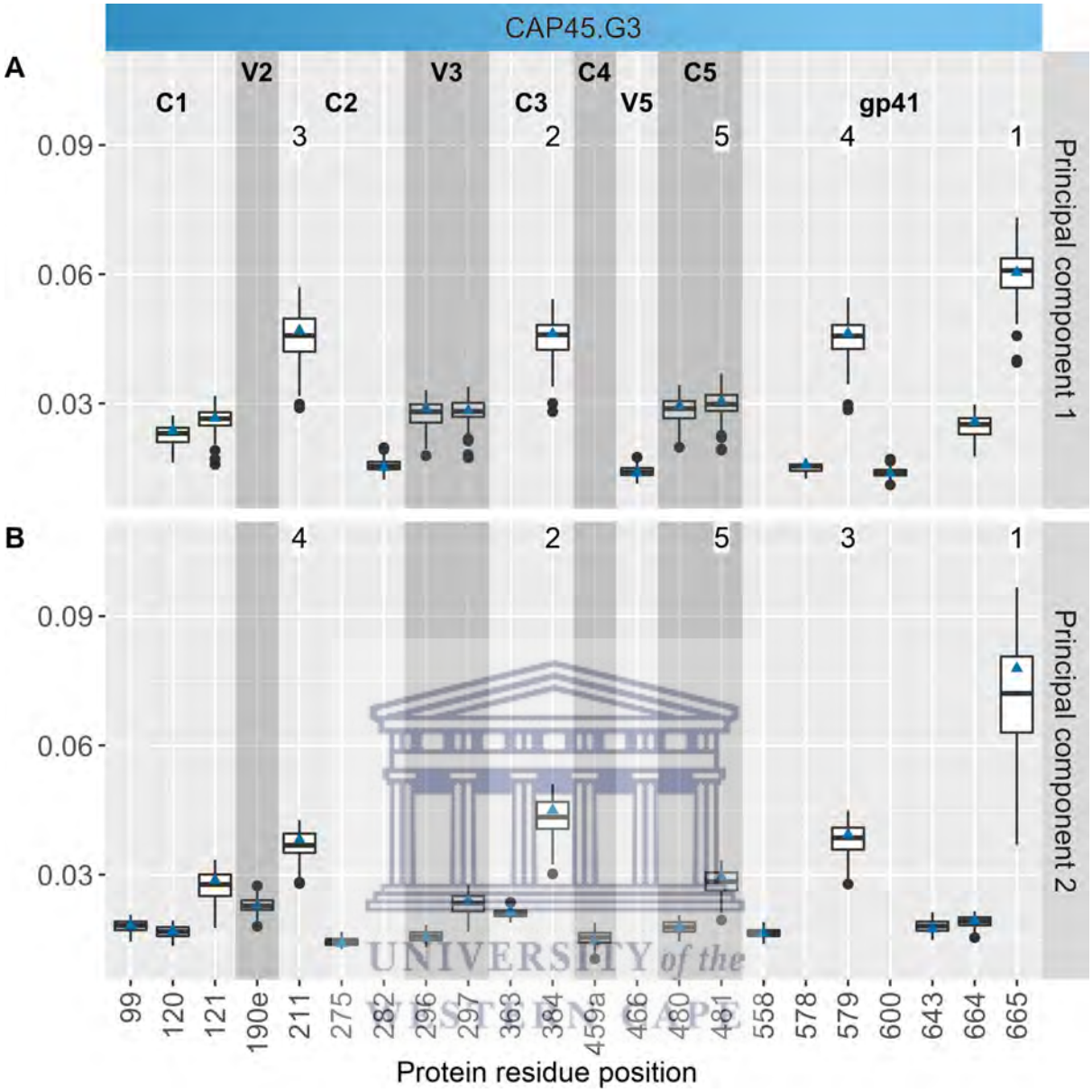

Figure 3.25: Factor loadings of the CAP45.G3 wild-type protomer comparison focussing on protein residues

Box and whisker plots of the factor loadings of protein residues for the (A) first and (B) second principal components with a median value above a threshold (0.1) such that only the top protein residues are displayed. Blue triangles represent the factor loadings calculated using the simulated trajectories. The rankings based on median factor loadings are show for the top five protein residues.

3.3.7 The locations of exposed surface regions are associated with the unique glycan conformational changes in response to the N301A mutation

The analyses thus far have demonstrated several interesting observations regarding the conformational heterogeneity between the simulations. Briefly, the PNGS locations

of the CAP45.G3 and Du156.12 wild-type models are considerably different. Consequently, the members of their respective N301-glycan clusters were also diverse. Nevertheless, three glycans (N156, N197 and N442) were N301-glycan cluster members across the protomers of both wild-type models. However, further investigation of these glycans revealed that not only were their conformations heterogeneous, but also that the landscapes around each of these glycans were distinct for each wild-type model. Subsequently, removing glycan N301 from each wild-type model resulted in different altered conformations for these glycans, especially glycans N197 and N442, since the glycan shields of each model responded to the change in specific ways. Therefore, it is likely that both the conformational differences of these glycans on the wild-type models, as well as the landscapes around each of these glycans, affected the ultimate changes in the glycan shields.

To assess the influence of these unique responses to the N301A mutation in a meaningful way across the protein, the AASA was calculated using a 10Å probe (as an approximation of the size of an antibody). When comparing the AASA between paired models, differences could be due to protein and/or glycan movement. To focus on the glycan changes specifically, it was therefore necessary to normalise the AASA values; thus, the AASA, per residue, of the glycosylated protein was divided by the AASA of the non-glycosylated protein, which represents the “maximum” possible AASA per residue. Therefore, a proportion of 20% indicates that only 20% of the surface area for that residue remains accessible when the glycans are present.

Subsequently, the protein residues of the N301A mutant models with statistically significant increases in their average AASA measures compared to the wild-type models, were identified (Figure 3.26, blue and orange regions). Furthermore, these result-sets of protein residues were further classified to distinguish residues with a substantial increase in the average AASA measure, i.e. an average increase of 10% or greater (Figure 3.26, blue regions). The location, and number, of clusters of these residues on the protein surface were compared between the CAP45.G3 and Du156.12 N301A mutant models.

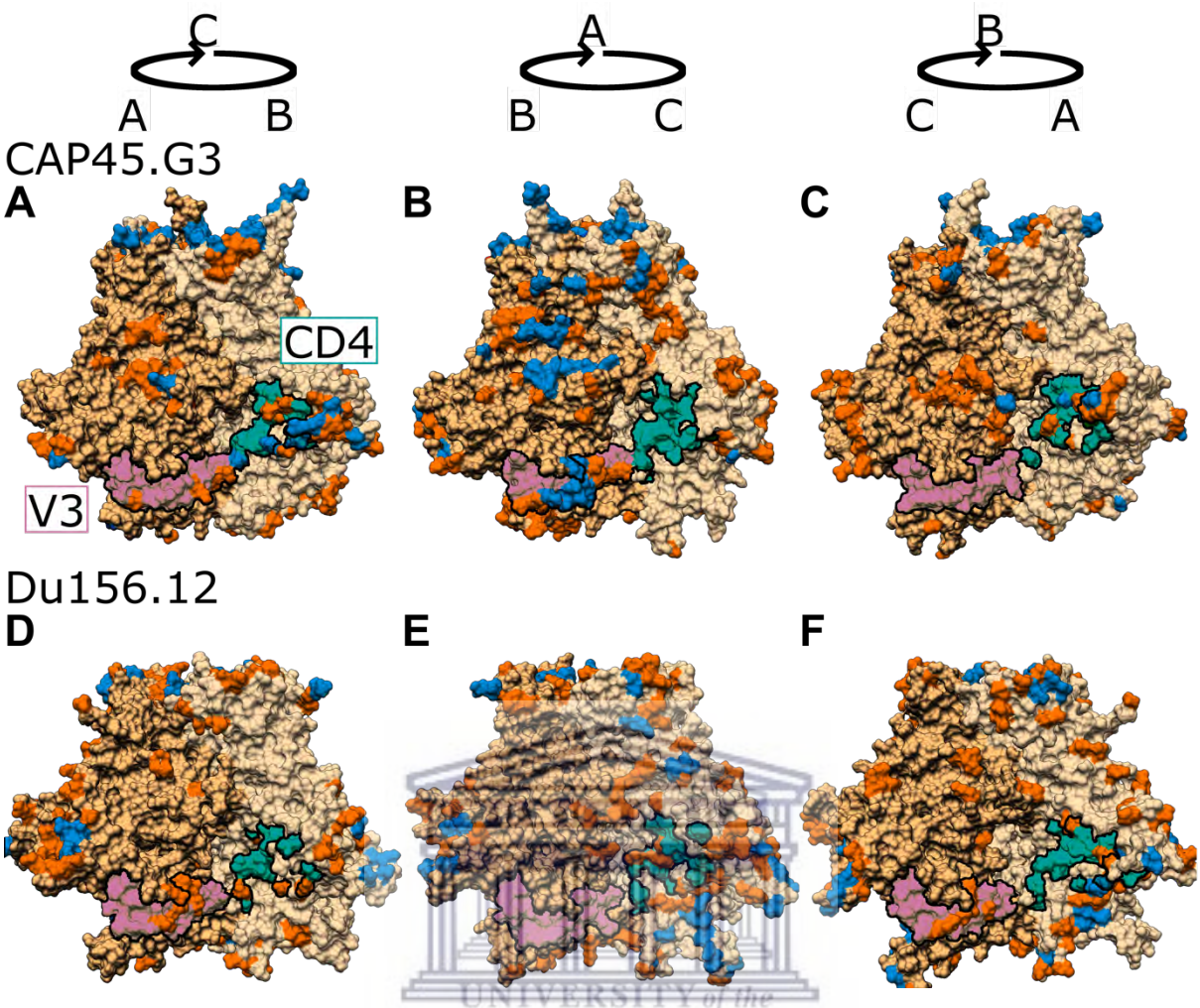

Figure 3.26: AASA for the N301A mutant models

The N301A mutant models of (A, B, C) CAP45.G3 and (D, E, F) Du156.12 with different orientations (A – C, 120° increments) to show each protomer (different shades of tan) and the residues with statistically significant increases in their average AASA ratios relative to their wild-type counterparts (difference in average AASA <10% is orange, ≥10% is blue). A statistically significant increase was evaluated at a 5% significance level using a bootstrap approach (see Methods). The V3-loop (pink) and CD4 (green) regions are outlined.

Although the number of statistically significant residues with a substantial increase were roughly equal for each N301A mutant model (157 for CAP45.G3 and 187 for Du156.12), there were clear differences in the clustering of these residues between the CAP45.G3 and Du156.12 N301A mutant models (Figure 3.26, blue regions). For the CAP45.G3 N301A mutant model, the identified residues were relatively spread out, with the exception of protomer B where a cluster, in either the C1/C2, V3 or C3/V4 regions, was observed. On the Du156.12 N301A mutant model, however, the identified residues were clustered in the C3/V1/V4 regions on all the protomers. This

suggests that removing a glycan from the glycan shield not only uniquely impacted the conformations of the surrounding glycans, but that these differences may have further implications regarding the specific regions exposed after the N301A mutation was introduced. This is an important point, since predicting the regions exposed after the removal of a glycan may offer clues on where antibodies can bind. Moreover, the location of these regions may further indicate the relative ease with which antibodies can bind, since some regions on Env are generally more accessible than others (Supplementary Figure S3.9).

3.4 Discussion

Traditionally, detailed inspection of crystallised glycoprotein structures has allowed for certain observations regarding the conformations of N-linked glycans. However, apart from the protein-proximal GlcNAc residues, it is difficult to resolve the conformations of HIV-1 N-linked glycans unless the glycans are stabilised[83–92] (section 1.1.1). This is thought to be due to the conformational heterogeneity of glycans[83–92]. Molecular dynamics simulations present an alternative approach for studying glycan conformations, since a converged simulation provides an ensemble of 3D structures from which it is possible to compute the relative populations of each conformational state observed[120]. These simulations further allow comparison between different glycoprotein structures, and, therefore, a comparison between HIV-1 strains, which was the focus of this chapter.

In order to understand whether the conformations of equivalent glycans (i.e. shared HXB2 numbering) vary between wild-type models, and additionally if these conformations were impacted by the removal of a glycan, the conformational differences during the molecular dynamics simulations of four models, two wild-type and two N301A mutant models, were investigated in detail.

The comparison between the wild-type models revealed that equivalent glycans had a wide range of conformational differences and diverse glycan-glycan interactions (sections 3.3.3 and 3.3.4). Since glycan conformations are constrained by neighbouring glycans, variation in the locations of neighbouring glycans led to this

observed diversity of glycan conformations and glycan-glycan interactions. This finding is in contrast to the observation from Steward *et al.*[88], who compared the glycan conformations of three crystal structures and where a conserved structural motif of the protein-proximal GlcNAc residues was reported. Here, however, the conformational diversity was not restricted to the mannose branches, but also observed for the protein-proximal GlcNAc residues. This difference is likely attributable to the relatively larger disparity between the locations of neighbouring glycans between the sequences considered here compared to the Steward *et al.*[88] study, where strains with more similar PNGS profiles were investigated. The differences, and similarities, between PNGS profiles, which provide an overview of potential glycan neighbours, is therefore an important factor when extrapolating information about glycan conformations to other HIV-1 Env glycoprotein structures (and thus HIV-1 strains). Importantly, equivalent HXB2 numbering does not necessarily translate to equivalent conformational characteristics.

The next part of the investigating focused on the paired-model comparison, where the wild-type models were each compared to their N301A mutant model counterparts. This analysis revealed that at least one of the glycans neighbouring glycan N301 displayed altered conformations when comparing the wild-type and N301A mutant models. Due to the extra space generated by removing a glycan, it was expected that the glycan conformations of the N301A mutant models would exhibit greater overall differences compared to those of the wild-type models. However, this was not always the case since certain glycans established “new” equilibrium positions, which were less flexible and more consistent across the protomers (e.g. glycan N446, section 3.3.5.1.1, Figure 3.11). Extending the analysis using a different approach, beyond only the neighbouring glycans, revealed that while glycans around the N301A mutation were among those with relatively greater variation between paired models, further glycans also displayed relatively large variations. This suggests that there were farther reaching consequences to removing a glycan than were initially expected.

Altogether, the results from the wild-type and paired-model comparisons established that the distinct wild-type conformations, and implicitly the landscapes, were linked to the observed changes and differences after removing a glycan. This was demonstrated by distinct glycan conformations not only emerging between paired

models, but also between the N301A mutant models. The distinct glycan conformations, in turn, impacted the location of the areas likely to be exposed, if any, due to the N301A mutation. Thus, these results again demonstrated that equivalent glycans are not as equivalent as implied by the identical HXB2 numbering, and that the landscape surrounding each glycan is an important consideration. Therefore, HIV-1 Env glycans should not be considered in isolation, but in their glycan shield "context".

A caveat for the comparison of glycan conformations from molecular dynamics simulations is that the observed differences may be due to distinct equilibrium glycan conformations sampled on relatively short time scales, whereas the average properties of glycan conformations across an unlimited time scale are the same. This "random" variation was clearly seen when glycans were compared between the protomers of the same models. In this case, observed variations in glycan conformations were due to glycans occupying different equilibrium conformations and one would expect, given a long enough time scale, that these glycans would exhibit similar average properties due to their near identity. However, observing enough of the sample space during molecular dynamics simulations, i.e. a long enough trajectory, is a key, and ongoing, problem for molecular dynamics research of large structures. Since extending the simulations to an indefinite time scale was impractical, an alternative approach was considered here, i.e. using the variation between protomers as proxy for quantifying random variation (sections 3.3.5.3 and 3.3.6). Therefore, only where the variations between models exceeded that which was observed between protomers, were these variations considered as substantial differences. While this may have excluded glycans with small, yet statistically significant differences, the aforementioned, more cautious, approach was preferred during this investigation.

Several studies have focussed on the changes to exposed HIV-1 Env surface areas when different mutations are introduced[53–55,66–71]. While this remains an important question, a further consideration relates to how the changes in glycan conformations impact the efficacy of glycan-binding antibodies. This particular line of enquiry was previously investigated by Kwong *et al*.[51] using crystal structures of HIV-1 Env bound to different glycan-binding antibodies; these crystal structures contained a number of resolved glycans conformations that were compared. The authors speculated that the

initial orientation of an entire glycan might not be critical for glycan-binding antibodies that bind and recognise only a contiguous region on a single glycan[51]. However, for glycan-binding antibodies that bind multiple glycans, or multiple glycans and the protein surface, the initial orientation of the entire glycan, including its degree of flexibility, may be important[51]. This is due to the additional constraints on the angle of approach that is required for successfully binding to occur between the glycan-binding antibody and its target[51,162]. Therefore, the conformations of glycans can play a notable part in the ability of glycan-binding antibodies to recognise a glycan and are also likely to affect the elicitation of the particular antibodies[51]. Thus, the glycan conformations should be considered along with the glycan-binding antibody. Unlike the antibody-bound crystal structures studied by Kwong *et al.*[51], the glycan conformations during the molecular dynamics simulations described here, allow for the additional investigation of the conformations of these glycans when unbound to the glycan-binding antibody.

Thus, this chapter demonstrates that equivalent glycans can adopt conformations distinct to specific wild-type Env models, in this case the CAP45.G3 and Du156.12 models. Moreover, these model-specific conformations also emerged when only a single glycan was removed from the glycan shield, and these differences were not limited to glycans in the immediate vicinity of the removed glycan. Additionally, the model-specific glycan conformations that emerged after glycan N301 was removed, were unique to each of the N301A mutant models, which was consequently directly related to whether certain protein residues were exposed. However, due to random variation implicit to macromolecular structures, there remained a degree of uncertainty as to whether the conformational heterogeneity observed for glycans located farther from the N301A mutation were a direct result of the mutation. Therefore, associating the conformational heterogeneity of a glycan to the original mutation was the focus for the subsequent investigation (Chapter 4).

Chapter 4 Structural Rearrangements Maintain the Glycan Shield of an HIV-1 Envelope Trimer After the Loss of a Glycan[h]

As highlighted in Chapter 1, the HIV-1 Env glycans shield the underlying protein from the immune system and by removing a glycan, a hole, or gap, is created in the glycan shield. However, as described in Chapter 3, distinct glycan conformations were observed for the two HIV-1 Env wild-type subtype C isolates and their corresponding N301A mutant counterparts. These distinct conformations provide one potential explanation as to why one of these N301A mutant isolates not only remained protected against frequently elicited neutralising antibodies, but also exhibited increased resistance to a CD4-binding site broadly neutralising antibody despite the missing glycan. However, drawing a direct association between the conformational heterogeneity of any glycan and the loss of glycan N301, required a more detailed analysis. Therefore, in order to further investigate the resistant strain and directly associate the conformational heterogeneity of its glycans with the N301A mutation, a more detailed comparison between this wild-type model and its N301A mutant counterpart was carried out and is presented in this chapter, as published online in Scientific Reports (DOI:10.1038/s41598-018-33390-2).

4.1 Abstract

The HIV-1 envelope (Env) glycoprotein is the primary target of the humoral immune response and a critical vaccine candidate. However, Env is densely glycosylated and thereby substantially protected from neutralisation. Importantly, glycan N301 shields V3 loop and CD4 binding site epitopes from neutralising antibodies. Here, we use molecular dynamics techniques to evaluate the structural rearrangements that maintain the protective qualities of the glycan shield after the loss of glycan N301. We examined a naturally occurring subtype C isolate and its N301A mutant; the mutant not only remained protected against neutralising antibodies targeting underlying

[h] *This chapter, from the Abstract onwards, is included verbatim as published in Scientific Reports (DOI:10.1038/s41598-018-33390-2). Where necessary, minor changes, e.g. Figure numbering, were made to ensure reasonable flow and consistency between chapters of this thesis.*

epitopes, but also exhibited an increased resistance to the VRC01 class of broadly neutralising antibodies. Analysis of this mutant revealed several glycans that were responsible, independently or through synergy, for the neutralisation resistance of the mutant. These data provide detailed insight into the glycan shield's ability to compensate for the loss of a glycan, as well as the cascade of glycan movements on a protomer, starting at the point mutation, that affects the integrity of an antibody epitope located at the edge of the diminishing effect. These results present key, previously overlooked, considerations for HIV-1 Env glycan research and related vaccine studies.

