Natural Polymorphisms at *gag* cleavage sites and their potential impact on the substrate envelope structure of HIV-1 Subtype C

By

Laurinda Mqhaba

Submitted in fulfilment of the requirements for the degree of Master of
Medical Science in Molecular Virology

Nelson R Mandela School of Medicine, School of Laboratory Medicine and Medical Sciences, University of KwaZulu-Natal

2025

Preface

The experimental work described in this dissertation was carried out at the Nelson R. Mandela School of Medicine, University of KwaZulu-Natal, Durban, from March 2020 to June 2025 under the supervision of Professor Michelle Lucille Gordon.

These studies represent original work by the author and have not otherwise been submitted in any form for any degree or diploma to any other University. Where use has been made of the work of others, it is duly acknowledged in the text.

Signed: _____ [redacted] _____ Date: _____

Laurinda Mqhaba (Candidate)

Signed: _____ Date: _____

Professor Michelle Lucille Gordon (Supervisor)

Plagiarism declaration

I, Laurinda Mqhaba, declare that:

i. The research reported in this dissertation, except where otherwise indicated, is my original work.

ii. This dissertation has not been submitted for any degree or examination at any other university.

iii. This dissertation does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.

v. This dissertation does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:

a. their words have been re-written but the general information attributed to them has been referenced;

b. Where their exact words have been used, their writing has been placed inside quotation marks, and referenced.

v. Where I have reproduced a publication of which I am an author, co-author or editor, I have indicated in detail which part of the publication was actually written by myself alone and have fully referenced such publications.

vi. This dissertation does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the dissertation and in the Reference sections.

Signed:  Date: 07 July 2025

Laurinda Mqhaba (Candidate)

Signed:  Date: 07 July 2025

Professor Michelle Lucille Gordon (Supervisor)

Ethical Approval

Full ethical approval, was obtained for this study from the Biomedical Research Ethics Committee of the Nelson R. Mandela School of Medicine, University of KwaZulu-Natal (Protocol reference number: 00014274).

Dedication

To my mother,
whose love and strength carry me through every challenge.
Your faith lights my path and inspires my journey,
this work is dedicated to you.

“Just like moons and like suns,
With the certainty of tides,
Just like hopes springing high,
Still I’ll rise.”
— Maya Angelou, *Still I Rise*

Acknowledgements

I would like to extend my sincere gratitude to Professor Michelle Gordon for her invaluable contribution to this study. Her guidance in the conception and design of the project, as well as her support throughout the writing of this dissertation, has been instrumental. This work would not have been possible without her insight, encouragement, and dedication.

I would also like to sincerely thank Dr Verrona Marie for her insightful guidance and for helping clarify key concepts that were essential to shaping the foundation of this study. Her support and input had a meaningful impact on the direction and development of this research.

I would like to thank the University of Kwazulu-Natal and the National Research Foundation for their support and assistance throughout the course of this research. The resources, opportunities, and funding they provided were instrumental in enabling the successful completion of this study.

Lastly, I would like to sincerely thank my family and partner for their unconditional love, support, and encouragement throughout this journey. Their belief in me and their constant presence gave me the strength to persevere and complete this dissertation. I am truly grateful for everything they have done for me.

Abstract

Limited studies have investigated the natural variations within the *gag* gene of HIV-1 subtype C, particularly at the cleavage sites (CSs), with most existing research focusing on subtype B. This study extended prior findings by comparing the natural variability at the CSs between HIV-1 subtypes B and C, extending the analysis from 5AA to 10AA and 15 AA on either side of the scissile bond, highlighting differences that may contribute to protease (PR)-substrate interactions and viral fitness. In addition, this study provided a more comprehensive understanding of how natural polymorphisms at the CSs (5AA) influence the substrate envelope, the substrate's consensus volume, which serves as a template that the PR uses to recognize and bind to a specific CS. The findings revealed distinct patterns of CS variability between subtypes B and C. Notably, subtype C sequences exhibited high variability at the P2/NC and P1/P6 CSs. The P2/NC CS showed the highest variability, with 100% of sequences in subtype C being polymorphic at this site. Furthermore, the study demonstrated that the increase in sequence length from 5AA to 15AA amplified the variability, particularly at the P2/NC and P1/P6 sites. While this was expected, it was interesting to note that the greatest variability was seen where the extended sites overlapped. This suggests that subtype C may have a more diverse and mutable PR CS profile. However, this requires further investigation.

The structural analysis of the CSs showed that strong binding affinities were linked to extensive hydrogen bonding and π-alkyl interactions, often involving conserved residues, while unfavourable interactions such as steric clashes weakened binding. Subtype B generally had more diverse and distributed interactions, including extensive hydrophobic contacts (e.g., Val32, Ile50), salt bridges, and favourable hydrogen bonds involving the D25, Asp29, and Asp30 residues. Subtype C often formed fewer but stronger hydrogen bonds (shorter distances), with specific π interactions (e.g., with Val82), but also displayed unfavourable donor–donor clashes, especially in MA/CA and NC/P1 complexes. For P2/NC, subtype B had a wider interaction network, while subtype C relied on localized binding. Although subtype C sometimes showed slightly higher binding affinities (e.g., -8.3 kcal/mol), subtype B's interactions were more varied and involved more structural and catalytic residues, suggesting potentially more stable binding overall.

In conclusion, natural polymorphisms at the *gag* CSs impacted the structure of the substrate envelope of HIV-1 subtype C which could impact the cleavage by PR. These findings emphasize the importance of understanding the distinct mutation profiles of HIV-1 subtypes B versus C, which is important for the advancement of effective therapeutic strategies to combat HIV-1 globally.

Contents

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

2D	Two-dimensional
3D	Three-dimensional
AA	Amino acids
ADP	Adenosine Diphosphate
AI	Artificial Intelligence
AIDS	Acquired immune deficiency syndrome
ALIX	ALG-2 interacting protein 1/X
AMBER	Assisted model building with energy refinement
AM1-BCC	Austin Model 1 – Bond Charge Corrections
ATI	Antiretroviral treatment interruption
CA	Capsid; P24
CDC	Centre for Disease Control and Prevention

Cpptraj	C++ Trajectory Processor
CRF	Circulating recombinant form
CRF01_AE	Circulating recombinant form combination of genetic material from subtypes A and E
CRF02_AG	Circulating recombinant form combination of genetic material HIV-1 subtypes A and G
Cryo-EM	Cryo-electron microscopy
CS	Cleavage Site
CSs	Cleavage Sites
CTD	C-terminal domain
ESCRT	Endosomal Sorting Complex Required for Transport
GA	Genetic algorithm
Gag	group-specific antigen
GAFF	General Amber Force Field
GB	generalized Born
Gbind	binding free energy
GPU	Graphics processing unit
gRNA	genomic RNA
HIV	Human immunodeficiency virus
HIV-1	Human immunodeficiency virus type 1
HIV-2	Human immunodeficiency virus type 2
HTLV-III	Human T-lymphotropic virus type III
LANL	Los Alamos National Laboratory
LAV	Lymphadenopathy-associated virus
LEaP	Linking, Editing, Analysing, and Parameterizing
LTR	Long terminal repeat
MA	Matrix; P17
MD	Molecular dynamics
MM-GBSA	Molecular mechanics-generalized born surface area
MM-PBSA	Molecular mechanics-Poisson Boltzmann surface area
MMPBSA.PL	is a Perl script that was originally part of the AMBER suite used to perform Molecular Mechanics Poisson–Boltzmann Surface Area (MM-PBSA) calculations
MMPBSA.PY	is a Python-based tool included in the AMBER suite designed to calculate binding free energies using the Molecular Mechanics Poisson–Boltzmann Surface Area (MM-PBSA) or Generalized Born Surface Area (MMGBSA) methods
Mol2	SYBYL Mol2 file format
MPI	Message passing interface

NC	Nucleocapsid; p7
nm	nanometre
NMR	Nuclear Magnetic Resonance
NPT	Constant pressure and normal temperature
ns	Nanoseconds
ntb	Nonbonded type for boundary conditions
ntc	Number of total constraints on bonds involving hydrogen atoms
ntt	Thermostat type
NTD	N-terminal domain
NVT	Constant volume and normal temperature
P1	Spacer peptide one
P2	Spacer peptide two
PCP	*Pneumocystis carinii pneumonia*
PDB	Protein Databank
PDB2PQR	Protein Data Bank to PQR, where PQR is a file format used in computational chemistry
pH	potential of hydrogen
PI	Protease inhibitor
PIC	Pre-integration complex
pKa	The negative base-10 logarithm of the acid dissociation constant (Ka)
PME	Particle mesh ewald
Pmemd	Particle Mesh Ewald Molecular Dynamics
Pmemd.CUDA	Particle Mesh Ewald Molecular Dynamics. Compute Unified Device Architecture
PPI	Protein-protein interaction
ps	Picoseconds
PTMs	Posttranslational modifications
ptraj	Processing of Trajectory
PR	Protease
QM/MM	Quantum mechanics/molecular mechanics
RC	Replicative capacity
REMD	Replica exchange molecular dynamics
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation
SAR	Structure-activity relationships
SIV	Simian immunodeficiency virus
SIVcpz	Simian immunodeficiency virus chimpanzee
SIVgor	Simian immunodeficiency virus gorilla

SIVsmm	Simian immunodeficiency virus sooty mangabeys
SmFRET	Single-molecule fluorescence resonance energy transfer
SQM	Semiempirical quantum mechanics
TIP3PBOX	Transferable Intermolecular Potential with 3 Points in a Box
TLEaP	Terminal LEaP
Tsg101	Tumour susceptibility gene 101
UCSF	University of California, San Francisco
UNAIDS	Joint United Nations Programme on HIV/AIDS
vdW	van der Waals
WHO	World Health Organization
WT	Wildtype
XLEaP	xleap is a GUI version of AMBER's LEaP program

Chapter 1: Literature Review

1.1 Introduction

The Gag protein is a 500-amino acid (AA) precursor that plays a critical role in the assembly and maturation of HIV-1 subtype C virions (Sundquist and Krausslich, 2012). In the late phase of the HIV-1 replication cycle, Gag is sequentially cleaved by the HIV-1 aspartyl protease (PR) enzyme comprising 99 AAs per monomer, into the matrix (MA), capsid (CA), nucleocapsid (NC), and P6 domains, along with two spacer peptides (P1 and P2) to generate a mature, infectious virion (Bell and Lever, 2013). Interestingly, the Gag CSs exhibit minimal AA sequence identity (Prabu-Jeyabalan et al., 2002). Polymorphisms within the Gag protein of HIV-1 subtype C have been shown to affect cleavage efficiency by the viral PR, potentially influencing the kinetics of virion maturation and the efficacy of antiretroviral therapies targeting the PR enzyme (Velazquez-Campoy et al., 2001).

Despite the variability at the CSs, certain features such as the exclusion of β-branched AA from the P1 site, and preference for hydrophobic AA along the scissile bond (P1/P1'), remain consistent (Pettit et al., 1991). These AA arrangements contribute to a conserved structural motif known as the substrate envelope, shaped through substrate and enzyme coevolution (Kurt Yilmaz et al., 2016; Özen et al., 2011; Prabu-Jeyabalan et al., 2000, 2002). Gag mutations near the CSs have been shown to restore Gag-protease binding by introducing new chemical interactions and inducing subtle conformational changes that compensate for the reduced binding affinity caused by mutations in HIV-1 PR (Özen et al., 2014).

In addition to mutations near the protease CSs, Gag also contains mutations in regions distant from these sites, many of which have been shown to directly contribute to protease inhibitor (PI) resistance (Gatanaga et al., 2002; Myint et al., 2004; Sutherland et al., 2015). The functional significance of these non-CS mutations remains largely unresolved; however, emerging evidence indicates they may participate in long-range allosteric communication within the Gag-protease complex (Su et al., 2018). These mutations influence the conformational dynamics of the Gag polyprotein, impacting the accessibility of the PR enzyme to the CS (Li et al., 2013; Mariani et al., 2014).

Therefore, this study extended previous studies by investigating the variability at the CSs spanning up to 15AA on either side of the scissile bond, which could have implications for cleavage efficiency. In addition, the binding affinity of the *gag* CS variants and their substrate envelopes were compared between subtypes B and C.

1.2 Epidemiology

The global AIDS pandemic remains a significant public health challenge. As of 2024, an estimated 39.9 million people worldwide are infected with HIV, along with 1.3 million novel infections and 630,000 HIV-associated deaths reported annually (UNAIDS/WHO estimates, 2024). The distribution of HIV infections varies significantly across different regions and populations worldwide. Sub-Saharan Africa is the most affected region, with the highest prevalence and incidence rates of HIV infections (UNAIDS/WHO estimates, 2024). In contrast, other regions such as North America and Europe have seen stabilization in infection rates due to effective prevention and treatment efforts (UNAIDS/WHO estimates, 2024). The manner in which HIV spreads includes sexual transmission, blood-borne transmission by exposure to infected blood through shared injection equipment, tainted transfusions, or transmission from an infected mother to her baby during pregnancy, childbirth, or breastfeeding (Sharp and Hahn, 2011).

1.3 History and Classification of the Human Immunodeficiency Virus (HIV)

In 1981, the Centre for Disease Control and Prevention(CDC) documented cases of *Pneumocystis carinii pneumonia* (PCP) among previously healthy homosexual men, marking the initial recognition of AIDS. Subsequent reports identified AIDS cases in diverse populations, including recipients of blood transfusions and injection drug users (Goedert and Gallo, 1985). Revolutionary research in 1983 by Françoise Barré-Sinoussi and Luc Montagnier led to the detection of a retrovirus in a lymph node tissue sample from a French homosexual individual. It was provisionally termed the lymphadenopathy-associated virus (LAV) (Barré-Sinoussi et al., 1983). Concurrently, Robert Gallo and colleagues isolated a similar virus from multiple AIDS patients, designating it human T-lymphotropic virus type III (HTLV-III) (Gallo et al., 1984).These discoveries paved the way for identifying the virus responsible for AIDS, which was officially named the human immunodeficiency virus (HIV) in 1986 (Coffin J, 1986).

HIV is classified into two main types: HIV-1 and HIV-2, both falling under the Lentivirus genus within the *Retroviridae* family (Williams and Burdo, 2009). HIV-1 is further categorized into groups M (main), N (non-M; non-O), O (outlier), and P, with group M comprising nine subtypes (A–D, F–H, J, K) and numerous circulating recombinant forms (Lau and Wong, 2013). Among these, HIV-1 subtype C is the most dominant globally, particularly affecting sub-Saharan Africa (Castley et al., 2017). HIV-2, originally endemic to West Africa, is classified into groups A–H, with groups A and B

being the predominant pathogenic types (Hahn et al., 2000). Both HIV-1 and HIV-2 stem from zoonotic transmissions of simian immunodeficiency viruses (SIVs) from non-human primates to humans. HIV-1 groups M and N originated from chimpanzees (SIVcpz), while HIV-1 group O and HIV-2 likely emerged from gorillas (SIVgor) and sooty mangabeys (SIVsmm), respectively (Bailes et al., 2003).

1.4 Geographic distribution of HIV-1

Globally, HIV-1 subtypes exhibit substantial geographic variation, reflecting regional disparities in infection rates. Subtype C dominates in Eastern and sub-Saharan Africa, Brazil, and the Indian Pacific, while subtype B is prevalent in the US, Japan, Europe, and Northern Africa (Su et al., 2018; Teto et al., 2017). CRF01_AE and CRF02_AG are prominent in Southeast Asia and West Africa, contributing to the overall HIV-1 diversity (Olesen et al., 2018; Visseaux et al., 2016). These patterns are shaped by historical factors such as the founder effect, where initial virus introductions spread and were established within populations (Buonaguro et al., 1995; McCutchan et al., 1992; Myers, 1994).

1.5 The structure of HIV-1

Retroviruses, including HIV-1, share fundamental physical and genetic characteristics that distinguish them from other types of viruses. The HIV-1 virion is spherical in shape and about 100 nm in diameter (Charneau et al., 1994). A defining feature is their genome, consisting of two positive-sense RNA strands organized into nine genes across three reading frames (Chameettachal et al., 2023; Kieken et al., 2002). These genes are essential for viral replication and pathogenesis, with *gag, pol*, and *env* encoding the virus's main structural and enzymatic components, while regulatory genes such as *tat, rev, vif, vpu, vpr*, and *nef* govern viral gene expression and evasion strategies (Watts et al., 2009).The *gag* gene directs the synthesis of MA, CA and NC proteins which are critical for virion assembly and genome packaging, while *pol* encodes enzymes like PR, reverse transcriptase, and integrase necessary for viral replication (Watts et al., 2009). The *env* gene produces glycoproteins gP120 and gp41, forming the viral envelope's spikes that facilitate entry into host cells (Gelderblom et al., 1987).

1.6 The HIV Replication Cycle

The replication cycle of HIV-1 involves a sequence of events divided into early and late phases (Nisole and Saïb, 2004). Initially, the virus binds to host CD4+ cells through engaging with its receptor and a co-receptor (Wilen et al., 2012). This interaction causes conformational shifts in the viral envelope, allowing fusion with the host cell membrane (Kirchhoff, 2013). Reverse transcription

is catalysed by the viral enzyme reverse transcriptase, converting the viral RNA genome into double-stranded DNA (Engelman and Cherepanov, 2012). The viral DNA is later fused into the host cell genome through the action of integrase. Integration via integrase involves processing viral DNA ends and inserting them into the host chromosomal DNA (Krishnan and Engelman, 2012).

Once integrated, the viral DNA can either remain dormant (lysogenic replication) or become active (lytic replication). During replication, the host cell's mechanisms are exploited to transcribe viral RNA and synthesize polyproteins, including the structural proteins encoded by the *gag* gene (Kuzembayeva et al., 2014) . These polyproteins assemble at the host cell membrane, forming immature viral particles. (Rossi et al., 2021). The mature virions bud from the host cell membrane, acquiring an envelope containing viral glycoproteins gp120 and gp41 (Gelderblom et al., 1987). This budding process completes the late phase of the HIV-1 replication cycle, releasing infectious virus particles capable of infecting new host cells.

1.7 PR

The HIV-1 PR is a homodimeric enzyme belonging to the family of aspartic PRs, with each subunit composed of 99 AAs arranged into a single α-helix and nine anti-parallel β-sheets (Brik and Wong, 2003; Davies, 1990). These β-sheets form a conserved dimer interface involving residues 1–4 and 96–99, crucial for the structural integrity of PR (Todd et al., 1998). The PR enzyme is initially part of the Gag-Pol polyprotein, where it triggers auto-processing for virion maturation (Louis et al., 1999b). Upon release, PR molecules catalyse further cleavage events, accelerating their own liberation through positive feedback loops. Within its dimeric structure, PR's catalytic triad D25-T26-G27 from both monomers shapes a hydrophobic active site cavity essential for enzymatic function (Miller et al., 1989).

The active site of PR is dynamically regulated by β-hairpin structures, known as flaps, which play a pivotal role in substrate recognition and catalysis (Groves et al., 1998). Without a substrate, PR adopts a semi-open conformation, whereas binding with a ligand triggers a closed conformation essential for proteolysis (Duan et al., 2003). Two models have been proposed to explain the flap dynamics mechanism: one posits that the flaps open completely upon collision with the substrate, later shifting to a closed conformation as the substrate nears (Scott and Schiffer, 2000), while the other proposes an initial semi-open state (Collins et al., 1995).

Functionally, PR drives the maturation of HIV-1 by catalysing the proteolytic processing of Gag and Gag-Pol polyproteins (Chou, 1996). It exhibits broad substrate specificity, targeting 12 canonical CSs within Gag precursors. This generates 66 distinct molecular substrates, intermediates, and products

concurrently within the virion (Beck et al., 2002). Similar to non-viral PRs, PR employs an acid-base mechanism facilitated by two critical aspartic acid residues in its active site (Brik and Wong, 2003).

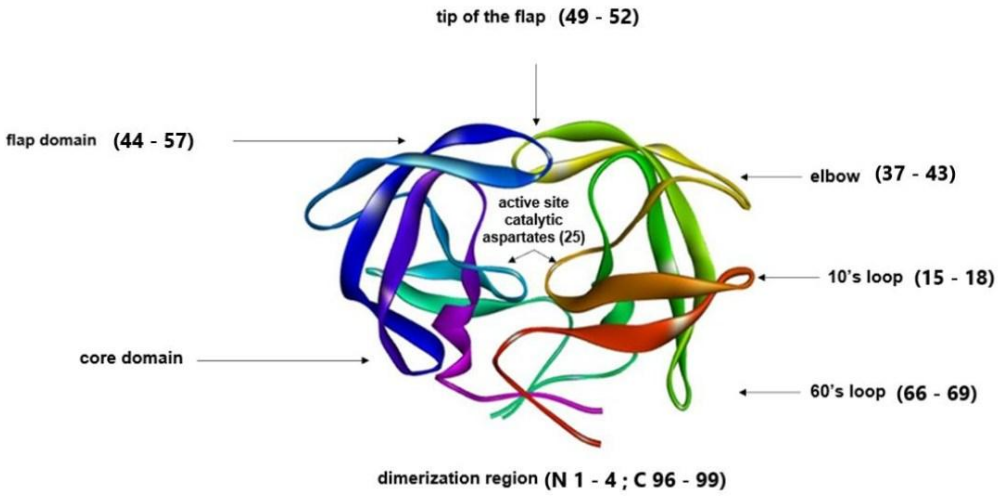


Figure 1: Structural depiction of HIV-1 subtype C PR enzyme generated in Biovia Discovery Studio (Systèmes, 2021)

1.8 The Function of the Gag Polyprotein during the HIV-1 Replication Cycle

The HIV-1 Gag protein plays a crucial role in virus replication and infectivity through its diverse functional and structural contributions (Marie and Gordon, 2022). Upon transcription of proviral DNA by host RNA polymerase II, the resultant RNA serves dual roles: as mRNA for translation and as a template for generating Gag and Gag-Pol polyproteins and packaging the viral genome (Ni et al., 2011; Ocwieja et al., 2012). Gag, a 55 kDa precursor protein, is produced in the host cell cytosol and subsequently undergoes cleavage to yield individual proteins—MA, CA, NC, a small P6 domain and two spacer peptides (P1 and P2) (Freed, 2015). This polyprotein is pivotal in viral assembly due to its ability to interact with genomic RNA (gRNA) through cis-acting packaging signals located in the 5' untranslated region (UTR) of Gag (Chamanian et al., 2013; Miyazaki et al., 2011).

Once synthesized, the Gag-RNA complex is transported through the cytoplasm to specific assembly sites at the plasma membrane, marked by a dense lattice of actin filaments that facilitate virion budding (Gladnikoff et al., 2009). At the plasma membrane, Gag orchestrates the rapid accumulation of viral components, initiates RNA dimerization, and ensures the proper balance of gRNA dimers in

the cytoplasm (Dubois et al., 2018). This intricate orchestration highlights Gag's indispensable role in the HIV-1 lifecycle, from the initial stages of genome packaging to the assembly and release of mature virions.

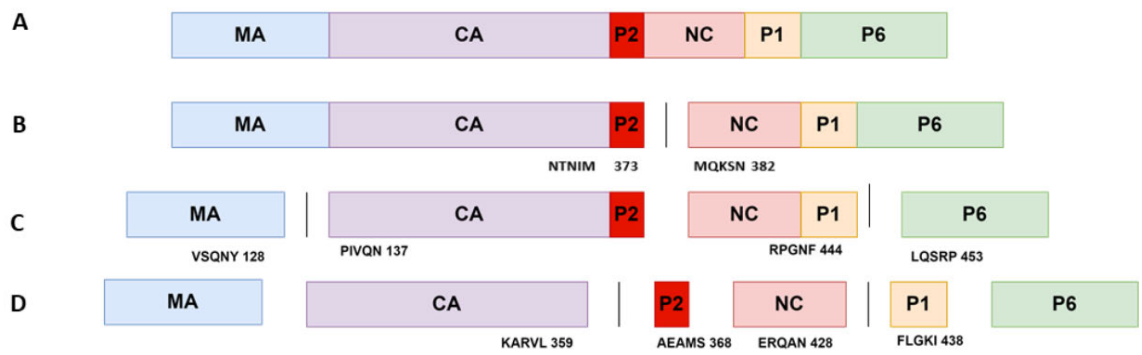


Figure 2: : Schematic diagram of Gag viral maturation in an immature viral particle, showing the 5 CSs

Following cleavage of (A) p55 Gag polyprotein by PR into: (B) P2/NC which is the first site to be cleaved and is the rate determining step, (C) MA/CA and P1/P6 are cleaved second, (D) CA/P2 and NC/P1 which are cleaved last resulting in fully cleaved Gag protein (MA, CA, P2, NC, P1 and P6.) Abbreviations: MA = matrix, CA = capsid and NC = nucleocapsid, P1=Spacer peptide 1, P2=Spacer Peptide 2.

1.8.1 MA

Comprising of 128 AA , MA is integral to the structural integrity of HIV-1 and its ability to infect host cells (Massiah et al., 1994; Massiah et al., 1996). Post-translation, MA undergoes lipidation at its N-terminus through the attachment of a saturated long-chain fatty acid, crucial for virion construction and interaction with the host cell membrane (Chukkapalli et al., 2010). This interaction is critical for assembling Gag polyproteins into higher ordered structures such as hexamers, which form atop capsid hexamers during virion formation (Alfadhli et al., 2007; Checkley et al., 2011). Once cleaved, MA remains bound to the virion's lipid envelope, further underscoring its importance in the structural integrity of mature virions (Datta et al., 2007; Hurley et al., 2010).

Additionally, MA's role in directing *gag* to the plasma membrane involves host cell proteins such as ADP ribosylation factor and Golgi-localized γ-ear-binding proteins, which aid in its trafficking and positioning during virion assembly (Joshi et al., 2008). The formation of MA hexamers of trimers not only regulates envelope integration into virions but also influences virion infectivity through steric trapping mechanisms and interactions with the cytoplasmic tail of gp41 (Buttler et al., 2018; Zhu et al., 2003). Its conservation across HIV-1 subtypes makes MA an attractive target for antiviral drug

development, particularly given its role in directing viral and cellular components to assembly sites and facilitating envelope glycoprotein incorporation into budding viruses (Jouvenet et al., 2006).

1.8.2 CA

The CA region plays a pivotal role in viral replication and is highly conserved across HIV-1 strains (Rihn et al., 2013). CA consists of two distinct domains connected by a proline-rich linker: the N-terminal domain (NTD) and the C-terminal domain (CTD), both comprising of alpha helices crucial for core assembly and viral transmissibility (Jiang et al., 2011). During maturation, the cleavage of MA and CA sites leads to the formation of a beta-hairpin structure stabilized by a salt bridge, which is integral for the spherical fullerene core formation of the mature virion (Ganser-Pornillos et al., 2008; Kelly et al., 2006). The NTD and CTD interact to form homodimers that link polymer rings into a triple-axis network, essential for higher ordered viral structures (Pornillos et al., 2011). The maturation process orchestrated by viral PR leads to the formation of a hexagonal conoidal structure with pentameric rings at its ends, ensuring the structural integrity of the mature core (Briggs and Kräusslich, 2011). Given its critical role and conservation, CA is an attractive target for therapeutic intervention. Inhibitors designed to disrupt CA-CA interactions in the Gag lattice or the mature core have shown promise in preventing viral maturation and replication, highlighting CA as a viable target for antiviral strategies (Stremlau et al., 2004).

1.8.3 P2

The P2 spacer peptide, located between the MA and NC regions of the HIV-1 Gag polyprotein, consists of 14 AA and forms a six-helix bundle that stabilizes Gag hexamers in immature virions (Bell and Lever, 2013; Wright et al., 2007) . Cleavage of P2 from NC during viral maturation is essential for the formation of ribonucleoprotein complexes and condensation of the CA domain (Shehu-Xhilaga et al., 2001). Additionally, delayed cleavage at the CA/P2 site facilitates proper virion morphogenesis through sequential CA-CA interactions, highlighting the importance of the intact CA/P2 segment in immature particle assembly (Gross et al., 2000).