4.2 Introduction

A key scientific challenge in the field of HIV-1 vaccine development is the design of immunogens that elicit antibodies capable of neutralising the wide range of HIV-1 isolates in circulation, despite its immense global diversity[7]. Due to immune-mediated selection pressure, the majority of this diversity is in the viral *envelope* gene that encodes the Env proteins on the surface of a virion[171]. The Env proteins facilitate viral entry to target cells and are formed by gp120/gp41 heterodimers that non-covalently associate, forming a trimer of heterodimers[172,173]. Even though the majority of HIV-infected individuals mount an immune response targeting these Env trimers, within-host diversity ensures that certain strains continue to evade recognition and neutralisation[56,174–176].

Large sections of the Env trimers are covered by dense glycosylation and roughly half of its molecular mass is made up by glycans[177,178], which have been suggested to protect the virus from antibody binding and neutralisation[42,44,53–55]. Changes in these glycosylation patterns can therefore have a large impact on its ability to escape from immune attack. Once HIV infiltrates the host cells, it takes advantage of host cellular biosynthetic pathways for its own benefit, which includes protein glycosylation as one of the main post-translational modifications[45]. N-linked glycosylation occurs in the endoplasmic reticulum and Golgi apparatus, where glycans are attached to asparagine residues within an Asn-X-Ser/Thr motif (X is any amino acid except proline[45]). The attached glycans, initially assumed to be immunologically inert "self"

molecules, were until recently considered a largely insurmountable challenge for antibody recognition; hence termed, the “glycan shield”[56,57].

However, some HIV-1 infected individuals develop potent and broadly neutralising antibodies (bNAbs) that specifically target, or find ways to bypass, the glycan shield[39–43]. These bNAbs are characterised by their target region, and are generally defined by particular monoclonal antibodies that target specific regions: the CD4 binding site[59], the membrane proximal external region of gp41[60], the glycan outer domain (typified by mAb 2G12)[61], the V1V2 apex region around glycan N160[58], the V3 base around glycans N301 and N332[62], and the gp120/gp41 interface[63]. Despite the presence of such bNAbs in the serum of infected individuals, circulating plasma viruses generally escape, resulting in continued infection[64,65]. This escape from bNAbs has been linked to shifting glycosylation sites or mutations in the protein sequence surrounding specific glycans[42,44]. For example, Lynch *et al.*[44] showed that the mutation introducing a glycan at position N276 lead to partial VRC01 resistance, and Moore *et al.*[42] reported that the shift of glycan N332 to position N334 resulted in PGT128 resistance.

To quantify the impact of glycans on the effectiveness of neutralising antibodies, previous *in vitro* studies have used targeted de-glycosylation to compare the neutralisation of a range of viral strains with and without a specific glycan[53–55,70]. For example, the removal of glycan N301 (HXB2 numbering throughout), which is highly conserved[47,179] amongst HIV strains, has been shown to expose V3 loop and CD4 binding site epitopes[66–69,71]. However, Moyo *et al.*[53] described a subtype C strain, CAP45.2.00.G3 (referred to herein as CAP45.G3), in which removal of the glycan at position 301 unexpectedly did not result in increased sensitivity to to a large proportion of sera (61/64 panels) from chronically infected individuals[53], despite its central role in protecting other isolates from neutralisation[39,53,54,66,67,69,71]. Furthermore, the N301A mutant had increased resistance to the CD4 binding site bNAb VRC01, and other VRC01-like bNAbs, when the neutralisation profile was compared to that of the wild-type[53]. The authors suggested that this virus typified a subset of viruses that could tolerate the loss of glycan N301 while largely maintaining the protective qualities of the glycan shield. The mechanism of compensation for the loss of a glycan, in this case glycan N301, as well as the development of increased resistance to VRC01, was not understood.

Here, we explored glycan conformations *in silico* to explain these findings. We analysed two molecular dynamics simulations of glycosylated Env trimers: the CAP45.G3 wild-type and the CAP45.G3 N301A mutant, which removes the glycosylation site at residue 301, to establish whether the models replicated the *in vitro* compensation of the glycan shield observed previously[53]. Subsequently we describe, in detail, the structural changes of glycans N442, N446 and N262 that bear the burden of compensation, how this burden is distributed, and the differences observed between the protomers of each model. We show that by determining the glycan nearest to each protein residue over time, we can clearly illustrate how changes in glycan conformations impact their ability to protect certain residues of the underlying protein. Finally, our study demonstrates how a cascade of events could contribute towards the increased resistance to antibodies targeting an epitope distal to the point mutation, in this case the N301A mutation and the VRC01 epitope. These *in silico* data provide a detailed investigation of the glycan shield's ability to compensate for the loss of a glycan as well as the associated cascade of events that affect a distal epitope, which provides further important considerations and avenues of exploration for vaccine studies focussing on the HIV-1 Envelope.

4.3 Methods

4.3.1 Structural modelling and molecular dynamics simulations:

System preparation

We used MODELLER[136,137] to generate the protomer structures. The CAP45.G3 (Genbank accession number DQ435682) sequence was used as the target, and three reference structures, PDB IDs 4NCO[91], 4TVP[87] and 2B4C[139] were used as starting templates. The modelling was repeated ten times and models were ranked according to their DOPE scores[140]. The model with the lowest DOPE score was then selected, triplicated, and the three copies were aligned to the protomers of the 4NCO trimer structure to generate the trimeric model. Potential N-linked glycosylation sites (PNGSs) were determined by identifying the NXT/S motifs, where X is any amino acid except a proline. For each PNGS on the trimer homology model, we attempted to

attach a Man9GlcNAc2 glycan (Manα1-2Manα1-6[Manα1-2Manα1-3]Manα1-6[Manα1-2Manα1-2Manα1-3]Manβ1-4GlcNAcβ1-4GlcNAcβ1-). The glycans were generated using the carbohydrate builder available at Glycam-Web[145] and were attached using a prototype tool under development for the Glycam-Web suite of web tools that explores the most populated rotamers of the Asn-GlcNac linkage[180] and then, if clashes are observed between previously added glycans or the protein, iteratively rotates the Asn sidechain and glycosidic linkages within reasonable bounds. Two of the glycosylation sites, N335 and N678, were not glycosylated; site N335 occurs in an NNST motif and we were specifically interested in the positioning of glycan N334 in the context of bNAbs[42] and due to its relative abundance compared to N335[47], and N678 falls outside the modelled domain. For the remaining 81 PNGSs, 79 were computationally glycosylated. During the first round, glycosylation of sites N160 (protomer C), N399 (protomer B), and N386 (protomer C) failed. However, after a 30 ns simulation (according to the steps described below), we retried glycosylating these three sites; only site N386 (protomer C) was successfully glycosylated, whereas sites N160 (protomer C) and N399 (protomer B) remained unglycosylated. It is unknown whether, but unlikely that, the three protomers of an Env trimer will always have matching glycosylation profiles; although position N160 is widely known as a conserved PNGS, researchers have used mass spectrometry to quantify the presence of glycans at PNGSs of 94 cross-clade HIV-1 gp120 proteins and the results show that glycosylation at N160 is not absolute[169]. Furthermore, the dependence of a glycan at this position for effective neutralisation has been shown to be variable[42], and the particular binding characteristics of bNAb PG9 and PG16 to glycan N160 does not necessitate trimer specific glycosylation, i.e. the antibodies recognise a single N160 glycan[163]. Therefore, although we could not glycosylate each PNGS of the model (positions N335 and N678, and positions N160 on protomer C and N399 protomer B), each viral clone will have varied glycosylation patterns and our 'fully' glycosylated model represents one form of the CAP45.G3 Env glycoprotein. We continued with this fully glycosylated model to create the N301A mutant by replacing the asparagine residue at position 301 (HXB2 numbering) with an alanine residue and removing the glycan.

The systems were created using the tLEaP package contained in AmberTools 14[123]. The ff14SB[121] force field was used for the protein and the GLYCAM06j-1[122] force field

for the glycans. The wild-type and N301A mutant systems were immersed in a truncated octahedron water box containing TIP5P[124] water molecules, since TIP5P was found to produce the best quantitative agreement with experimental data[125]. The box size was set such that all protein and glycan atoms were 15 Å from the edge of the box. Chloride ions were added to neutralise the system.

4.3.2 Simulation

The molecular dynamics simulations were produced using AMBER 14[123]. The systems were minimised by running 10,000 steps of steepest descent and 10,000 steps of conjugate gradient minimisation. During minimisation restraints were placed on all non-hydrogen protein and glycan atoms. The systems were equilibrated by running 0.4 ns simulations under nPT (1 bar, 300 K) on a CPU cluster. During the first equilibration stage, Cartesian restraints (5.0 kcal/mol) were placed on all non-hydrogen protein and glycan atoms. The restraints were removed, and the equilibration was extended by another 1 ns on a GPU cluster to ensure stability across clusters before the production run started on the GPU cluster. The 520 ns production runs were generated on a GPU cluster using AMBER GPU acceleration pmemd[123] and 0.002 ps time steps, with coordinates written to the trajectory file every 10,000 steps.

4.3.3 Analyses

4.3.3.1 Root Mean Square Deviation to determine system equilibrium

The conformational stability of the protein during the simulation was assessed by calculating the root mean squared deviation (RMSD) between the protein backbone atoms (C, C-alpha, N, O) and the starting structure of the production run. Conformational stability was achieved after 20 ns (data not shown) and this section of each trajectory was therefore discarded; the remaining 500 ns simulations were used for further analyses.

4.3.3.2 Antibody-accessible surface area for different regions and residues

To investigate the efficiency with which the glycan shield compensates for the loss of glycan N301, we calculated the antibody accessible surface area (AASA) using a 10 Å probe (as an approximation of the size of an antibody[87]) with Naccess[146] for both the wild-type and N301A mutant simulations. The van der Waals radii of the glycan atoms were defined as described for the GLYCAM06j-[122] force field. The AASA was calculated for 2,500 evenly spaced frames across the 500 ns trajectory.

Differences between the wild-type and N301A mutant AASA values could be attributed to either the protein and/or glycan movements and we, therefore, normalise the values to remove changes in the AASA due to protein movements. To do this, we determined the "base/maximum" AASA by removing the glycans and re-calculating the AASA for these non-glycosylated frames. The final AASA ratio, per frame, is the ratio of the glycosylated and non-glycosylated AASA values, calculated by dividing the AASA of the glycosylated protein by the AASA of the non-glycosylated protein. Averages were calculated using these AASA ratio over time.

We compared the mean AASA ratio, per residue, of the N301A mutant simulation to that of the wild-type to determine whether there was a statistically significant increase. The AASA ratio distribution of each residue under the null-hypothesis (wild-type and N301A mutant means are equal) was assessed by using 100 moving-blocks bootstrap replicates. The AASA ratio datasets for each model were mean normalised to satisfy the null-hypothesis before bootstrap replicates were drawn. The 50 ns blocks were defined such that each block could only start on whole nanoseconds for adequate sampling of the whole trajectory. There were, therefore, 450 possible blocks, with ten random blocks required for each bootstrap replicate. Variance normalisation was not performed due to numerous residues with average AASA ratio over time equal to zero. We further filtered the significant residues to include only those where the difference between the average AASA ratio for the wild-type and N301A mutant over time was 10% or greater. We opted for moving-blocks bootstrap above a normal bootstrap approach to conserve the correlation between sequential observations.

4.3.3.3 Nearest glycan calculation for all protein residues

Since the AASA results revealed that some residues have increased AASA ratios in the N301A mutant when compared to the wild-type simulation, we wanted to understand the shielding capacity of the surrounding glycans and determine which glycans are the most likely to affect the AASA ratio of particular protein residues. To achieve this, we calculated, for each residue, the distances between its atoms and all the atoms of the glycans. Each residue is subsequently assigned a nearest glycan based on the calculated distances. The nearest glycan for each residue was calculated for 25 evenly spaced frames across the 500 ns trajectory. The glycan that was most frequently the nearest to a particular residue throughout the trajectory, was defined as that residue's "nearest glycan".

4.3.3.4 Overlap between wild-type glycans and specific glycans from the N301A mutant model

The nearest glycan calculation revealed which glycans were nearest to the protein residues that had increased AASA ratios, and showed conformational changes, during the N301A mutant simulation compared to that of the wild-type. In order to determine if these conformational changes were specifically due to the additional space generated from the absence of glycan N301, we investigated whether the changes were viable when glycan N301 was present, i.e. whether or not the identified glycans caused large conformational clashes with any of the wild-type glycans. Hence, we iteratively removed the identified nearest glycans from the wild-type model trajectory and replaced each with the N301A mutant model trajectory equivalent. One glycan was analysed at a time and then restored before the next N301A mutant glycan's overlap calculations were performed. All the frames of the wild-type and N301A mutant models were aligned to the first frame of the wild-type trajectory before the analysis was carried out. The overlap was calculated using UCSF Chimera[168] for 2,500 evenly spaced frames across the 500 ns trajectory. The default values were used for the van der Waals overlap (0.6 Å) and potential hydrogen bonding between clashing pairs (subtract 0.4 Å).

4.3.3.5 Tree representation of hydrogen bonds between glycans to illustrate the cascade of events after the loss of a glycan

In order to determine which further glycans were affected by the absence of glycan N301, but not immediately adjacent to the N301A mutation, we calculated and compared the hydrogen bonds that formed between glycans in the N301A mutant and wild-type simulations. The number of hydrogen bonds that formed between any two glycans was calculated for all 25,000 frames. The proportion of the total frames in which there was a hydrogen bond between any two glycans was calculated and rounded to the nearest percentage. The analysis was performed using the hbond function in cpptraj[166] contained in AmberTools 14, using the default cut-off values for the distance between the heavy atoms (3 Å) and the angle between the acceptor and donor atom (135 degrees). The glycans that showed a change of 10% or greater in the frequency of the interaction between the wild-type and N301A mutant simulations were represented on tree graphs. The analyses tree graphs were generated using R[147] and, where necessary, the trajectory data was parsed into R using the Bio3D package[181]. Visual representation of the protein structures were generated using UCSF Chimera[168].

4.4 Results

The CAP45.G3 subtype C virus was used to model the wild-type and N301A mutant trimer structures, comprising of protomers A, B and C, for molecular dynamics (MD) simulations. The CAP45.G3 Env sequence has 29 potential N-linked glycosylation sites (PNGSs, Figure 4.1) and although we could not computationally glycosylate each PNGS, given the extent of variation seen during glycan occupancy studies[49,72–82,95], the generated model represents one possible form of the wild-type glycosylated model (Figure 4.2). The N301A mutant model matched the wild-type model, except the asparagine at position 301 was replaced with alanine and the glycan excluded. The systems were equilibrated for the first 20 ns of the simulation and analyses were performed on the remaining 500 ns.

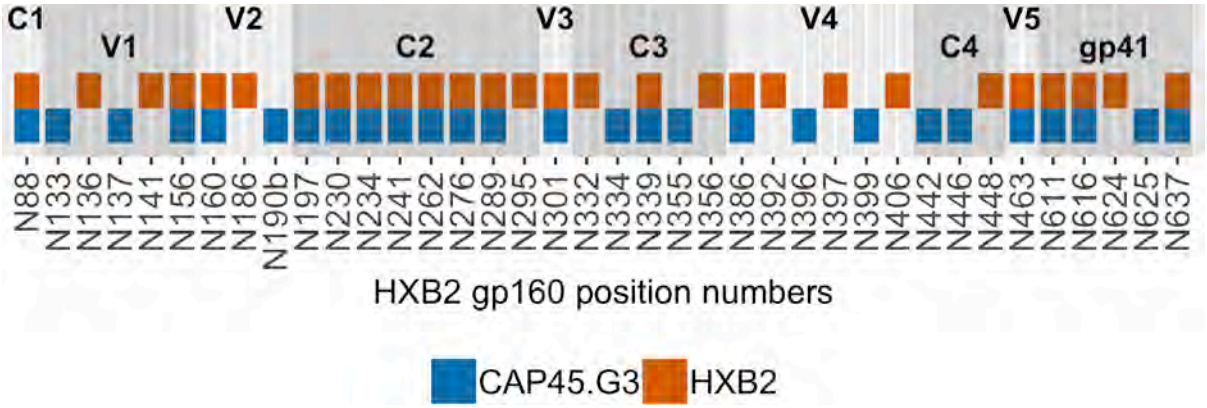

Figure 4.1: Distribution of PNGSs of the CAP45.G3 strain

Distribution of PNGSs for the CAP45.G3 strain (blue) relative to the PNGSs of the HIV-1 reference strain, HXB2 (orange). The gp160 conserved (C1-C4), variable (V1-V5) and gp41 sequence regions are labelled and shaded to indicate the borders of each region.

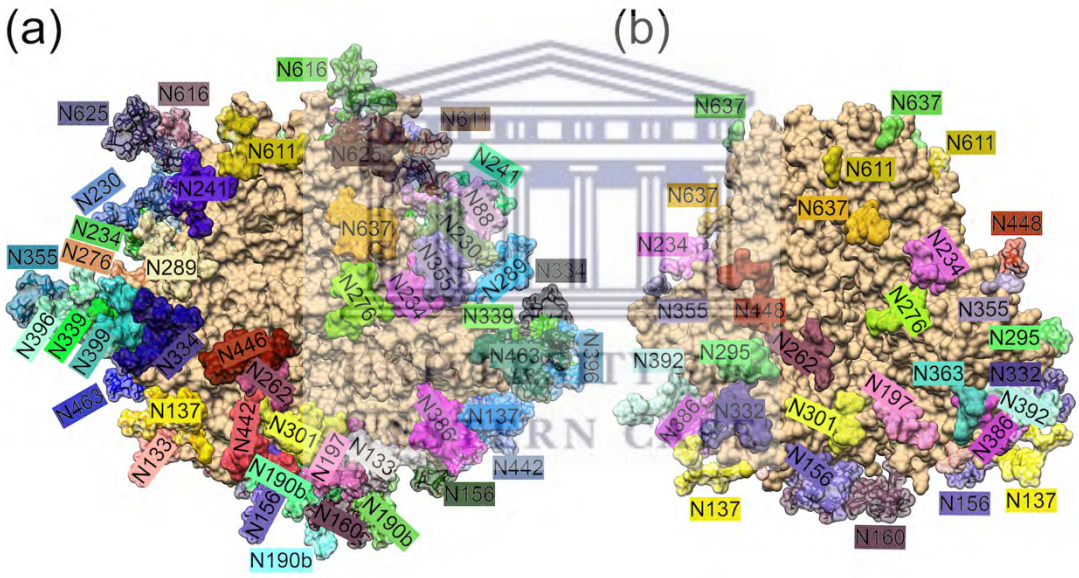

Figure 4.2: The CAP45.G3 computationally determined model vs. a crystal structure

3D representation of the N-linked glycosylation sites of the (a) CAP45.G3 strain (computationally determined) and the (b) clade A BG505 strain[88] (crystal structure). The protein and glycan residues are shown as surfaces and the glycans are labelled according to HXB2 numbering. The depicted orientation is such that the lipid membrane is located at the top and the V1/V2-loop regions are at the bottom of the figure.

4.4.1 Region-specific changes in the average antibody-accessible surface area (AASA) between the wild-type and N301A mutant viruses

To investigate how the glycan shield compensates for the loss of glycan N301 in CAP45.G3, we calculated the antibody accessible surface area (AASA) of the wild-type and N301A mutant models using a 10 Å probe (as an approximation of the size of an antibody[87]) with Naccess[146]. When directly comparing AASA values between the wild-type and mutant simulations, there are differences due to underlying protein movement (particularly in the loop regions) and as a result of the loss of a glycan. In order to isolate the effect of the glycan loss, we divided the AASA of each residue by its AASA when all glycans are removed from the same simulation. The non-glycosylated value represents the "maximum" possible AASA for a residue. We next compared these AASA ratios between the wild-type and N301A mutant simulations using a bootstrap approach (see Methods section) to calculate which residues displayed a statistically significant (5% significance level) increase in the average AASA ratios between the wild-type and N301A mutant simulations.

From the statistically significant subset (Figure 4.3, blue and orange residues), we distinguished those residues with a considerable increase in their AASA ratios, i.e. an average increase of 10% or greater, which we define as 'substantial' (Figure 4.3, blue residues), to further isolate those residues that were most affected by the loss of glycan N301. Residues identified in this manner were observed across all protomers and a total of 14, 42 and 18 residues were found for protomers A, B and C, respectively. Two clusters, one within the V3 region (Figure 4.3, pink and outlined) and one within the CD4 binding site region (Figure 4.3, green and outlined), were apparent.

These clusters (Figure 4.3 (a) and (b), blue clusters) with a substantial increase in their AASA ratios (average increase of 10% or greater) surround the N301A mutant site, and since the clusters fall within the V3 loop and CD4 binding site regions, we defined each cluster as the "V3 sub-region" (including residues 303-305, 321-322, 323, and 440) and "CD4 sub-region" (including residues 198, 365, 367, 368, 460, 462, and 464) (Table 4.1).

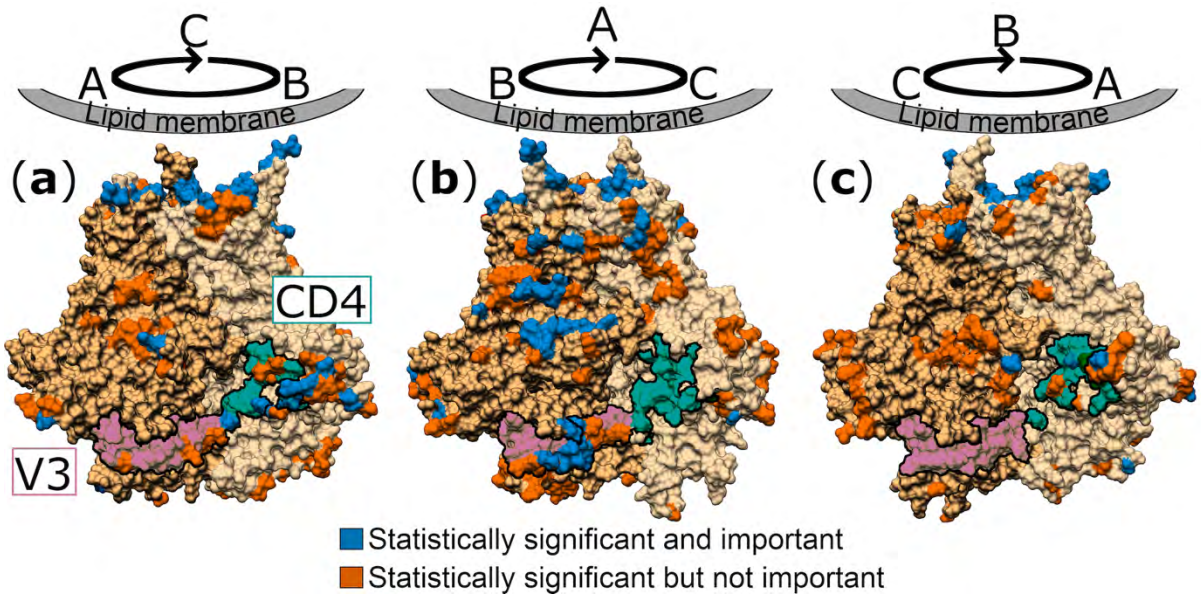

Figure 4.3: AASA for the CAP45.G3 N301A mutant model

N301A mutant model with different orientations (A – C, 120° increments) to show each protomer (different shades of tan) and the residues with statistically significant increases in their average AASA ratios relative to the wild-type simulation (difference in average AASA <10% is orange, ≥10% is blue). A statistically significant increase was evaluated at a 5% significance level using a bootstrap approach (see Methods). The V3 (pink) and CD4 (green) regions are outlined.

Table 4.1: Residues with significantly different AASA values between the wild-type and N301A mutant simulations.

SUB-REGION	[ENV REGION] RESIDUES
V3	[V3] 303-305, 321-322, 323, [C4] 440
CD4	[C2] 198, [C3] 365, 367, 368, [V5] 460, 462, 464

For these V3 and CD4 sub-regions, we calculated the average AASA ratios over time (glycosylated region total, over non-glycosylated total averaged over time) for the protomers of the wild-type and N301A mutant models ((Table 4.2). Apart from the CD4 sub-region on protomer A, the average AASA values are substantially higher for the N301A mutant sub-regions on protomer B (33% for the CD4 sub-region and 14% for the V3 sub-region, N301A mutant virus) than for the other N301A mutant or wild-type protomers.

Table 4.2: Average AASA ratios (%) of the residues for the wild-type (WT) and N301A mutant (M) viruses that form part of the CD4 sub-region and V3 sub-region.

	Protomers					
	A		B		C	
	WT	**M**	**WT**	**M**	**WT**	**M**
CD4 sub-region	30	30	8	33	20	4
V3 sub-region	<1	0	<1	14	<1	<1

Because the laboratory study showed that the N301A mutant strain had increased resistance to VRC01, and VRC01-like, antibodies compared to the wild-type strain, we also calculated the AASA ratios for the residues that form part of this epitope (residues [C1] 123, 128, 129, [C2] 276, 278-283, [C3] 365-368, 371, [C4] 427-430, 455-459, [V5] 460-461, 463, 465-467, 469, 471, [C5] 472, 474, 476[182]). Apart from protomer C, where there was a substantial decrease in the AASA ratio for the N301A mutant model (wild-type: 11% and N301A mutant: 2%), the wild-type and N301A mutant ratios remained relatively unchanged for protomers A (wild-type: 11% and N301A mutant: 16%) and B (wild-type: 10% and N301A mutant: 14%).

4.4.2 Glycan conformational changes around the N301A mutation

As shown in previous studies[42,44,53–55], and evident from the AASA ratio results focussing on glycan N301, glycosylation shields the underlying protein. To fully understand the shielding capacity of any glycan, it is important to establish which glycans are capable, and most likely given their proximity, to affect the AASA of a particular protein residue. To achieve this, we determined the nearest glycan to each protein residue by calculating the distances between its atoms and all the atoms of each of the glycans for the wild-type and N301 mutant trimer models. Since the systems are dynamic, the nearest glycan can vary between frames and we defined each glycan's 'neighbourhood' as the region that includes all the protein residues nearest to that glycan in the majority of the frames. For example, the protein residues surrounding glycan N442 form the neighbourhood of glycan N442 and can be represented on the 3D structure. These calculations allow us to determine which

glycans shield the residues of the sub-regions identified during the AASA calculations (Table 4.1).

We specifically focus on comparing the neighbourhoods of the N301A mutant protomers to their wild-type counterparts, since the starting structures are identical and thus the differences in glycan neighbourhoods are likely to be due to the loss of glycan N301. To further visualise and interpret the 3D representation of the glycan neighbourhoods, we also calculate the average centre of mass position (to illustrate the directionality of the change), throughout the simulation, for each glycan. The results are described for the two sub-regions surrounding the N301A mutation, CD4 and V3, identified during the AASA calculations.

CD4 sub-region: Our neighbourhood calculation determined which glycans were nearest to each of the residues within each sub-region. For the C3 residues (CD4 sub-region, Table 1), glycan N386 was the nearest to these residues for all wild-type protomers, as well as for protomers A and B of the N301A mutant, whereas glycan N197 was the nearest on protomer C of the N301A mutant (Table 4.3). To further investigate how glycan N197 replaces glycan N386 as nearest glycan to the C3 residues, we represent their 3D neighbourhoods on all protomers (Figure 4.4). There is a clear difference in the conformation of glycan N197 (protomer C; N301A mutant) when compared to its wild-type counterpart, and this conformational change results in a shift in the residues contained in its neighbourhood (Figure 4.4 (a) and (b), protomer C). A large proportion (29%) of the neighbourhood of the wild-type glycan N197 (protomer C) contains residues from the V3 region of protomer B. However, the neighbourhood of glycan N197 of the N301A mutant includes a large proportion (31%) of CD4 binding residues from protomer C and five fewer V3 residues of protomer B than the wild-type.

Table 4.3: The glycans nearest to the two protein clusters, with a statistically significant and substantial (≥10%) increase in their average AASA ratio (CD4 and V3 sub-region), for the three protomers (A, B and C).

	Protomers		
	A	B	C
CD4 sub-region:			
Wild-type	N386, N463	N386, N355	N276, N386, N463
N301A mutant	N276, N386	N386, N463	N197, N276, N463
V3 sub-region:			
Wild-type	N156, N262, N442	N156, N197, N301, N442	N156, N197, N262, N301
N301A mutant	N156, N262, N442	N156, N442	N156, N262

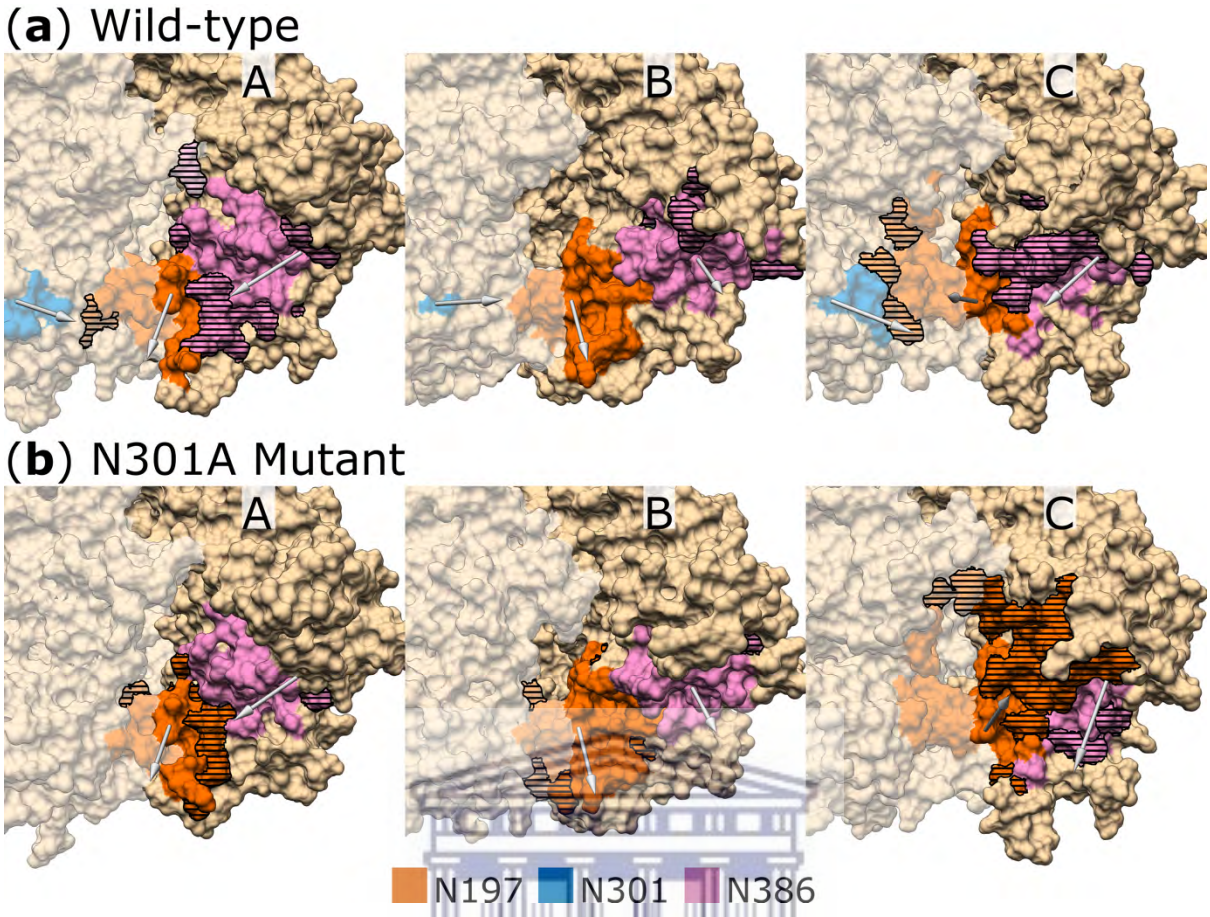

Figure 4.4: Neighbourhood of the CD4 sub-region

Neighbourhoods of the (A) wild-type and (B) N301A mutant glycans nearest to the CD4 sub-region for the three protomers (A, B, and C): glycans N197 (orange), N301 (blue, only present in the wild-type), and N386 (pink). Horizontal lines represent residues that are in different neighbourhoods when comparing the wild-type and N301A mutant. The arrows originate from the Cα atom of the Asn and end at the average centre of mass position for the selected glycans. The representations are cropped around the CD4-binding site of each protomer and the surface representation of residues on adjacent protomers are shown as opaque.

Also, for protomer C of the N301A mutant, a change in the conformation of glycan N386 is associated with the change in conformation of glycan N197 (Figure 4.4 (b) protomer C), which causes the exclusion of any CD4 binding residues from glycan N386's neighbourhood. Conversely, glycan N386 of the N301A mutant protomer A also undergoes a conformational change (Figure 4.4 (b) protomer A), but this change results in the inclusion of an equivalent number of CD4 binding residues as its wild-type counterpart.

V3 sub-region: The neighbourhoods of the glycans nearest to the V3 sub-region, N156, N262 and N442 (Table 4.3), are shown in Figure 4.5. Additionally, the neighbourhood of glycan N446 is included because its neighbourhood is itself enclosed by the abovementioned V3 sub-region glycan neighbourhoods (Figure 4.5 (a) protomer B). There are marked differences between the conformations of glycans N442 and N446 of the N301A mutant model on protomer A, and to a lesser extent those on protomers B and C, when compared to their wild-type counterparts (Figure 4.5).

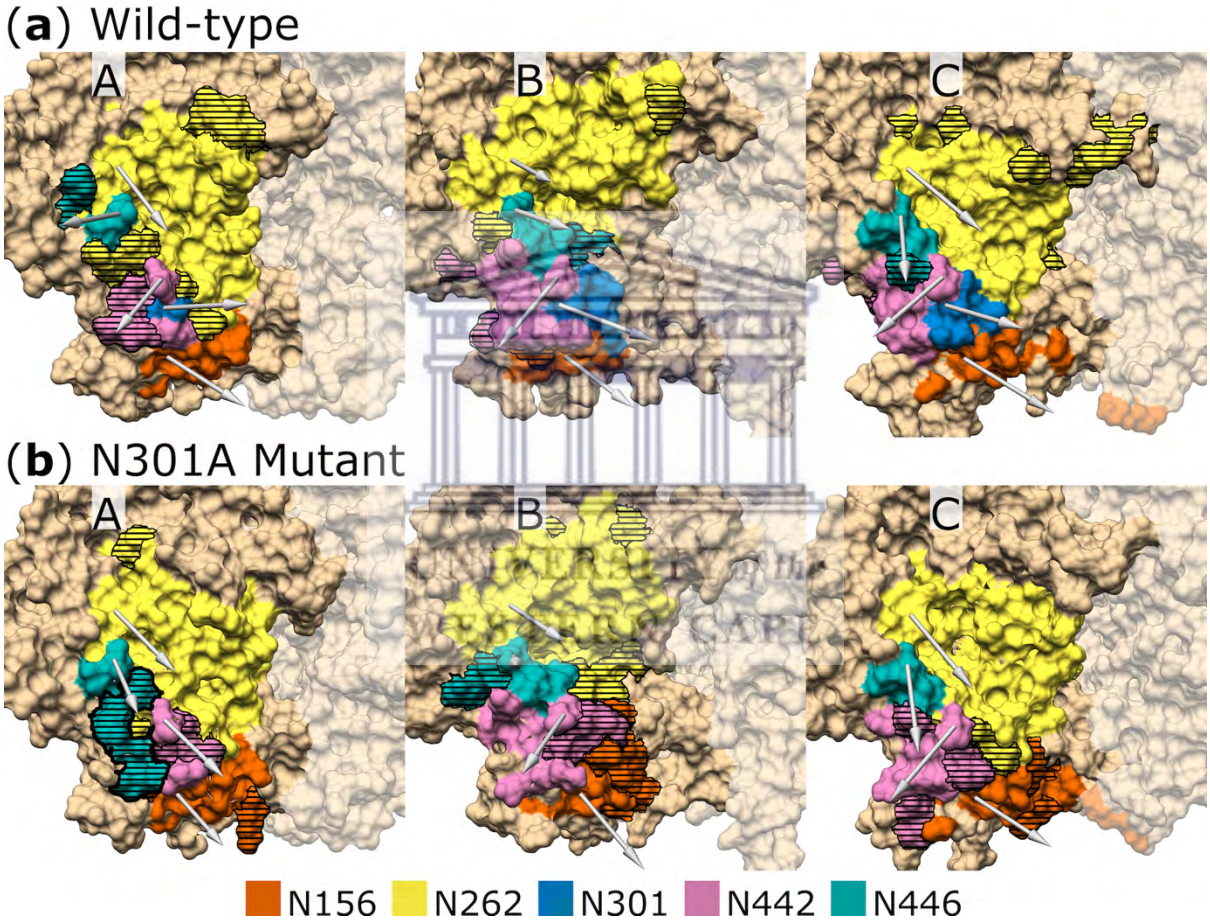

Figure 4.5: Neighbourhood of the V3 sub-region

Neighbourhoods of the (A) wild-type and (B) N301A mutant glycans nearest to the V3 sub-region for the three protomers (A, B, C): glycans N156 (orange), N262 (yellow), N301 (blue, only present in the wild-type), N442 (pink), and N446 (green). Horizontal lines represent residues that are in different neighbourhoods when comparing the wild-type and N301A mutant. The arrows originate from the Cα atom of the Asn and end at the average centre of mass position for the selected glycans. The representations are cropped around the V3 region of each protomer and the surface representation of residues on adjacent protomers are shown as opaque.

Glycan N442 (N301A mutant) re-orientates towards the N301A site in all the protomers to varying degrees, with the largest shift occurring on protomer A (Figure 4.5 (b) protomer A) and only a slight movement on protomer C (Figure 4.5 (b) protomer C). Comparatively, glycan N446 (N301A mutant) re-orientates towards the N301A site on protomers A and C only (Figure 4.5 (b) protomers A and C), while for protomer B, the glycan shifts away from the N301A site (Figure 4.5 (b) protomer B). Finally, glycan N262 on protomer C (N301A mutant) also re-orientates towards the N301A site (Figure 4.5 (b) protomer C).

4.4.2.1 Protomer-specific glycan conformational changes that compensate for the loss of glycan N301

Since glycans N197, N262, N386, N442 and N446 were identified as the glycans nearest to the protein residues that had increased AASA ratios and showed conformational changes, we investigated whether these conformations were viable when glycan N301 was present, i.e. whether or not they caused large conformational clashes with any of the wild-type glycans, which would imply that the N301A mutant glycan conformations are impossible in the wild-type model. Hence, we superimposed the N301A mutant glycan conformations of each protomer onto its wild-type counterpart and directly compared each time point. For each of these N301A mutant glycans, we calculated the proportion of frames where an atomic overlap was observed with any of the wild-type glycans, as well as the average number of atoms with an atomic overlap with wild-type glycans (Table 4.4). Due to differences between the results, each protomer is described separately.

Table 4.4: Overlap calculations between the N301A mutant glycan substitute and any wild-type glycan.