1.8.4 NC

The NC protein, spans 55 AAs and possesses two zinc fingers (Eswar et al., 2006). It is located at the C-terminal section of the Gag polyprotein and plays a pivotal role in HIV-1 replication by facilitating RNA chaperone activity and genome packaging (Levin et al., 2010). The zinc fingers are pivotal in HIV-1 replication, with each finger playing distinct roles: one facilitates genomic RNA encapsulation

and helix destabilization, while the other regulates virion quality assessment (Guo et al., 2000). Disruption of these zinc finger motifs leads to the production of immature and non-infectious viral particles (Houzet et al., 2008).

NC also contributes to virus-cell communication, establishing viral synapse formation and aiding in the congregation of Gag proteins (Llewellyn et al., 2010). Its role extends to the uncoiling of viral DNA during reverse transcription and facilitating nucleic acid rearrangements essential for viral replication (Muriaux and Darlix, 2010). Mutations within NC, such as the N17K variant, can significantly enhance viral transduction by promoting increased RNA packaging and mature virion formation (Chien et al., 2006). NC is a multifunctional protein critical for HIV-1 replication and virion assembly (Levin et al., 2010). Its precise interactions with RNA and cellular factors underscore its importance in the viral life cycle, making it a potential target for therapeutic interventions aimed at disrupting viral replication (Freed, 2015).

1.8.5 P1

The P1 spacer, located between NC and P6 in Gag, contains conserved prolines (P439, P445) vital for Gag-Pol incorporation (Bell and Lever, 2013; Hill et al., 2002). It overlaps the frameshift site, making it sensitive to mutations. Proline-to-leucine substitutions reduce NC-RNA stability and infectivity (Hill et al., 2002). While NC/P1 mutations have little effect, those generating P15 (NC/P1/P6) or p8 (P1/P6) peptides impair proviral integration, indicating distinct roles for these regions (Coren et al., 2007).

1.8.6 P6

The P6 domain spans 52 AAs at the C-terminal end of Gag. It is responsible for recruiting the endosomal sorting complex required for transportation (ESCRT) and incorporation of *vpr* into the virion (Freed, 2015). P6 exhibits significant polymorphisms, particularly influencing virus release through its interaction with the ESCRT machinery via the PTAP motif (Göttlinger et al., 1991). This motif specifically binds to Tsg101, a key host factor essential for viral budding (Raiborg and Stenmark, 2009). Mutations or truncations within the PTAP motif lead to defective budding and accumulation of non-infectious virions at the cell surface, highlighting the essential role of P6 in viral replication (Martin-Serrano and Neil, 2011). Moreover, P6 contains another late assembly domain characterized by the YPLASL sequence, which interacts with ALIX (ALG-2 interacting protein 1/X) to facilitate viral budding independently of Tsg101 (Strack et al., 2003; Usami et al., 2009). This redundancy in budding pathways underscores the importance of P6 in ensuring efficient viral release under varying cellular conditions. Structural studies often overlook P6 due to its inherent disorder and flexibility (Xue et al., 2011). Nevertheless, its absence, as observed in truncated Gag polyproteins

lacking P1 and P6 domains, disrupts inter-domain interactions crucial for virion maturation (Deshmukh et al., 2015). Specifically, P6 influences the dynamic assembly of Gag, particularly in forming the hexameric MA/CA structure at the membrane interface (Fossen et al., 2005). P6 of HIV-1 integrates multiple late assembly domains crucial for coordinating viral budding, making it a potential target for therapeutic interventions aimed at disrupting viral assembly and release (Chen and Wang, 2024). Its intricate interactions with cellular factors underscore its pivotal role in the HIV-1 replication cycle (Bell and Lever, 2013; Lever and Lever, 2011).

Table 1: Wild-type HIV-1 subtype C *gag* CS sequences (Oliveira et al., 2003)

CLEAVAGE SITES	SEQUENCE
MA/CA	VSQNY/PIVQN
CA/P2	KARVL/AEAMS
P2/NC	NTNIM/MQKSN
NC/P1	ERQAN/FLGKI
P1/P6	RPGNF/LQSRP

1.9 The Substrate Envelope Hypothesis

The Substrate Envelope Hypothesis proposes that approximately 1000 Å^2 of the PR surface becomes buried upon binding to its substrate (Prabu-Jeyabalan et al., 2002). This binding induces asymmetry in the enzyme, while maintaining intact water molecules and hydrogen bonds. This asymmetry is crucial for cleavage, as it depends more on the overall shape of the substrate rather than a specific residue pattern. The substrate's consensus volume, known as the substrate envelope, serves as a template that the PR uses to recognize and bind a specific CS (Chellappan et al., 2007; Nalam et al., 2010; Prabu-Jeyabalan et al., 2002; Shen et al., 2013). Within the substrate, the P3-P1' region plays a critical role in stable binding and specificity for the PR (Huang and Chen, 2013). The P1 loop acts like a guide, providing necessary positional cues to the P1 and P1' residues during flap movements. When the PR flaps close over the substrate, there is a reduction in van der Waals forces, indicating stronger interactions between P1 and P1' residues during early substrate recognition compared to when the flaps are closed (Prabu et al., 2006).

1.9.1 Bi-functional S2 and S2' subsites

HIV-1 protease cleaves substrates with either a proline or hydrophobic residue at P1' (Pettit et al., 1991). A P1' proline redirects P2 to the hydrophilic S2 and P2' to the hydrophobic S2' subsite, while its absence favours polar interactions at S2' (Griffiths et al., 1992; Potempa et al., 2018). This flexibility suggests PR's subsites adapt to different substrate types (Desantis et al., 2022).

1.10 Protein folding

A polypeptide chain must undergo a complicated folding process to take on a distinctive three-dimensional structure (Crick, 1958; Mayer, 2006). The protein's AA sequence encodes the information necessary for the conformation that is biologically native (Alberts B, 2002; Anfinsen, 1973; Crick, 1958). When denatured ribonuclease was added to a physiological buffer solution, Anfinsen and colleagues' research in the 1960s demonstrated that the enzyme refolded to its original conformation and recovered nearly all its biological activity. This fundamentally proved that the AA sequence carries the information necessary to fold a protein into its 3D active configuration (Anfinsen, 1972, 1973). Multiple intramolecular noncovalent interactions help the slow but spontaneous transition from an unfolded to a folded state (Anfinsen, 1972). A protein's native conformation is a thermodynamic state where the protein molecule's total free energy is as low as it can be due to the optimization of several favourable interactions and the minimization of unfavourable ones (Anfinsen, 1972; Branden and Tooze, 2012; Dill and MacCallum, 2012). Two types of intramolecular interactions can be identified as contributing to protein folding: (a) interactions that originate from forces inherent in the protein molecule and (b) interactions that are influenced by the solvent in the surrounding molecule. The former includes van der Waals and steric interactions, while the latter includes hydrogen bonding, electrostatic, and hydrophobic interactions (Kendrew et al., 1958; Pauling et al., 1951; Richards, 1974).

Protein folding remains a profound puzzle in molecular biology, where the linear sequence of AAs dictates the intricate 3D structure essential for function. The process involves a delicate interplay of thermodynamic forces, including hydrophobic interactions, hydrogen bonding, van der Waals forces, and electrostatic interactions, all navigating a complex energy landscape towards the native folded state (Anfinsen, 1973; Bryngelson and Wolynes, 1987; Kauzmann, 1959; Richards, 1974; Tanford, 1962). Recent advances in computational modelling (Hollingsworth and Dror, 2018) and experimental techniques such as single-molecule fluorescence resonance energy transfer (smFRET) (Sasmal et al., 2016) have provided unprecedented insights into folding pathways, intermediate states, and the role of molecular chaperones in ensuring fidelity. However, challenges such as predicting folding kinetics and the impact of cellular environment on folding dynamics persist (Gershenson and Gierasch, 2011).

1.11 Protein Structure

One important aspect of protein folding involves long-range interactions, which are critical for the proper folding and stability of proteins (Gromiha and Selvaraj, 1999). Interactions of this kind occur

between residues that are far apart in the primary sequence but are positioned close together in the 3D structure (Anfinsen, 1973; Richardson, 1981). For example, in a protein consisting of 10 AAs, folding may involve the formation of local secondary structures, with interactions occurring between residues that are up to 10 AAs apart in the sequence. In a slightly larger protein of 15 AAs, folding dynamics become more complex, with longer-range interactions becoming more pronounced (Anfinsen, 1973; Debrunner et al., 1969; Dill and Chan, 1997; Levinthal, 1969). These interactions can include the formation of more extensive secondary structures and interactions between distant regions of the protein chain, contributing to the overall stability and function of the protein (Anfinsen, 1973; Dill, 1990; Pauling and Corey, 1951). Understanding these long-range interactions is essential for accurately predicting protein structures and for gaining insights into their biological functions (Cao, 2020; Chung and Eaton, 2018; Jumper et al., 2021).

Proteins exhibit diverse structures at four hierarchical levels—primary, secondary, tertiary, and quaternary—each essential for their function.

1.11.1 Primary Structure

A protein's primary structure consists of its linear chain of AAs, linked through covalent bonds formed between the amino group of one AA and the carboxyl group of the next (Sanger, 1952; Sanger and Tuppy, 1951). While the backbone structure remains consistent across all AAs, it's the unique side chains that differentiate them (Fischer and Fischer, 1909; Hageman, 1977; Voet and Voet, 2010). The specific sequence of AAs determines how the protein will fold and the conformations its side chains will adopt, influencing its stability and function (Anfinsen, 1972, 1973).

1.11.2 Secondary Structure

Secondary structures arise from repetitive patterns of hydrogen bonding along the polypeptide chain (Pauling and Corey, 1951). The main secondary structures include α-helices and β-sheets, formed by specific arrangements of peptide bonds. α-Helices are helical coils stabilized by hydrogen bonds, with approximately 3.7 AA residues per turn, while β-sheets consist of strands linked by turns, allowing for different hydrogen bond orientations—parallel, anti-parallel, or mixed (Pauling et al., 1951).

1.11.3 Tertiary Structure

Tertiary structure refers to the 3D arrangement of a single polypeptide chain, where secondary structural elements and loops fold into a compact, functional shape (Branden and Tooze, 2012; Creighton, 1993). This folding process is driven by interactions such as hydrophobic interactions

(Kauzmann, 1959; Tanford, 1962) among AA side chains and non-covalent forces like hydrogen bonds (Pace et al., 2014; Pauling and Corey, 1951) and van der Waals interactions (Newberry and Raines, 2019; Richards, 1974). Loops, characterized by high flexibility and fewer hydrogen bonds, contribute to the overall flexibility of the protein (Kumar and Bansal, 1996; Lesk and Chothia, 1980). Domains within proteins are independent folding units that often perform specific functions and contribute to the overall tertiary structure (Alberts B, 2002; Wetlaufer, 1973).

1.11.4 Quaternary Structure

The quaternary structure describes the spatial arrangement formed when various polypeptide chains (subunits) merge to form a functional protein structure. These subunits are held together by non-covalent bonds like salt bridges and disulfide bonds, allowing for reversible dissociation of the complex (Alberts B, 2002; Lehninger et al., 2005; Perutz et al., 1960). Quaternary structures enable proteins to perform specialized functions, create specific substrate binding sites, and organize enzymatic activities within a defined spatial arrangement (Alberts B, 2002; Lehninger et al., 2005).

Understanding these hierarchical levels of protein structure is essential for comprehending their diverse functions within biological systems. Each level—from primary sequence to quaternary structure—dictates how proteins interact with their environment and fulfil their biological roles effectively.

1.12 Computational Methods

1.12.1 Molecular Docking

Molecular docking is an integral part of drug development, modification and reconfiguration (Meng et al., 2011; Yang et al., 2012). The process is optimal, quick and provides an opportunity for extensive evaluation of broad databases of molecular compounds (Ferreira et al., 2015). Molecular docking involves an in-silico procedure used to predict the binding modes of molecules that are connected to a receptor and to determine their molecular activity (Fadahunsi et al., 2024; Meng et al., 2011). It has two main objectives. Firstly, it is to accurately forecast and pinpoint the most favourable binding mode of a ligand in its active site or a protein's binding pocket (Lengauer and Rarey, 1996). The second is to classify a family of ligands according to their experimentally determined binding affinities (Vajda and Guarnieri, 2006). Molecular docking is founded on the notion of "lock and key" (Fischer, 1894). Molecular recognition occurs when the protein receptor is structurally and chemically compatible with the ligand. Binding is influenced by a combination of enthalpic and entropic factors,

including hydrogen bonding, van der Waals interactions, electrostatic forces, and changes in system disorder (Du et al., 2016; Freire, 2008; Meng et al., 2011). The high-throughput category of docking known as virtual or in silico screening, intends to collect a small catalogue of probable active compounds for downstream experimental assessment from a database of millions of compounds (Kitchen et al., 2004).

The protocols for docking have two crucial parts namely a scoring system and a superior positioning algorithm (Kitchen et al., 2004; Torres et al., 2019). Docking demands the large- scale sampling of configured space for a ligand in the binding pocket of a protein and therewith produces a broad number of potential positions that align a ligand in the active site. An adequate positioning algorithm tests all feasible binding modes, however the scoring system categorizes all the solutions and recognizes the most probable binding mode of the ligand (Huang and Zou, 2010; Meng et al., 2011). Docking algorithms can calculate the appropriate ligand's binding mode by evaluating various modes (Grinter and Zou, 2014; Miteva et al., 2011). Combining molecular docking and molecular dynamics simulations can provide a more precise and definitive protein-ligand complex (De Vivo et al., 2016; Santos et al., 2019).

1.12.2 Autodock and Autodock Vina

Autodock and Autodock Vina are prominent software tools utilized in computational biology to forecast the binding orientations and strengths of small molecules (ligands) to macromolecular receptors, such as proteins. These tools play a crucial role in various aspects of drug discovery and structural biology by providing insights into molecular interactions that govern biological processes (Agu et al., 2023; Forli et al., 2016; Sarkar et al., 2024; Trott and Olson, 2010).

1.12.3 Autodock

Autodock, developed by the Olson laboratory, employs a Lamarckian genetic algorithm (GA) coupled with an empirical free energy force field to perform molecular docking simulations (Morris et al., 1998). The Lamarckian GA approach involves iteratively optimizing ligand conformations based on the receptor's binding site, allowing flexibility in accommodating different receptor conformations and ligand orientations during docking simulations (Morris et al., 1998). Key features of Autodock include grid-based energy evaluations which assess various interactions between ligands and receptors, such as steric clashes, hydrogen bonds, electrostatic interactions, and desolvation effects (Goodsell et al., 1996). These calculations are essential for predicting binding poses and estimating binding affinities, crucial for rational drug design and virtual screening campaigns (Spassov, 2024).

1.12.4 Autodock Vina

Autodock Vina represents an advancement over Autodock, developed by Oleg Trott and Arthur J. Olson, aimed at enhancing efficiency and user-friendliness in molecular docking studies (Trott and Olson, 2010). It introduced a new scoring function and optimization algorithm that significantly improved the speed and accuracy of docking calculations. The scoring function in Autodock Vina strikes a balance between computational efficiency and accuracy, making it suitable for large-scale virtual screenings of compound libraries against molecular targets (Trott and Olson, 2010). By employing a hybrid search algorithm that combines global and local optimization techniques, Autodock Vina efficiently explores ligand conformational space and predicts optimal binding modes (Sarkar et al., 2024).

Both Autodock and Autodock Vina have revolutionized computational biology and drug discovery by enabling researchers to explore molecular interactions in detail. These tools have been extensively applied in pharmaceutical research to identify potential drug candidates, optimize lead compounds, and understand structure-activity relationships (SAR) (Morris et al., 1998). Their impact extends to various fields, including medicinal chemistry, structural biology, and biophysics, where they facilitate the elucidation of protein-ligand interactions crucial for therapeutic intervention. Researchers use these tools to predict binding energies, analyse ligand-protein complexes and guide experimental studies aimed at developing new treatments for diseases (Du et al., 2016). Autodock and Autodock Vina remain indispensable tools in computational biology, offering robust methodologies for studying molecular recognition and binding mechanisms (Forli et al., 2016). Their continuous development and application contribute significantly to advancing our understanding of biological processes and accelerating drug discovery efforts.

1.12.5 Homology Modelling

Homology modelling, frequently referred to as comparative modelling, represents a computational methodology employed to forecast the 3D conformation of a protein by utilizing its AA sequence as a basis and leverages the known structure of a homologous protein (Chothia and Lesk, 1986; Fuller et al., 2009; Waterhouse et al., 2018a). This approach is essential for filling the gap between well-characterized protein sequences and their corresponding theoretical structures. The process of homology modelling typically involves several key steps: template selection, sequence alignment, model generation, and model evaluation (Eswar et al., 2006).

1. Template Selection: An appropriate template protein with a known structure that exhibits similar evolutionary homology with the target protein is chosen. The quality and reliability of the resulting model heavily depends on the similarity between the target and template sequences.
2. Sequence Alignment: The AA target protein's sequence is compared and aligned with that of the template protein. This alignment is crucial as it dictates how the residues in the target protein correspond to those in the template, ensuring correct spatial arrangement in the model.
3. Model Generation: Using the aligned sequence, a 3D structure of the target protein is generated based on the spatial restraints provided by the template structure (Eswar et al., 2003). MODELLER, for example, uses a satisfaction of spatial restraints approach, incorporating bond distances and angles to construct the model (Webb and Sali, 2016).
4. Model Evaluation: The validity of the generated model is assessed using various validation methods. This includes checking stereochemical quality (e.g., Ramachandran plots), clash scores, and assessing the overall fit of the model to experimental data or known functional sites (Bagaria et al., 2012; Ravikumar et al., 2019) (Saravanan and Selvaraj, 2017).

Beyond the basic steps, homology modelling can incorporate advanced techniques such as loop refinement algorithms, which improve accuracy in flexible regions (Sali and Blundell, 1993). Integration of experimental data from techniques like nuclear magnetic resonance (NMR) or cryo-electron microscopy (cryo-EM) further enhances model reliability by refining local structures and conformations. Popular tools like MODELLER and SWISS-MODEL facilitate these processes. MODELLER utilizes iterative optimization to refine models, while SWISS-MODEL automates template selection and alignment, making it accessible for users without extensive modelling expertise (Waterhouse et al., 2018b; Webb and Sali, 2016).

Homology modelling faces several key limitations, including difficulties in accurately predicting regions with low sequence similarity between the target and template (Forrest et al., 2006). These regions, such as loops or disordered areas, often do not align well, resulting in structural inaccuracies. Additionally, homology modelling struggles with incorporating post-translational modifications (PTMs) like phosphorylation or glycosylation, which are essential for protein function (Krieger et al., 2003). Standard templates usually lack information on these modifications, making it harder to model their effects. Another challenge is the choice of template; selecting an appropriate template is crucial for generating accurate models, and when suitable templates are scarce, the model's reliability can decreases, particularly in regions where protein structure is sensitive to sequence changes. Despite these challenges, future directions in homology modelling are promising (Fiser, 2010). Machine learning, especially deep learning techniques such as AlphaFold, are expected to improve the accuracy of models, even in regions of low sequence similarity. Integrating personalized medicine with

homology modelling will also enhance disease-specific structural predictions, aiding in the development of individualized therapies (Chen et al., 2024; Nussinov et al., 2022). Thus future directions include integrating machine learning for improved accuracy and expanding applications in personalized medicine and structure-based drug design (Vyas et al., 2012)

Furthermore, homology modelling's role in structure-based drug design is expanding, allowing for the rational design of drugs targeting undruggable proteins and optimizing drug efficacy (Wu et al., 2023). Looking ahead, homology modelling will likely benefit from combining techniques such as protein-protein interaction (PPI) modelling, cryo-EM data integration, and dynamic simulations (Carter et al., 2019) . The integration of protein dynamics through methods like molecular dynamics (MD) will help capture the conformational flexibility of proteins, making models more reflective of biological function (Childers and Daggett, 2017). The future of homology modelling lies in creating more accurate and dynamic models that not only predict static structures but also provide insights into how proteins function and interact in their natural environments. These advancements, combined with machine learning and artificial intelligence (AI), will revolutionize drug discovery, allowing for the development of better-targeted therapies and personalized treatments, ultimately making homology modelling a more powerful tool in understanding complex biological systems (Liang et al., 2024; Qiu et al., 2024).

1.13 MD simulations

MD simulations are fundamental computational tools that replicate the movements and interactions of molecules within a system (Karplus and Petsko, 1990). These simulations are pivotal in elucidating the atomic-level characteristics and behaviours of biomolecules, particularly proteins, at various time points. In MD, every atom in a molecule undergoes continuous movement, governed by the principles of classical mechanics and statistical thermodynamics (Karplus and McCammon, 2002). Through the numerical solution of Newton's equations of motion, MD simulations yield detailed information about the dynamic behaviour of biomolecules under varying conditions such as temperature, pressure, and solvent environment (Adcock and McCammon, 2006; Hollingsworth and Dror, 2018).

MD simulations are versatile, offering predictive capabilities for understanding molecular responses to mutations, environmental changes or interactions with ligands (Hollingsworth and Dror, 2018). They play an important role in verifying and refining theoretical models of biomolecular structures generated through techniques like homology modelling or experimental methods (Vyas et al., 2012). Advanced MD techniques include enhanced sampling methods like replica exchange molecular dynamics (REMD) which improve simulation efficiency and capture rare events (Chen et al., 2024). These methods are particularly useful for studying conformational changes, protein folding dynamics,

and binding or unbinding processes. MD simulations can integrate experimental data from techniques such as X-ray crystallography, NMR spectroscopy or cryo-EM, enhancing accuracy and reliability in predicting molecular structures and interactions (Son et al., 2024). Future developments in MD aim to enhance realism by incorporating quantum mechanical effects and capturing longer timescales relevant to biological processes. Furthermore, advancements in computing power and algorithms continue to expand the scope of MD simulations in drug discovery, enzyme mechanisms, and personalized medicine (Borhani and Shaw, 2012).

MD simulations provide a powerful framework for studying biomolecular systems at atomic resolution, offering insights into their function, dynamics, and interactions. By bridging theoretical models with experimental observations, MD simulations contribute significantly to advancing our understanding of complex biological phenomena (Hollingsworth and Dror, 2018).

1.13.1 AMBER (Assisted Model Building and Energy Refinement)

AMBER encompasses a suite of integrated codes designed for establishing, executing, and analyzing MD simulations (Madhusudhan et al., 2005). This comprehensive software package facilitates three primary steps: system construction, simulation and analysis of simulation results (Case et al., 2005). The initial phase of using AMBER involves system construction, which is crucial for setting up molecular models for simulation. This process is supported by standalone modules within AMBER that handle tasks such as creating coordinate and topology files, essential for defining the molecular system's structure and parameters. Key preparatory programs include Antechamber and LEaP (Case et al., 2023; Case et al., 2005).

Antechamber is used alongside AMBER for parameterizing ligands using the General Amber Force Field (GAFF). It assigns atom types and parameters to ligands, enabling their integration into molecular simulations (Wang et al., 2004; Wibisono and Suhartanto, 2012) LEaP: This module constructs biopolymer structures from individual residues, integrates molecular systems, and generates a detailed list of force field terms and their corresponding specifications. LEaP is available in two versions: TLEap for command-line operations and xleap for interactive command-line editing, offering flexibility in system setup (Wibisono and Suhartanto, 2012). AMBER's suite of programs and force fields, complemented by advanced tools like Pmemd, Ptraj and MM/PBSA, forms a robust framework for conducting MD simulations and analysing complex biomolecular systems. The integration of GPU acceleration through Pmemd.CUDA further enhances computational efficiency, underscoring AMBER's pivotal role in computational biophysics and chemistry research (Madhusudhan et al., 2005).

Sander serves as the core MD program within AMBER, implemented in Fortran 90. It interprets user-defined variables using a label-value pair syntax, facilitating efficient MD simulations of biomolecular systems (Peramo, 2016). Pmemd is an optimized version of Sander, leveraging the message passing interface (MPI) for parallel execution across multiple processors. It excels in handling large-scale solvated systems over extended simulation periods, enhancing computational efficiency (Peramo, 2016). Pmemd.CUDA builds upon the capabilities of Pmemd since Pmemd.CUDA utilizes CUDA kernels tailored for GPU acceleration. This harnesses the high computational power and memory bandwidth of GPUs, making it particularly advantageous for intensive calculations in MD (Peramo, 2016; Salomon-Ferrer et al., 2013).

Ptraj is AMBER's primary trajectory analysis program, capable of processing and compiling trajectory files from simulations. It supports various analyses related to energies and molecular structures, making it indispensable for interpreting simulation results (Seabra et al., 2007; Watson et al., 2022). Cpptraj is written in C++ and complements Ptraj by providing additional functionalities to analyse trajectory files with varying topologies within a single job. Together, Ptraj and Cpptraj facilitate comprehensive post-simulation analysis (Seabra et al., 2007). MM/PBSA calculates binding free energies as a post-MD analysis tool, utilizing snapshots from simulation trajectories. It employs methods such as Generalized Born surface area (GBSA) calculations, implemented in MM/PBSA.PL (Perl) or MM/PBSA.PY (Python), to evaluate the energy differences between two states (Srinivasan et al., 1998).

AMBER incorporates several molecular mechanics force fields tailored for simulating biomolecules and experimental fragments (Madhusudhan et al., 2005). The force fields, such as ff10, encompass diverse nucleic acid and AA parameters and are compatible with explicit solvent models like the TIP3P water model (Vlachakis et al., 2014). AMBERTools, an extension of AMBER, hosts specialized simulation codes that cater to specific research needs: Sqm is an independent semiempirical quantum chemistry software originally part of Sander's (Quantum Mechanics/Molecular Mechanics) QM/MM capabilities (Case et al., 2023). Sqm is integral to Antechamber for calculating AM1-BCC charges and serves as a reference for quantum mechanics in simulations conducted by Sander (Xu et al., 2013). QM/MM Algorithms are algorithms that ensure energy conservation during lengthy simulations by combining quantum mechanical and molecular mechanical calculations (Senn and Thiel, 2009; Warshel and Levitt, 1976). They facilitate comprehensive structural modifications that enhance semiempirical performance while supporting parallel computing capabilities (Seabra et al., 2007). AMBER supports both explicit and implicit solvation models (Case et al., 2005).

Explicit Solvation utilizes Particle Mesh Ewald (PME) algorithms for handling long-range electrostatic interactions in QM/MM simulations, ensuring accuracy in explicit solvent environments (Seabra et al., 2007). Implicit Solvation employs AMBER's Generalized Born (GB) models to approximate solvent effects in QM/MM simulations, providing efficient and reliable solutions for computational studies (Seabra et al., 2007). AMBER's versatile suite of tools and algorithms makes it a cornerstone in computational biology and chemistry research (Madhusudhan et al., 2005).

1.13.2 Binding-free energy calculations

The Molecular Mechanics with Generalized Born and Surface Area solvation (MM/GBSA) method calculates the binding free energies of the receptor-ligand complexes. It evaluates the receptor-ligand interaction using basic scoring functions to classify the various binding poses of a ligand to determine its most stable binding orientation. It is also used to compare the energetic interactions of the ligands with the same receptor (Genheden and and Ryde, 2015; Homeyer and Gohlke, 2012; Massova and Kollman, 2000; Wang et al., 2017). Free energy binding can be split into entropic and enthalpic reactions which are calculated as a total of gas-phase molecular interaction energy of the ligand and protein (ΔE_{MM}), solvation free energy (ΔG_{solv}) and the configurational changes of entropy linked to the ligand binding ($-T\Delta S$) according to the equation (Deng and Roux, 2009) :

$$\Delta G_{bind} = \Delta E_{MM} + \Delta G_{solv} (-T\Delta S)$$

The ΔE_{MM} consists of the differences in the internal energies ΔE_{int}, which is consistent with dihedral, angle, bond, electrostatic ΔE_{elec} and van der Waals energies (ΔE_{vdW}) (Srinivasan et al., 1998).

$$\Delta E_{MM} = \Delta E_{int} + \Delta E_{ele} + \Delta E_{vdW}$$

The total electrostatic solvation energy is represented by ΔG_{solv} and a non-polar ΔG_{SA} is determined via the solvent accessible surface area ($\Delta G_{SA} = \gamma SASA$) between the solute and the solvent (Gilson et al., 1988; Gilson and Zhou, 2007; Sitkoff et al., 1994; Wang et al., 2019)

$$\Delta G_{solv} = \Delta G_{GB} + \Delta G_{SA}$$

1.14 Visualization programs

UCSF Chimera (Meng et al., 2006) is a versatile molecular visualization program renowned for its all-inclusive toolkit designed for analysing sequences and structures in tandem. It facilitates advanced investigations into biomolecular interactions and dynamics (Goddard et al., 2018; Pettersen et al., 2004). The Multi-align Viewer in UCSF Chimera enables simultaneous visualization of sequence alignments and their corresponding structures. This feature automatically links sequences with their

structures, providing insights into sequence adjustments and structural superimpositions (Meng et al., 2006; Pettersen et al., 2004). Utilizing the Match->Align extension, UCSF Chimera aligns structures based on sequence data, enhancing comparative structural analysis (Yang et al., 2012).