N301A mutant glycan substitute	Wild-type glycans that clash with the N301A substitute	**Protomers** A	B	C
	N133		53%[a]; 45[b]	
	N156[c]	**86%; 58**	**68%; 46**	**<1%; 2**
N197	N160		33%; 30	
	N301	**21%; 37**	**10%; 25**	**12%; 27**
	N386	55%; 25	<1%; 6	94%; 115
	N301	77%; 22	3%; 7	97%; 58
N262	N334		<1%; 1	16%; 6
	N442	90%; 42	12%; 13	67%; 26
	N446	6%; 9	99%; 175	85%; 23
	N137	9%; 5	17%; 67	
N386	N197	11%; 13	1%; 14	6%; 7
	N463	13%; 33	22%; 20	8%; 43
	N133			34%; 14
	N137		13%; 49	59%; 43
N442	N156	21%; 33	7%; 43	
	N262	52%; 9	2%; 13	1%; 6
	N301	99%; 226	29%; 28	19%; 14
	N446		31%; 45	36%; 19
	N137	15%; 13		1%; 12
	N262	>99%; 80	88%; 86	24%; 9
N446	N301	22%; 18	<1%; 16	66%; 36
	N334		24%; 56	3%; 49
	N442	95%; 148	6%; 16	96 %; 125

[a]The proportion of frames with a overlap (%).

[b]The average number of overlapping atoms in those frames with a overlap. Wild-type glycans are omitted if the proportion of frames with an overlap is below 10% on all protomers.

[c]Wild-type glycans on a different protomer from that of the N301A mutant glycan are shown in bold.

Protomer A: The glycan with the largest and most frequent overlap with glycan N301 (protomer A; wild-type) was glycan N442 (N301A mutant). In 99% of the frames, glycan N442 occupies the same space as glycan N301 does in the wild-type simulation with an average of 226 overlapping atoms per frame (Table 4.4). A Man-9 glycan (described in the Methods section) has a total of 244 atoms; 226 clashes therefore imply that almost all of the atoms overlap with glycan N301 and that glycan N442 (N301A mutant) occupies most of the space vacated by glycan N301 for practically the entire duration of the trajectory. Similarly, glycan N262 (protomer C; N301A mutant) occupies some (22 clashes per frame) of the space vacated by glycan N301 for a large (77%) proportion of the trajectory (Table 4.4).

Concurrently, the space vacated by glycans N442 and N262, is in turn occupied by glycan N446 (N301A mutant). This is evident from the large proportion of frames with an atomic overlap, as well as a large average number of overlapping atoms per frame (Table 4.4), observed between glycan N446 and glycans N442 (95%, 148 overlapping atoms) and N262 (>99%, 80 overlapping atoms).

Protomer B: None of the glycans on protomer B have noticeably large overlap with glycan N301 on the wild-type protomer B, which suggests that the wild-type glycan conformations are maintained on the N301A mutant protomer B.

Protomer C: The largest overlap with glycan N301 were observed for glycan N262 (N301A mutant), where 97% of the frames have an atomic overlap with glycan N301, with an average of 58 overlapping atoms per frame, and for glycan N446 (N301A mutant; 66%, 36 overlapping atoms; Table 4.4).

Lastly, the nearest glycan results highlighted the conformational change of glycan N197 (protomer C; N301A mutant). This shifted conformation appeared to occupy the space where glycan N386 of the wild-type is located, which is confirmed here by the large overlap observed between glycan N197 (N301A mutant) and glycan N386 (wild-type). Only a small overlap is observed for glycan N386 (protomer C; N301A mutant) with glycan N301, which suggests that the adjusted conformation for this glycan is viable in the presence of glycan N301. The conformation of glycan N197 (protomer C; N301A mutant), however, has moderate overlap with glycan N301 (12%, 27

overlapping atoms). Further investigation revealed that glycan N197 overlaps with glycan N301 at two specific time intervals (14-35 ns and 346-500 ns) and not throughout the simulations. This suggests the 'new' conformation for glycan N197 can coexist with glycan N301.

4.4.3 Domino effect of glycan conformational changes: how changes propagate and taper off

During the investigation of the protomer specific conformational changes, we observed that the space generated by removing glycan N301 is occupied by another glycan, which in turn leaves a new space that is then occupied by yet another glycan. To extend the investigation of glycan conformational changes beyond those glycans nearest to the N301A mutation site, while maintaining the relevance of the changes (i.e. the changes associated with the loss of the N301 glycan), we compare the frequency of the hydrogen bonds that form between glycans in the wild-type and N301A mutant. Substantial changes in the hydrogen bond network allude to changes in the conformations of glycans adjacent to those glycans near the N301A mutation site. For each of the identified glycans, we calculated the average centre of mass position, throughout the simulation, for the wild-type (Figure 4.6 (a), orange arrows) and N301A mutant (Figure 4.6 (a), blue arrows).

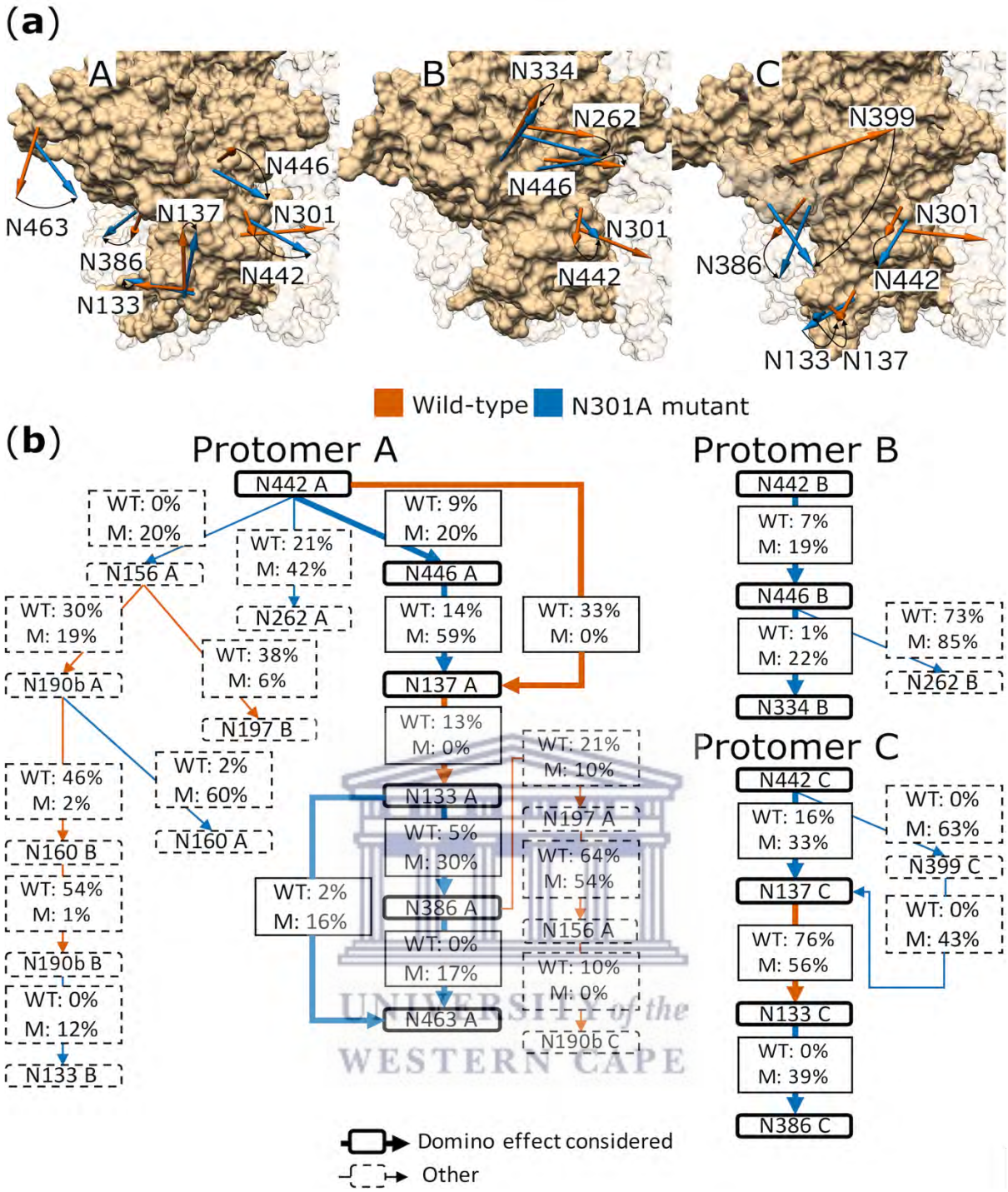

Figure 4.6: Domino effect of glycan conformational changes on the three protomers

(A) Structures illustrating the location of glycans. The arrows originate from the Cα atom of the Asn and end at the average centre of mass position for the selected wild-type (orange) and N301A mutant (blue) glycans identified by the hydrogen bond analysis for protomers A, B and C. Curved arrows indicate the degree of change in directionality, e.g. the large movement for glycan N399 on protomer C. The representations are cropped around the V3 region of each protomer and the surface representation of residues on adjacent protomers are shown as opaque. (B) Each schematic – protomer A, B and C - shows the hydrogen bond network starting from glycan N442 (first node) on protomer A, B and C, respectively. Solid borders around glycan nodes and broad arrows represent glycans and interactions that form part of the domino effect considered, whereas dotted borders and fine arrows indicate any glycans and interactions outside the considered domino effect. The arrows connecting glycan nodes distinguish interactions that are either greater on the wild-type (WT; orange) or on the N301A mutant (M; blue).

Protomer A: The conformational change domino effect is clear on protomer A, where the conformations of six glycans (N442, N446, N137, N133, N386 and N463) are affected. In the wild-type, these glycans are interlinked by sequential hydrogen bonds. However, the interlinked nature of these glycans is broken in the absence of glycan N301, and two glycan clusters exist on the N301A mutant. The first cluster includes glycans N442, N446 and N137, and the second glycans N133, N386 and N463.

Glycan N442 occupies the space vacated by glycan N301 (Figure 4.6 (a) protomer A). In turn, the space vacated by glycan N442 is occupied by glycans N446 and N137 (Figure 4.6 (a) protomer A). These rearrangements result in an increased interaction between glycans N442 and N446 (9% wild-type vs. 20% N301A mutant; Figure 4.6 (b) Protomer A). Glycan N137 on the other hand, no longer interacts with glycan N442 (33% wild-type vs. 0% N301A mutant; Figure 4.6 (b) Protomer A) but instead has an increased interaction with glycan N446 (14% wild-type vs. 59% N301A mutant; Figure 4.6 (b) Protomer A). The second cluster is separated from the first by the change in conformation of glycan N137 (Figure 4.6 (a) protomer A), which eliminates the interaction between glycans N137 and N133 (13% wild-type vs. 0% N301A mutant; Figure 4.6 (b) Protomer A). This reduced structural hindrance near glycan N133 may allow for its re-orientation (Figure 4.6 (a) protomer A) and subsequent increased interaction with glycan N386 (5% wild-type vs. 30% N301A mutant; Figure 4.6 (b) Protomer A), as well as with glycans N463 (2% wild-type vs. 16% N301A mutant; Figure 4.6 (b) Protomer A). There is also an increased interaction between glycan N386 and N463 (0% wild-type vs. 17% N301A mutant; Figure 4.6 (b) Protomer A). Other changes in the hydrogen bond network quickly taper off to glycans located at the protein apex (Figure 4.6 (b) Protomer A).

Protomer B: Apart from glycans N156, N197, N262, N386, N442 and N446, which were previously investigated on protomer B during the neighbourhood and subsequent overlap analyses, the hydrogen bond network further identifies interaction with glycan N334 (N301A mutant; Figure 4.6 (b) Protomer B). However, glycan N334 shows no movement when compared to its wild-type counterpart (Figure 4.6 (a) protomer B), which indicates that there was almost no domino effect.

Protomer C: The domino effect was similar to that of protomer A, however, glycans N446 and N463 (N301A mutant) are not implicated, whereas glycan N399 is. The hydrogen bonds between glycans N137 and N133 persist, although at a reduced scale compared to the wild-type (76% wild-type vs. 56% N301A mutant; Figure 4.6 (b) Protomer C). The hydrogen bonds between glycans N133 and N386 were not present in the wild-type and are substantial in the N301A mutant (0% wild-type vs. 39% N301A mutant; Figure 4.6 (b) Protomer C). The persistent interaction between glycans N133 and N137 results in all these interlinked glycans collectively moving closer to the gap left by the N301A mutation on the N301A mutant model (Figure 4.6 (a) protomer C). In our model, this movement, coupled with the increase in interaction between glycans N133 and N386, likely contributes to the 'new' conformation of glycan N197 (protomer C; N301A mutant).

4.5 Discussion

Here, we analysed two 500 ns molecular dynamics simulations (CAP45.G3 wild-type and CAP45.G3 N301A mutant) and show that the systems imitate *in vitro* neutralisation data – the glycan shield restores itself and retains its ability to protect key epitopes after the removal of glycan N301[53]. This was in contrast to a second isolate, Du156.12, where the laboratory results showed that this N301A mutant virus had increased sensitivity to a panel of sera from chronically infected individuals[53], and where the *in silico* simulations of the Du156.12 wild-type and N301A mutant models were vastly different to that of the CAP45.G3 models (Supplementary Figure S4.1). The key observation during our initial, *in silico*, comparative analysis was that the conformational differences of the glycans on the wild-type models, as well as the landscapes around each of these glycans, likely affected the ultimate changes in the glycan shields. The importance of this collective behaviour of glycans on the Env surface was previously noted by Lemmin *et al.*[86], describing several, relatively stable, glycan microdomains[86]. Since Moyo *et al.* [53] speculated that the CAP45.G3 virus typified a subset of viruses where the loss of glycan N301 was tolerated, i.e. the protective qualities of the glycan shield, or perhaps the glycan microdomain, were retained, the focus of our manuscript was on providing a thorough account for the suggested compensation. Therefore, for the CAP45.G3 modelled systems, we

describe, in detail, the change in the glycan landscape, and cascade of events, that may have contributed towards the maintenance of the glycan shield for this viral isolate. The only structural difference between the model simulations is the presence or absence of the N-linked glycan at position 301 and the stability of the distinct glycan interactions throughout each simulation suggests that the observed differences are meaningful.

We observed clear evidence of compensation for two of the three N301A mutant protomers (A and C) within the trimeric protein structures. The antibody accessible surface area (AASA) ratios of these two protomers are equivalent to their wild-type counterparts, despite the lack of glycan N301. However, unexpectedly, there are differences in the particular glycans compensating for the loss. Glycan N442 bears the bulk of the compensation burden on protomer A, whereas glycans N446 and N262 share the burden on protomer C. However, even though glycan N442 bears the compensation burden on protomer A, its 'new' conformation was considerably influenced by glycan N446 on the same protomer. The spatial pressure from glycan N446 resulted in a glycan N301-like conformation observed for glycan N442 (protomer A; N301A mutant). Therefore, this suggests that in our model, glycan N446 is essential for maintaining the glycan shield by not only directly bearing a part of the compensation burden, but also through its substantial influence on the conformation of glycan N442. Although we are currently unaware of studies describing the influence of glycan N446 on epitopes accessibility, there have been reports on glycan N442 and its role in shielding epitopes of the CD4 and gp41 regions[55]. This suggests that glycan N446 may also influence the availability of epitopes in these regions.

In contrast to protomers A and C, we found ambiguous evidence of compensation for protomer B (N301A mutant), where increased AASA ratios were observed for various residues. The apparent lack of compensation for protomer B can be, in part, attributed to the lack of conformational change in glycan N442, as well as the substantial change in the conformation of glycan N197 on protomer C (adjacent to protomer B), which is commonly reported as shielding the CD4 and V3 regions[54,55].

We also identified specific glycan conformational changes that have a direct impact on the accessibility of the modelled protein surface. The conformational change of

glycan N197 (protomer C; N301A mutant) has drastic implications for the AASA ratios of the residues of the CD4 sub-region and VRC01 epitope on protomer C. Both the residues in the CD4 sub-region (Table 4.1) and the residues of the VRC01 epitope show decreased AASA ratios for protomer C of the N301A mutant model. Glycan N386 (protomer A; N301 mutant) also plays a role in maintaining the AASA ratio of residues that form part of the CD4 sub-region (Table 4.1); however, in comparison, the conformational change of glycan N386 does not lead to a decrease in the AASA ratio of the VRC01 epitope, but instead, the removal of glycan N301 leads to a slight increase in AASA for this region (protomer A). Thus, based on the AASA results, glycan N386 may not contribute to the increased resistance observed to the VRC01 bNAb (unlike glycan N197), which is contrary to evidence of this glycan shielding the CD4 binding site[70,183].

Despite the popularity of calculating the AASA[86,87,89,90] as a predictor of whether residues are accessible by antibodies, it remains to be seen whether they are useful, and accurate, for describing the contact between HIV-1 gp160 and antibodies. For example, there are considerable differences between the conformation of particular HIV-1 envelope glycans co-crystalised with VRC01 compared to an unliganded crystal structure[88]. Additionally, Stewart-Jones *et al.*[88] showed that substantial overlap occurred between broadly neutralising antibodies and one or more glycan/s throughout 500 ns trajectories, suggesting that known broadly neutralising antibodies likely need to accommodate at least one glycan during the binding process. Given this evidence, the passive nature of AASA calculations is likely to result in an underestimation of the accessibility of residues in regions where the “shielding” glycans can be flexible. This may be the case for glycan N386 (protomer A; N301A mutant), where the flexibility of glycan N133 (protomer A; N301A mutant) increased the structural hindrance near glycan N386. This, in turn, reduced the freedom of movement of glycan N386 resulting in slightly higher AASA ratios around this glycan in our model, but potentially still leading to increased VRC01 resistance due to its reduced flexibility to accommodate the antibody. However, complete resistance to VRC01 was not observed for the CAP45.G3 N301A mutant[53] and it is unknown if glycan N386 plays a role, or whether the conformation of glycan N197 is solely responsible for, the observed increase in VRC01 resistance.

Due to the potential caveats associated with AASA, we extended the analyses to examine the capacity for each glycan to act as a shield to its surrounding protein residues by determining which glycan was the nearest to each protein residue across time. A benefit of this analysis is the ability to generate leads in the cases where the binding site of a particular antibody is known and point mutations that would reduce resistance are sought. The analysis also provides an intuitive and easy representation of the glycan positions over time. For example, the large change in the conformation of glycan N197 (protomer C; N301A mutant) is apparent and easily related to a change in the conformation of glycan N386 (protomer C; N301A mutant; Figure 4.4 (b)).

The nearest glycan approach also presents certain disadvantages; small differences are easily overlooked and, since the crowded nature of the HIV glycan shield can result in large conformational differences that change the glycan neighbourhood only slightly, small changes could be an important consideration. Additionally, the nearest glycan may not necessarily be the shielding glycan, which is the case when glycans are buried and therefore relatively inflexible. Glycan N262, where a large neighbourhood was observed, presents one such case. Visual inspection suggested that other glycans, only slightly further away, may have greater capacity to shield certain residues assigned to glycan N262. However, due to its static, buried, nature, these residues were consistently assigned to glycan N262, and therefore removing it prior to the analysis might be prudent.

The length of the trajectory (MD simulation), which relates to the uncertainty associated with observing enough of the sample space during the simulation is a key, and ongoing, problem of molecular dynamics research of large structures. We would expect that, on a long enough time scale, the glycans would populate similar volumes across the protomers. However, the glycans of each protomer adopt distinct conformations that do not interconvert during our simulation. Unlike the protomer scissoring reported by Lemmin *et al.*[86], we did not observe this effect for either the wild-type or N301 mutant model simulations, but did identify small scale, protomer specific, protein movements (data not shown). Nonetheless, despite the differences, the data for each protomer is self-consistent and stable; i.e. although the effect of glycosylation on the AASA is different across protomers, the effect is constant in each protomer throughout the simulation. A possible solution to this issue is to use replicate

exchange (which should explore greater proportions of the potential energy surface) but requires more study. Yang *et al*.[89] used replicate exchange and described enhanced sampling for glycosylated Env trimers; it would be worthwhile for future studies to compare the degree of convergence in the observed glycan shapes between different protomers in the different approaches. The advantage of the molecular dynamic protocol used here is that it clearly shows that there are glycan-glycan and glycan-protein interaction networks that extend across the Env surface and persist on at least a 500 ns timescale. Although the lack of convergence between protomers is concerning for drawing statistically strong conclusions, it is interesting, and important, to note that the interactions driving the observed differences occur at longer time scales than 500 ns.

Finally, our modelling results suggest that the loss of a glycan, due to a point mutation, can result in a cascade of events on the same protomer (intra-protomer) that could contribute towards increased resistance to epitopes distal to the location of the initial sequence mutation causing the cascade of events – in this case the VRC01 epitope. We initially speculated that this epitope would be influenced directly by the N301A mutation via cross-protomer interactions (through the additional space created), since it has been shown that the N301 glycan overlaps with VRC01 on an Env crystal structure bound to this antibody[88]. However, here we present evidence for intra-protomer conformational rearrangements of specific glycans, which we believe contributed to the increased resistance of the CAP45.G3 N301A mutant to VRC01, and other VRC01-like, antibodies.

It is unlikely that the glycan shield of the CAP45.G3 isolate is unique in its ability to compensate for a loss of a glycan, however, the particular glycan distribution and clustering meant that the absence of glycan N301 was not crucial for the maintenance of its protective qualities as a whole. The further implication is that different viral Env glycoproteins will likely each have their own set of glycans that are, individually, either dispensable or indispensable in forming, and maintaining, the glycan shield. This argument extends to the asymmetrical effects seen across protomers; the degree of glycan conformity, both in terms of site occupancy and glycan type, between the three protomers of an Env trimer is currently unknown and largely unacknowledged in HIV-1 Env studies. It is completely plausible that the HIV-1 glycans of Env trimers vary

across their protomers and that these differences affect the neutralisation efficiency of glycan-dependent antibodies in different ways. This ties in with the knowledge that antibodies are not always present on all three protomers, and that glycan heterogeneity is one likely cause of this finding[115,184].

Despite the caveats associated with molecular dynamics simulations, as well as the immense potential variation in the Env glycan shields of HIV-1 isolates, this study provides a detailed investigation of how the loss of a single HIV-1 Env glycan does not result in a hole, but rather results in a cascade of events that may have led to the maintenance of the glycan shield and increased resistance to a broadly neutralising antibody observed for the viral isolate. Given the focus on Env glycans within HIV-1 vaccine research, and the importance of these glycans for bNAb binding, we hope the techniques and results presented here will encourage further in-depth consideration of the virus-specific glycan landscapes. Future investigations, both *in vitro* and *in silico*, focussing on different glycan point mutations and including systems composed of a variety of glycan forms, will demonstrate to what extent these results translate, and are predictive, across viral isolates and subtypes.

Chapter 5 Conclusion and Future Work

This manuscript focussed on examining the HIV-1 Env glycan shield using molecular dynamics simulations. These molecular dynamics simulations were carried out on the computationally determined, fully glycosylated, Env structures of two wild-type HIV-1 subtype C isolates and their paired N301A mutant variants, i.e. where glycan N301 was removed. These particular strains were used since *in vitro* data showed that the absence of glycan N301 had a substantially different effect on the neutralisation sensitivity of the two strains, despite the large (~90%) identity between the sequences. Therefore, the project focussed on using computational techniques to describe the differences between the glycan shields of these two isolates and to identify the particular glycans that contributed toward maintaining the glycan shield of the resistant strain (to both frequently elicited antibodies and a bNAb), despite the missing glycan. However, the literature on the analysis of molecular dynamics simulations in this context is limited and, consequently, an additional goal for this study emerged very quickly, which was to determine what analyses were not only feasible, but also provided useful features for comparison.

Initially, in order to understand the potential impact of removing glycan N301, static state representations of the fully glycosylated Env trimer structures were compared (Chapter 2). These results suggested that the removal of glycan N301 likely exposed a greater surface area on one model compared to the other and in that way affected the neutralisation sensitivity of each strain uniquely, which was in agreement with the *in vitro* study. However, since molecular molecules are dynamic and change conformation over time, the static state representations offered only limited information for carrying out comparative studies.

Therefore, the next step was to carry out and compare molecular dynamics simulations of the modelled structures (Chapter 3). The results of these comparisons revealed glycan position-specific differences. For example, the N442 and N197 glycans were more likely to vary between models than the N156 glycans, irrespective of which models were compared. Nevertheless, the comparisons also demonstrated that not only can equivalent glycans adopt distinct conformations specific to each wild-

type model, but that model-specific conformations emerged when glycan N301 was removed from the glycan shield. Moreover, the N301A mutant model-specific glycan conformations were directly associated with the protein residues ultimately exposed. The difference in the locations of these exposed regions on the N301A mutant models may explain the varied resistant to neutralising antibodies observed, *in vitro*, after glycan N301 was removed. Furthermore, the results also demonstrated that each glycan shield is conformationally unique, regardless of large sequence identity (e.g. between the wild-type models) or where the glycan shields differ by only a single glycan (e.g. between a wild-type model and its N301A mutant counterpart).

However, due to the random variation implicit in macromolecular structures, a degree of uncertainty remained regarding whether the conformational heterogeneity observed for glycans located farther from the N301A mutation were a direct result of the N301A mutation. Additionally, these result did not provide substantial insight on how one of the strains remained resistant to the frequently elicited neutralising antibodies and had increased resistance to a CD4-binding site bNAb despite the loss of glycan N301. Therefore, due to this remaining uncertainty, a detailed comparison between the wild-type simulation and its N301A mutant counterpart was carried out on the resistant strain (Chapter 4). The results of the detailed computational analysis showed that the glycan shield of this strain retained its ability to shield the protein residues even after the removal of glycan N301. Moreover, the change in the glycan landscape, and cascade of events, that contributed towards the maintenance of this glycan shield and the increased resistance to a bNAb epitope located at the edge of the cascade, were discernible.

This thesis therefore provides a detailed analysis of the HIV-1 Env glycans and their capacity to protect the underlying protein and, in some cases, shield the virus against frequently elicited neutralising antibodies. During the study, several key considerations, and important opportunities for further explorations, for vaccine and treatment studies focussing on HIV-1 Env were highlighted. These include:

- PNGS profiles can vary substantially despite large sequence identity

- The conformations of equivalent glycans (glycans that share an HXB2 reference sequence position) can be distinct between strains and, therefore, this should be considered during all Env glycan shield comparisons
- The conformation of a glycan is dependent on the conformations of the surrounding glycans
- Protomers may present alternate equilibrium states
- Glycans move and adjust to accommodate changes in the surrounding landscape
- Glycan adjustments, after a glycan is removed, could lead to the maintenance of the glycan shield

These considerations came to light throughout this project where, broadly speaking, the analyses always focussed on one of three glycan features, i.e. the glycan's surroundings, its conformation, or its function (shield and/or target). These glycan features are intrinsically related since a change in one likely impacts the others. For example, removing a glycan changes the surroundings of the neighbouring glycans, which themselves could change conformations as a result, and this could affect the region they shield. Therefore, a single change in the glycan shield can affect the region each glycan is capable of shielding, i.e. its fundamental function within the glycan shield. However, it is important to note that an altered feature does not necessarily imply that the glycan shield is impaired. In fact, glycans appear to be capable of altering their features over time, and these changes may even be distinct across protomers, as demonstrated in this thesis where unique glycan features were observed for the different protomers of the same model. Thus, it is important that glycan features are not considered in isolation at a single time point, but rather as an integrated network that changes over time.

However, this type of analysis is complex, since determining a fully glycosylated HIV-1 Env trimer structure and carrying out molecular dynamics simulations are non-trivial tasks. The glycosylated structures can either be obtained from deposited X-ray crystal structures or computationally determined, as was done here. The considerable differences between these types of structures (Figure 4.2) demonstrate the advantages of computationally determined structures, which allow the composition of the system to be defined in much more detail. Together, the established methods that

predict protein structures from template crystal structures, the higher resolution of available HIV-1 Env crystal structures, and the development of tools that facilitate computational glycosylation of these models, will inevitably streamline the creation of computationally-determined glycosylated models and alleviate a substantial portion of the burden associated with performing HIV-1 Env molecular dynamics simulations.

It is important to reiterate that, although this study has offered one of the current most systematic analysis of the molecular dynamics simulations of HIV-1 Env and its glycans, it was conducted on only 500 ns simulations of glycosylated Env trimers that lacked the MPER regions and where the glycan shield comprised of only Man9GlcNAc2 glycans. In the context of a virion, however, the Env trimer is membrane-bound with variations in occupancy and glycoforms at N-linked glycan sites, as well as glycoform micro-heterogeneity at the same N-linked glycan site (i.e. a wide spectrum of similar, but not identical, glycoforms have been identified at the same site). These features give rise to additional questions regarding the differences between glycan characteristics of Env glycoproteins. Therefore, future work will include investigating glycan features over longer time scales, the effect of various glycoforms at the same or different positions, and models that include both MPER and a membrane in order to determine the impact of membrane proximity on the surrounding glycans.

In conclusion, the current study has demonstrated that the landscape of the glycan shield contains immense diversity, and that glycans that share an HXB2 reference sequence position will not necessarily conform in shape, dynamics and function between different strains. These conformational differences ultimately determine whether or not the glycan shield can compensate for the loss of a glycan. This is a key, and too often ignored, detail, since it directly relates to the main function of the glycans in their ability to shield and protect the virus from immune surveillance. The results from the analysis of the simulations in this thesis further highlight the remarkable potential of molecular dynamics simulations, which can be used as a powerful predictive technique to facilitate, and direct, laboratory studies focussing on vaccine research.

Supplementary Materials

Chapter 3

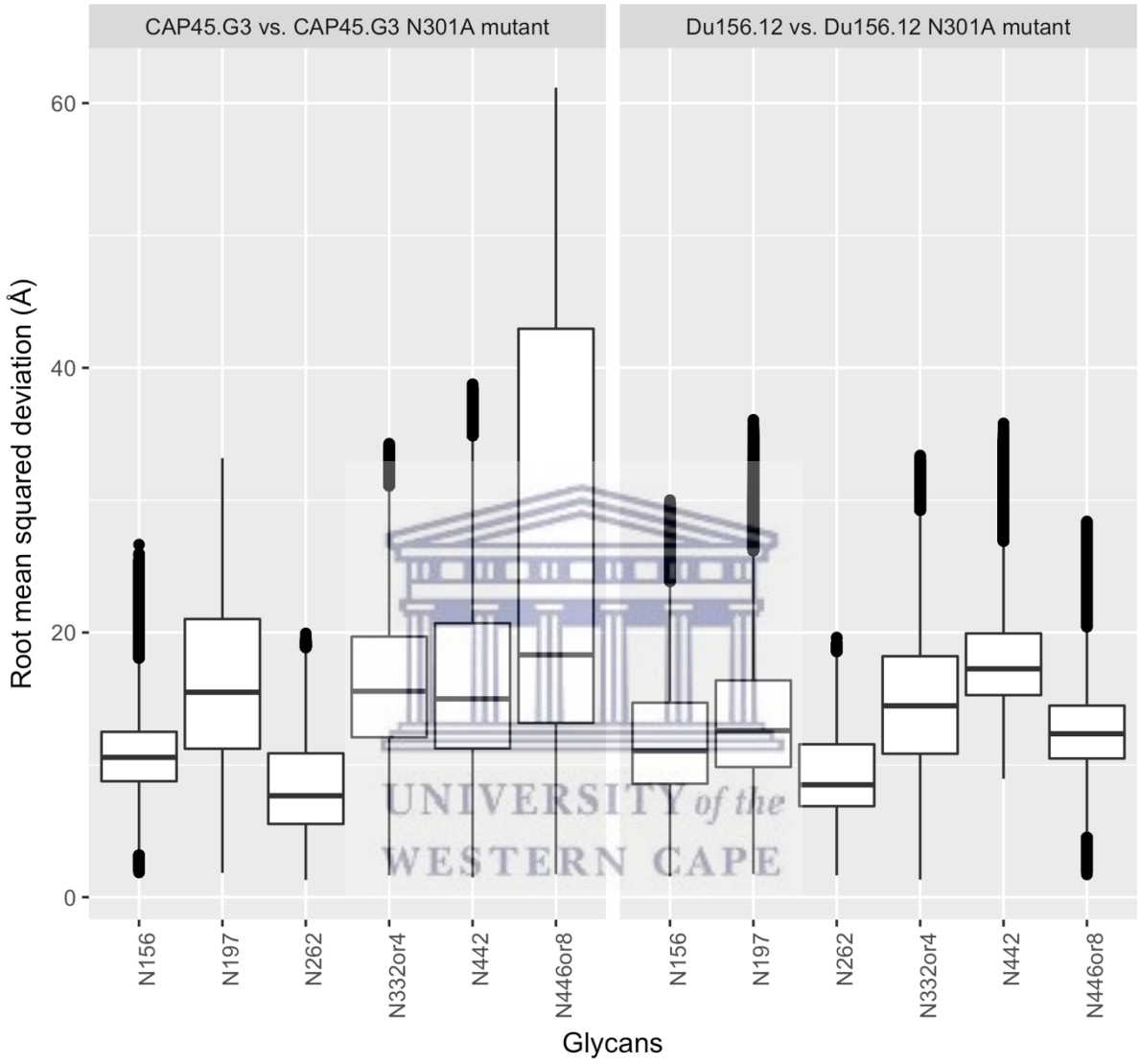

Figure S3.1: Conformational heterogeneity between wild-type glycan siblings and the N301A mutant siblings.

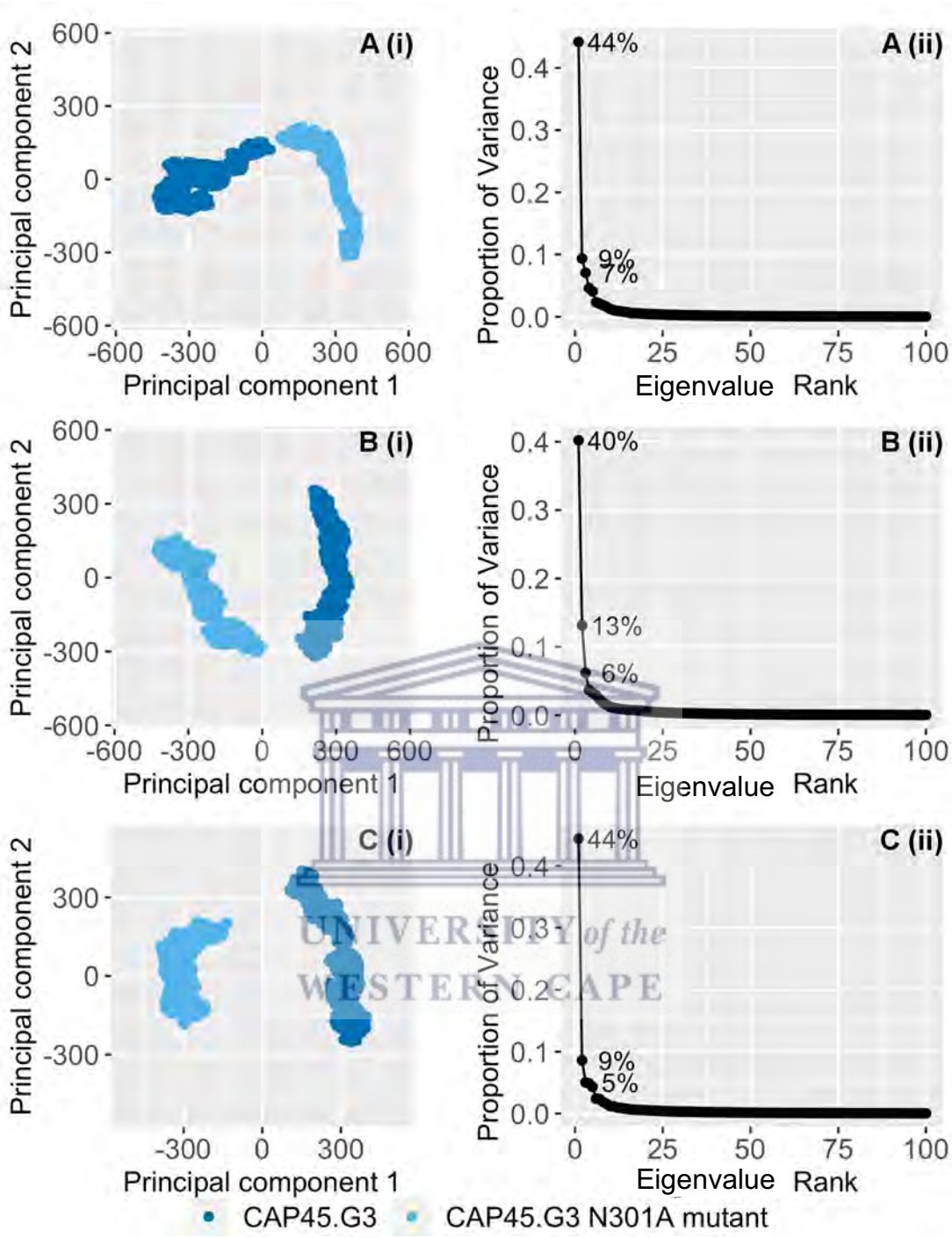

Figure S3.2: Standard PCA plots for the CAP45.G3 glycans

The standard PCA plots for the CAP45.G3 paired-model comparison performed on the glycans of protomer A (A), B (B) and C (C). Demonstrating the projection of the frames onto the (i) first and second principal components as well as (ii) the proportion of the variance explained by the first three principal components.

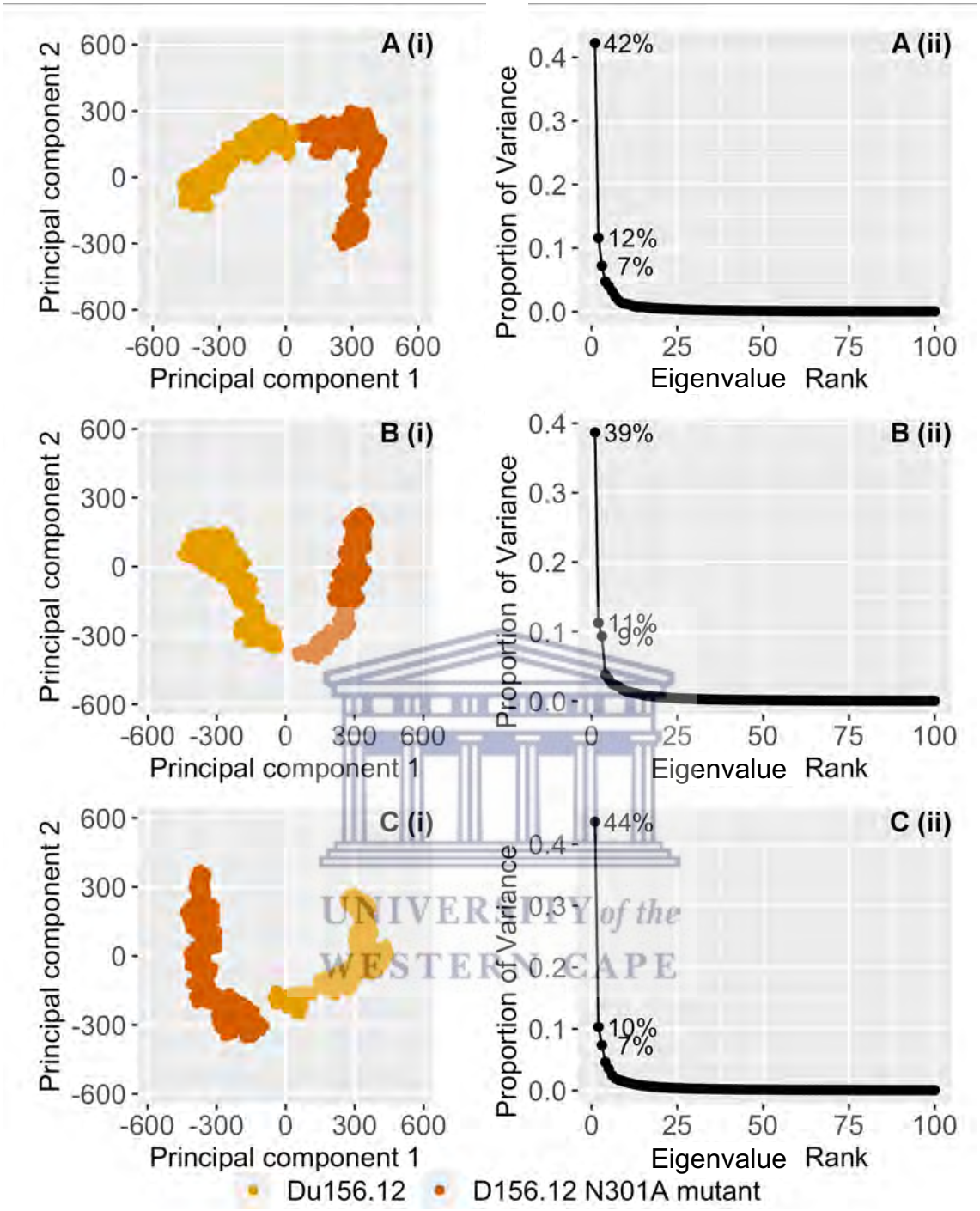

Figure S3.3: Standard PCA plots for the Du156.12 glycans

The standard PCA plots for the Du156.12 paired-model comparison performed on the glycans of protomer A (A), B (B) and C (C). Demonstrating the projection of the frames onto the (i) first and second principal components as well as (ii) the proportion of the variance explained by the first three principal components.