Biovia Discovery Studio is a powerful, commercial-grade software tailored for molecular modelling, visualization, and analysis of protein and small molecule data (Systèmes, 2021). It offers extensive capabilities for 3D visualization, enabling researchers to explore complex molecular structures with precision and clarity. One of the stand-out features of Biovia Discovery Studio is its ability to calculate relative binding free energies. This functionality is akin to performing competitive binding assays in silico, providing insights into molecular interactions and energetics (Cournia et al., 2017). The software incorporates refined force field parameters for accurate representation of bond angles and distances, crucial for simulating MD and interactions (Systèmes, 2021).

While UCSF Chimera excels in sequence-structure alignments and comparative modelling through its Multialign Viewer and alignment tools (Meng et al., 2006); Biovia Discovery Studio offers robust capabilities in 3D visualization and sophisticated energy calculations (Systèmes, 2021). Together, these programs cater to diverse needs in molecular modelling, from structural elucidation to dynamic simulations and interaction analyses. UCSF Chimera and Biovia Discovery Studio represent cutting-edge tools in the field of molecular visualization and modelling. Their functionalities empower researchers to explore complex biological systems, analyse molecular structures, and predict molecular interactions with high accuracy and efficiency (Cavasotto et al., 2018; Meng et al., 2006; Pettersen et al., 2004; Wang et al., 2017).

1.15 Aim

The aim of this study was to investigate the natural polymorphisms at *gag* CSs and analyse their potential impact on the substrate envelope structure in HIV-1 subtype C.

1.16 Objectives

To achieve this aim, the study objectives were:

- To determine the sequence variability at the five *gag* CSs in HIV-1 subtype C.
- To compare the AA sequences found at the CSs in HIV-1 subtype B and subtype C.
- To dock and determine the binding affinities of *gag* CSs to protease in HIV-1 subtype B and C.

- To use MD simulations to characterize the substrate envelope structure in HIV-1 subtype B and C, elucidating how natural polymorphisms at *gag* CSs may influence its conformation.

1.17 Project Rationale

The *gag* gene in HIV-1 encodes structural proteins that are essential for viral assembly, maturation, and infectivity (Balasubramaniam and Freed, 2011; Freed, 2015; Sundquist and Krausslich, 2012). These proteins are initially synthesized as a Gag polyprotein, which must be cleaved at specific sites by the viral PR to form mature, infectious virions (Freed, 2015). This proteolytic process is highly ordered and temporally regulated, involving multiple sequential cleavage events at defined junctions between Gag domains (e.g., MA/CA, CA/P2, P2/NC, NC/P1, and P1/P6) (Pettit et al., 2005). HIV-1 protease recognizes and cleaves a range of different AA sequences at these sites with high specificity, suggesting that recognition involves not just the linear sequence but also the 3D conformation and context of the CS (Könnyű et al., 2013; Prabu-Jeyabalan et al., 2002).

Despite substantial sequence variability across HIV-1 subtypes and even within individual patients—cleavage of Gag polyproteins remains efficient. This suggests that HIV-1 PR has evolved to accommodate a degree of structural flexibility, allowing it to maintain processing fidelity across diverse viral substrates (Berkhout, 1999). However, this plasticity also presents a challenge: naturally occurring polymorphisms within or near CSs can subtly alter the cleavage efficiency, modify substrate recognition, and influence the order of processing events. These changes can impact viral fitness and significantly affect the susceptibility to PIs (Dam, Quercia et al., 2009; Clavel and Mammano, 2010; Santos, Tebit et al., 2012).

Given that the emergence of drug resistance remains a major obstacle in antiretroviral therapy, a deeper understanding of how PR interacts with variable Gag sequences is critical. The substrate envelope theory describes the substrate's consensus volume (Prabu-Jeyabalan et al., 2002). It has been suggested that PR uses this to recognize and bind specific CSs (Chellappan et al., 2007; Nalam et al., 2010; Prabu-Jeyabalan et al., 2002; Shen et al., 2013).

This study investigated subtype-specific Gag CS variations, particularly between Subtypes B and C and how these variations influenced the substrate envelope and subsequent protease-ligand interactions. The findings may inform the design of next-generation inhibitors for the improved clinical management of HIV/AIDS.

2 Chapter 2:Variability at *gag* CS

2.1 Introduction

Chapter 2 delved into the variability at *gag* CSs by investigating the frequency of polymorphic sequences per CS. The CSs, which occur at the scissile bond, are essential regions within the Gag polyprotein that are crucial for proper PR function and subsequent viral maturation. To comprehensively analyse these regions, sequences flanked by 5AA, 10AA, and 15AA on either side of the CSs were examined. This approach allowed for an in-depth understanding of the immediate and surrounding sequence context, providing insights into how variations might influence cleavage efficiency and viral maturation.

HIV-1 exhibits significant genetic diversity, particularly at its Gag CSs, which are essential for viral replication, maturation, and fitness (Sundquist and Krausslich, 2012). While most studies have focused on subtype B, this dissertation examined the natural polymorphisms and variability at the Gag CSs in HIV-1 subtype C, the most dominant subtype globally. The study by de Oliveira *et al.* (2003) showed that there are significant variations in PR CSs both between different subtypes (inter-subtype) and within the same subtype (intra-subtype), particularly at the P2/NC site. However, their study compared only 5AA on either side of the CS and did not consider the effect of the variability (if any) of AA beyond this. Therefore, in this chapter, a comprehensive analysis of the CSs extending up to 15AA on either side of the scissile bond was conducted to provide a detailed view of variability across the extended CSs. The variability seen in subtype C was also compared to that observed in subtype B CSs.

The results from this chapter provided a foundational understanding of the natural variability at *gag* CSs in HIV-1 subtype C. This knowledge is crucial for the subsequent chapter, where the focus will shift to the structural and functional implications of these polymorphisms, particularly their effects on the substrate envelope structure.

2.2 Methods

2.2.1 Sequence data and CSs

A total of 1000 subtype B and 1000 subtype C *gag* sequences were downloaded from the Los Alamos database (LANL) (https://www.hiv.lanl.gov/). The sequences were selected based on the patient being naïve to PR inhibitor treatment. Multiple sequence alignments were done using the software tool ClustalW (Thompson et al., 1994) in Bioedit (Hall, 1999). Following successful alignment, the ElimDupes tool (https://www.hiv.lanl.gov/content/sequence/elimdupesv2/elimdupes.html) was used to remove duplicated sequences. The nucleotides and AA were numbered according to the HXB2

subtype B reference sequence. The *gag* CSs (MA/CA, CA/P2, P2/NC, NC/P1 and P1/P6) were identified and the percentage frequency was calculated for each residue. Additionally, the motif pattern in each sequence was counted to determine which sequence appeared more often and was possibly the recognition pattern that PR identified for cleavage.

2.2.2 Statistical test

This study employed the chi-squared test for independence to compare mutation frequencies at *gag* CSs between HIV-1 subtype B and subtype C. Degrees of freedom were set to 1, which aligned with the comparison of two categorical variables (subtypes B and C), with a significance level of 0.05.

2.3 Results

Figure 3 showed that most sequences (97.5%) originated from African countries, with South Africa accounting for the largest percentage (61.5%), followed by Zambia (31.7%) and Tanzania (4.3%). The remaining 2.5% of sequences originated from non-African countries, with India and Canada having the highest percentage (0.7%), followed by, Cyprus, Spain, Argentina and China, each accounting for less than 1% of the total sequences.

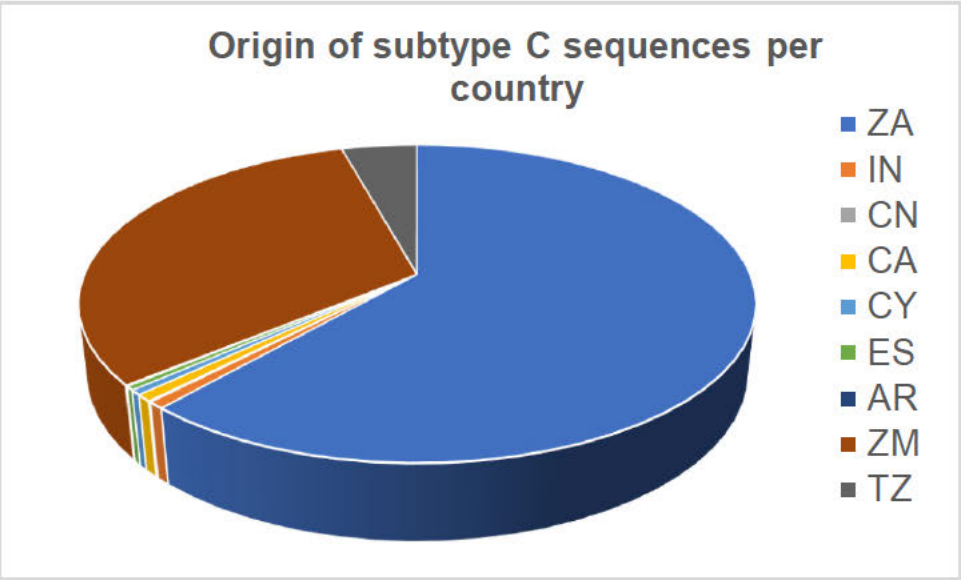


Figure 3: : Subtype C sequence origin per country.

The majority (61.5%) of the sequences were isolated from South Africa(ZA) and 31.7% were from Zambia(ZM). The last 2.5% belonged to Argentina (AR), Canada (CA), China(CN), Cyprus(CY), India(IN), Spain(ES) and Tanzania(TZ).

Table 2: Frequency of the subtype C CSs commonly occurring sequences at 5AA, 10AA and 15AA per CS

AA	CS	Commonly occurring sequence	Frequency of commonly occurring sequence	Frequency of polymorphic sequences
5AA	MA/CA	VSQNY/PIVQN	918	82
10AA		SKQQQVSQNY/PIVQNLQGQM	571	429
15AA		EEQNVSKQQQVSQNY/PIVQNLQGQMVHQAI	239	761
5AA	CA/P2	KARVL/AEAMS	966	34
10AA		GGPGHKARVL/AEAMSQA~NN	123	877
15AA		ACQGVGGPGHKARVL/AEAMSQA~NNANIMM	40	960
5AA	P2NC	NANIMMQRSN	48	952
10AA		SQA~NNANIMMQRSNFKGSK	9	991
15AA		LAEAMSQA~NNANIMMQRSNFKGPKRIVKC	5	995
5AA	NC/P1	ERQAN/FLGKI	848	152

10AA		MKDCTERQAN/FLGKIWPSHK	568	432
15AA		KEGHQMKDCTERQAN/FLGKIWPSHKGRPGN	502	498
5AA	P1P6	RPGNF/LQSRP	480	520
10AA		PSHKGRPGNF/LQSRPEPTAP	281	719
15AA		LGKIWPSHKGRPGNF/LQSRPEPTAPPAESF	234	766

Table 3: Variability at *gag* CS in HIV-1 subtype C

CS	CS Sequence											Frequency
Position	**128**	**129**	**130**	**131**	**132**	**/**	**133**	**134**	**135**	**136**	**137**	
MA/CA	**V**	**S**	**Q**	**N**	**Y**	**/**	**P**	**I**	**V**	**Q**	**N**	918
	.	.	.	.	F	/	.	.	.	.	.	19
	I	.	.	.	.	/	.	.	.	.	.	9
	.	.	H	.	.	/	.	.	.	.	.	6
	.	.	.	.	.	/	V	V	.	.	.	6
	A	.	.	.	.	/	.	.	.	.	.	5
	.	.	K	.	.	/	.	.	.	.	.	4
	T	.	.	.	.	/	.	.	.	.	.	3
	.	.	L	.	.	/	.	.	.	.	.	3
	.	.	.	.	.	/	.	.	I	.	.	3
	I	.	.	.	.	/	.	V	.	.	.	2
	.	.	.	.	H	/	.	.	.	.	.	2
	.	.	.	.	.	/	.	.	.	T	.	2
	.	.	.	.	.	/	.	.	.	R	D	2
Position	**359**	**360**	**361**	**362**	**363**	**/**	**364**	**365**	**366**	**367**	**368**	
CA/P2	**K**	**A**	**R**	**V**	**L**	**/**	**A**	**E**	**A**	**M**	**S**	966
	.	.	.	I	.	/	.	.	.	.	.	10

	.	.	.	.	.	/	.	.	.	.	N	4
	.	.	.	.	.	/	.	.	.	.	Q	3
	.	.	K	.	.	/	.	.	.	.	.	3
	.	.	.	.	.	/	.	.	.	.	I	2
	.	.	.	.	.	/	.	G	.	.	.	2
	E	.	.	.	.	/	.	.	.	.	.	2
Position	**373**	**374**	**375**	**376**	**377**	**/**	**378**	**379**	**380**	**381**	**382**	
P2/NC	**N**	**A**	**N**	**I**	**M**	**/**	**M**	**Q**	**R**	**S**	**N**	48
	.	T	.	.	.	/	.	.	K	.	.	27
	.		.	.	.	/	.	.	.	G	.	24
	.	S	.	.	.	/	.	.	K	.	.	23
	S	.	.	.	.	/	.	.	.	.	.	22
	.	T	.	.	.	/	.	.	.	.	.	21
	.	T	.	.	.	/	.	.	.	G	.	15
	.	S	.	.	L	/	.	.	.	.	.	15
	.	.	.	.	.	/	.	.	K	.	.	14
	S	T	.	.	.		.	.	.	G	N	14
	.	V	.	.	.	/	.	.	.	.	.	13
	.	I	.	.	.	/	.	.	K	.	.	13
	S	T	.	.	.	/	.	.	.	S	.	13
	.	T	.	.	.	/	.	.	.	N	.	13
	S	T		.	L	/	.	.	.	.	.	11
	.	N	.	.	.	/	.	.	.	.	.	11
	S	.	.	.	.	/	.	.	.	G	.	11
	.	G	.	.	.	/	.	.	.	.	.	10
	.	S	.	.	.	/	.	.	.	G	.	10
	S	.	T	.	.	/	.	.	.	G	.	9
	.	.	.	.	.	/	.	.	.	N	.	8
	.	M	.	.	.	/	.	.	.	.	.	8
	S	.	.	.	.	/	.	.	K	.	.	8
	S	V	.	.	.	/	.	.	.	.	.	8
	S	N	.	.	.	/	.	.	.	.	.	7
	.	.	.	.	.	/	I	.	.	.	.	7
				V	.	/	.	.	.	.	.	6
	.	.	T	.	.	/	.	.	.	G	.	6

	.	G	.	.	.	/	.	.	.	.	.	6
	.	I	.	I	.	/	.	.	.	G	.	6
	.	I	.	I	.	/	.	.	.	.	.	6
	.	V	.	.	.	/	.	.	.	G	.	6
	.	V	.	.	.	/	.	.	.	.	.	6
	S	T	.	.	L	/	.	.	.	G	.	6
	S	T	.	.	L	/	.	.	.	.	.	6
	S	V	.	.	.	/	.	.	.	G	.	6
	S	G	T	.	.	/	.	.	.	.	.	6
	S	G	T	.	.	/	.	.	.	.	.	6
Position	**428**	**429**	**430**	**431**	**432**	**/**	**433**	**434**	**435**	**436**	**437**	
NC/P1	**E**	**R**	**Q**	**A**	**N**	**/**	**F**	**L**	**G**	**K**	**I**	848
	.	.	.	.	.	/	.	.	.	R		48
	.	.	.	.	.	/	.	.	.	.	L	24
	.	.	.	.	.	/	.	.	.	.	V	14
	.	.	.	.	.	/	.	.	.	R	L	10
	.	.	.	.	.	/	.	.	.	K	F	8
	G	.	.	.	.	/	.	.	.	.	.	7
	.	.	.	.	.	/	.	.	.	R	F	6
	.	.	.	.	.	/	.	.	.	N	I	5
	.	K	.	.	.	/	.	.	.	R	.	3
	.	.	R	.	.	/	.	.	.	.	.	3
	D	.	.	.	.	/	.	.	.	.	.	2
	.	K	.	.	.	/	.	.	.	.	.	2
	.	.	.	.	.	/	.	.	.	.	A	2
	.	.	.	.	.	/	.	.	.	G		2
	.	.	.	.	.	/	.	.	.	.	M	2
	.	.	.	.	.	/	.	.	.	.	N	2
	.	.	.	V	.	/	.	.	.	.	.	2
	G	.	.	.	.	/	.	S	.	.	.	2
Position	**444**	**445**	**446**	**447**	**448**	**/**	**449**	**450**	**451**	**452**	**453**	
P1/P6	**R**	**P**	**G**	**N**	**F**	**/**	**L**	**Q**	**S**	**R**	**P**	480
	.	.	.	.	.	/	.	.	N	.	.	349
	.	.	.	.	.	/	.	.	N	.	L	21

	.	.	.	.	.	/	P	.	.	.	L	18
	.	.	.	.	.	/	I	.	.	.	.	16
	.	.	.	.	.	/	.	.	.	.	L	11
	.	.	.	.	.	/	I	.	N	.	L	9
	.	.	.	.	.	/	.	.	.	N	A	9
	.	.	.	.	.	/	.	.	E	P	.	5
	.	.	.	.	.	/	.	.	.		A	5
	.	.	.	.	.	/	I	.	.	.	.	4
	.	.	.		.	/	.	.	R	.	.	4
	.	.	.	.	.	/	P	.		.	.	4
	.	.	.	.	.	/	.	.	N	.	V	3
	.	.	.	.	.	/	.	.	.	.	I	3
	G					/						2
	.	.	.	.	.	/	F		N	.	L	2
	.	.	.	.	.	/	H	.	.	.	L	2
	.	.	.	.	.	/	I	.	.	.	I	2
	.	.	.	.	.	/	.	.	G	.	.	2
	.	.	.	.	.	/	.	.	N	.	I	2
	.	.	.	.	.	/	.	.	N	.	S	2
	.	.	.	.	.	/	M	.	.	.	L	2
	.	.	.	.	.	/	P	.	.	.	T	2

When the variability at subtype C CSs were compared to subtype B (Figure 4), overall, subtype B appeared to be more variable overall across the CS, especially in MA/CA and NC/P1 where the differences in frequency are more pronounced (although only significantly different for the 10AA comparison for MA/CA (p=0.005). The P2/NC CS showed the highest variability in both subtypes, with variability increasing as sequence length extended from 5AA to 15AA. Notably, the CA/P2 site displayed a stark contrast between the two subtypes, with subtype B showing significantly higher variability at 5AA (p=0.007) , while subtype C showed significantly higher variability at 10AA (p=0.013). This trend continued for 15AA, although it was not significant (p =0.116).

The P1/P6 CS also demonstrated notable differences, where subtype C showed a slightly higher frequency than subtype B at 5AA, however this switched as the sequence length increased to 10AA and 15AA, with subtype B then showing slightly higher variability.

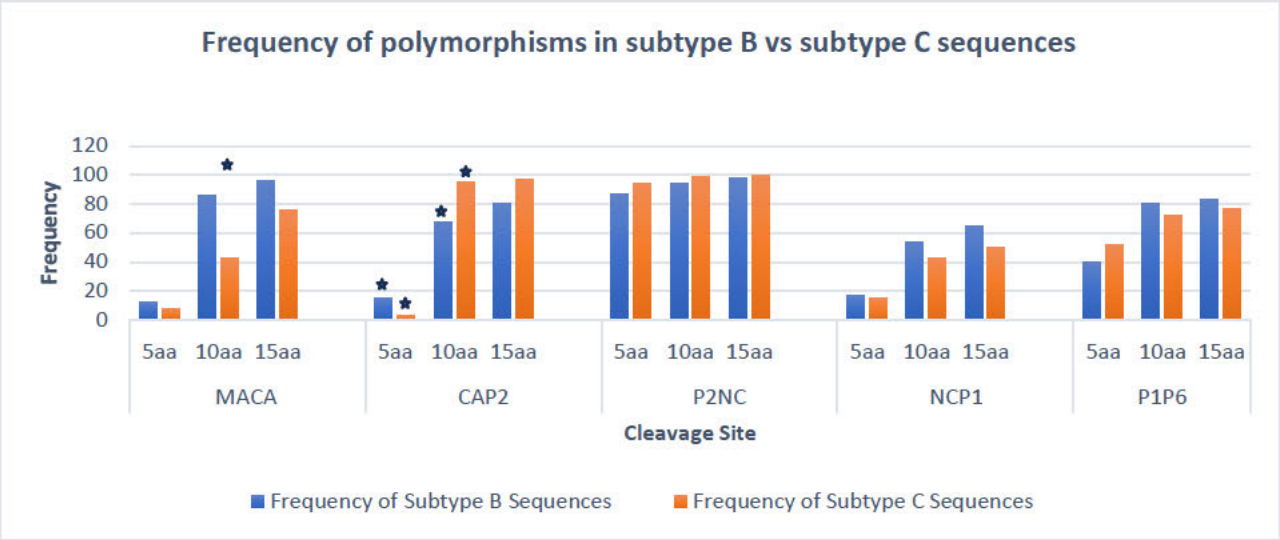


Figure 4: Polymorphic sequences per CS in subtype B versus subtype C.

The CS occurs at the scissile bond and is flanked by 5,10,15 AA on either side. Amino acids= aa. (Bars marked with * indicate significance of P<0.05)

Taking a cut-off of 1% variability for categorising as a polymorphism, as shown in figure 5, Y132F in subtype B was 3.7% vs 2.1 % in subtype C. The Q130H mutation occurred in 2.7% of the subtype B sequences, while it was at a very low frequency in subtype C. Multiple variants at codon V128 were seen at low levels for both subtypes, with subtype B favouring the V128A/G variants and subtype C the V128I variant (1.3%).

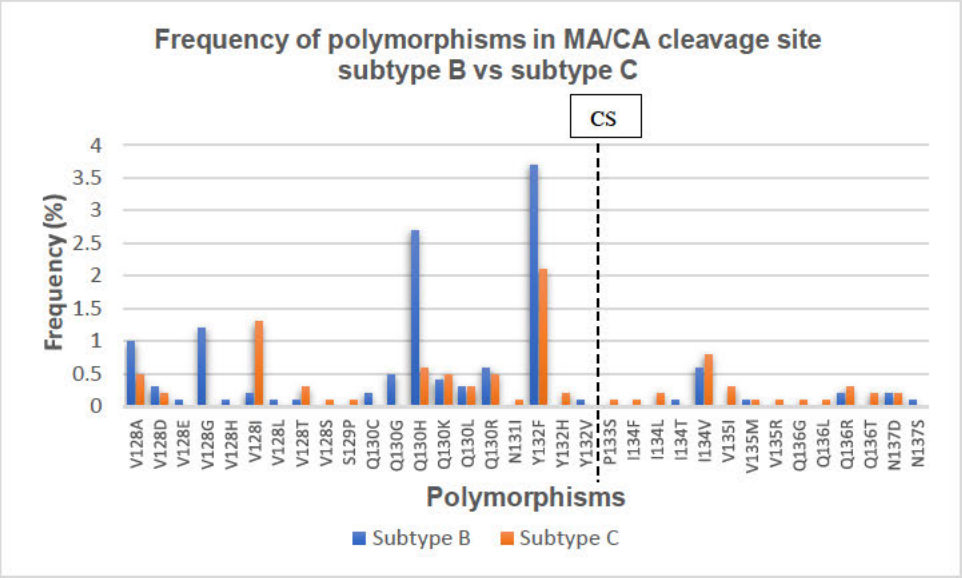


Figure 5: Mutations at the MA/CA *gag* CS in subtype B vs subtype C

In figure 6, HIV subtype B showed a high frequency of the V362I (13.5%) mutation while this was only seen in 1% of subtype C sequences.

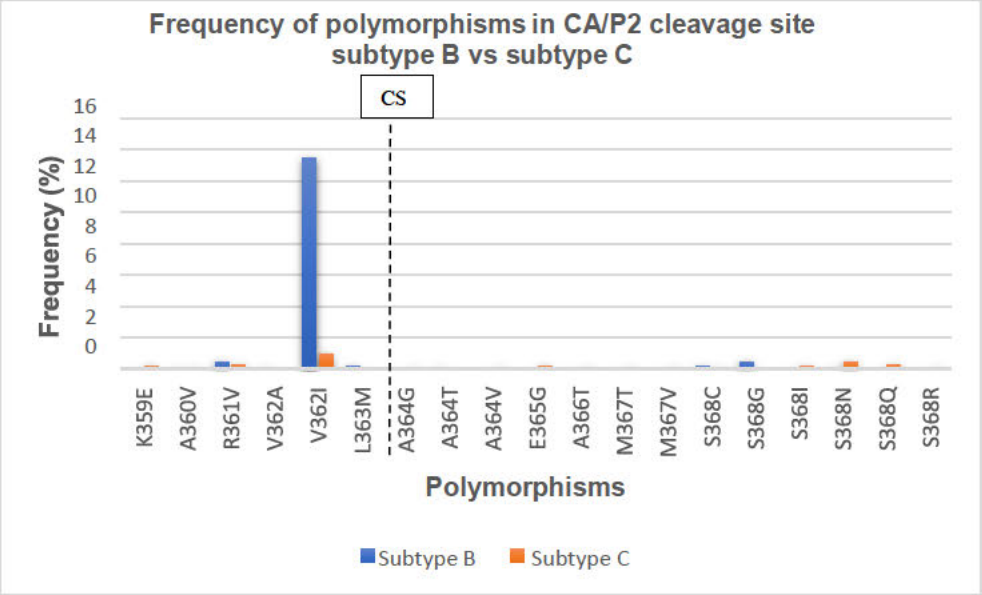


Figure 6: Mutations at the CA/P2 *gag* CS in subtype B vs subtype C

In figure 7, HIV subtype B: R380K (38.1%; p=0.0092), S373P (22.4%; p=2E-6), G381G (26.8%; p=2E-07), A374T (18.5%; p=2E-05), and T375A (11.1%; p=0.009) are the most frequent mutations. In subtype C, the highest occurrences are A373T (26.6%; p=2E-07), N375T (23.1%; p=1E-06), R380K (18.4%; p=0.009) and G381N (14.4%; p=0.0176). Some mutations, such as S373P, G381G, and T375A, are exclusive to subtype B, while A373T, A373N(13.2%;p=0.0003), and N375T are unique to subtype C. I376V (8.3% in B, 8.6% in C) occurs at similar frequencies in both subtypes.

Several mutations, including R380G, N382K, and A374N (1.4%; p=0.0354), show moderate presence in both subtypes but are more dominant in one. R380G (2.7% in B) and N382K (2.4% in B) are more dominant in subtype B, whereas A374N (7.8% in B vs. 1.4% in C) were higher in subtype B. Conversely, G381N (4.2% in B vs. 14.4% in C; p=0.0176) is more dominant in subtype C. Other statistically significant mutations include: S373T (p=0.046), A373G (p=0.0177), A373I (p=0.0061), A373S (p=0.005), A373V (p=0.0106), A374P (p=0.0076), T375N (p=0.008), T375S (p=0.0137), N375S (p=0.0356) and M377L (p=0.007).

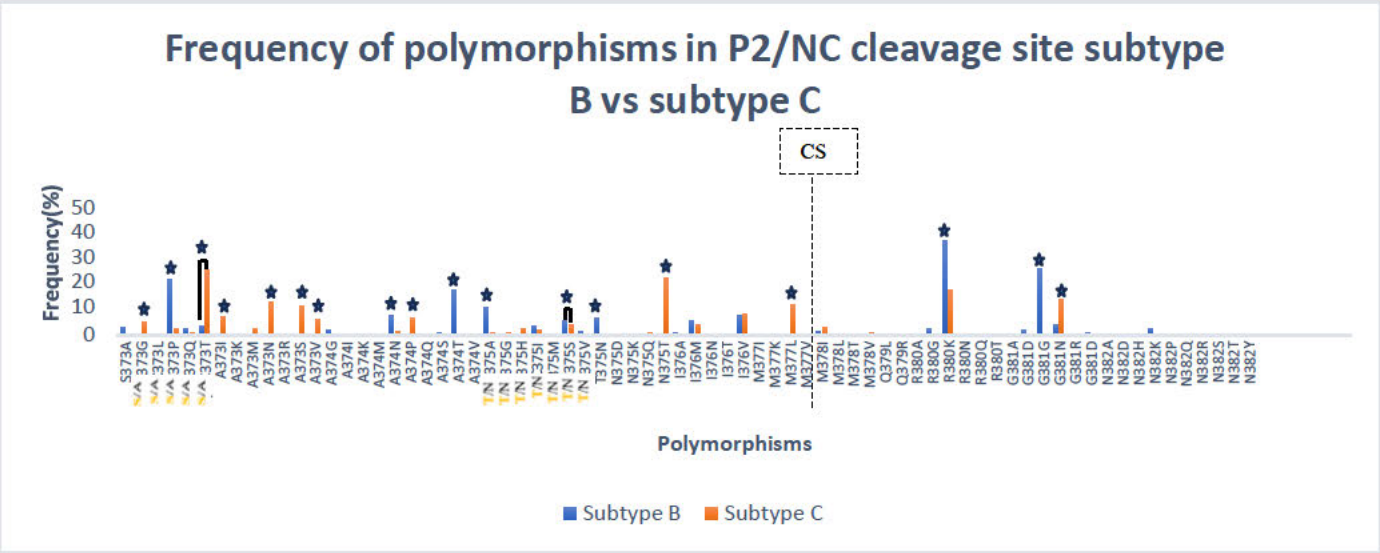


Figure 7: Mutations at the P2/NC *gag* CS in subtype B vs subtype C

Subtype B WT AA are indicated in orange (Bars marked with ⋆ indicate significance of P<0.05)

As shown in figure 8, E428G was higher in subtype B (1.4%) vs in subtype C (1%). R429G (2.5%) was only found in subtype B, and R429K was 3.1% in subtype B. For K436R subtype B showed a frequency of 4.5% vs 6.8% in subtype C. I437L showed a frequency of 4.2% in subtype B vs 3.5% in subtype C. I437V was found at 1.5% in subtype C. I437F was only seen in subtype C (1.5%).

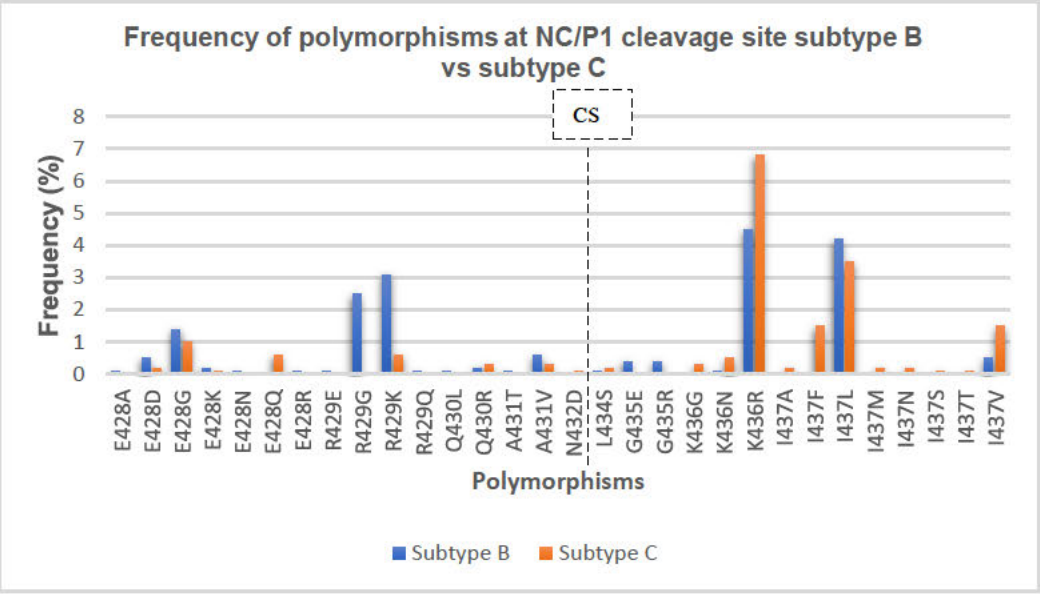


Figure 8: Mutations at the NC/P1 *gag* CS in subtype B vs subtype C

In figure 9, the dataset presents mutation prevalence in HIV P1/P6 CS subtypes B and C, highlighting distinct differences between the two. Subtype B shows a diverse range of mutations with notable prevalence, including R451P (15.2%), P452E (16.4%), and P453L (7.5%). In contrast, subtype C primarily shows no significant mutations, except for S451N, which has a striking prevalence of 41.8%.

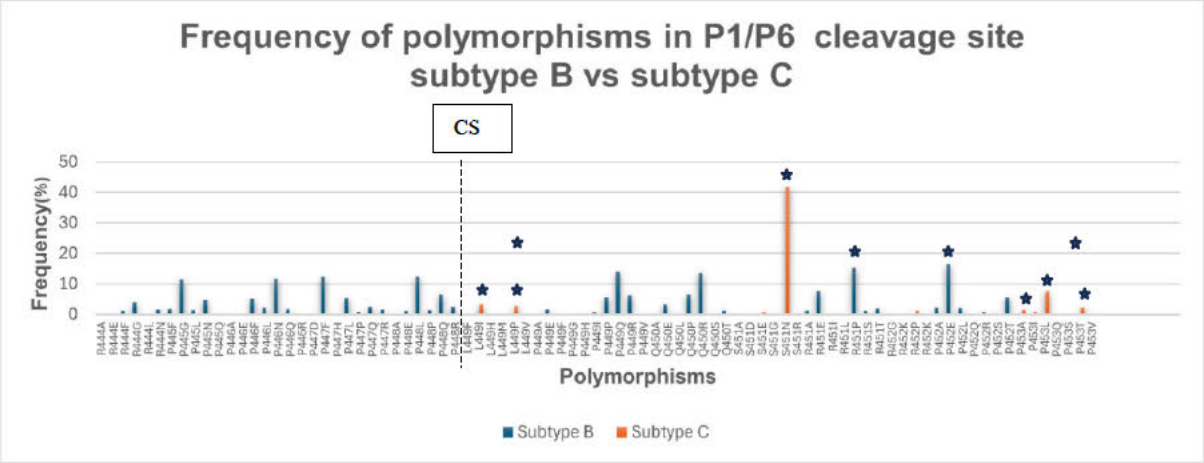


Figure 9: Mutations at the P1/P6 *gag* CS in subtype B vs subtype C

(Bars marked with * indicate significance of P<0.05)

2.4 Discussion

There have been limited studies investigating the natural variation within the *gag* gene of HIV-1 subtype C, particularly at the CSs (Oliveira et al., 2003). Most existing research has focused on subtype B (Barrie et al., 1996; Fun et al., 2012). This study extended prior findings by comparing the natural variation at the CSs between HIV-1 subtypes B and C, highlighting differences that may contribute to viral fitness, PR-substrate interactions and resistance to PI's. Importantly, while previous research typically focused on sequences spanning 5 AA on either side of the scissile bond, this study extended the analysis to 15 AA on either side. This extension allowed for a more comprehensive understanding of how natural polymorphisms influence the substrate envelope, particularly regarding PR cleavage and viral maturation. The findings revealed distinct patterns of CS variability between subtypes B and C. Overall, when the number of polymorphisms appearing at different AA positions was calculated, subtype C exhibited greater variability, particularly at the P2/NC CS, which is consistent with previous reports (Fun et al., 2012; Oliveira et al., 2003; Torrecilla et al., 2014).

However, when the frequency of mutations at a particular site was calculated, overall, subtype B showed a few highly dominant mutations but with fewer overall sites affected.

Mutations in the MA region have been linked to changes in viral assembly, efficiency, budding, as well as the virus' ability to evade host immune responses (Schweighardt et al., 2010). In particular, mutations at the CS have been shown to directly affect the interaction of the substrate with PI's (Fun et al., 2012). In this study, the most frequent mutations observed at the CS were V128A/G which occurred in 1.1% and 1.2% of subtype B sequences, V128I which occurred in 1.3% of subtype C sequences, Q130H in 2.7% of subtype B sequences and Y132F found in 3.7% and 2.1% in subtype B and C respectively. In the study by Hunter *et al*. (2022), the Y132F mutation was noted as one of the mutations that remained after treatment interruption as part of an analytical antiretroviral treatment interruption (ATI) study. In another study on PI failures, V128I was found in 10.42% of patients and Y132F, at the P1 position in the CS, was found in 11.46% of patients (Alencar et al., 2024).

Overall, in this study, the MA/CA sequence was very conserved which could be linked to the critical role of P1-P1' positioning in PR-substrate interactions as mutations at this site could potentially affect drug binding and the development of drug resistance (Özen et al., 2011; Özen et al., 2014). The V128G mutation has been associated with diminished cytotoxic T cell response and is known as an escape mutation (Carlson et al., 2012). Batonick *et al.* (2005) showed that Y132 and V135 near the MA–CA junction are key for Gag–clathrin-associated adaptor complex AP-2 binding, which is important in the later stages of the viral cycle. The Gag precursor is cleaved between Y132 and P133 during budding, showing this interaction isn't involved in viral entry. The Y132F MA mutation impairs HIV infection in non-dividing cells by disrupting PIC formation and nuclear import, despite allowing viral DNA circle formation. Serine phosphorylation may partially compensate by aiding MA transfer to the viral core (Gallay et al., 1995; Goldfarb, 1995).

The CA/P2 CS is the last of the *gag* CS to be cleaved during the maturation process and mutations at positions within this CS have been reported to cause immature and non-infectious virions (Pettit et al., 2002). In this study, the V362I mutation at position P3 of the CA/P2 CS showed a notably higher frequency in subtype C (13.5%) compared to subtype B (1%). A study by Daugherty *et al.* (2011) demonstrated that differences in capsid mutations among subtypes could result in varied susceptibilities to host restriction factors. Margot *et al.* (2010) demonstrated for the first time that V362I was a major mutation conferring resistance to bevirimat and Dicker *et al.* (2019), reported that V362I, and A364V were selected in vitro against GSK3532795, a novel maturation inhibitor. However, V362I required secondary substitutions to reduce susceptibility, some within the Cyclophilin A binding domain. Ross *et al*.(2021), suggested that altered interactions with Cyclophilin

A could affect viral replication efficiency. V362I has also been shown to be selected by WT viruses *in vivo*, but more often by viruses with mutated PRs (Verheyen et al., 2010). The mutation has also been associated with faster CA/P2 cleavage and greater PR efficiency, suggesting a mechanism for the development of drug resistance (Fun et al., 2011). Interestingly, the P1' to P5' AA sequence (P2) was more conserved with variants in this region having a frequency of less than 1%.

The P2/NC CS showed the most variability out of all the CSs and this is aligned with earlier research that has reported that mutations occur at this site in both drug naïve and drug-experienced patients (Ghosn et al., 2010; Teto et al., 2017). In this study, the highest level of conservation was seen at the P1 and P1' positions (MM). The greatest variation was seen at the P4 and P5 positions followed by the P3 and P4' positions. As the number of AA in the analysis increased, more variability was seen at the C terminal region of the CS. When comparing the two subtypes, the signature consensus sequence in subtype B was SATIM/MQRGN while subtype C was NANIM/MQRSN, with greater variability seen in subtype C. Studies have shown that mutations in the P2 region can significantly alter the kinetics of polyprotein cleavage by HIV PR, which is critical for viral maturation (Fun et al., 2012). Taken together, this suggests that higher variability in subtype C could drive preferential binding of the natural substrate over the inhibitor (Muzammil et al., 2003; Velazquez-Campoy et al., 2003; Wensing et al., 2014). The S373P mutation has been associated with a lower virological response to saquinavir/ritonavir (Malet et al., 2007; Verheyen et al., 2010), whereas A374S/P and T375A are more commonly found in individuals who have experienced prior PI treatment (Prado et al., 2002).

The NC/P1 CS showed a similar pattern, with greater variability seen in subtype C. Interestingly, the variability was mainly seen on the P1 end (primed) of the CS, with NC more conserved. Several studies have linked NC mutations to altered interactions with viral RNA and the *gag*-pol polyprotein, which may affect the efficiency of viral assembly (Ott et al., 2009). P1 has been reported to play an important role in the processing of the Gag polyprotein and mutations like K436R, which were more frequent in subtype C (6.8%) vs subtype B (4.5%), have been linked to PI resistance (Fun et al., 2011).

Finally, at the P1/P6 CS, P1 showed minimal polymorphisms in both subtypes. In contrast, the P6 region revealed high variability in both subtypes, with subtype B showing slightly higher overall frequencies of polymorphisms. Several mutations show marked variations between subtypes B and C. S451N is present at 41.8% of subtype C sequences, while it was absent in subtype B. P447F, P448L, P450R, and P453E all have high frequencies in subtype B (12.3%, 12.3%, 13.7%, and 20.9%, respectively) but are absent in subtype C. These variations could indicate key functional differences between the subtypes. The P6 domain is involved in the recruitment of the ESCRT machinery for viral budding (Bieniasz, 2006; Morita and Sundquist, 2004; Votteler and Sundquist, 2013).

Mutations in the Gag CS, particularly at P1/P6, have been linked to drug resistance or the restoration of replicative capacity (RC), both in the absence (Nijhuis et al., 2007) and presence (Bally et al., 2000; Côté et al., 2001; Doyon et al., 1996) of PI resistance-associated mutations.
Mutations in this region, such as L449P, have been associated with increased viral release efficiency and resistance to host restriction factors (Stern A, 2016). The study by Myint *et al*. (2004) found that L449F improved replication of viruses with severely impaired PR activity (D30N, N88D, L90M mutations). S451N has previously been linked to increased PI exposure in non-B clade HIV-1 subtypes (Liu et al., 2014; Sutherland et al., 2015). A study by Climaco-Arvizu (2022) found that the P6 protein exhibited the highest degree of variation among Gag proteins (which this study agrees with), with the P453L/T mutation being the most frequently observed. Notably, mutations at the P453 residue were associated with increased resistance to PIs (Kolli et al., 2009; Maguire Michael et al., 2002). The presence of known resistance-associated mutations in the drug-naïve subtype C population is a cause for concern.

In general, when the length of the sequence alongside the scissile bond was increased, the frequency of polymorphisms increased. The most dramatic effect was seen at the CA/P2 CS which changed from 3,5% to 96%. For the MA/CA CS, the frequency of polymorphic sequences increased from 8% for the 5 AA sequence to 76% for the 15 AA sequence. The P2/NC CS maintained variability at around 95% irrespective of the increase to 15AA in the analysis. Where the CS contained a spacer peptide (P1 or P2), this resulted in an overlap of the recognition sequences. This overlap may indicate functional redundancy or flexibility in protease recognition, potentially allowing the virus to tolerate a wider range of sequence variation without loss of cleavage efficiency (Fehér et al., 2002). Interestingly, in these instances, the frequency at 15AA changed to that observed for the 5 AA length when the spacer peptide was on the amino end of the CS. For example, at P2/NC the variability for the 5AA length was 95% and at CA/P2, the 15AA length increased to 96%. Similarly, at P1/P6 the variability for the 5AA length was 52% and at NC/P1 the 15AA length increased to 50%. It is unclear whether the amino end of the CS therefore determines the limit of variability that can be tolerated by the CSs and warrants further investigation.

Overall, subtype C exhibited higher variability at the *gag* CS compared to subtype B and this suggested that subtype C may have a more diverse and mutable PR cleavage profile, potentially contributing to differences in viral evolution, adaptability, and resistance patterns by altering enzyme-substrate interactions. The further investigation of enzyme-substrate interactions could identify residues that could be targeted for the design of next-generation inhibitors that exhibit reduced sensitivity to the development of drug resistance.

In conclusion, the comparative analysis of *gag* CS mutations highlights the differences between HIV-1 subtypes B and C and understanding these differences is crucial for developing effective treatment regimens and informing public health strategies to combat HIV-1 globally. Future research should continue to monitor mutation patterns and their impact on viral fitness and drug resistance, ensuring that therapeutic interventions remain robust against the evolving landscape of HIV-1 diversity.

3 Chapter 3: The effects of CS variability on the substrate envelope structure

3.1 Introduction

Chapter 2 analysed the variability at the *gag* CS in HIV-1 subtypes B and C. With this knowledge in hand, chapter 3, focused on understanding how these polymorphisms might influence the interaction between viral PR and the various *gag* substrates, since alterations in these positions would likely have profound effects on enzyme-substrate interactions (Chellappan et al., 2007; Leidner et al., 2021; Prabu-Jeyabalan et al., 2002; Ragland et al., 2017; Zephyr et al., 2021).

Prabu-Jeyabalan *et al*.(2000) proposed the idea of a substrate envelope, describing the shared 3D space occupied by natural substrates within the active site of a target, thereby highlighting the key interaction region essential for binding (Prabu-Jeyabalan et al., 2000). It was shown that the substrate envelope could be used to design robust inhibitors against rapidly mutating targets like HIV-1 PR (Nalam et al., 2013; Özen et al., 2011; Prabu-Jeyabalan et al., 2002; Shen et al., 2013). The inhibitors that conformed to the substrate envelope were less prone to resistance mutations (Nalam et al., 2010; Ragland et al., 2014) since resistance mutations often arose in regions where inhibitors extended beyond the substrate envelope (Prabu-Jeyabalan et al., 2000; Spielvogel et al., 2023). This concept has also been generalized to other systems where resistance is driven by evolutionary pressure, integrating structural biology and drug design to predict and circumvent resistance effectively (Ghosh et al., 2016; Weber and Agniswamy, 2009).

The natural variability in amino AA sequences at the CS could potentially alter the substrate envelope and consequently, susceptibility to PIs (Clavel and Mammano, 2010; Fun et al., 2012; Marie and Gordon, 2019; Nalam and Schiffer, 2008; Wensing et al., 2017). For example, HIV-1 subtype C, which predominates in regions like sub-Saharan Africa, has been shown to harbour mutations substitutions at PR CSs that enhance viral fitness under immune or drug pressure (Arts and Hazuda, 2012; Sankaran et al., 2024; Sutherland et al., 2016; Venkatachalam et al., 2023). Considering these differences are essential for designing inhibitors that are effective across subtypes, as relying solely on subtype B based models may not account for the structural and functional nuances of other subtypes (Isaacs et al., 2020; Poon et al., 2019; Sanches et al., 2007). Integrating subtype-specific data into drug design, such as ensuring inhibitors fit the substrate envelopes of diverse subtypes, helps create therapies that are broadly effective and less prone to resistance in diverse global populations (Matthew et al., 2021; Nalam and Schiffer, 2008).

To investigate the impact of polymorphisms on substrate-envelope interactions within HIV-1 subtype C PR, MD simulations were performed on all CSs and their variants in subtype C including 5 AA flanking the scissile bond. The structure of the common sequence for each CS was then compared to the structure of the common sequence found in subtype B. By examining patterns of polymorphisms within both PR and substrate sequences across diverse HIV-1 subtype C isolates, we elucidated the potential consequences of genetic variation for the substrate envelope structure and function. We identified polymorphisms that induced conformational changes that affected substrate-binding affinity, thereby altering the stability and effectiveness of enzyme-substrate interactions.

The peptide sequences selected for molecular modelling in this chapter were based on the initial statistical analysis of CS variability presented in Chapter 2. However, following further data refinement and recalculation of sequence frequencies, updated results revealed that several of the originally identified sequences were no longer the most statistically prevalent variants. As a result, some of the peptides modelled in this chapter differ from the corrected dominant sequences. Despite these discrepancies, the modelled peptide–protease complexes remain structurally valid and informative. They offer valuable insight into protease–substrate interactions, residue-specific binding behaviour, and structural determinants of cleavage, particularly within HIV-1 subtype C.

3.2 Methods

3.2.1 Modelling the *gag* CS

The MODELLER v10.1 (Sali, 1989-2022) homology modelling software was used to generate the *gag* CS structures. The modelling steps to construct suitable ligand structures were as follows: template selection, alignment of the sequence of interest to the template sequence, model creation and evaluation of the experimental model. The following scripts were run: alignment.py→ build_profile.py →model.py→evaluate_template.py→evaluate_model.py→plot_profiles.py (see Appendix).

3.2.2 Preparation of the structures for MD simulations

Following modelling, the structures were uploaded to the PDB2PQR online web server (http://nbcr-222.ucsd.edu/pdb2pqr_2.1.1/) where hydrogens were added to them. Briefly, the newly modelled theoretical CS structure's PDB file was uploaded onto the server. The Amber force field and output naming scheme selected was 'Amber'. New atoms were not built too close to selected atoms and the hydrogen bonding network was optimized. PROPKA was selected to assign protonation states at a pH of 7 for the pKa options. Thereafter the job was run and the output files were downloaded. The newly added hydrogens were checked in UCSF chimera to see that they were correctly added using:

‘tools’→ , ‘structure’.→ ‘analyses.→ , ‘FindH-bond’→ The file was saved in mol2 format. For the PR, the atomic coordinates of HIV-1 subtypes B (2p3b) and C (2r5q) WT PR’s were downloaded from the PDB (Coman et al; 2008) database. The PR PDB files were edited to remove heteroatoms and inhibitors attached to them.

3.2.3 Molecular Docking

Autodock Vina was used to dock the ligand to its PR receptor molecule. The process included assigning bonds, hybridisation, assigning charges and detecting flexible torsions in ligands (Goodsell et al., 1996). Autodock vina provides docked conformations for molecules with approximately 12 torsional degrees of freedom. It can predict free energies to within 3 kcal/mol in a system where the protein motion is insignificant (Huey et al., 2007; Sriramulu and Lee, 2021). The docked system with the lowest binding affinity and an RMSD value of zero was selected as these had the most stable energy. Further minimization and geometrical conversions were performed in antechamber. Autodock Vina evaluated 9 binding poses. The models with the lowest value were selected for further analyses.

3.2.4 MD simulations

Antechamber, within AmberTools was used for the stereochemical and geometrical conversion of the ligands using the sqm chemistry software to compute the AM1-BCC charges. The allocation of atom types and missing parameters was achieved by applying the General Amber Force Field (GAFF) to the system. The Leap module was used to create the topology and coordinate files by applying a force field of ff03.r1 and combining the system. The ff03.r1 force field is recommended for all new simulations. It optimizes backbone torsions for proteins and diminishes the inclination for helical conformations. It derives its charges from quantum calculations using a continuum dielectric to simulate solvent polarization (Case et al., 2022; Case and Rutgers; Duan et al., 2003). Counter ions were added to neutralize the system and explicitly solvate the TIP3PBOX within 12Å and 0.75 closeness. Covalent hydrogen bonds were restricted using the SHAKE algorithm.

Following structural optimization of the ligand, MD simulations were carried out using Sander and Pmemd and Pmemd.Cuda modules in Amber 14 (Case et al., 2012). The minimization was executed in two stages namely: minimization of waters calculation to relax the water around the solute molecule. The waters were minimized with 3000 steps of steepest-descent and a conjugated gradient of 2000 steps. Then the entire system was minimized with 7000 steps of steepest-descent and a conjugated gradient of 3000 steps. Thereafter the system was gradually heated from 0K to 300K at a fixed volume and normal temperature dynamics (constant number, volume and temperature) over 50

picoseconds. Equilibrium was run for 1000000 steps with a time step of 2 fs (dt=0.002). The periodic boundary with constant volume ntb=1 was used and the bonds involving hydrogen were restrained at ntc=2 using the SHAKE algorithm. The temperature was regulated via the Langevin dynamics thermostat (ntt=3) with a collision frequency of gamma_ln=2.0`. The pressure relaxation time was set to default of 1.0. The atomic residues were constrained at 2.0kcal/mol. The production simulation was conducted for 100 nanoseconds, corresponding to 50 million steps with a 2 fs timestep. Snapshots were taken every 1000 steps.

3.2.5 Analysis and visualisation of the peptide molecule

UCSF Chimera (Meng et al., 2006; Yang et al., 2012) and Biovia Discovery Studio (Systèmes,2021) were used to analyse the sequences and structures. The Molprobity (http://molprobity.biochem.duke.edu/) structure validation web server was used to validate the structure by evaluating the model quality, the hydrogen positioning and all-atom contact analysis. Briefly, the PDB structure was uploaded to the server where hydrogen atoms were added and improved and the Asn/Gln/His 180º flips were rectified (Word JM, 1999). The results were verified and communicated in charts. Ramachandran and summarised output files were downloaded. The Biovia Discovery studio Visualiser version 21 by Dassault (Systèmes, 2021) program was used to create the 2D AA receptor-ligand complex interaction maps.

3.2.6 Post MD simulation analyses

Cpptraj in Amber was used for post MD simulation analyses. The temperature and pressure, as well as the RMSD and RMSF were plotted using xmgrace. The lowest energy structure was extracted for visualisation by identifying the lowest energy minima over the production stage. The binding-free energies were calculated using the (MM/GBSA) method program in Amber over the final 2000 snapshots of the production simulations, time ranging from 80-100 nanoseconds of the trajectory simulations.

3.3 Results

3.3.1 Prediction of theoretical ligand structures

The following templates were shown to have a high sequence similarity for each CS and were considered acceptable for molecular modelling: MA/CA=1kj4, CA/P2=1f7a, P2/NC=1kj7 and P1/P6=1kjf (Prabu-Jeyabalan et al., 2002). These models had no gaps thus requiring no manual adjustments or gap penalties to be made.

3.3.2 Assessment of model refinement and optimal binding poses

The RMSD calculation was carried out for each CS. The models for the CS reached equilibrium and stabilised (Figure 10). The NC/P1 site showed the most stable energy with a lower RMSD as compared to the other models. This was followed by P1/P6, CA/P2, MA/CA and P2/NC. The structure of the consensus AA sequence for each CS is shown in Figure 11 and their binding poses are shown in Figure 12.