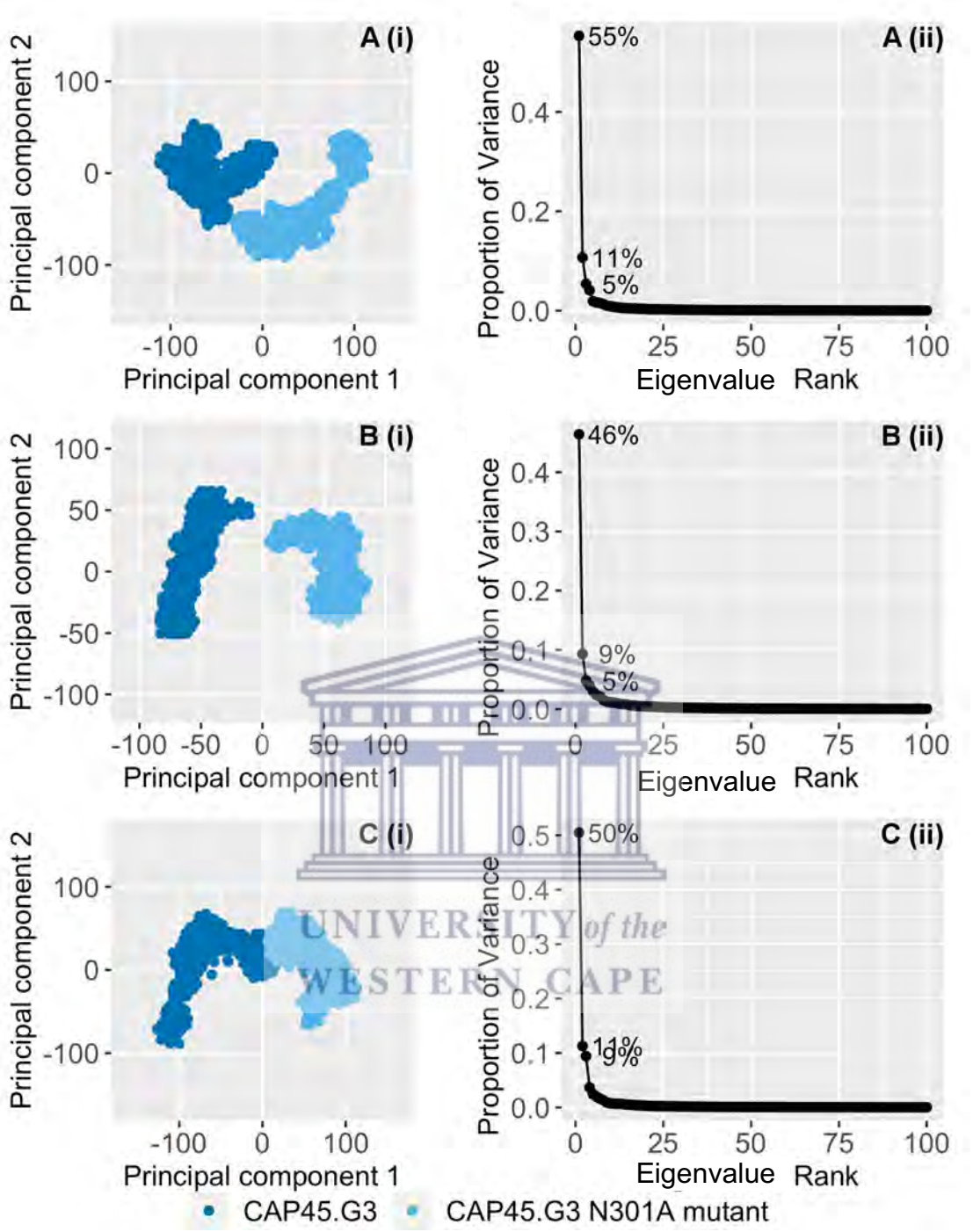

Figure S3.4: Standard PCA plots for the CAP45. G3 protein residues

The standard PCA plots for the CAP45.G3 paired-model comparison performed on the glycans of protomer A (A), B (B) and C (C). Demonstrating the projection of the frames onto the (i) first and second principal components as well as (ii) the proportion of the variance explained by the first three principal components.

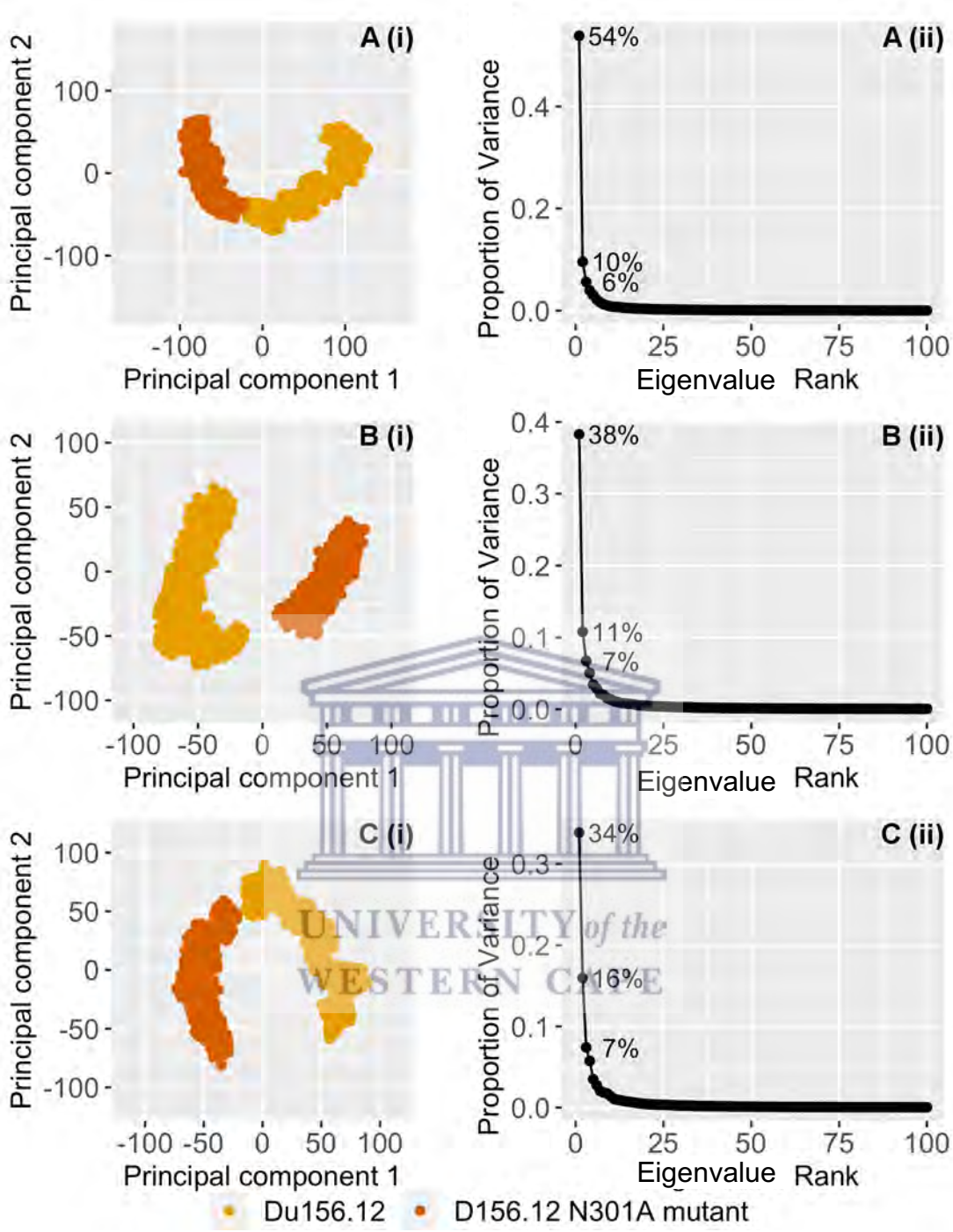

Figure S3.5: Standard PCA plots for the Du156.12 protein residues

The standard PCA plots for the Du156.12 paired-model comparison performed on the glycans of protomer A (A), B (B) and C (C). Demonstrating the projection of the frames onto the (i) first and second principal components as well as (ii) the proportion of the variance explained by the first three principal components.

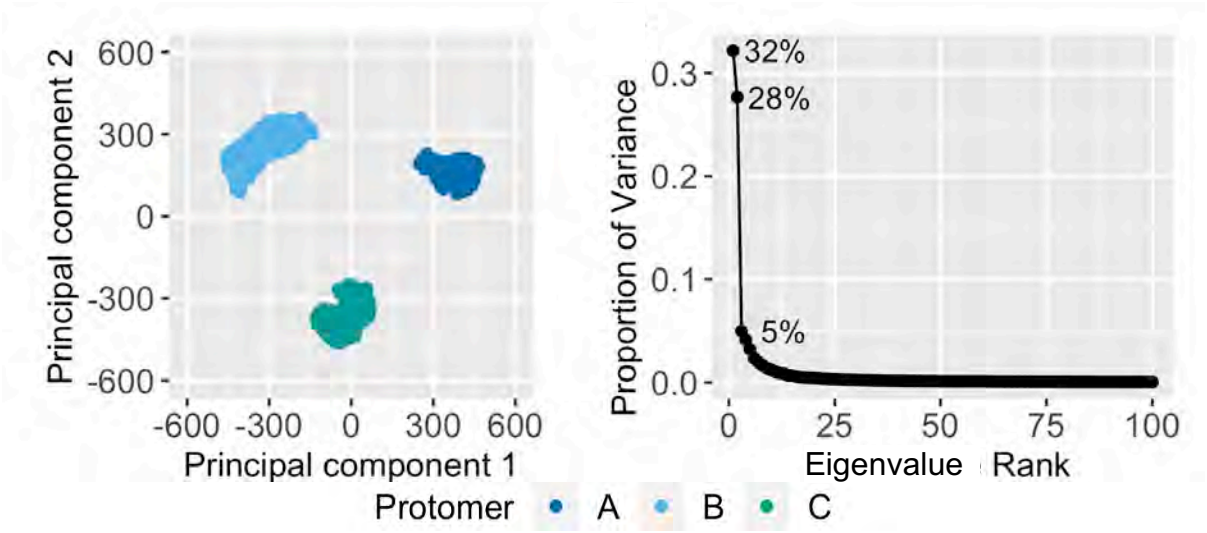

Figure S3.6: Standard PCA plots for the CAP45.G3 protomer glycans

The standard PCA plots for the CAP45.G3 protomer comparison performed on the glycans. Demonstrating the projection of the frames onto the first and second principal components as well as the proportion of the variance explained by the first three principal components.

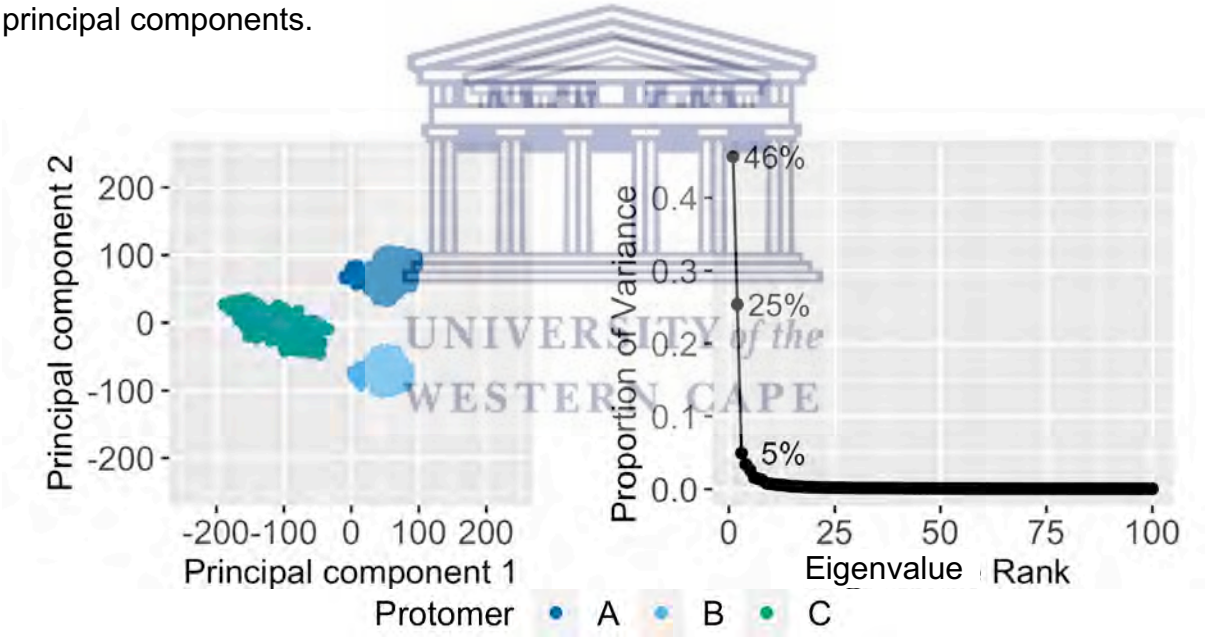

Figure S3.7: Standard PCA plots for the CAP45.G3 protomer protein residues

The standard PCA plots for the CAP45.G3 protomer comparison performed on the protein residues. Demonstrating the projection of the frames onto the first and second principal components as well as the proportion of the variance explained by the first three principal components.

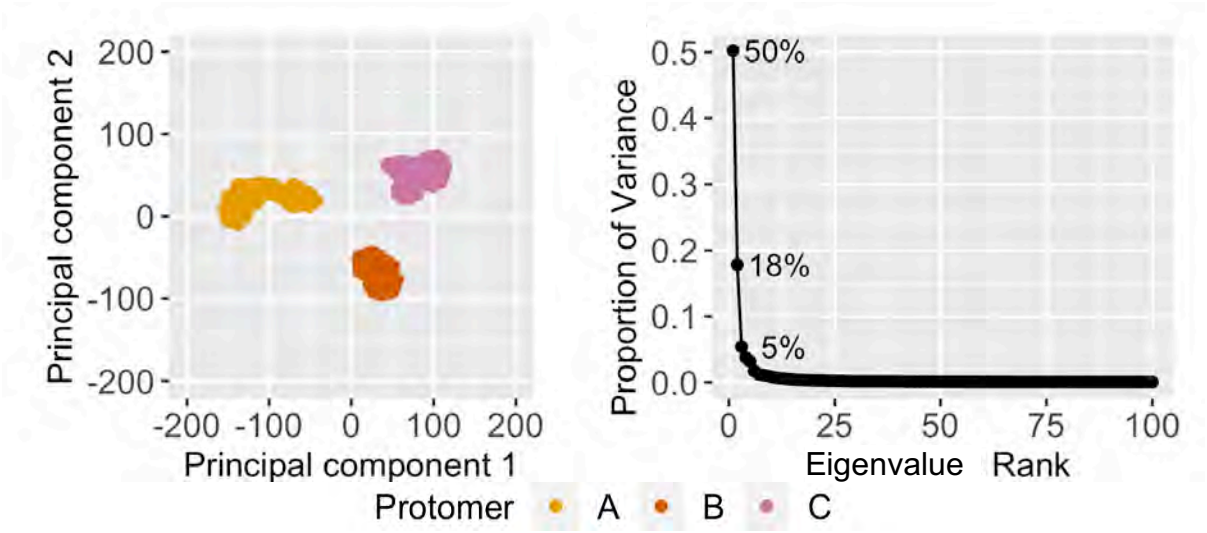

Figure S3.8: Standard PCA plots for the Du156.12 protomer protein residues

The standard PCA plots for the Du156.12 protomer comparison performed on the protein residues. Demonstrating the projection of the frames onto the first and second principal components as well as the proportion of the variance explained by the first three principal components.

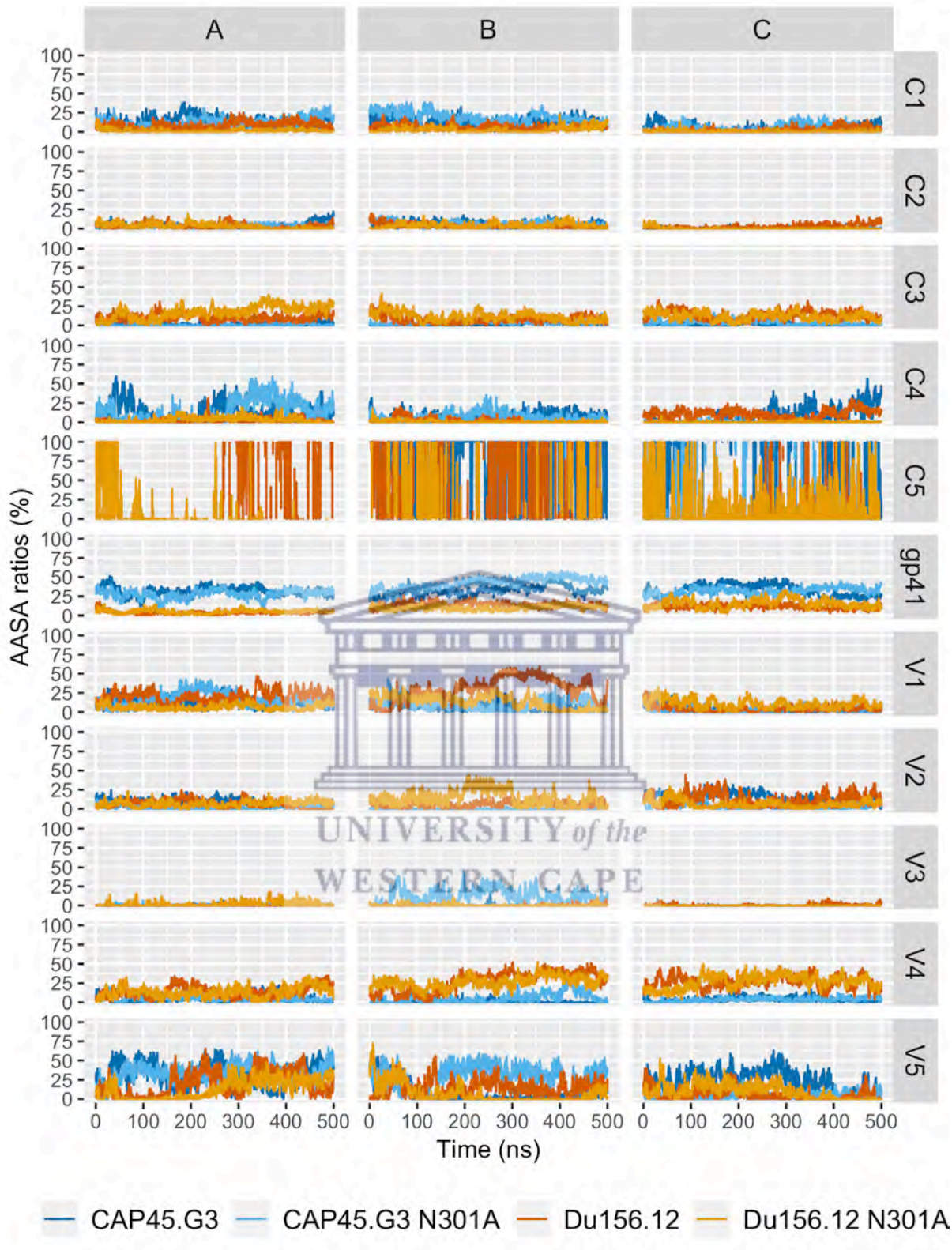

Figure S3.9: AASA ratios for the Env regions

AASA ratios (%), using a 10Å probe radius, of the Env region (C1-C5, gp41, and V1-V5) for the three protomers A, B and C. The ratios are shown for each wild-type and N301A mutant simulations.