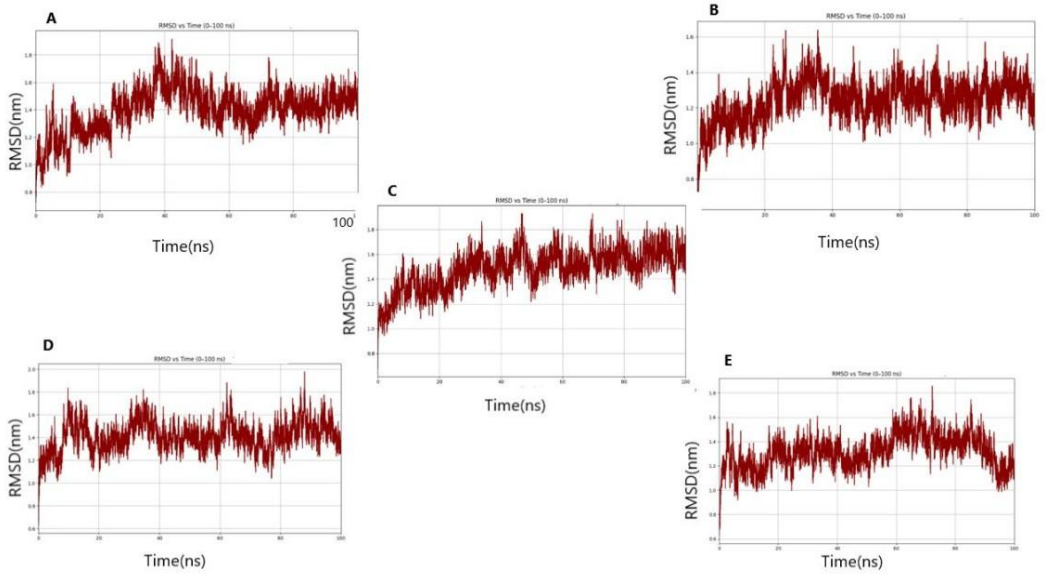


Figure 10: Assessment of HIV-1 *gag* CS sequence models

A) MA/CA = VSQNY/PIVQN, B) CA/P2 = KARVL/AEAMS, C) P2/NC = NNNIM/MQKSN, D) NC/P1 = ERQAN/FLGKI, E) P1/P6 = RPGNF/LQSRP

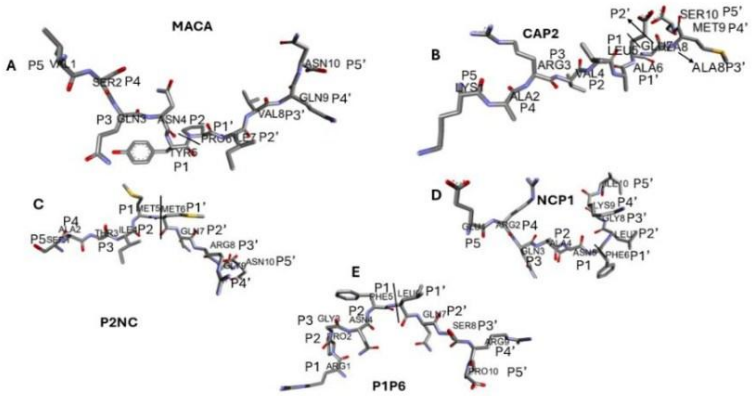
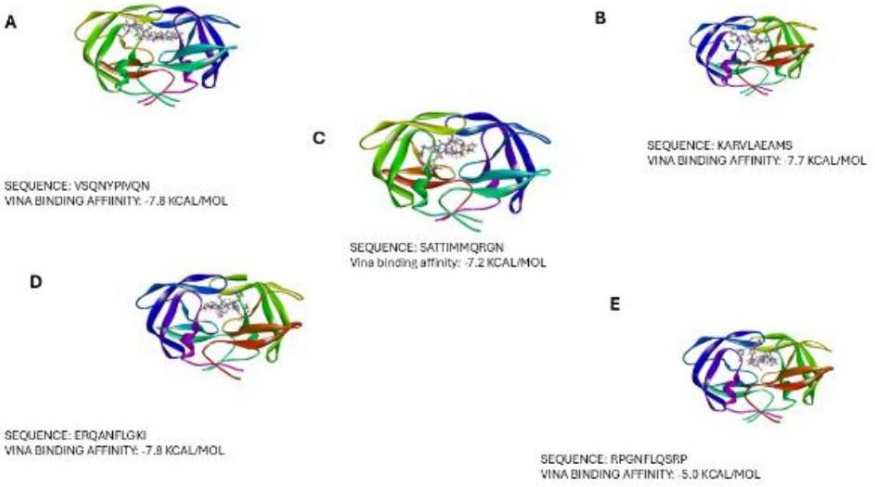


Figure 11: HIV-1 subtype C most common structure at each of the *gag* CSs

A) MA/CA, B) CA/P2, C) P2/NC, D) NC/P1, and E) P1/P6. The scissile bond is the site of proteolytic cleavage represented by the black line drawn down the middle between the P1 and P1' positions. The AA residues and their positions are indicated for each CS, with P5, P4, P3, P2, P1, P1', P2', P3', P4', and P5' labelled accordingly.

3.3.3 Binding poses generated for HIV-1 Gag ligands bound to PR

Subtype B:



Subtype C:

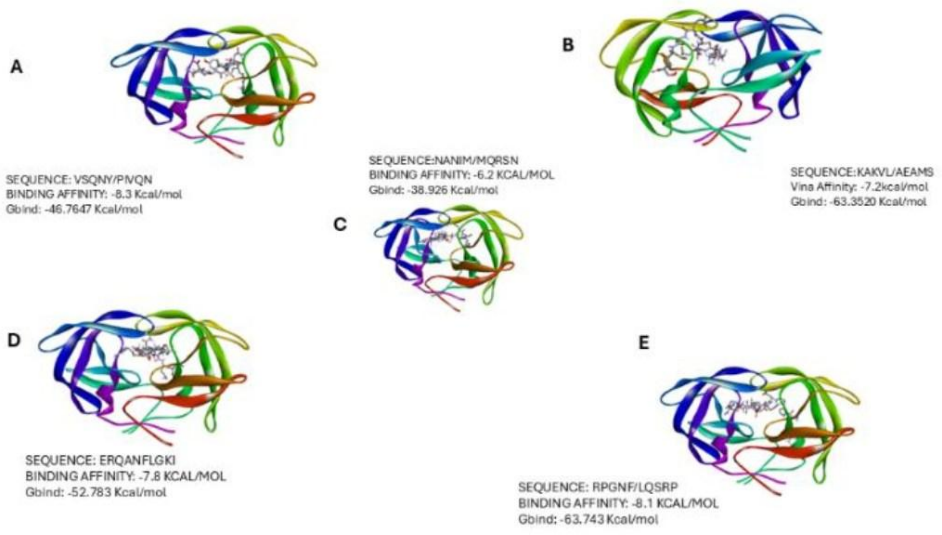


Figure 12: Binding poses generated for HIV-1 subtype B and C ligands bound to PR

Protein-ligand complexes, each associated with a specific peptide sequence, along with corresponding binding afinity values (in kcal/mol) and Gibbs change in free energy (Δ_{Gbind}) in kcal/mol values. A) MA/CA, B) CA/P2, C) P2/NC, D) NC/P1, E) P1/P6.

3.3.4 PR-ligand Interactions:

2D interaction maps were constructed for each complex to analyse the interactions between PR and the ligand.

3.3.4.1 MA/CA

The MA/CA cleavage site (Figure 11A) contained a combination of hydrophobic and polar residues. Valine at position P5 and isoleucine at P2' were key hydrophobic residues with hydrophobicity values of 4.2 and 4.5, respectively, enhancing internal non-polar interactions. Hydrophilic residues, such as asparagine at P2 and P5', along with glutamine at P3 and P4', suggested the involvement of polar interactions and potential hydrogen bonding due to their amide groups. Tyrosine at P1, was suitable for ionizable interactions, while proline at P1' provided rigidity to the ligand's conformation. Glycine at P4' showed a high average isotopic displacement (253.38), indicating flexibility.

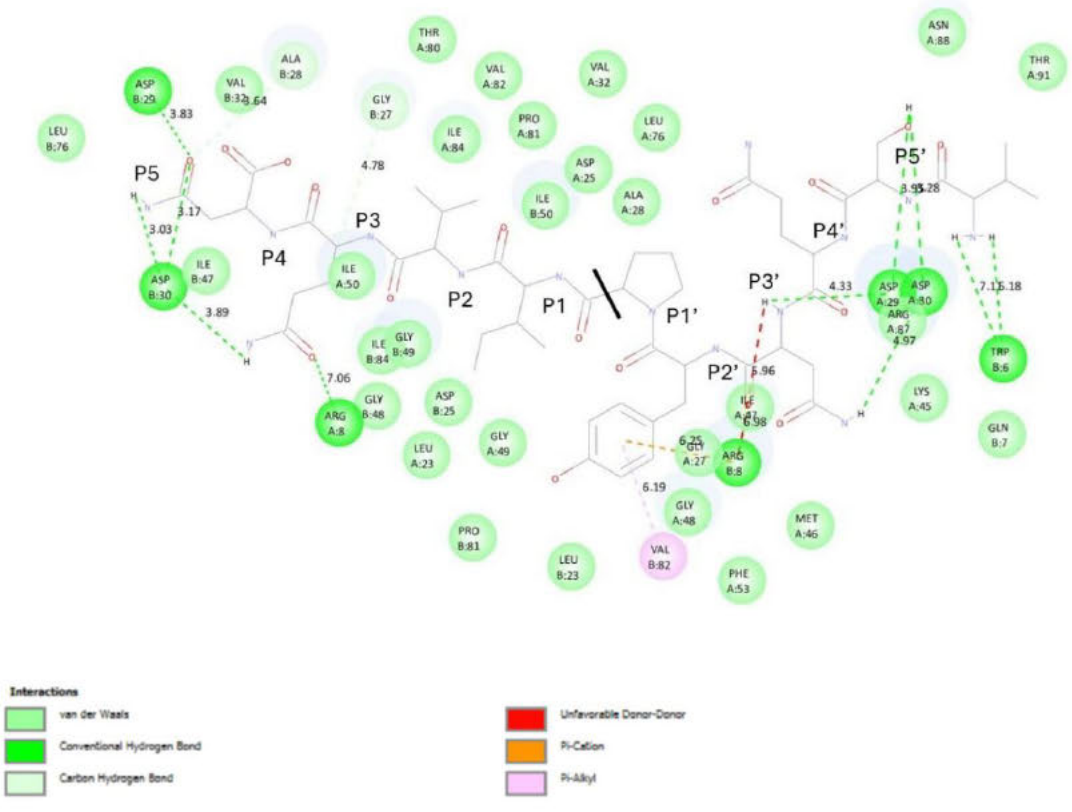


Figure 13: AA Interaction map of VSQNY/PIVQN cleavage ligand site complexed with WT PR

The interaction profile of the MA/CA CS variant: VSQNY/PIVQN complexed with HIV-1 PR (Figure 13) reveals a network of stabilizing forces, contributing to its binding affinity of -8.3 kcal/mol. A notable π-alkyl interaction occurs between the benzene ring and Val82 at the P1' position proline. Additionally, a π-cation interaction is observed between the benzene ring and catalytic Gly27 (chain A) at P1' proline. However, two unfavourable donor-donor interactions involving Arg8 and Ile47 at the P2' isoleucine position could introduce some instability in the binding pose.
The complex forms nine conventional hydrogen bonds with PR which significantly contributes to the binding. Asp29 (Chain B) forms a bond in the P5 valine position, while Asp30 (Chain B) contributes two bonds in the P4 serine and P5 valine positions. Arg8 (Chain B) interacts at the P3 glutamine position and Asp29 (Chain A) forms bonds in the P4' glutamine and P2' isoleucine positions. Further interactions are noted with Arg87 at the P2' isoleucine position, Asp30 (Chain A) at the P4' glutamine position and Trp6 at the P5' asparagine position.

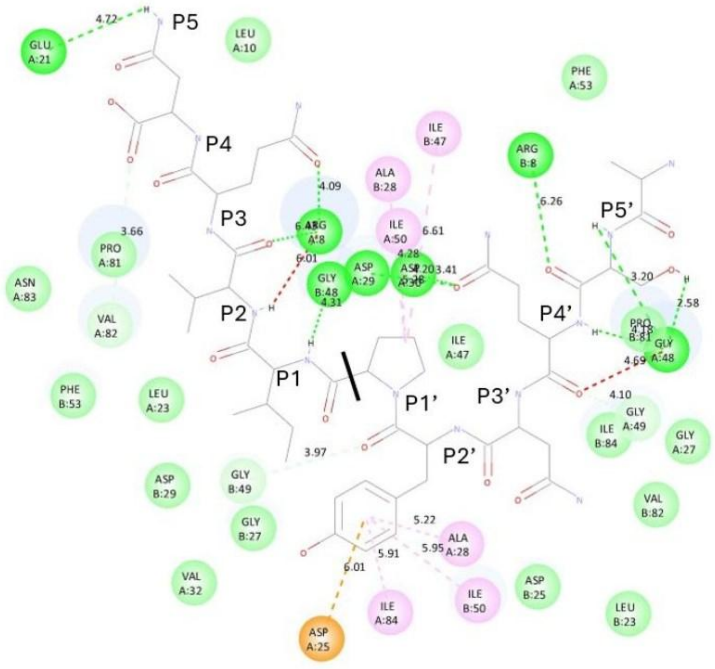


Figure 14: AA Interaction map of ASQNY/PIVQN cleavage ligand site complexed with WT PR

The ASQNY/PIVQN-PR complex (Figure 14) exhibited a binding affinity of -7.5 kcal/mol, slightly lower than that of the VSQNY/PIVQN sequence. The isoleucine forms at P1' alkyl interactions with Ala28, Ile50 and Ile84. Compared to the π interactions observed in the VSQNY/PIVQN site, these alkyl interactions appear less effective in stabilizing the complex. The π-alkyl interactions of isoleucine at P2' with Ile50 and Ala28 (chain B) contribute to the stabilization of the complex, though with reduced intensity. Additionally, a π-anion interaction with isoleucine at P2' with catalytic Asp25 was identified, but the interaction distance of 6.01 Å indicates a weaker electrostatic component to the binding.

Unfavourable acceptor-acceptor interactions with Gly48 and glutamine, donor-donor interactions with Arg8 and serine at P2 were also observed, with distances extending up to 6.01 Å, potentially contributing to reduced binding efficiency. Despite these unfavourable interactions, conventional hydrogen bonds were formed with Glu21 at P5 valine position, Arg8 at P2 serine position, and Gly48 at P1 tyrosine, with distances ranging from 2.31 Å to 6.26 Å.

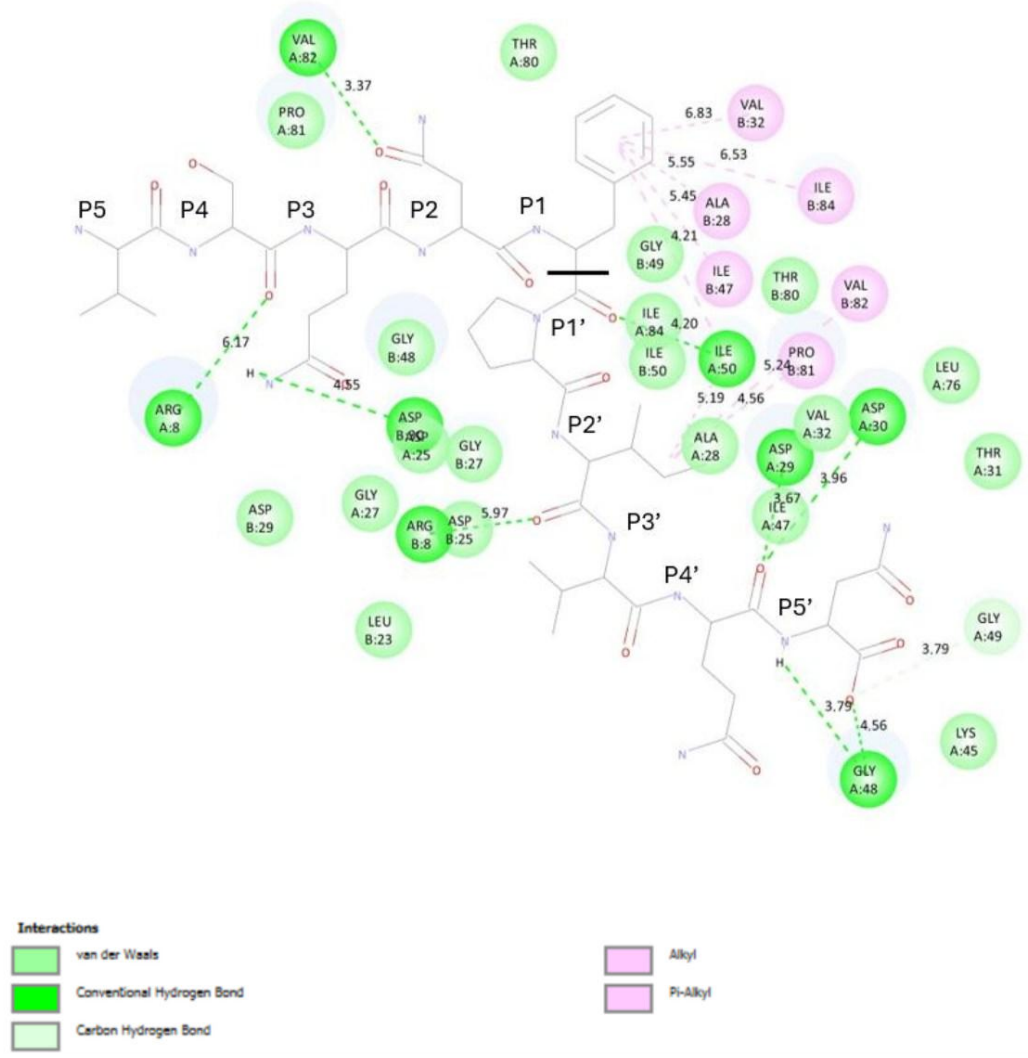


Figure 15: AA Interaction map of VSQNF/PIVQN cleavage ligand site complexed with WT PR

The VSQNF/PIVQN-PR complex (Figure 15) had a binding affinity of -7.4 kcal/mol. Alkyl interactions of Val82, Pro81, and Ile50 with isoleucine at P2' position, along with π-alkyl interactions of Val32, Ile84, Ala28, Ile84, Ile47 with P1 phenylalanine contributed to the binding, though the distances suggest a somewhat less stable interaction compared to other MA/CA sites. The hydrogen bond with Asp30 at P3 glutamine and conventional hydrogen bonds with residues such as Val82 with P1 phenylalanine and Arg8 with P3 glutamine indicate a robust network of interactions, albeit with some variability in binding strength.

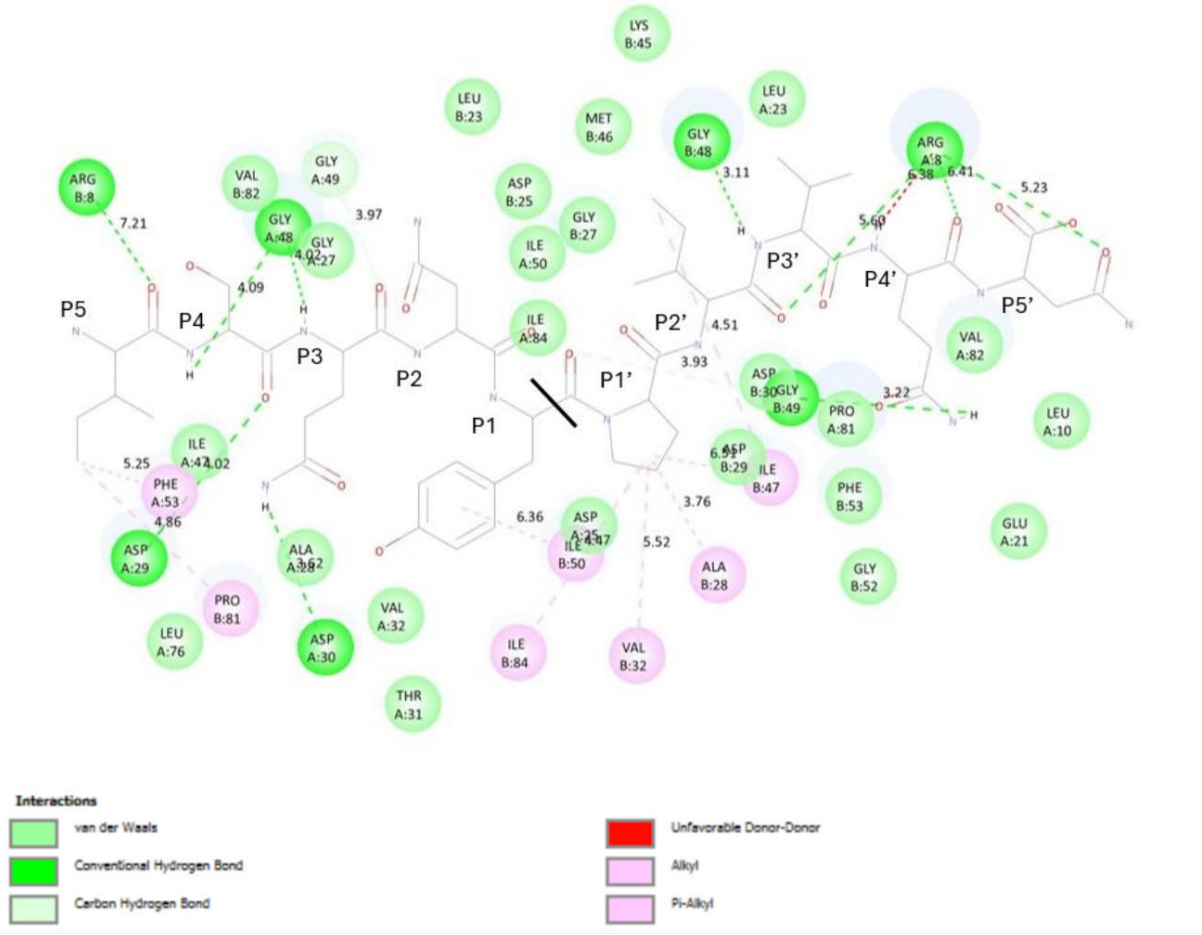


Figure 16: AA Interaction map of ISQNY/PIVQN cleavage ligand site complexed with WT PR

The ISQNY/PIVQN-PR complex (Figure 16) sequence exhibited a binding affinity of -7.3 kcal/mol, forming multiple interactions with the PR. Alkyl interactions with Phe53 at P3 glutamine and Pro81 at P5 isoleucine, as well as π-alkyl interactions with Ile47, Ile50, Ile84, Val32 and Ala28 at the P1' proline which contributed to the overall binding stability. However, the presence of an unfavourable donor-donor interaction with Arg8 at the P4' glutamine and variability in hydrogen bond distances, reaching up to 6.41 Å, indicated a slightly less favourable binding profile compared to other CSs.

Key interactions involve alkyl and π-alkyl interactions, which primarily stabilize the enzyme-substrate complex. For instance, residues such as Phe53, Pro81, and Ile47 contribute to alkyl interactions at P5 isoleucine, P4 serine and P2' isoleucine respectively, where Ile47 notably participated in both P2' and P1' positions, thereby reinforcing the substrate binding. The π-alkyl interactions were also substantial, with residues like Ala28, Val32 and Ile84 anchoring the substrate at critical positions P1 tyrosine and P1' proline.

Furthermore, unfavourable donor-donor interactions, specifically involving Arg8 at P4’ glutamine residue, exhibited a significant impact at the P4' position. This unfavourable interaction suggests potential steric hindrance or electrostatic repulsion, which could influence the cleavage efficiency. On the other hand, hydrogen bonding interactions were more favourable, with key residues such as Gly49 and Asp30 forming critical bonds at P2 asparagine and P1' proline positions, respectively. The presence of multiple conventional hydrogen bonds, notably involving Arg8 at P4’ glutamine, Gly48 with serine and glutamine, and Asp30 with asparagine, underscored the significance of these interactions in maintaining the stability of the substrate in PR's active site.

3.3.4.2 CA/P2

The CA/P2 cleavage site (Figure 11B) integrated both hydrophobic and polar residues. Hydrophobic residues such as valine (P2) and leucine (P1) with hydrophobicity values of 4.2 and 3.8, respectively, established a stable hydrophobic core. Alanine residues at P4 and P1’ also supported this trend with values of 1.8. Polar residues like lysine at P1 and arginine at P3 were basic and positively charged, facilitating electrostatic interactions. Glutamic Acid at P2’, with a pKa of 4.3, suggested ionic interactions. Methionine at P4’ contributed to hydrophobic interactions at the terminal region. Overall, the CA/P2 ligand balanced hydrophobic and polar residues to maintain stability and accommodate electrostatic and ionic interactions.

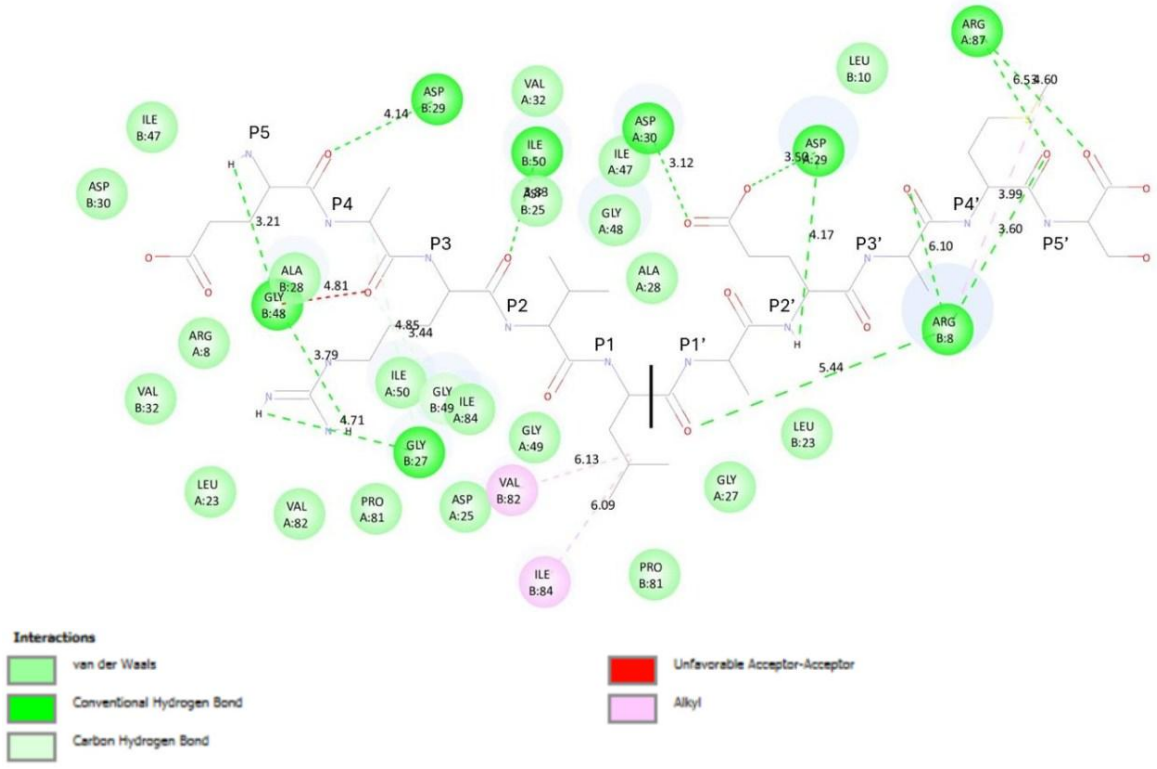


Figure 17: AA Interaction map of EARVL/AEAMS CS ligand complexed with WT PR

The CA/P2 CS variant: EARVL/AEAMS complexed with PR (Figure 17) demonstrated a varied interaction profile with a binding affinity of -8.2 kcal/mol. The sequence exhibited significant alkyl interactions of leucine at P1 with Val82, Ile84, and Arg8 with P1', P3' alanine and P5' serine. While the interactions with Val82 and Ile84 were relatively longer, indicating less optimal binding, the shorter distance with Arg8 enhanced the binding. However, Gly48's unfavourable acceptor-acceptor interaction with P3 arginine at 4.81 Å pointed to potential steric or electrostatic repulsion, reducing overall interaction efficiency. The hydrogen bonding network is robust, featuring two hydrogen bonds of Gly49 and Ile50 with P3 arginine . The conventional hydrogen bonds involved residues such as catalytic Gly27 with P2 alanine, Gly48 with P2 alanine and P5 glutamic acid, and Asp29 with P2' glutamic acid. These bonds, with distances ranging from 3.12 Å to 6.53 Å, contributed to a well-established and stable binding environment.

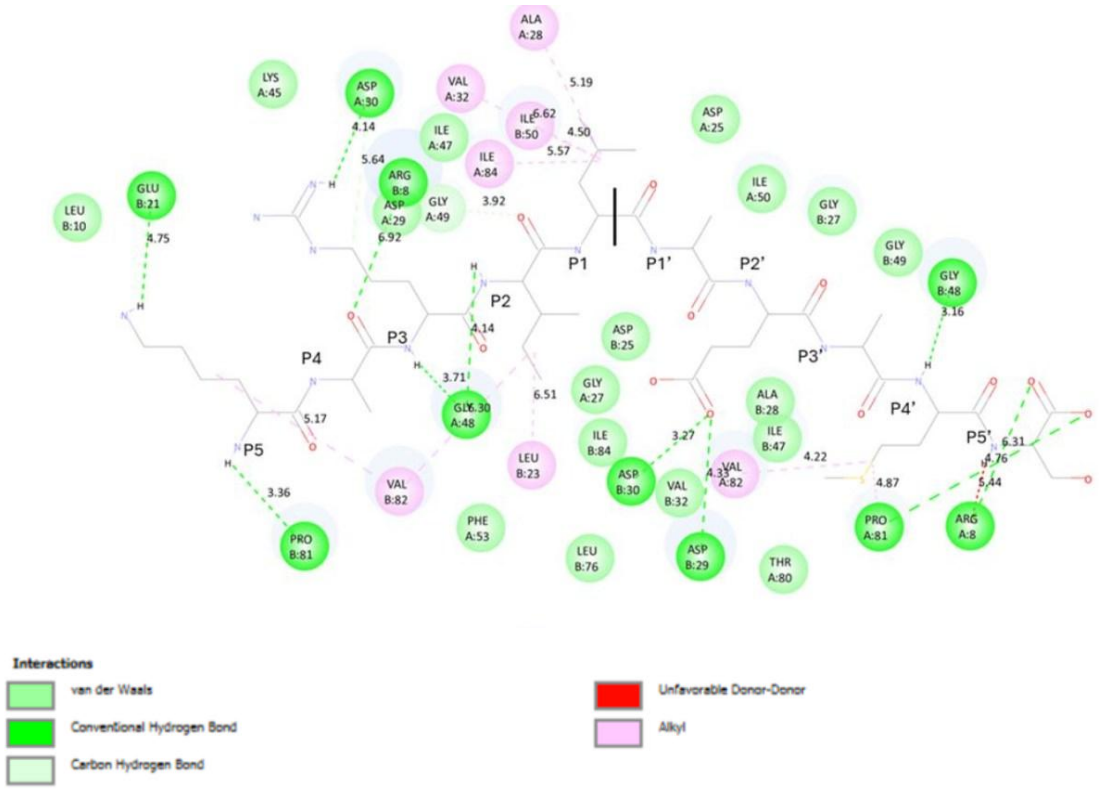


Figure 18: AA Interaction map of KARIL/AEAMS CS ligand complexed with WT PR

The KARIL/AEAMS-PR complex (Figure 18) exhibited a strong binding profile with a binding affinity of -7.5 kcal/mol. Alkyl interactions were prominent, involving Val32, Ile84, Ile50, and Ala28 all with P1 leucine. The shorter distance of Ile50 suggested a more favourable interaction compared to the others. The unfavourable donor-donor interaction of Arg8 with P5' serine at 5.44 Å indicated potential electrostatic repulsion or steric hindrance. Two hydrogen bonds of Gly49 with P1 leucine and Asp30 (chain A and B) with P2 alanine and P2' glutamic acid contributed to the binding stability,

with Gly49 showing a more favourable distance. The conventional hydrogen bonds are numerous, totalling eleven, and involved residues such as Glu21 with P4 alanine, Pro81 with P5 lysine and Asp30. These interactions, with distances from 3.27 Å to 6.92 Å, ensured a strong and stable binding interface.

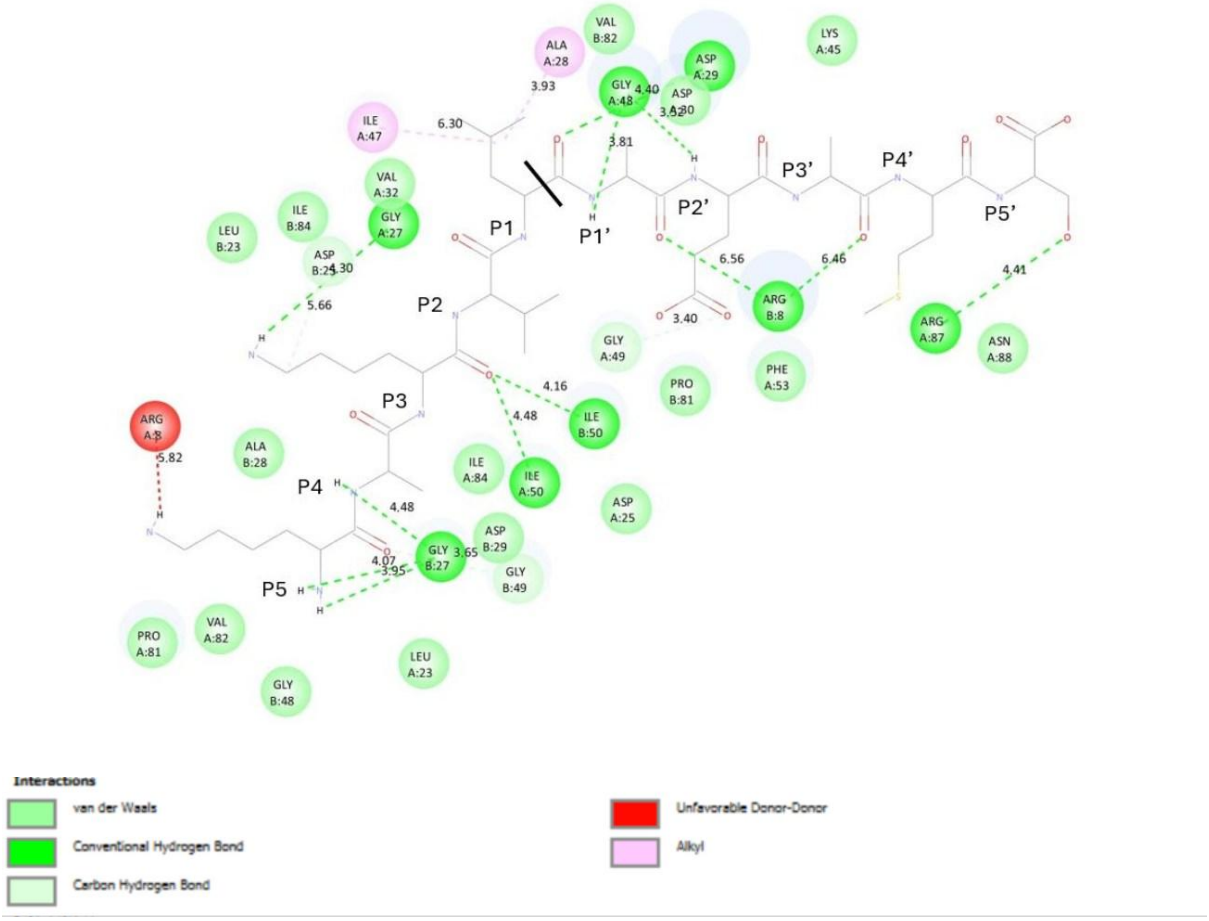


Figure 19: AA Interaction map of KAKVL/AEAMS CS ligand complexed with WT PR

The KAKVL/AEAMS-PR complex (Figure 19) demonstrated a binding affinity of -7.2 kcal/mol. The site exhibited notable alkyl interactions at P1 leucine with Ile47 and Ala28, although the interaction with Ile47 was at a longer distance (6.30 Å) compared to the shorter and more optimal binding distance with Ala28 (3.93 Å) at position P1. Unfavourable donor-donor interactions involving Arg8 at 5.82 Å at the P4 alanine position suggested potential steric or electrostatic repulsion, which may affect the interaction efficiency. Despite this, the hydrogen bonding network was prominently effective, with two key hydrogen bonds between Gly49 and the ligand at 3.40 Å, strengthening the binding at P1' alanine.

Conventional hydrogen bonds played an additional role in stabilizing the binding environment, with multiple interactions involving residues such as Catalytic Gly27 (Chain A) at P4 alanine and P5 lysine and Gly48. Specifically, catalytic Gly27 forms bonds at distances of 3.95 Å, 4.07 Å, and 4.48 Å at position P4, while Gly48 engages in multiple bonds at 3.81 Å and 4.40 Å at positions P1' alanine and P2' glutamic acid. Additionally, interactions with Ile50 (chain A & B) at P2 alanine and Arg87 at P5' serine further stabilized the complex, although Arg8 at 6.56 Å and 6.46 Å at P2' glutamic acid and P3' alanine, respectively, introduced less favourable interactions.

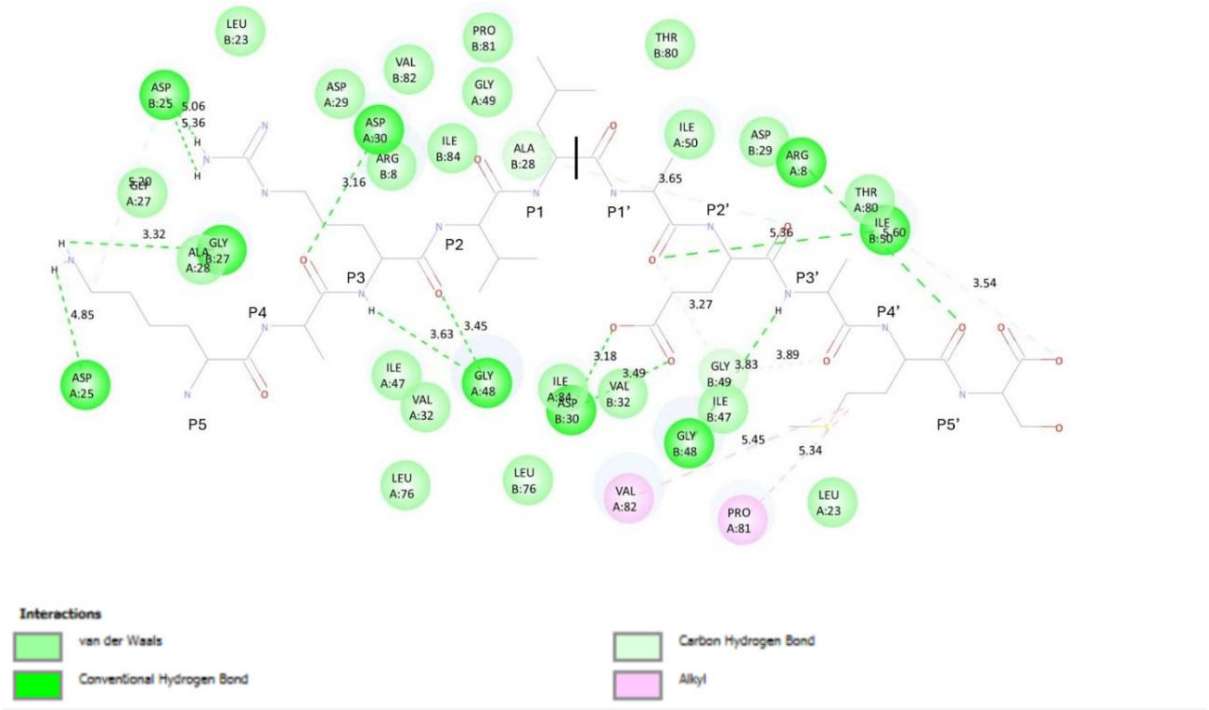


Figure 20: AA Interaction map of KARVL/AEAMS CS ligand complexed with WT PR

The KARVL/AEAMS-PR complex (Figure 20) had a binding affinity of -4.2 kcal/mol. Alkyl interactions with Val32 (5.45 Å) and Pro81 (5.34 Å) with P4' methionine were present, but their moderate distances suggested less favourable interactions compared to other sequences. The hydrogen bonding network included two bonds with Gly49 (3.83 Å) at P3' alanine and Thr80 (3.54 Å) with P5' serine, where Gly49 displayed a more favourable interaction. The conventional hydrogen bonds were numerous, totalling thirteen, and involved residues such as catalytic D25 (Chains A and B) with P3, P5 lysine and Gly48 with P2 alanine and P3 lysine. These bonds, with distances from 3.32 Å to 5.36 Å, contributed to a robust and stable binding interaction, despite the overall lower binding affinity. Key alkyl interactions included residues Val32 and Pro81 interacting with P4' methionine, although both were characterized by longer distances of 5.45 Å and 5.34 Å, respectively, which may have limited their optimal binding at position P4'. The hydrogen bonding network was robust, featuring significant interactions such as those between Gly49 with P3' lysine at 3.83 Å and Thr80 at 3.54 Å,

contributing to the stability at P4' methionine. Additionally, Gly49 (chain B) interacted at 3.89 Å with P3' arginine. Conventional hydrogen bonds played a crucial role in the stabilization of the CS. Notable interactions involved catalytic D25 (chain A) and catalytic Gly27 (chain B) with bond distances of 4.85 Å and 3.32 Å at position P5. Catalytic D25 (chain B) also formed bonds at 5.06 Å and 5.36 Å at position P2 alanine, while Asp30 contributed significantly at 3.16 Å to P3 arginine. Gly48 engaged at 3.45 Å and 3.63 Å at positions P2 alanine and P3 arginine, respectively. Additionally, Ile50 at 5.38 Å and Arg8 at 5.60 Å provided stabilization at positions P2' glutamic acid and P4' methionine, respectively. Further stabilizing interactions were observed with Asp30 (chain B) and Gly48 (chain B) at distances of 3.18 Å, 3.49 Å and 3.83 Å at positions P2' glutamic acid and P3' arginine.

3.3.4.3 P2/NC

The P2/NC cleavage site (Figure 11 C) showed polar residues, such as asparagine at P5 and P5', facilitated hydrogen bonding and electrostatic interactions, stabilizing the structure. Threonine at P4 enhanced ligand interactions through hydrogen bonding. Isoleucine at P2, with a hydrophobicity value of 4.5, played a key role in forming the hydrophobic core, enhancing stability. Methionine at P1 and P1', with a hydrophobicity value of 1.9, further reinforced the hydrophobic environment. Glutamine at P2' supported polar interactions and potential hydrogen bonding, which were essential for binding affinity. Serine at P4' was involved in hydrogen bonding (Gray and Matthews, 1984), adding to ligand stability and specificity.

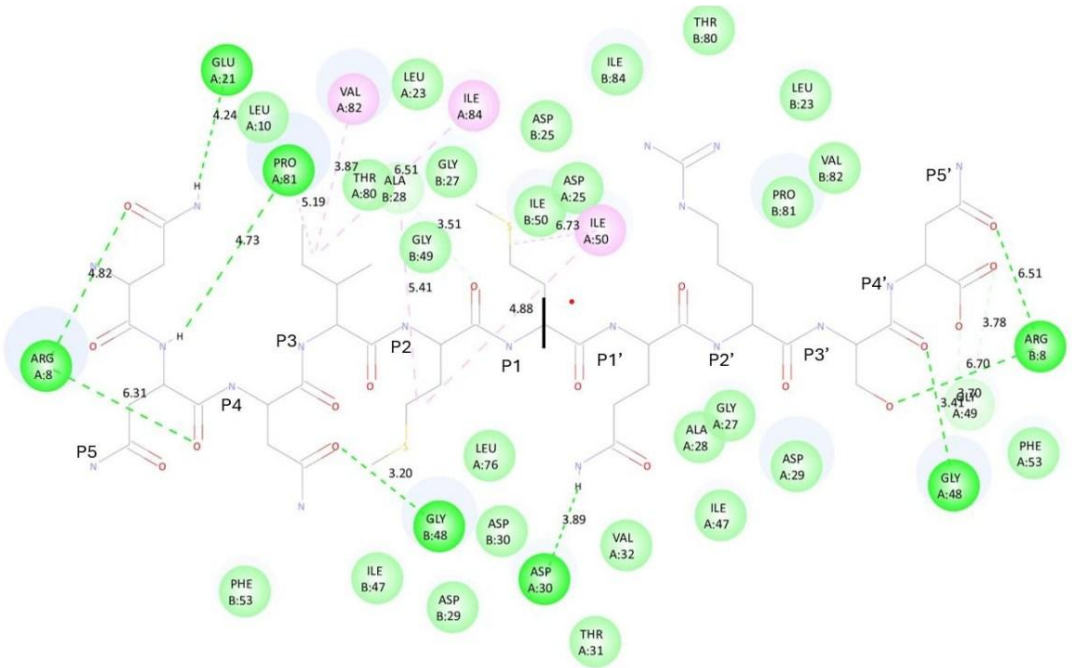


Figure 21: AA Interaction map of NNNIM/MQRSN CS ligand complexed with WT PR

The P2/NC CS variant, NNNIM/MQRSN complexed with PR (Figure 21) had a binding affinity of -7.1 kcal/mol. This sequence showed alkyl interactions with residues such as Val82, Ile84 at P2 isoleucine, and Ile50 with P1 methionine. Val82 and Ile84, with distances of 3.87 Å and 6.51 Å respectively, contributed to the binding with varying strengths. Notably, the closer interaction with Ile50 at 4.88 Å suggested a more stable binding. The site also featured three key hydrogen bonds: Ala28 with P2 isoleucine, Gly49 with P3' arginine, and Asp30 with P1' methionine. These interactions occurred at distances ranging from 3.41 Å to 3.78 Å, indicating a strong hydrogen bonding network. The conventional hydrogen bonds were extensive, totalling nine, and involved residues such as Arg8 with P4 asparagine and Pro81 with asparagine at P4 as well. These interactions, with distances from 3.20 Å to 6.70 Å, further stabilized the binding interface.

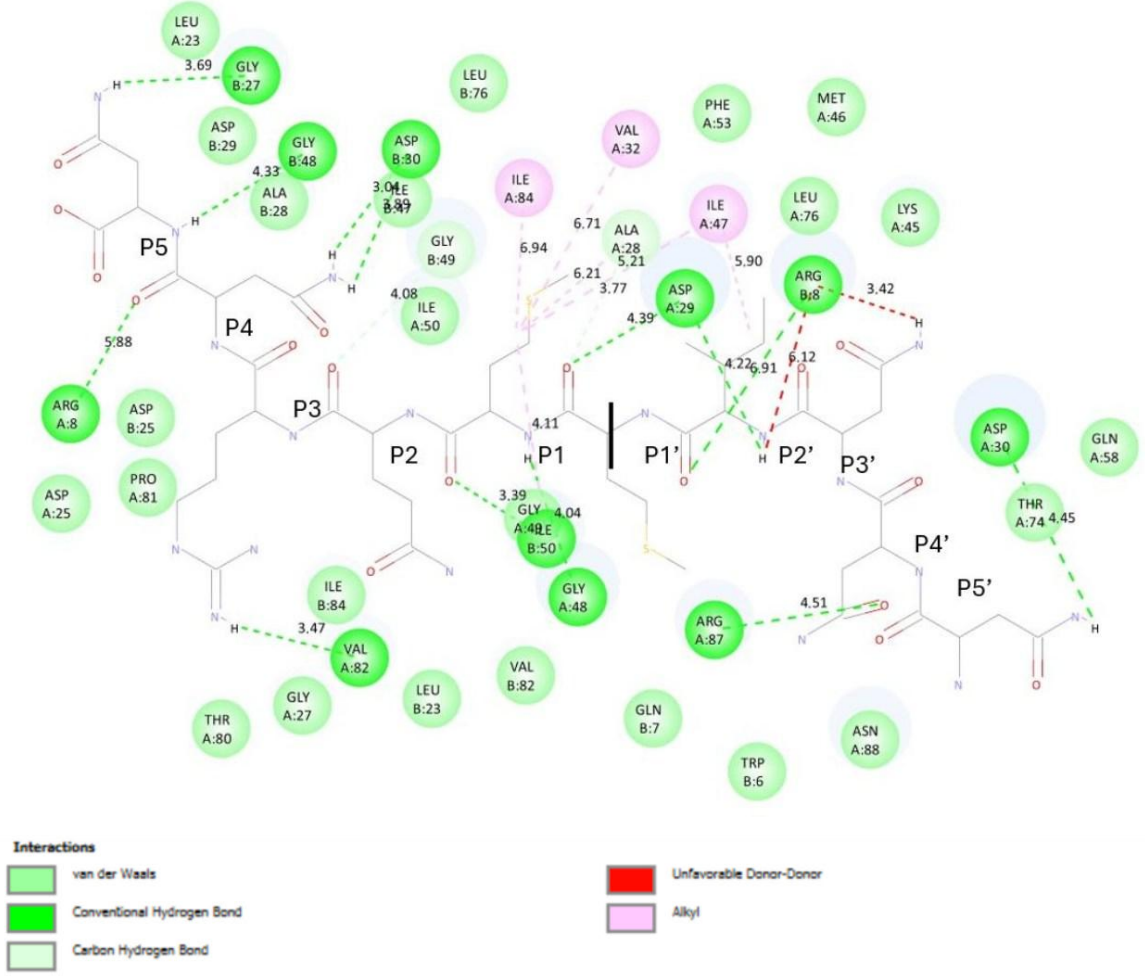


Figure 22: AA Interaction map of NNNIM/MQRGN CS ligand complexed with WT PR

For the NNNIM/MQRGN-PR complex (Figure 22), the binding affinity was -7.1 kcal/mol. The alkyl interactions observed with Ile84, Val32 and Ile47 with P1 methionine at distances of 6.24 Å and 6.71 Å respectively, exhibited moderate binding, while Ala28 at 6.21 Å provided additional stabilization. However, the unfavourable donor-donor interactions with Arg8 at distances of 3.42 Å and 6.12 Å suggested some repulsion or steric hindrance. The hydrogen bonding network included bonds with Gly84 and Ala28 at distances of 4.03 Å and 5.21 Å. Conventional hydrogen bonds were numerous, totalling twelve, involving residues such as Arg8 and Catalytic Gly27. These bonds, with distances ranging from 3.04 Å to 6.91 Å, contributed significantly to the overall stability of the binding interaction.

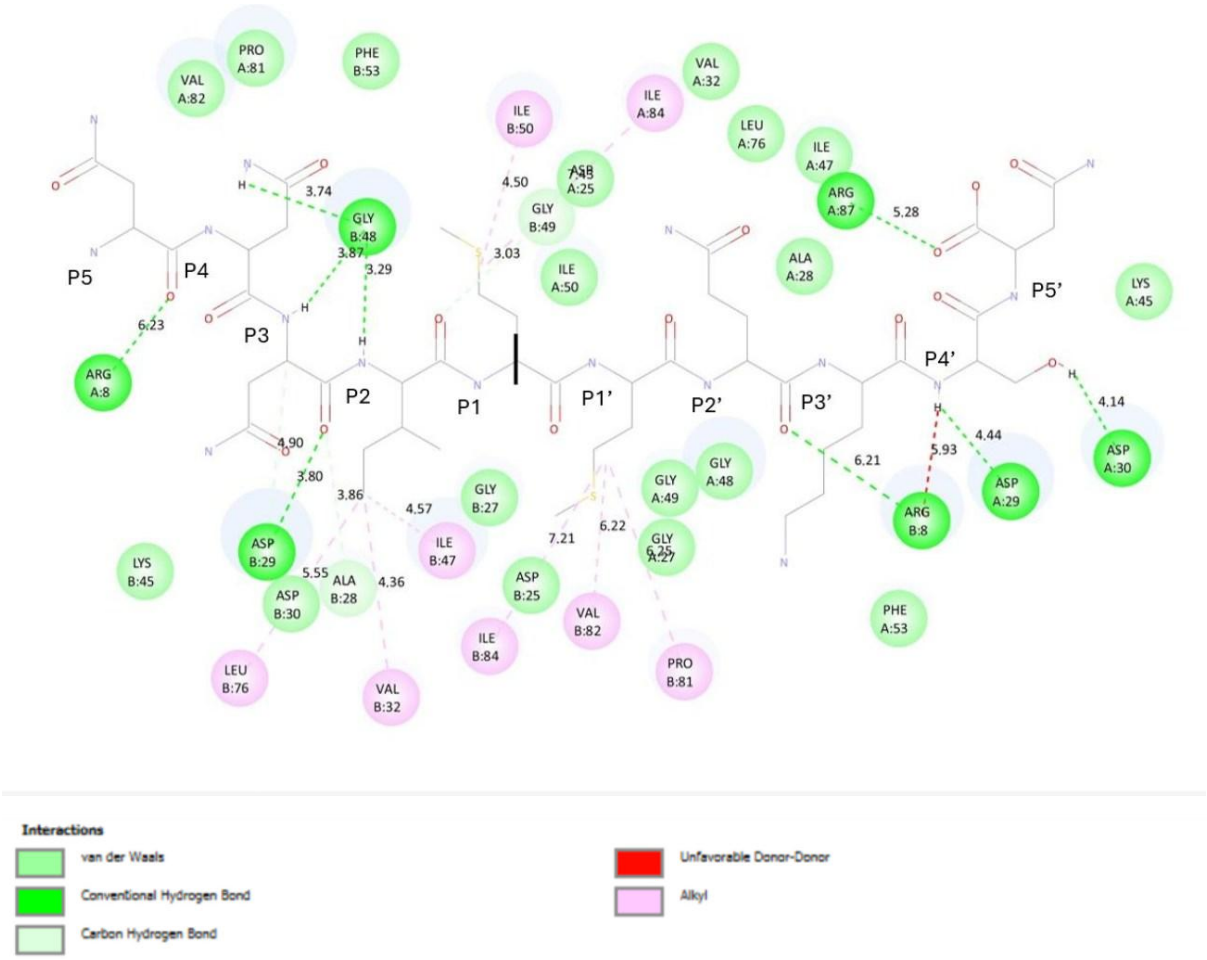


Figure 23: AA Interaction map of NNNIM/MQRNN CS ligand complexed with WT PR

The NNNIM/MQRNN-PR complex (Figure 23) had a binding affinity of -6.2 kcal/mol. Alkyl interactions were extensive, with Leu76, Val32, Ile47, Ile84, Val82, Pro81, Ile50 and Ile84A contributed to the binding. The distances for these interactions range from 4.36 Å to 7.45 Å, indicating varying levels of interaction strength. Notably, the unfavourable donor-donor interaction with Arg8 at 5.93 Å indicated potential repulsion. The hydrogen bonding network included bonds with Gly49, Ala28, and Asp29 at distances from 3.03 Å to 3.86 Å, providing strong stabilization. Conventional hydrogen bonds, totalling nine, involved residues such as Arg8 and Gly48. The distances for these bonds range from 3.29 Å to 6.23 Å, which ensured a robust binding interface despite the overall lower binding affinity.

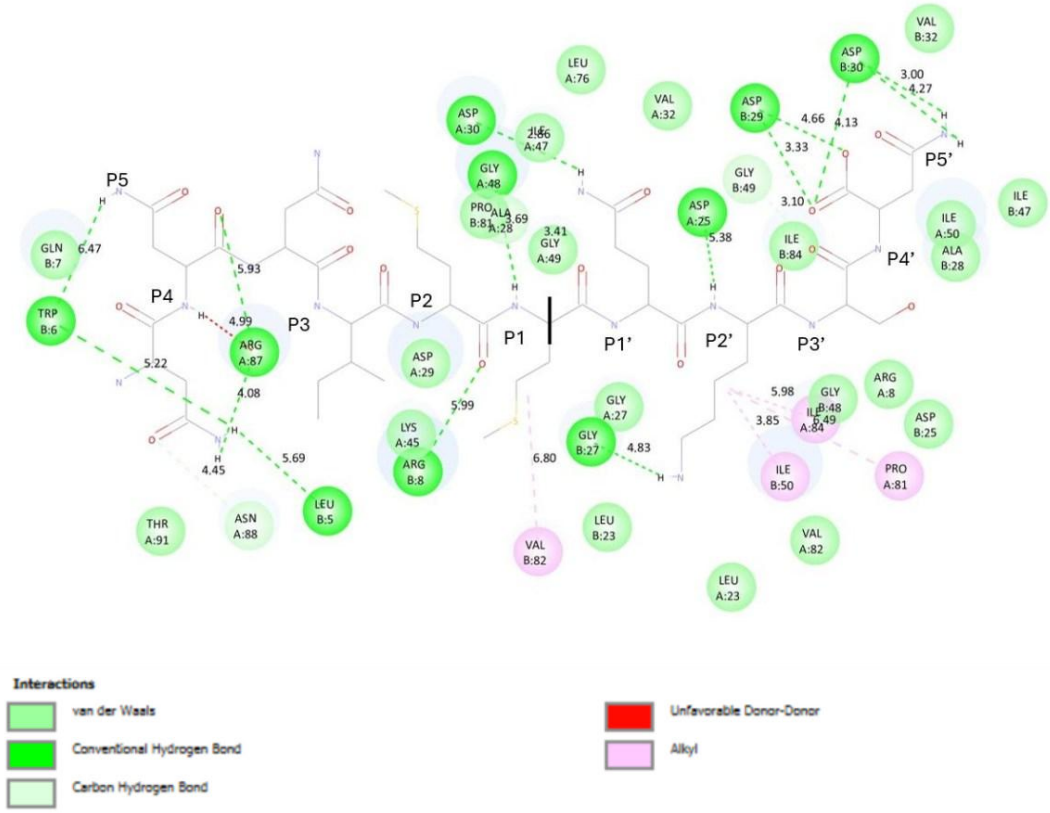


Figure 24: : AA Interaction map of NNNIM/MQKSN CS ligand complexed with WT PR

The NNNIM/MQKSN-PR complex (Figure 24) had a binding affinity of -6.1 kcal/mol. Alkyl interactions involved Val82, Ile50, Ile84, and Pro81. The distances for these interactions ranged from 3.85 Å to 6.80 Å, with Val82 and Ile50 providing strong interaction points. The unfavourable donor-donor interaction with Arg87 at 4.99 Å indicated potential steric or electrostatic repulsion. The hydrogen bond with Asn88 at 4.14 Å contributed to the binding stability. The conventional hydrogen bonds were extensive, totalling fifteen and involved residues such as Trp6, Arg87, and catalytic

Gly27. These interactions, with distances from 3.00 Å to 6.47 Å, provided significant stability and reinforced the binding interface.