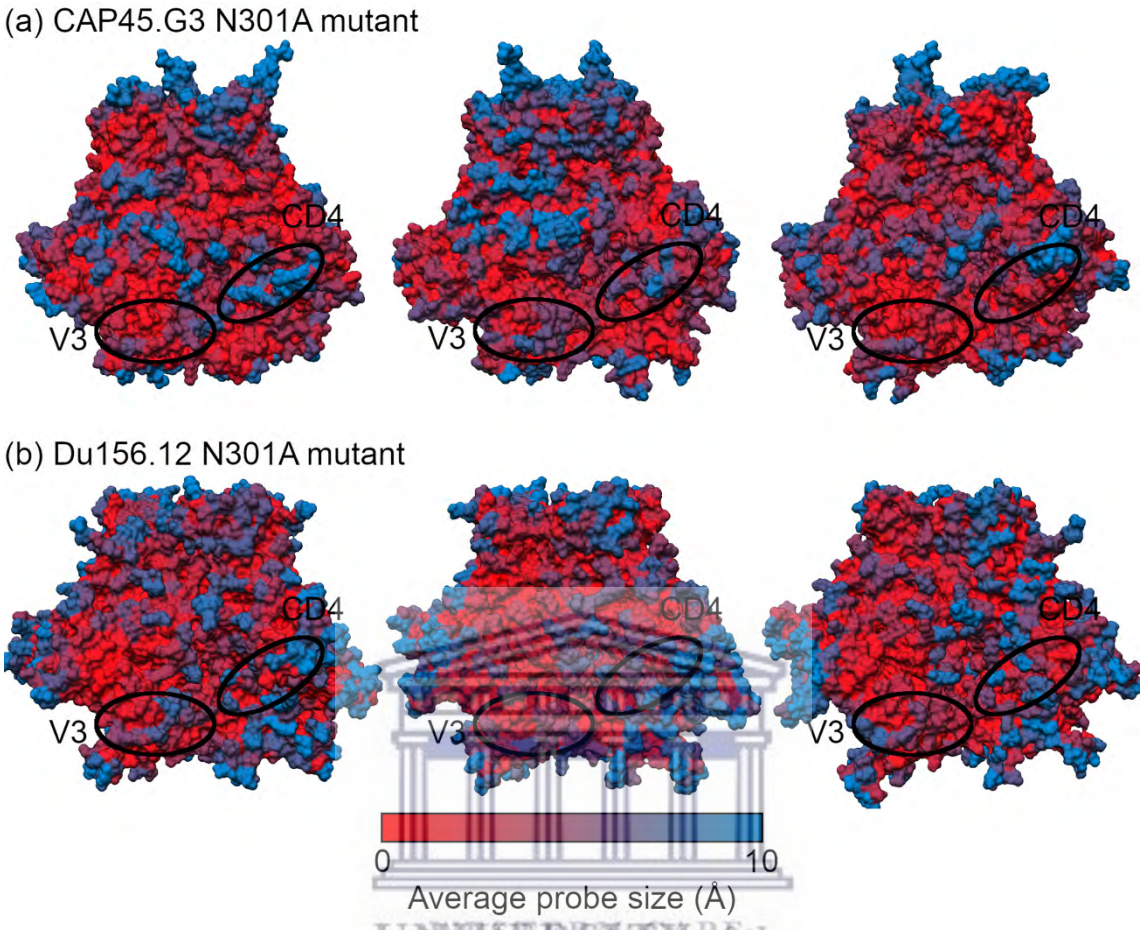

Figure S4.1: Residue accessibility for the CAP45.G3 and Du156.12 N301A mutant models.

The surface area accessibility for each residue on the (a) CAP45.G3 and (b) Du156.12 N301A mutant models is shown on a scale where bright red represents least accessible and blue indicates high accessibility. The V3-loop and CD4 binding site regions are circled and labelled.

References

1. AIDS info. People living with HIV and new HIV infections. (2017). Available at: http://aidsinfo.unaids.org/. (Accessed: 7th August 2018)
2. Burton, D. R. *et al.* A Blueprint for HIV Vaccine Discovery. *Cell Host Microbe* **12**, 396–407 (2012).
3. Burton, D. R. *et al.* HIV vaccine design and the neutralizing antibody problem. *Nat. Immunol.* **5**, 233–236 (2004).
4. Kim, J. H., Rerks-Ngarm, S., Excler, J.-L. & Michael, N. L. HIV vaccines: lessons learned and the way forward. *Curr. Opin. HIV AIDS* **5**, 428–434 (2010).
5. Kwong, P. D., Mascola, J. R. & Nabel, G. J. The changing face of HIV vaccine research. *J. Int. AIDS Soc.* **15**, 17407 (2012).
6. Kwong, P. D., Mascola, J. R. & Nabel, G. J. Rational design of vaccines to elicit broadly neutralizing antibodies to HIV-1. *Cold Spring Harb. Perspect. Med.* **1**, a007278 (2011).
7. Mascola, J. R. & Haynes, B. F. HIV-1 neutralizing antibodies: understanding nature's pathways. *Immunol. Rev.* **254**, 225–244 (2013).
8. Stamatatos, L., Morris, L., Burton, D. R. & Mascola, J. R. Neutralizing antibodies generated during natural HIV-1 infection: good news for an HIV-1 vaccine? *Nat. Med.* **15**, 866–870 (2009).
9. Vaccari, M., Poonam, P. & Franchini, G. Phase III HIV vaccine trial in Thailand: a step toward a protective vaccine for HIV. *Expert Rev. Vaccines* **9**, 997–1005 (2010).
10. Zolla-Pazner, S. Identifying epitopes of HIV-1 that induce protective antibodies. *Nat. Rev. Immunol.* **4**, 199–210 (2004).

11. Gao, F. *et al.* Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. *Nature* **397**, 436–441 (1999).

12. Gao, F. *et al.* Human infection by genetically diverse SIVSM-related HIV-2 in west Africa. *Nature* **358**, 495–499 (1992).

13. Hirsch, V. M., Olmsted, R. A., Murphey-Corb, M., Purcell, R. H. & Johnson, P. R. An African primate lentivirus (SIVsm) closely related to HIV-2. *Nature* **339**, 389–392 (1989).

14. Lemey, P. *et al.* Tracing the origin and history of the HIV-2 epidemic. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 6588–6592 (2003).

15. Murphey-Corb, M. *et al.* Isolation of an HTLV-III-related retrovirus from macaques with simian AIDS and its possible origin in asymptomatic mangabeys. *Nature* **321**, 435–437 (1986).

16. Santiago, M. L. *et al.* Simian immunodeficiency virus infection in free-ranging sooty mangabeys (Cercocebus atys atys) from the Taï Forest, Côte d'Ivoire: implications for the origin of epidemic human immunodeficiency virus type 2. *J. Virol.* **79**, 12515–12527 (2005).

17. Van Heuverswyn, F. *et al.* Human immunodeficiency viruses: SIV infection in wild gorillas. *Nature* **444**, 164 (2006).

18. Marlink, R. *et al.* Reduced rate of disease development after HIV-2 infection as compared to HIV-1. *Science* **265**, 1587–1590 (1994).

19. Hemelaar, J. The origin and diversity of the HIV-1 pandemic. *Trends Mol. Med.* **18**, 182–192 (2012).

20. Hemelaar, J., Gouws, E., Ghys, P. D., Osmanov, S. & WHO-UNAIDS Network for HIV Isolation and Characterisation. Global trends in molecular epidemiology of HIV-1 during 2000-2007. *AIDS Lond. Engl.* **25**, 679–689 (2011).

21. Taylor, B. S., Sobieszczyk, M. E., McCutchan, F. E. & Hammer, S. M. The challenge of HIV-1 subtype diversity. *N. Engl. J. Med.* **358**, 1590–1602 (2008).

22. Shaw, G. M. & Hunter, E. HIV transmission. *Cold Spring Harb. Perspect. Med.* **2**, (2012).

23. de Jong, M. A. W. P., de Witte, L., Taylor, M. E. & Geijtenbeek, T. B. H. Herpes simplex virus type 2 enhances HIV-1 susceptibility by affecting Langerhans cell function. *J. Immunol. Baltim. Md 1950* **185**, 1633–1641 (2010).

24. Gringhuis, S. I. *et al.* HIV-1 exploits innate signaling by TLR8 and DC-SIGN for productive infection of dendritic cells. *Nat. Immunol.* **11**, 419–426 (2010).

25. Laurence, J. Reservoirs of HIV infection or carriage: monocytic, dendritic, follicular dendritic, and B cells. *Ann. N. Y. Acad. Sci.* **693**, 52–64 (1993).

26. Simon, V. & Ho, D. D. HIV-1 dynamics in vivo: implications for therapy. *Nat. Rev. Microbiol.* **1**, 181–190 (2003).

27. Freed, E. O. HIV-1 replication. *Somat. Cell Mol. Genet.* **26**, 13–33 (2001).

28. Hallenberger, S. *et al.* Inhibition of furin-mediated cleavage activation of HIV-1 glycoprotein gp160. *Nature* **360**, 358–361 (1992).

29. Lifson, J. D. *et al.* Induction of CD4-dependent cell fusion by the HTLV-III/LAV envelope glycoprotein. *Nature* **323**, 725–728 (1986).

30. Deng, H. *et al.* Identification of a major co-receptor for primary isolates of HIV-1. *Nature* **381**, 661–666 (1996).

31. Arhel, N. Revisiting HIV-1 uncoating. *Retrovirology* **7**, 96 (2010).

32. Davies, J. F., Hostomska, Z., Hostomsky, Z., Jordan, S. R. & Matthews, D. A. Crystal structure of the ribonuclease H domain of HIV-1 reverse transcriptase. *Science* **252**, 88–95 (1991).

33. Engelman, A. & Cherepanov, P. The structural biology of HIV-1: mechanistic and therapeutic insights. *Nat. Rev. Microbiol.* **10**, 279–290 (2012).

34. Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M. & Ho, D. D. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science* **271**, 1582–1586 (1996).

35. Bebenek, K., Abbotts, J., Wilson, S. H. & Kunkel, T. A. Error-prone polymerization by HIV-1 reverse transcriptase. Contribution of template-primer misalignment, miscoding, and termination probability to mutational hot spots. *J. Biol. Chem.* **268**, 10324–10334 (1993).

36. Fairbanks, A., Fairbanks, D. J. & Andersen, W. R. *Genetics: The Continuity of Life*. (CENGAGE Learning, 2001).

37. Rambaut, A., Posada, D., Crandall, K. A. & Holmes, E. C. The causes and consequences of HIV evolution. *Nat. Rev. Genet.* **5**, 52–61 (2004).

38. Gao, F. *et al.* Unselected mutations in the human immunodeficiency virus type 1 genome are mostly nonsynonymous and often deleterious. *J. Virol.* **78**, 2426–2433 (2004).

39. Binley, J. M. *et al.* Comprehensive cross-clade neutralization analysis of a panel of anti-human immunodeficiency virus type 1 monoclonal antibodies. *J. Virol.* **78**, 13232–13252 (2004).

40. Gao, F. *et al.* Cooperation of B cell lineages in induction of HIV-1-broadly neutralizing antibodies. *Cell* **158**, 481–491 (2014).

41. Hraber, P. *et al.* Prevalence of broadly neutralizing antibody responses during chronic HIV-1 infection. *AIDS Lond. Engl.* **28**, 163–169 (2014).

42. Moore, P. L. *et al.* Evolution of an HIV glycan-dependent broadly neutralizing antibody epitope through immune escape. *Nat. Med.* **18**, 1688–1692 (2012).

43. Tomaras, G. D. *et al.* Polyclonal B cell responses to conserved neutralization epitopes in a subset of HIV-1-infected individuals. *J. Virol.* **85**, 11502–11519 (2011).

44. Lynch, R. M. *et al.* HIV-1 fitness cost associated with escape from the VRC01 class of CD4 binding site neutralizing antibodies. *J. Virol.* **89**, 4201–4213 (2015).

45. Kornfeld, R. & Kornfeld, S. Assembly of asparagine-linked oligosaccharides. *Annu. Rev. Biochem.* **54**, 631–664 (1985).

46. Johnson, W. E., Sauvron, J. M. & Desrosiers, R. C. Conserved, N-linked carbohydrates of human immunodeficiency virus type 1 gp41 are largely dispensable for viral replication. *J. Virol.* **75**, 11426–11436 (2001).

47. Travers, S. A. Conservation, compensation, and evolution of N-linked glycans in the HIV-1 group M subtypes and circulating recombinant forms. *ISRN AIDS* **2012**, (2012).

48. Ware, F. E. *et al.* The molecular chaperone calnexin binds Glc1Man9GlcNAc2 oligosaccharide as an initial step in recognizing unfolded glycoproteins. *J. Biol. Chem.* **270**, 4697–4704 (1995).

49. Behrens, A.-J. & Crispin, M. Structural principles controlling HIV envelope glycosylation. *Curr. Opin. Struct. Biol.* **44**, 125–133 (2017).

50. Rudd, P. M. & Dwek, R. A. Glycosylation: heterogeneity and the 3D structure of proteins. *Crit. Rev. Biochem. Mol. Biol.* **32**, 1–100 (1997).

51. Pantophlet *et al. HIV Glycans in Infection and Immunity*. (Springer-Verlag, 2014).

52. Harvey, D. J., Royle, L., Radcliffe, C. M., Rudd, P. M. & Dwek, R. A. Structural and quantitative analysis of N-linked glycans by matrix-assisted laser desorption ionization and negative ion nanospray mass spectrometry. *Anal. Biochem.* **376**, 44–60 (2008).

53. Moyo, T. *et al.* Chinks in the armor of the HIV-1 Envelope glycan shield: Implications for immune escape from anti-glycan broadly neutralizing antibodies. *Virology* **501**, 12–24 (2017).

54. Townsley, S., Li, Y., Kozyrev, Y., Cleveland, B. & Hu, S.-L. Conserved Role of an N-Linked Glycan on the Surface Antigen of Human Immunodeficiency Virus Type 1 Modulating Virus Sensitivity to Broadly Neutralizing Antibodies against the Receptor and Coreceptor Binding Sites. *J. Virol.* **90**, 829–841 (2015).

55. Wang, W. *et al.* A systematic study of the N-glycosylation sites of HIV-1 envelope protein on infectivity and antibody-mediated neutralization. *Retrovirology* **10**, 14 (2013).

56. Wei, X. *et al.* Antibody neutralization and escape by HIV-1. *Nature* **422**, 307–312 (2003).

57. Wyatt, R. *et al.* The antigenic structure of the HIV gp120 envelope glycoprotein. *Nature* **393**, 705–711 (1998).

58. Walker, L. M. *et al.* Broad and potent neutralizing antibodies from an African donor reveal a new HIV-1 vaccine target. *Science* **326**, 285–289 (2009).

59. Wu, X. *et al.* Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1. *Science* **329**, 856–861 (2010).

60. Muster, T. *et al.* A conserved neutralizing epitope on gp41 of human immunodeficiency virus type 1. *J. Virol.* **67**, 6642–6647 (1993).

61. Buchacher, A. *et al.* Generation of human monoclonal antibodies against HIV-1 proteins; electrofusion and Epstein-Barr virus transformation for peripheral blood lymphocyte immortalization. *AIDS Res. Hum. Retroviruses* **10**, 359–369 (1994).

62. Walker, L. M. *et al.* Broad neutralization coverage of HIV by multiple highly potent antibodies. *Nature* **477**, 466–470 (2011).

63. Falkowska, E. *et al.* Broadly neutralizing HIV antibodies define a glycan-dependent epitope on the prefusion conformation of gp41 on cleaved envelope trimers. *Immunity* **40**, 657–668 (2014).

64. Sather, D. N. *et al.* Factors associated with the development of cross-reactive neutralizing antibodies during human immunodeficiency virus type 1 infection. *J. Virol.* **83**, 757–769 (2009).

65. Wu, X. *et al.* Selection pressure on HIV-1 envelope by broadly neutralizing antibodies to the conserved CD4-binding site. *J. Virol.* **86**, 5844–5856 (2012).

66. Koch, M. *et al.* Structure-based, targeted deglycosylation of HIV-1 gp120 and effects on neutralization sensitivity and antibody recognition. *Virology* **313**, 387–400 (2003).

67. Li, Y. *et al.* Removal of a single N-linked glycan in human immunodeficiency virus type 1 gp120 results in an enhanced ability to induce neutralizing antibody responses. *J. Virol.* **82**, 638–651 (2008).

68. Malenbaum, S. E. *et al.* The N-terminal V3 loop glycan modulates the interaction of clade A and B human immunodeficiency virus type 1 envelopes with CD4 and chemokine receptors. *J. Virol.* **74**, 11008–11016 (2000).

69. McCaffrey, R. A., Saunders, C., Hensel, M. & Stamatatos, L. N-linked glycosylation of the V3 loop and the immunologically silent face of gp120 protects human immunodeficiency virus type 1 SF162 from neutralization by anti-gp120 and anti-gp41 antibodies. *J. Virol.* **78**, 3279–3295 (2004).

70. Sanders, R. W. *et al.* The carbohydrate at asparagine 386 on HIV-1 gp120 is not essential for protein folding and function but is involved in immune evasion. *Retrovirology* **5**, 10 (2008).

71. Zolla-Pazner, S. *et al.* Structure/Function Studies Involving the V3 Region of the HIV-1 Envelope Delineate Multiple Factors That Affect Neutralization Sensitivity. *J. Virol.* **90**, 636–649 (2015).

72. Cutalo, J. M., Deterding, L. J. & Tomer, K. B. Characterization of Glycopeptides From HIV-ISF2 gp120 by Liquid Chromatography Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **15**, 1545–1555 (2004).

73. Go, E. P. *et al.* Comparative Analysis of the Glycosylation Profiles of Membrane-Anchored HIV-1 Envelope Glycoprotein Trimers and Soluble gp140. *J. Virol.* **89**, 8245–8257 (2015).

74. Go, E. P., Hua, D. & Desaire, H. Glycosylation and disulfide bond analysis of transiently and stably expressed clade C HIV-1 gp140 trimers in 293T cells identifies disulfide heterogeneity present in both proteins and differences in O-linked glycosylation. *J. Proteome Res.* **13**, 4012–4027 (2014).

75. Go, E. P. *et al.* Characterization of host-cell line specific glycosylation profiles of early transmitted/founder HIV-1 gp120 envelope proteins. *J. Proteome Res.* **12**, 1223–1234 (2013).

76. Go, E. P. *et al.* Characterization of glycosylation profiles of HIV-1 transmitted/founder envelopes by mass spectrometry. *J. Virol.* **85**, 8270–8284 (2011).

77. Go, E. P. *et al.* Glycosylation site-specific analysis of clade C HIV-1 envelope proteins. *J. Proteome Res.* **8**, 4231–4242 (2009).

78. Go, E. P. *et al.* Glycosylation site-specific analysis of HIV envelope proteins (JR-FL and CON-S) reveals major differences in glycosylation site occupancy, glycoform profiles, and antigenic epitopes' accessibility. *J. Proteome Res.* **7**, 1660–1674 (2008).

79. Irungu, J. *et al.* Comparison of HPLC/ESI-FTICR MS versus MALDI-TOF/TOF MS for glycopeptide analysis of a highly glycosylated HIV envelope glycoprotein. *J. Am. Soc. Mass Spectrom.* **19**, 1209–1220 (2008).