3.3.4.4 *NC/P1*

The NC/P1 cleavage site (Figure 11D) exhibited a combination of hydrophobic and hydrophilic residues. Hydrophobic residues, such as leucine at P2' and phenylalanine at P1', with hydrophobicity values of 3.8 and 2.8, respectively, formed the hydrophobic core. In contrast, polar residues such as glutamic acid at P1 and arginine at P2, with pKa values of 4.3 and 12, respectively, were involved in ionic interactions. Glutamine at P3 and asparagine at P1 contributed to hydrogen bonding, while glycine at P3' provided flexibility. Lysine at P4' engaged in electrostatic interactions. The combination of hydrophobic and polar residues ensured the structural stability and functionality of the NC/P1 ligand through a range of interactions.

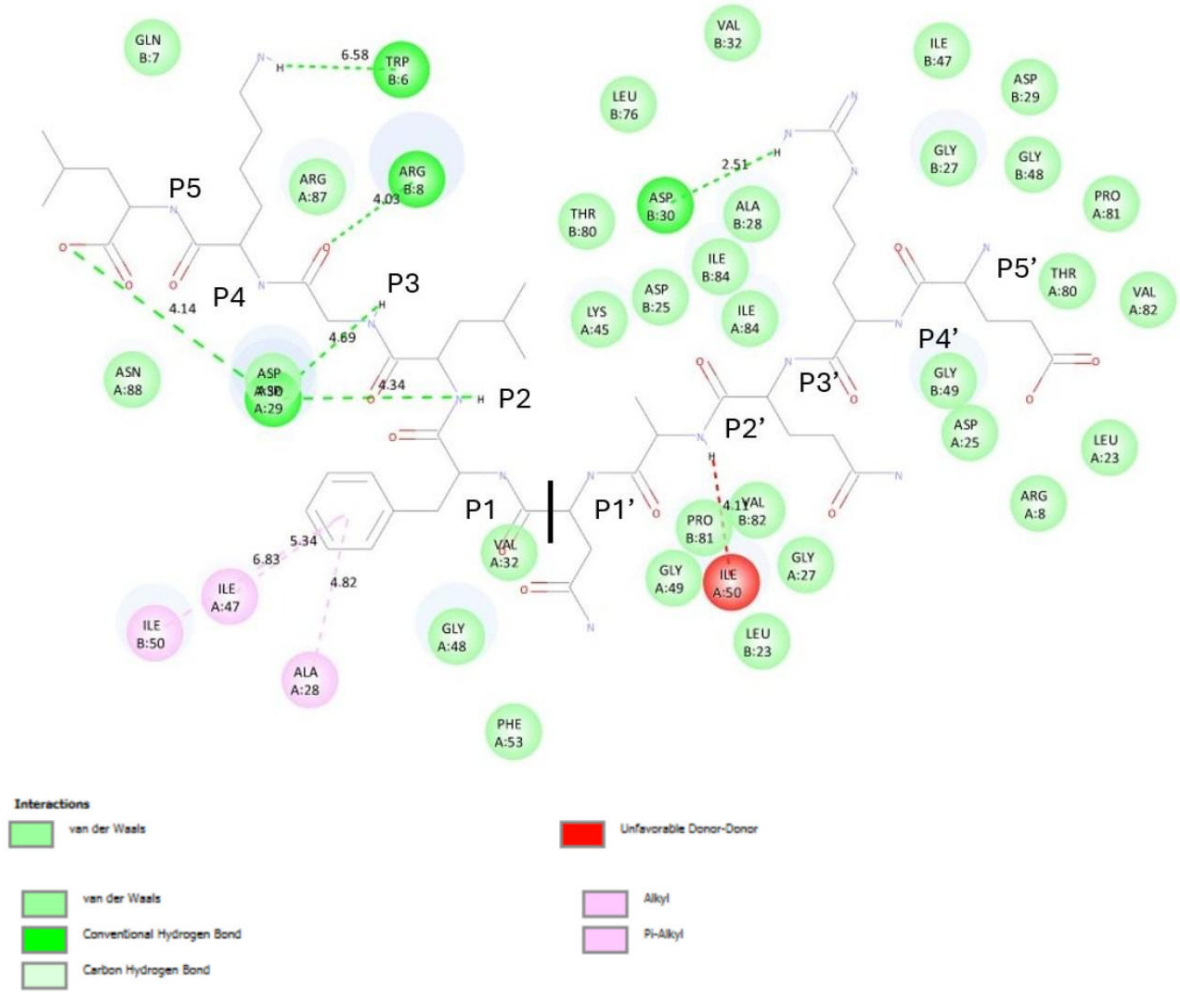


Figure 25: AA Interaction map of ERQAN/FLGKV CS ligands complexed with WT PR

The ERQAN/FLGKV-PR complex (Figure 25) had a binding affinity of -8.2 kcal/mol. This sequence was characterized by significant π-alkyl interactions involving Ile50, Ile47, and Ala28, with distances ranging from 4.82 Å to 6.83 Å. These interactions contributed substantially to the stability of the binding. However, there was an unfavourable donor-donor interaction between Ile50 (chain B) and residue P2' at 4.11 Å, which may have introduced some steric or electrostatic repulsion. Conventional hydrogen bonds were prominent, with six key interactions including Asp29, Arg8, and Trp6. These bonds spanned distances from 2.51 Å to 6.58 Å, providing strong stabilization across various residues.

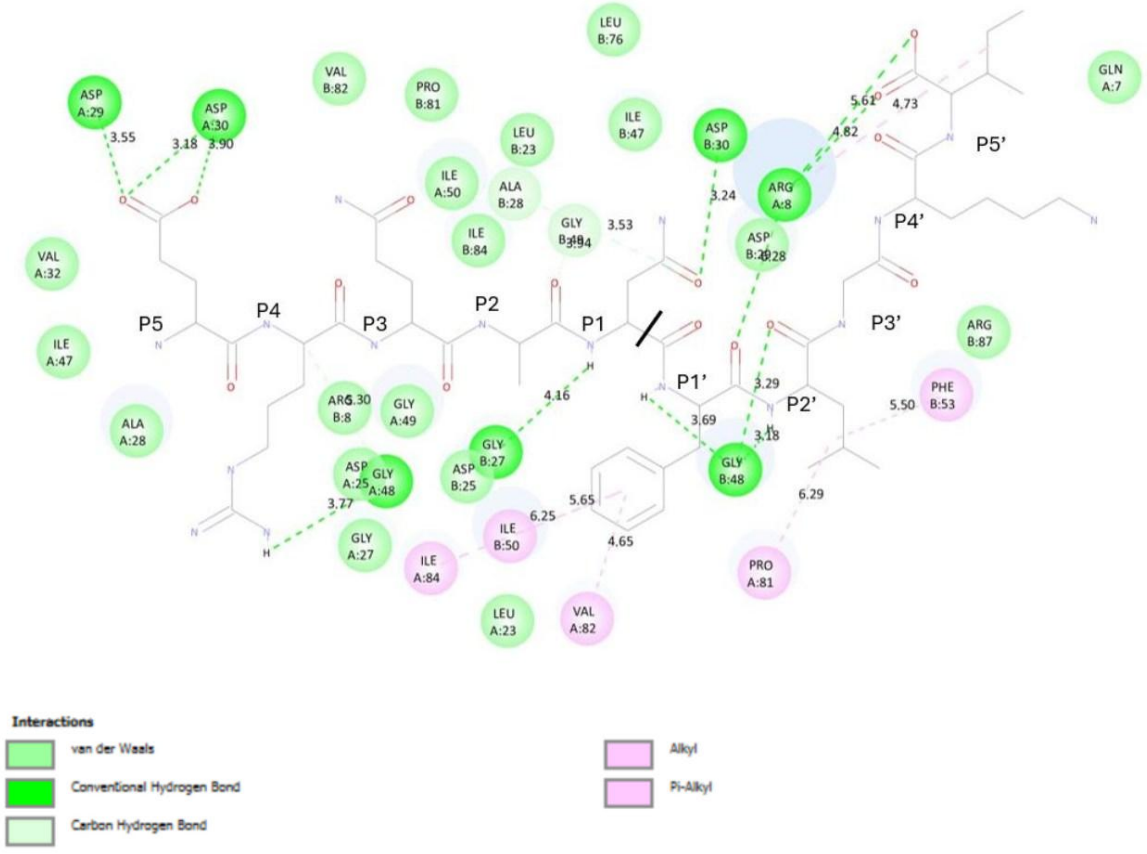


Figure 26: : AA Interaction map of ERQAN/FLGKI CS ligand complexed with WT PR

For the ERQAN/FLGKI-PR complex (Figure 26), the binding affinity was -7.8 kcal/mol. This site featured π-alkyl interactions with Val82, Ile84, and Ile50, with distances ranging from 4.65 Å to 6.25 Å. These interactions were important for the overall stability. Alkyl interactions with Pro81, Phe53, and Arg8 at distances between 4.73 Å and 6.23 Å further stabilized the binding. The hydrogen

bonding network included two notable bonds with catalytic D25 and Ala23, at distances of 3.53 Å and 5.32 Å respectively. The conventional hydrogen bonds were extensive, with ten interactions involving residues such as Asp29, Gly48, and Asp32, with distances from 3.10 Å to 5.61 Å. This extensive network contributed significantly to the binding affinity.

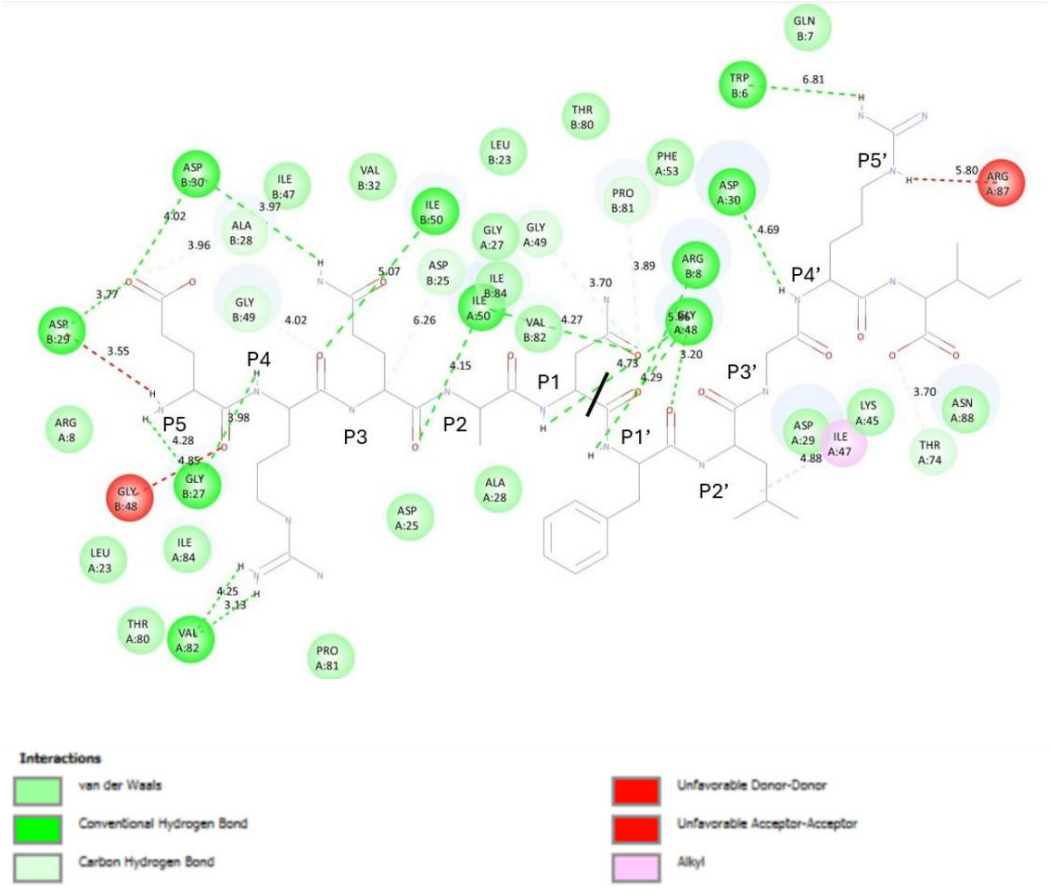


Figure 27: AA Interaction map of ERQAN/FLGRI CS ligand complexed with WT PR

The ERQANFLGRI-PR complex (Figure 27) had a binding affinity of -7.4 kcal/mol. This sequence was characterized by alkyl interactions with Ile47 at a distance of 4.88Å. Unfavourable interactions included acceptor-acceptor and donor-donor repulsions, with Gly41 and Arg82 showing potential steric hindrance. The hydrogen bonding network included six key interactions with residues such as Ala23, Gly41 (chain B), and catalytic D25, with distances from 3.70 Å to 6.26 Å. Conventional hydrogen bonds were extensive, totalling thirteen, and involved residues such as catalytic D25, Gly27, and Val32. The distances for these bonds range from 3.13 Å to 5.07 Å, indicating a robust interaction network despite some unfavourable interactions.

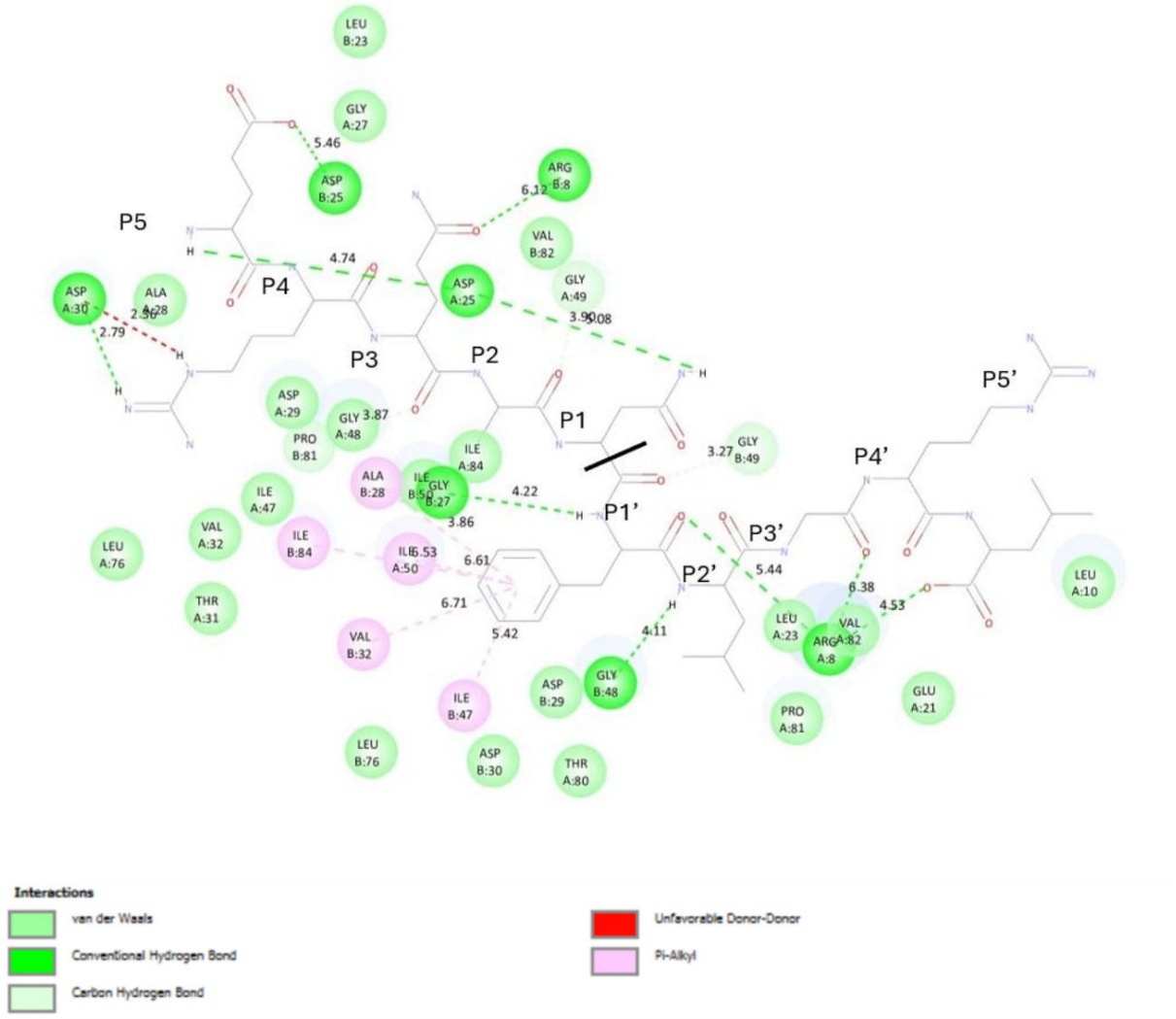


Figure 28: AA Interaction maps of ERQAN/FLGRL CS ligand complexed with WT PR

The ERQANFLGRL-PR complex (Figure 28) had a binding affinity of -6.5 kcal/mol. The π-alkyl interactions were prominent, involving residues such as Ala23,Val32,Ile47, Ile84, and Ile50, with distances ranging from 5.42 Å to 6.71 Å. These interactions provided substantial stabilization to the binding site. However, there was an unfavourable donor-donor interaction with Asp30 at a distance of 2.56 Å. The hydrogen bonding network included two key interactions with Gly41 and Gly49, with distances of 3.03 Å and 3.27 Å. Conventional hydrogen bonds totalled thirteen, where interactions involved residues such as Asp30, catalytic D25, and Arg8. The distances for these bonds ranged from 2.73 Å to 6.33 Å, ensuring a stable binding interface despite the presence of some unfavourable interactions.

3.3.4.5 P1/P6

The P1/P6 cleavage site (Figure 11E) included both hydrophobic and polar residues. Hydrophobic residues such as phenylalanine at P1 and leucine at P1', with hydrophobicity values of 2.8 and 3.8, respectively, contributed to the hydrophobic core. Polar residues, such as arginine at P5 and P4' with a pKa of 12, facilitated electrostatic interactions, while asparagine at P2 supported polar interactions. Proline at P4 and P5' provided rigidity, and glycine at P3 added flexibility. Serine at P3' indicated polar interaction involvement. The P1/P6 ligand balanced hydrophobic stability with polar and electrostatic interactions, maintaining its structural integrity.

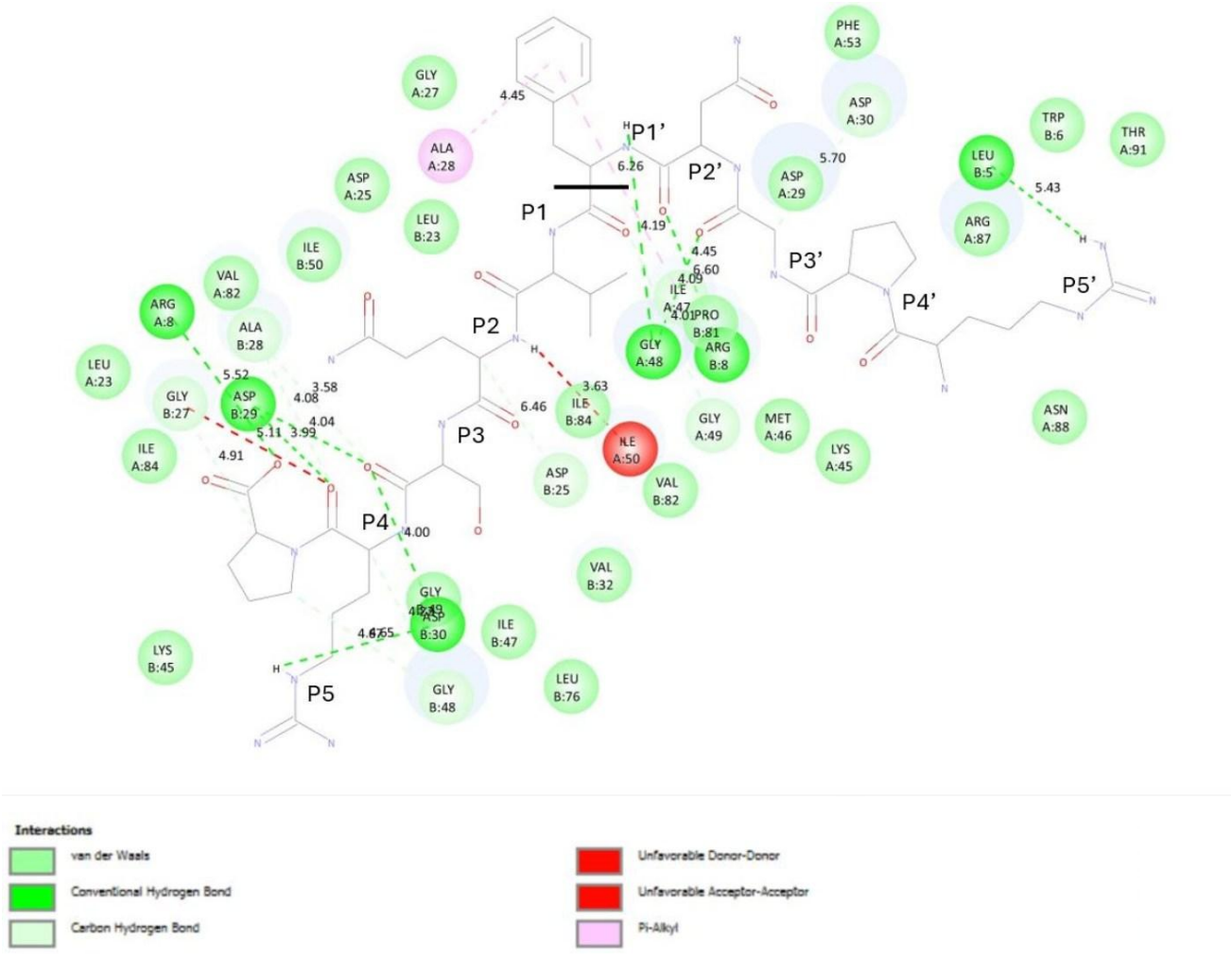


Figure 29: AA Interaction map of RPGNF/VQSRP CS ligands complexed with WT PR

The RPGNF/VQSRP-PR complex (Figure 29) had a binding affinity of -8.1 kcal/mol. This complex featured two π-alkyl interactions involving Ala28 and Ile47, with distances of 4.45 Å and 6.26 Å, respectively. There was an unfavourable donor-donor interaction with Ile50 at 3.63Å and an unfavourable acceptor-acceptor interaction with catalytic Gly27. Four hydrogen bonds were formed

with Gly49, catalytic D25, and Ala28 at distances between 3.58 Å and 6.46 Å. Nine conventional hydrogen bonds provided further stabilization, involving residues such as Arg8, Asp29, Gly29, Asp30, Gly48, Leu5, and Pro81, with distances ranging from 3.99 Å to 6.60

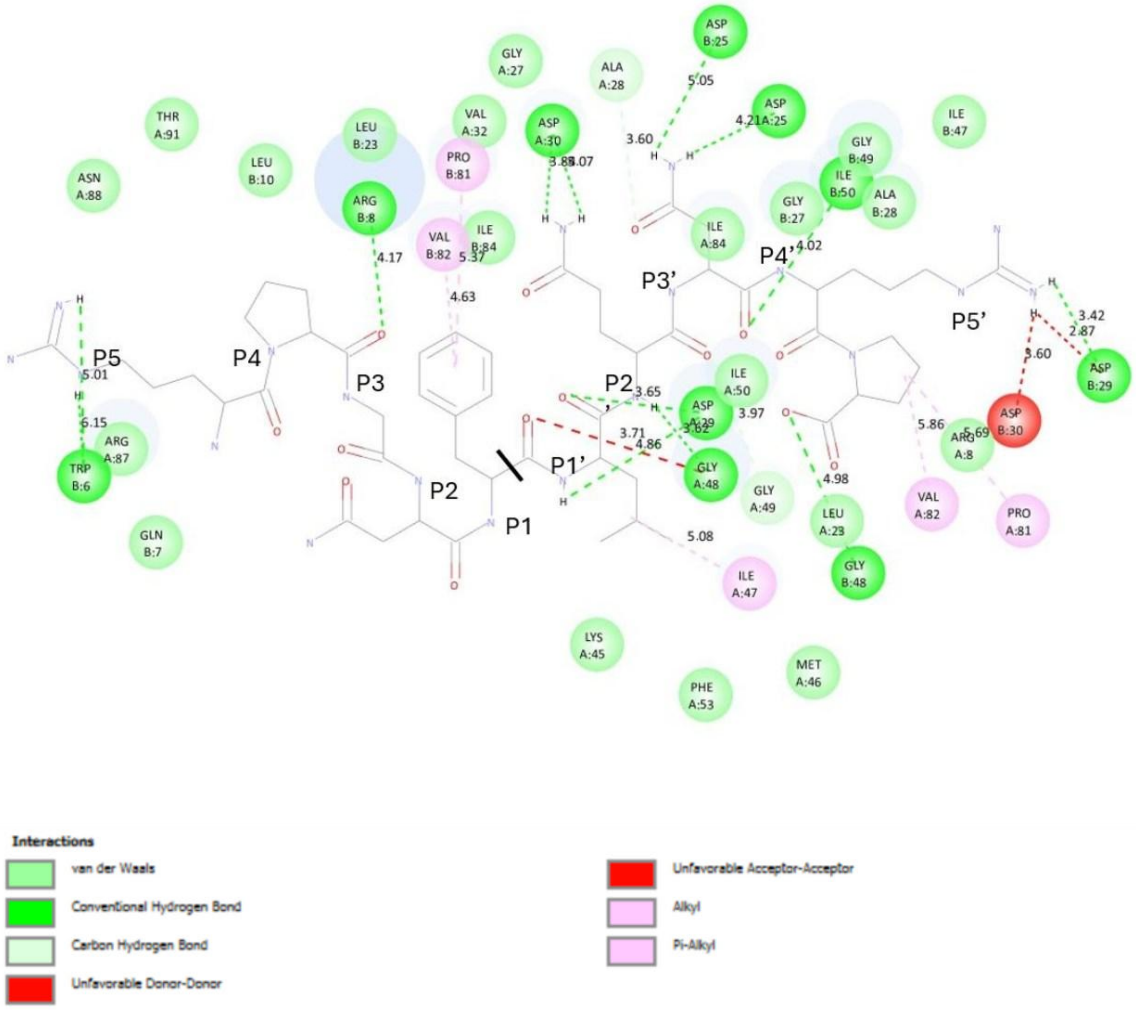


Figure 30: AA Interaction map of RPGNF/LQNRP CS ligand complexed with WT PR

The RPGNF/LQNRP-PR complex (Figure 30) had a binding affinity of -7.9 kcal/mol. This site included four π-alkyl interactions with Val82 and Pro81 and their respective chain A counterparts, at distances ranging from 4.63 Å to 5.95 Å. An alkyl interaction with Ile47 at 5.03 Å added to the stabilization. However, there were unfavourable acceptor-acceptor interactions with Gly48 at 3.21 Å and donor-donor interactions involving Asp30 and Asp23 at distances of 2.02 Å and 2.87 Å, respectively. Two hydrogen bonds with Ile50 and Ala28 at distances of 3.06 Å and 3.97 Å enhanced stability. Thirteen conventional hydrogen bonds involved residues such as Trp6, Arg8, Asp30, catalytic D25, Ile50, Asp29, Gly48, and Gly43, with distances ranging from 3.01 Å to 6.15 Å.