80. Pabst, M., Chang, M., Stadlmann, J. & Altmann, F. Glycan profiles of the 27 N-glycosylation sites of the HIV envelope protein CN54gp140. *Biol. Chem.* **393**, 719–730 (2012).

81. Yang, W. *et al.* Glycoform analysis of recombinant and human immunodeficiency virus envelope protein gp120 via higher energy collisional dissociation and spectral-aligning strategy. *Anal. Chem.* **86**, 6959–6967 (2014).

82. Zhu, X., Borchers, C., Bienstock, R. J. & Tomer, K. B. Mass spectrometric characterization of the glycosylation pattern of HIV-gp120 expressed in CHO cells. *Biochemistry* **39**, 11194–11204 (2000).

83. Calarese, D. A. *et al.* Antibody domain exchange is an immunological solution to carbohydrate cluster recognition. *Science* **300**, 2065–2071 (2003).

84. Kwong, P. D. *et al.* Structures of HIV-1 gp120 envelope glycoproteins from laboratory-adapted and primary isolates. *Struct. Lond. Engl. 1993* **8**, 1329–1339 (2000).

85. Kwong, P. D. *et al.* Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature* **393**, 648–659 (1998).

86. Lemmin, T., Soto, C., Stuckey, J. & Kwong, P. D. Microsecond Dynamics and Network Analysis of the HIV-1 SOSIP Env Trimer Reveal Collective Behavior and Conserved Microdomains of the Glycan Shield. *Struct. Lond. Engl. 1993* **25**, 1631-1639.e2 (2017).

87. Pancera, M. *et al.* Structure and immune recognition of trimeric pre-fusion HIV-1 Env. *Nature* **514**, 455–461 (2014).

88. Stewart-Jones, G. *et al.* Trimeric HIV-1-Env structures define glycan shields from clades A, B, and G. *Cell* **165**, 813–826 (2016).

89. Yang, M., Huang, J., Simon, R., Wang, L.-X. & MacKerell, A. D. Conformational Heterogeneity of the HIV Envelope Glycan Shield. *Sci. Rep.* **7**, 4435 (2017).

90. Zhou, T. *et al.* Quantification of the Impact of the HIV-1-Glycan Shield on Antibody Elicitation. *Cell Rep.* **19**, 719–732 (2017).

91. Julien, J.-P. *et al.* Crystal structure of a soluble cleaved HIV-1 envelope trimer. *Science* **342**, 1477–1483 (2013).

92. Gristick, H. B. *et al.* Natively glycosylated HIV-1 Env structure reveals new mode for antibody recognition of the CD4-binding site. *Nat. Struct. Mol. Biol.* **23**, 906–915 (2016).

93. Hossler, P., Khattak, S. F. & Li, Z. J. Optimal and consistent protein glycosylation in mammalian cell culture. *Glycobiology* **19**, 936–949 (2009).

94. Varki, A. Biological roles of glycans. *Glycobiology* **27**, 3–49 (2017).

95. Pritchard, L. K. *et al.* Glycan Microheterogeneity at the PGT135 Antibody Recognition Site on HIV-1 gp120 Reveals a Molecular Mechanism for Neutralization Resistance. *J. Virol.* **89**, 6952–6959 (2015).

96. Crooks, E. T., Tong, T., Osawa, K. & Binley, J. M. Enzyme Digests Eliminate Nonfunctional Env from HIV-1 Particle Surfaces, Leaving Native Env Trimers Intact and Viral Infectivity Unaffected▽. *J. Virol.* **85**, 5825–5839 (2011).

97. Moore, P. L. *et al.* Nature of Nonfunctional Envelope Proteins on the Surface of Human Immunodeficiency Virus Type 1. *J. Virol.* **80**, 2515–2528 (2006).

98. Schneider, J., Kaaden, O., Copeland, T. D., Oroszlan, S. & Hunsmann, G. Shedding and interspecies type sero-reactivity of the envelope glycopolypeptide gp120 of the human immunodeficiency virus. *J. Gen. Virol.* **67 (Pt 11)**, 2533–2538 (1986).

99. Petrescu, A.-J., Milac, A.-L., Petrescu, S. M., Dwek, R. A. & Wormald, M. R. Statistical analysis of the protein environment of N-glycosylation sites: implications for occupancy, structure, and folding. *Glycobiology* **14**, 103–114 (2004).

100. Lütteke, T. *et al.* GLYCOSCIENCES.de: an Internet portal to support glycomics and glycobiology research. *Glycobiology* **16**, 71R-81R (2006).

101. Lütteke, T. Analysis and validation of carbohydrate three-dimensional structures. *Acta Crystallogr. D Biol. Crystallogr.* **65**, 156–168 (2009).

102. Anthony, C. *et al.* Cooperation between Strain-Specific and Broadly Neutralizing Responses Limited Viral Escape and Prolonged the Exposure of the Broadly Neutralizing Epitope. *J. Virol.* **91**, (2017).

103. Apellániz, B. *et al.* Cholesterol-dependent membrane fusion induced by the gp41 membrane-proximal external region-transmembrane domain connection suggests a mechanism for broad HIV-1 neutralization. *J. Virol.* **88**, 13367–13377 (2014).

104. Gabriel, J. L. & Mitchell, W. M. Proposed atomic structure of a truncated human immunodeficiency virus glycoprotein gp120 derived by molecular modeling: target CD4 recognition and docking mechanism. *Proc. Natl. Acad. Sci. U. S. A.* **90**, 4186–4190 (1993).

105. Gossert, S. T., Parajuli, B., Chaiken, I. & Abrams, C. F. Roles of conserved tryptophans in trimerization of HIV-1 membrane-proximal external regions: Implications for virucidal design via alchemical free-energy molecular simulations. *Proteins* **86**, 707–711 (2018).

106. Hikichi, Y. *et al.* Increased HIV-1 sensitivity to neutralizing antibodies by mutations in the Env V3-coding region for resistance to CXCR4 antagonists. *J. Gen. Virol.* **97**, 2427–2440 (2016).

107. Hollingsworth, L. R., Lemkul, J. A., Bevan, D. R. & Brown, A. M. HIV-1 Env gp41 Transmembrane Domain Dynamics Are Modulated by Lipid, Water, and Ion Interactions. *Biophys. J.* **115**, 84–94 (2018).

108. Kong, R. *et al.* Fusion peptide of HIV-1 as a site of vulnerability to neutralizing antibody. *Science* **352**, 828–833 (2016).

109. Oakes, V. *et al.* Exposure of the HIV-1 broadly neutralizing antibody 10E8 MPER epitope on the membrane surface by gp41 transmembrane domain scaffolds. *Biochim. Biophys. Acta* **1860**, 1259–1271 (2018).

110. Sethi, A., Tian, J., Derdeyn, C. A., Korber, B. & Gnanakaran, S. A mechanistic understanding of allosteric immune escape pathways in the HIV-1 envelope glycoprotein. *PLoS Comput. Biol.* **9**, e1003046 (2013).

111. Wood, N. T. *et al.* The influence of N-linked glycans on the molecular dynamics of the HIV-1 gp120 V3 loop. *PloS One* **8**, e80301 (2013).

112. Yokoyama, M. *et al.* In silico Analysis of HIV-1 Env-gp120 Reveals Structural Bases for Viral Adaptation in Growth-Restrictive Cells. *Front. Microbiol.* **7**, 110 (2016).

113. Zhang, Y. *et al.* Understanding the molecular mechanism of the broad and potent neutralization of HIV-1 by antibody VRC01 from the perspective of molecular dynamics simulation and binding free energy calculations. *J. Mol. Model.* **18**, 4517–4527 (2012).

114. Collins, P. M., Collins, P. C., Ferrier, R. J. & Ferrier, R. *Monosaccharides: Their Chemistry and Their Roles in Natural Products*. (Wiley, 1995).

115. Lyumkis, D. *et al.* Cryo-EM structure of a fully glycosylated soluble cleaved HIV-1 envelope trimer. *Science* **342**, 1484–1490 (2013).

116. Lee, J. H., Ozorowski, G. & Ward, A. B. Cryo-EM structure of a native, fully glycosylated, cleaved HIV-1 envelope trimer. *Science* **351**, 1043–1048 (2016).

117. Wang, H. *et al.* Cryo-EM structure of a CD4-bound open HIV-1 envelope trimer reveals structural rearrangements of the gp120 V1V2 loop. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E7151–E7158 (2016).

118. Abraham, M. J., van der Spoel, D., Lindahl, E., Hess, B. & and the GROMACS development team. GROMACS user manual. (2018). Available at: www.gromacs.org.

119. D.A. Case *et al.* AMBER reference manual. (University of California, 2018).

120. Woods, R. J. & Tessier, M. B. Computational glycoscience: characterizing the spatial and temporal properties of glycans and glycan-protein complexes. *Curr. Opin. Struct. Biol.* **20**, 575–583 (2010).

121. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).

122. Kirschner, K. N. *et al.* GLYCAM06: a generalizable biomolecular force field. Carbohydrates. *J. Comput. Chem.* **29**, 622–655 (2008).

123. Case, D. *et al. AMBER 2014*. (University of California, 2014).

124. Mahoney, M. W. & Jorgensen, W. L. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J. Chem. Phys.* **112**, 8910–8922 (2000).

125. Sauter, J. & Grafmüller, A. Solution properties of hemicellulose polysaccharides with four common carbohydrate force fields. *J. Chem. Theory Comput.* **11**, 1765–1774 (2015).

126. Landais, E. *et al.* Broadly Neutralizing Antibody Responses in a Large Longitudinal Sub-Saharan HIV Primary Infection Cohort. *PLoS Pathog.* **12**, e1005369 (2016).

127. Mouquet, H. *et al.* Complex-type N-glycan recognition by potent broadly neutralizing HIV antibodies. *Proc. Natl. Acad. Sci. U. S. A.* **109**, E3268-3277 (2012).

128. Walker, L. M. *et al.* A limited number of antibody specificities mediate broad and potent serum neutralization in selected HIV-1 infected individuals. *PLoS Pathog.* **6**, e1001028 (2010).

129. Webb, N. E., Montefiori, D. C. & Lee, B. Dose-response curve slope helps predict therapeutic potency and breadth of HIV broadly neutralizing antibodies. *Nat. Commun.* **6**, 8443 (2015).

130. Mahalanabis, M. *et al.* Continuous viral escape and selection by autologous neutralizing antibodies in drug-naive human immunodeficiency virus controllers. *J. Virol.* **83**, 662–672 (2009).

131. Malherbe, D. C. *et al.* Sequential immunization with a subtype B HIV-1 envelope quasispecies partially mimics the in vivo development of neutralizing antibodies. *J. Virol.* **85**, 5262–5274 (2011).

132. Murphy, M. K. *et al.* Viral Escape from Neutralizing Antibodies in Early Subtype A HIV-1 Infection Drives an Increase in Autologous Neutralization Breadth. *PLoS Pathog.* **9**, (2013).

133. Pissani, F. *et al.* Motif-optimized subtype A HIV envelope-based DNA vaccines rapidly elicit neutralizing antibodies when delivered sequentially. *Vaccine* **30**, 5519–5526 (2012).

134. Moore, P. L. *et al.* Multiple pathways of escape from HIV broadly cross-neutralizing V2-dependent antibodies. *J. Virol.* **87**, 4882–4894 (2013).

135. Wibmer, C. K. *et al.* Viral escape from HIV-1 neutralizing antibodies drives increased plasma neutralization breadth through sequential recognition of multiple epitopes and immunotypes. *PLoS Pathog.* **9**, e1003738 (2013).

136. Webb, B. & Sali, A. Comparative protein structure modeling using MODELLER. *Curr. Protoc. Bioinforma.* **47**, 5.6.1-32 (2014).

137. Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).

138. Huang, C.-C. *et al.* Structures of the CCR5 N terminus and of a tyrosine-sulfated antibody with HIV-1 gp120 and CD4. *Science* **317**, 1930–1934 (2007).

139. Huang, C. *et al.* Structure of a V3-containing HIV-1 gp120 core. *Science* **310**, 1025–1028 (2005).

140. Shen, M.-Y. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci. Publ. Protein Soc.* **15**, 2507–2524 (2006).

141. Eisenberg, D., Lüthy, R. & Bowie, J. U. VERIFY3D: assessment of protein models with three-dimensional profiles. *Methods Enzymol.* **277**, 396–404 (1997).

142. Sippl, M. J. Recognition of errors in three-dimensional structures of proteins. *Proteins* **17**, 355–362 (1993).

143. Wiederstein, M. & Sippl, M. J. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res.* **35**, W407-410 (2007).

144. Topham, C. M., Srinivasan, N., Thorpe, C. J., Overington, J. P. & Kalsheker, N. A. Comparative modelling of major house dust mite allergen Der p I: structure validation using an extended environmental amino acid propensity table. *Protein Eng.* **7**, 869–894 (1994).

145. GLYCAM. Available at: http://glycam.org/. (Accessed: 10th January 2018)

146. Hubbard, S. & Thornton, J. *NACCESS*. (University College, London, Department of Biochemistry and Moelcular Biology, 1993).

147. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2016).

148. Binley, J. M. *et al.* Role of complex carbohydrates in human immunodeficiency virus type 1 infection and resistance to antibody neutralization. *J. Virol.* **84**, 5637–5655 (2010).

149. Rizzuto, C. & Sodroski, J. Fine definition of a conserved CCR5-binding region on the human immunodeficiency virus type 1 glycoprotein 120. *AIDS Res. Hum. Retroviruses* **16**, 741–749 (2000).

150. Rizzuto, C. D. *et al.* A conserved HIV gp120 glycoprotein structure involved in chemokine receptor binding. *Science* **280**, 1949–1953 (1998).

151. Derdeyn, C. A. & Hunter, E. Viral characteristics of transmitted HIV. *Curr. Opin. HIV AIDS* **3**, 16–21 (2008).

152. Haaland, R. E. *et al.* Inflammatory genital infections mitigate a severe genetic bottleneck in heterosexual transmission of subtype A and C HIV-1. *PLoS Pathog.* **5**, e1000274 (2009).

153. Rong, R. *et al.* Escape from autologous neutralizing antibodies in acute/early subtype C HIV-1 infection requires multiple pathways. *PLoS Pathog.* **5**, e1000594 (2009).

154. Korber, B. *et al.* Evolutionary and immunological implications of contemporary HIV-1 variation. *Br. Med. Bull.* **58**, 19–42 (2001).

155. Zhang, M. *et al.* Tracking global patterns of N-linked glycosylation site variation in highly variable viral glycoproteins: HIV, SIV, and HCV envelopes and influenza hemagglutinin. *Glycobiology* **14**, 1229–1246 (2004).

156. Pejchal, R. *et al.* A potent and broad neutralizing antibody recognizes and penetrates the HIV glycan shield. *Science* **334**, 1097–1103 (2011).

157. Blattner, C. *et al.* Structural delineation of a quaternary, cleavage-dependent epitope at the gp41-gp120 interface on intact HIV-1 Env trimers. *Immunity* **40**, 669–680 (2014).

158. Scanlan, C. N. *et al.* The broadly neutralizing anti-human immunodeficiency virus type 1 antibody 2G12 recognizes a cluster of alpha1-->2 mannose residues on the outer face of gp120. *J. Virol.* **76**, 7306–7321 (2002).

159. Scharf, L. *et al.* Antibody 8ANC195 reveals a site of broad vulnerability on the HIV-1 envelope spike. *Cell Rep.* **7**, 785–795 (2014).

160. Garces, F. *et al.* Structural evolution of glycan recognition by a family of potent HIV antibodies. *Cell* **159**, 69–79 (2014).

161. Huang, J. *et al.* Broad and potent HIV-1 neutralization by a human antibody that binds the gp41-gp120 interface. *Nature* **515**, 138–142 (2014).

162. Kong, L. *et al.* Supersite of immune vulnerability on the glycosylated face of HIV-1 envelope glycoprotein gp120. *Nat. Struct. Mol. Biol.* **20**, 796–803 (2013).

163. McLellan, J. S. *et al.* Structure of HIV-1 gp120 V1/V2 domain with broadly neutralizing antibody PG9. *Nature* **480**, 336–343 (2011).

164. Pancera, M. *et al.* Structural basis for diverse N-glycan recognition by HIV-1-neutralizing V1-V2-directed antibody PG16. *Nat. Struct. Mol. Biol.* **20**, 804–813 (2013).

165. Freund, N. T. *et al.* A New Glycan-Dependent CD4-Binding Site Neutralizing Antibody Exerts Pressure on HIV-1 In Vivo. *PLoS Pathog.* **11**, e1005238 (2015).

166. Roe, D. R. & Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **9**, 3084–3095 (2013).

167. Tarr, G. *edgebundleR*. (2018).

168. Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).

169. Yu, W.-H. *et al.* Exploiting glycan topography for computational design of Env glycoprotein antigenicity. *PLoS Comput. Biol.* **14**, e1006093 (2018).

170. Kong, L., Wilson, I. A. & Kwong, P. D. Crystal structure of a fully glycosylated HIV-1 gp120 core reveals a stabilizing role for the glycan at Asn262. *Proteins* **83**, 590–596 (2015).

171. Starcich, B. R. *et al.* Identification and characterization of conserved and variable regions in the envelope gene of HTLV-III/LAV, the retrovirus of AIDS. *Cell* **45**, 637–648 (1986).

172. Robey, W. G. *et al.* Characterization of envelope and core structural gene products of HTLV-III with sera from AIDS patients. *Science* **228**, 593–595 (1985).

173. Veronese, F. D. *et al.* Characterization of gp41 as the transmembrane protein coded by the HTLV-III/LAV envelope gene. *Science* **229**, 1402–1405 (1985).

174. Moore, P. L. *et al.* The c3-v4 region is a major target of autologous neutralizing antibodies in human immunodeficiency virus type 1 subtype C infection. *J. Virol.* **82**, 1860–1869 (2008).

175. Richman, D. D., Wrin, T., Little, S. J. & Petropoulos, C. J. Rapid evolution of the neutralizing antibody response to HIV type 1 infection. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 4144–4149 (2003).

176. Sagar, M., Wu, X., Lee, S. & Overbaugh, J. Human immunodeficiency virus type 1 V1-V2 envelope loop sequences expand and add glycosylation sites over the course of infection, and these modifications affect antibody neutralization sensitivity. *J. Virol.* **80**, 9586–9598 (2006).

177. Lasky, L. A. *et al.* Neutralization of the AIDS retrovirus by antibodies to a recombinant envelope glycoprotein. *Science* **233**, 209–212 (1986).

178. Montagnier, L. *et al.* Identification and antigenicity of the major envelope glycoprotein of lymphadenopathy-associated virus. *Virology* **144**, 283–289 (1985).

179. Sok, D. *et al.* Promiscuous glycan site recognition by antibodies to the high-mannose patch of gp120 broadens neutralization of HIV. *Sci. Transl. Med.* **6**, 236ra63 (2014).

180. Petrescu, A., Milac, A., Petrescu, S., Dwek, R. & Wormald, M. Statistical analysis of the protein environment of N-glycosylation sites: implications for occupancy, structure, and folding. *Glycobiology* **14**, 103–114 (2004).

181. Grant, B. J., Rodrigues, A. P. C., ElSawy, K. M., McCammon, J. A. & Caves, L. S. D. Bio3d: an R package for the comparative analysis of protein structures. *Bioinforma. Oxf. Engl.* **22**, 2695–2696 (2006).

182. Zhou, T. *et al.* Structural basis for broad and potent neutralization of HIV-1 by antibody VRC01. *Science* **329**, 811–817 (2010).

183. Dunfee, R. L. *et al.* Loss of the N-linked glycosylation site at position 386 in the HIV envelope V4 region enhances macrophage tropism and is associated with dementia. *Virology* **367**, 222–234 (2007).

184. Ward, A. B. & Wilson, I. A. The HIV-1 Envelope Glycoprotein Structure: Nailing down a Moving Target. *Immunol. Rev.* **275**, 21–32 (2017).