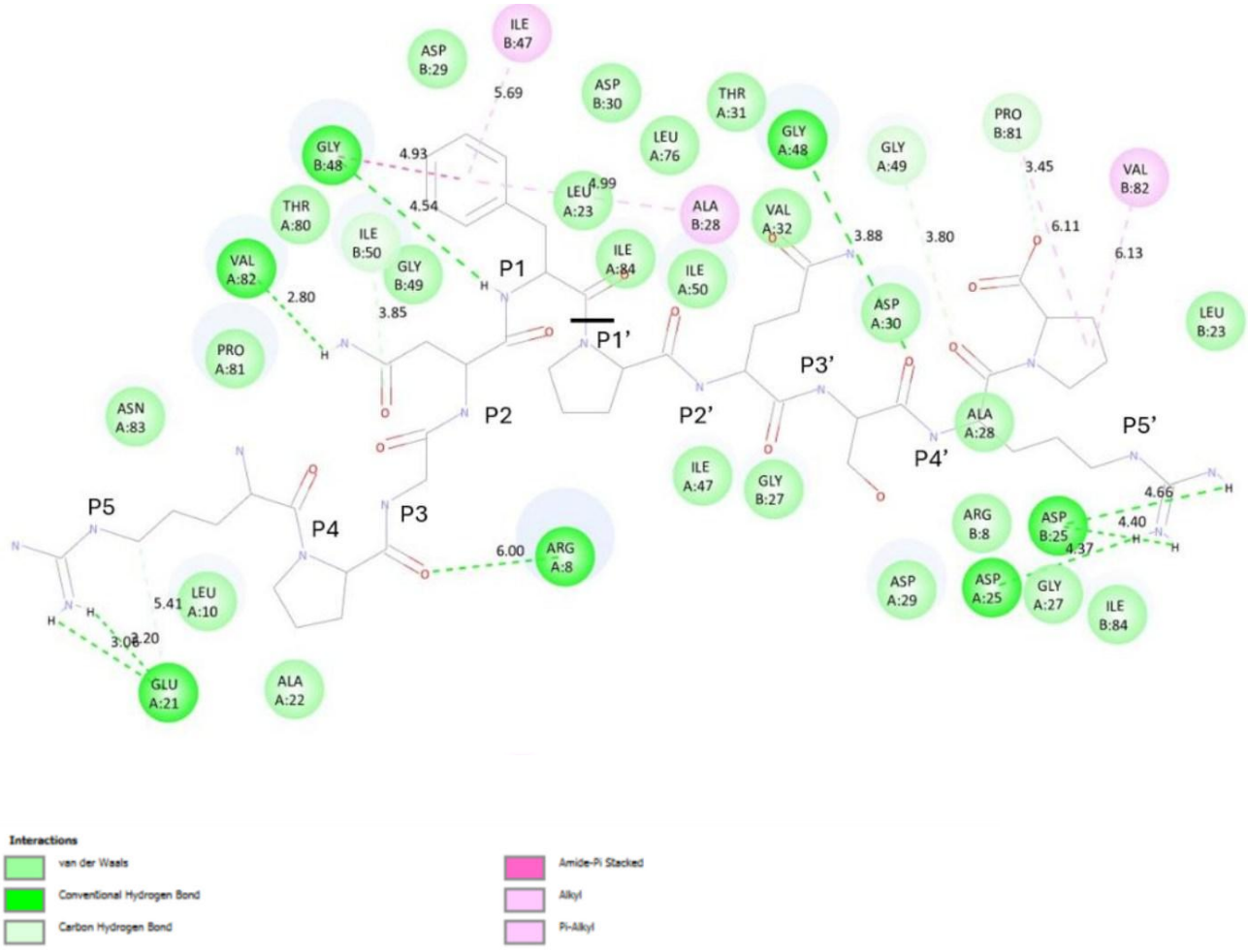


Figure 31: AA Interaction map of RPGNF/PQSRP CS ligand complexed with WT PR

The RPGNF/PQSRP-PR complex (Figure 31) had a binding affinity of -7.1 kcal/mol. This sequence featured an amide-π stacked interaction with Gly48 at 4.93 Å. Two alkyl interactions with Val82 and Pro81 at around 6.11 Å provided additional stabilization. π-Alkyl interactions with Ile47 and Ala28 at distances of 4.99 Å to 5.69 Å further contributed to the binding strength. Three hydrogen bonds with Ile50, Gly49, and Glu21 at distances between 3.80 Å and 5.41 Å enhanced the stability. Eight conventional hydrogen bonds involved residues such as Glu21, Arg8, Val82, Gly48 and the catalytic D25, with distances from 2.20 Å to 6.00 Å, providing substantial stabilization.

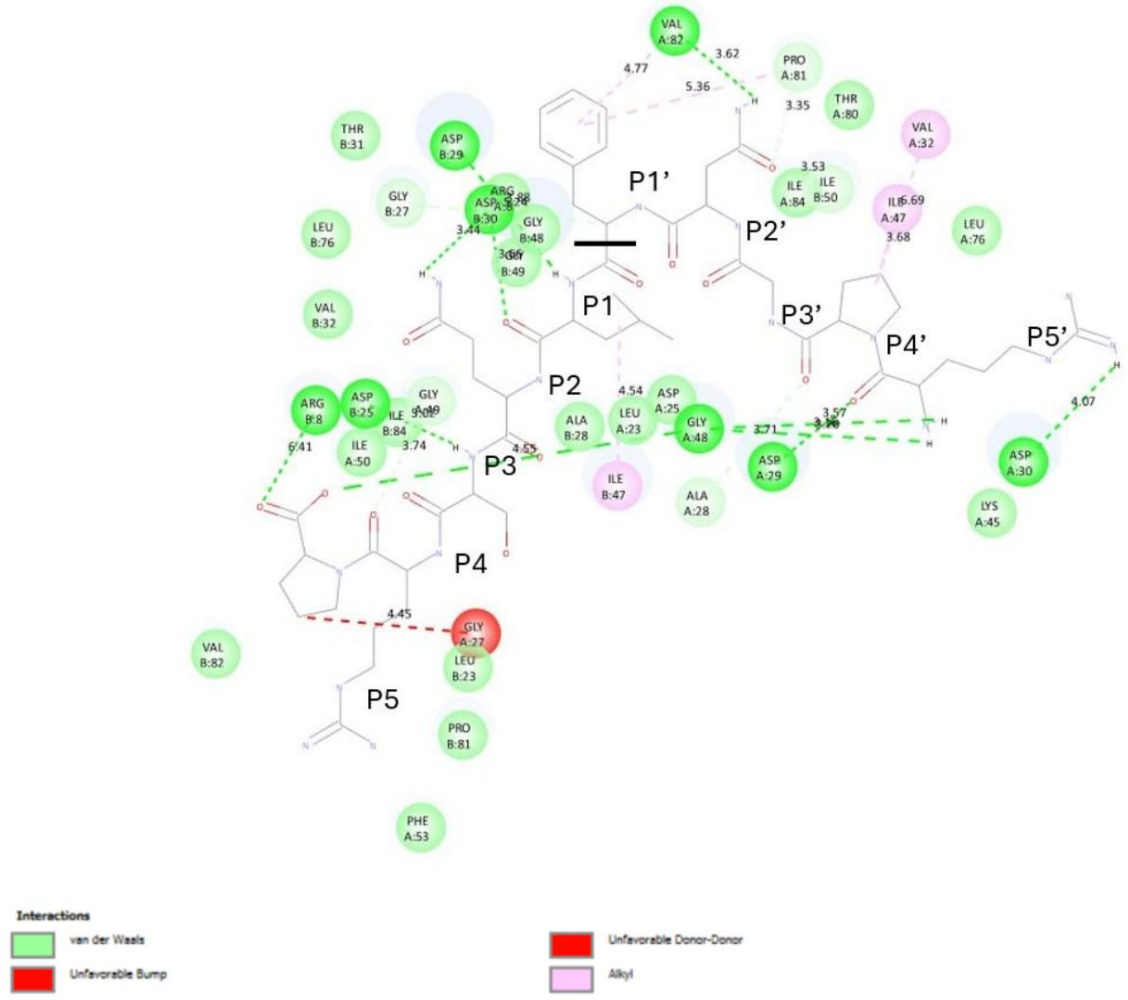


Figure 32: AA Interaction map of RPGNF/LQSRP CS ligand complexed with WT PR

The RPGNF/LQSRP-PR complex (Figure 32) has a binding affinity of -5.0 kcal/mol. This sequence was characterized by two π-alkyl interactions with Val82 and Pro81, with distances of 4.77 Å and 5.36 Å. Three alkyl interactions involve Val32, Ile87, and Ile47, at distances between 3.06 Å and 6.09 Å, further stabilizing the binding. However, there is an unfavourable bump with Gly22 at 4.65 Å. Three hydrogen bonds with Pro81, catalytic Gly27, and Ala28, ranging from 3.35 Å to 5.71 Å, contributed to the stability. Eleven conventional hydrogen bonds involved residues such as Arg8, catalytic D25, Gly48, Asp29, and Asp30, with distances from 3.54 Å to 6.41 Å, providing significant stabilization despite some unfavourable interactions.

3.3.5 Interactions between *gag* CS and PR in HIV-1 subtype B

Next, the interaction of the common *gag* CS sequences with PR in HIV-1 subtype B was examined and then compared to the respective common sequence in HIV-1 subtype C.

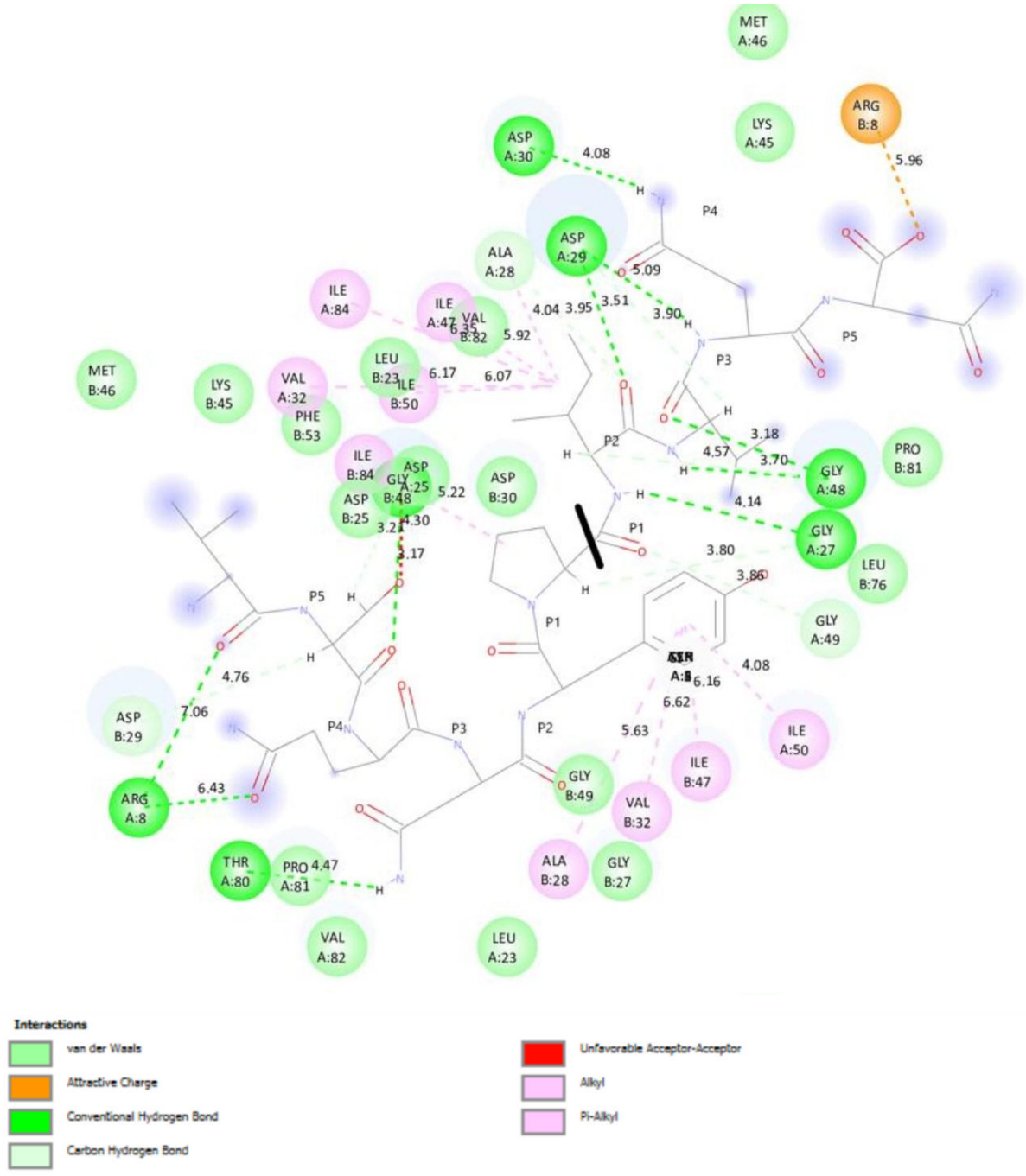


Figure 33: AA Interaction Map of the VSQNY/PIVQN CS in subtype B complexed with WT PR

3.3.5.1 MA/CA interactions with PR in HIV-1 Subtype B:

In subtype B (VSQNY/PIVQN-PR) has a binding affinity of -7.8 kcal/mol. The interaction (Figure 33) revealed a complex network of bonding types and affinities. The most prominent bond type observed was the hydrogen bond, which included both conventional and catalytic interactions. Specifically, there were a total of fifteen hydrogen bonds with varying distances and binding affinities. Notably, Asp29 and Ala28 were involved in multiple hydrogen bonds. The catalytic residues Gly27 and Asp29 demonstrated critical binding with distances as short as 3.21 Å and 3.80 Å, respectively.

In addition to hydrogen bonds, the interaction featured an attractive charge bond involving Arg8 with a distance of 5.96 Å and a binding affinity of -7.8 kcal/mol, suggesting a significant electrostatic interaction. However, an unfavourable acceptor-acceptor interaction was also observed with Gly48 at a distance of 3.17 Å, which could imply steric hindrance or repulsion in the binding site.
The alkyl interactions were also notable, with five residues (Val32, Ile50, Ile84, Ile47, Ala28) participating in these hydrophobic contacts. These interactions varied in distance from 4.04 Å to 6.35 Å, with a general trend indicating a moderate to strong affinity, particularly between Ala28 and the other alkyl residues. Finally, π-alkyl interactions further contributed to the binding, with Ala28, Val32, Ile47, Ile50, and Ile84 showing distances ranging from 4.08 Å to 6.62 Å. The presence of these non-polar interactions underscored the importance of hydrophobic forces in stabilizing the MA/CA complex. Overall, the results illustrated a multifaceted binding landscape characterized by a combination of hydrogen bonds, electrostatic interactions, and hydrophobic contacts, each contributing to the stability and specificity of the protein interaction.

3.3.5.1.1 Comparing MA/CA-PR in Subtypes B vs C

Both maps highlighted the importance of hydrogen bonds in the interaction. Subtype B indicated a significant number of hydrogen bonds, with distances ranging from 3.18 Å to 4.76 Å, while subtype C also had hydrogen bonding with distances between 2.04 Å and 3.10 Å. Notably, residues like Asp29 and Arg8 featured prominently in both datasets, indicating their central role in stabilizing the interaction. Additionally, both subtypes had multiple conventional hydrogen bonds involving residues such as Asp29 and Asp30, demonstrating the continued relevance of these bonds in stabilizing the complex. In subtype C, the distances were notably shorter, suggesting stronger interactions.

Subtype C introduced additional bond types that were not present in subtype B, which highlighted key differences in the interaction profiles between the two. One major difference is the presence of multiple π-alkyl interactions in subtype B vs C, which was characterized by a single π-alkyl involving a benzene ring and Val82 and having notable binding affinity of -8.3 kcal/mol at a distance

of 4.71 Å. The inclusion of these π-alkyl interactions suggested a significant role for aromatic residues in this new binding context, contributing to the stability and specificity of the interaction. Additionally, subtype C featured a π-cation interaction involving a benzene ring and the catalytic residue Gly27 at a distance of 4.16 Å. This newly introduced bond type indicated the presence of a specific electrostatic interaction that was not previously observed in subtype B, further differentiating the binding characteristics of subtype C.

Subtype C had an unfavourable donor-donor interaction involving Arg8 and Ile47 with distances of 2.22 Å and 2.48 Å, respectively. These interactions, which were not present in subtype B, suggesting potential steric clashes or repulsive forces that may influence the overall binding stability. Subtype B introduced interactions with additional residues like Arg87 and Trp6, which were not present in subtype C. The distances for these new interactions were also shorter, suggesting potentially more favourable or competitive binding scenarios.

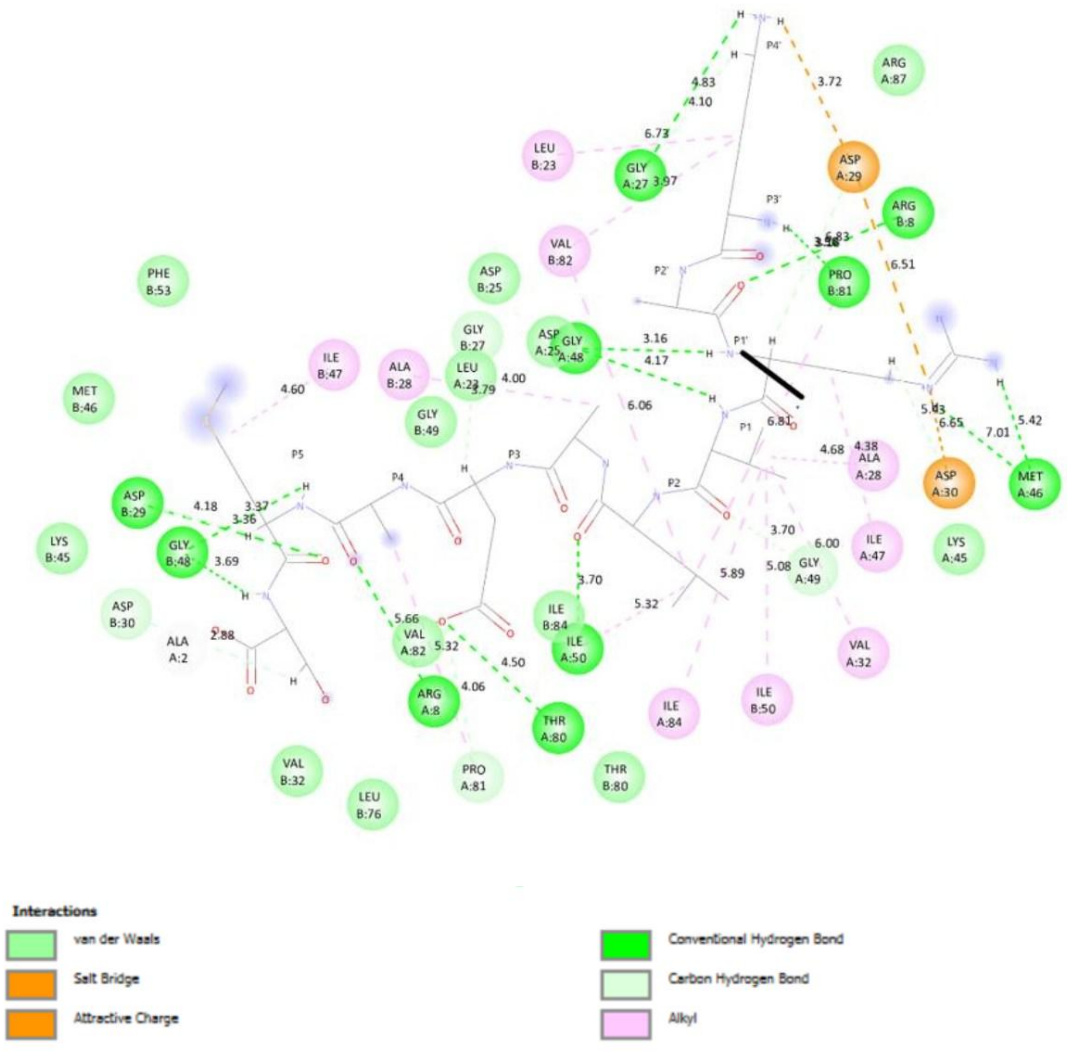


Figure 34: AA Interaction map of the KARVL/AEAMS CS in subtype B complexed with WT PR

3.3.5.2 *CA/P2 interactions with HIV-1 PR in Subtype B*

In subtype B (KARVL/AEAMS-PR) (Figure 34), has a binding affinity of -7.7 kcal/mol. A total of eleven alkyl interactions were observed, highlighting the significant role of hydrophobic contacts in subtype B. Key residues such as Ile50, Ile84, Val32, Ile47, Ala28, and Pro81 were involved, with distances ranging from 3.79 Å to 6.73 Å. The binding affinity for these interactions, specifically with Ile50 and Ile47, indicated their crucial role in anchoring the proteins together, with Ile50 and Pro81 showing notable binding strengths. The extensive network of alkyl interactions underscored the importance of hydrophobic effects in stabilizing the protein complex.
There were seven hydrogen bonds with distances between 2.88 Å and 5.43 Å. Prominent interactions involved Asp29 and Gly49, where Asp30 (Chain B) formed a particularly short and potentially strong bond at 2.88 Å. This suggested a highly favourable hydrogen bonding environment. The presence of residues such as Ala28 and Leu23 further contributed to the stability, indicating that hydrogen bonding was a critical component of the interaction.

There were fifteen conventional hydrogen bonds with distances ranging from 3.16 Å to 7.01 Å. Notably, Gly48 and Asp29 played significant roles in these interactions, with Gly48 forming bonds at shorter distances (as low as 3.16 Å), suggesting strong binding. The interactions with residues such as Ile50 and Met46, though at longer distances, still contributed to the overall stability. The presence of longer-range interactions with residues such as Pro81 and Arg8 indicated a diverse bonding landscape that supported the integrity of the complex. Two attractive charge interactions were identified involving Asp29 and Asp30. With distances of 4.52 Å and 5.43 Å, these electrostatic attractions further enhanced the binding affinity, complementing the hydrophobic and hydrogen bonding interactions.

3.3.5.2.1 Comparing CA/P2-PR in Subtypes B vs C

Subtype C showed fewer alkyl interactions (two in total) compared to B which had a broader range of alkyl interactions. Val32 and Pro81 were common in both maps, indicating their recurring role in hydrophobic interactions. The Subtype C map showed a range of hydrogen bonds with distances from 3.54 Å to 5.20 Å, with some overlaps in residues like Gly49. Both maps emphasized hydrogen bonding as a crucial interaction, though subtype C featured fewer hydrogen bonds in total. Subtype C featured thirteen conventional hydrogen bonds with distances ranging from 3.16 Å to 5.60 Å. This map included specific residues such as catalytic D25, Gly48, and Asp30, which were also present in

subtype B. The distances in subtype C were slightly shorter on average compared to B, indicating potentially stronger or more stable interactions.

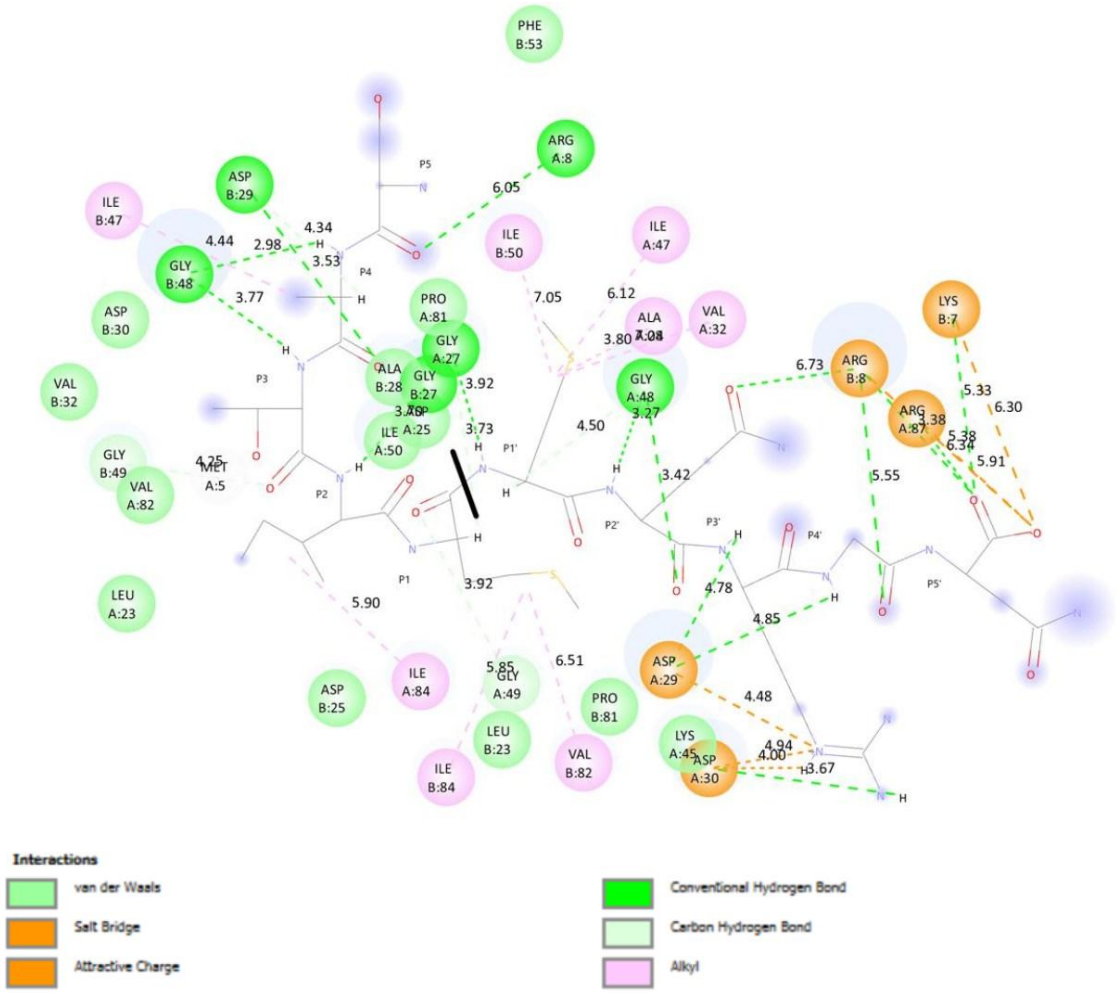


Figure 35: AA Interaction map of the SATIM/MQRGN CS in subtype B complexed with WT PR

3.3.5.3 P2/NC interactions with HIV-1 PR in Subtype B

In subtype B (SATTIM/MQRGN-PR) (Figure 35), has a binding affinity of -7.2kcal/mol. A prominent salt bridge involving Asp29 with a distance of 4.48 Å was observed. This interaction suggested a key electrostatic contribution to the stability of the complex. In addition, three attractive charge interactions were identified: Asp30 at 4.00 Å, Arg8 at 5.38 Å, and Arg87 at 5.91 Å. These residues created electrostatic attractions that stabilized the interaction through ionic interactions between negatively charged and positively charged residues.

Hydrogen bonding was a critical component, with four key hydrogen bonds identified. These included Asp29 (4.34 Å), Gly49 ,catalytic Gly27and Gly48 .These hydrogen bonds involved both backbone and side-chain interactions, and their distances suggested strong but varied bonding strengths, enhancing the overall stability and precision of the interaction.

Eight significant alkyl interactions were observed, emphasizing the role of hydrophobic interactions in stabilizing the complex. The distances for these alkyl interactions spanned from 3.80 Å to 7.05 Å, indicating a range of hydrophobic contact strengths that contributed to the overall binding affinity and specificity. A comprehensive network of twelve conventional hydrogen bonds were identified, which underscored the value of these interactions in strengthening the complex. The varying distances of these conventional hydrogen bonds, from as short as 2.98 Å to longer distances of 6.73 Å, reflected both strong and moderate binding interactions that collectively contributed to the stability of the complex. Salt bridges and attractive charge interactions provided significant electrostatic stabilization, while hydrogen bonds and alkyl interactions contributed to the structural integrity of the complex. The network of conventional hydrogen bonds further supported the binding, with a range of distances indicating diverse interaction strengths.

3.3.5.3.1 Comparing P2/NC-PR in Subtypes B vs C

Subtype B exhibited a greater number of alkyl interactions (eleven in total) with a broader range of distances. The residues involved are similar but include additional ones such as Ile50, Ile47, and Val32. Subtype B showed seven hydrogen bonds with distances ranging from 2.88 Å to 5.43 Å, which involved residues such as Asp29 and Gly49. Subtype B featured fifteen conventional hydrogen bonds with a wide range of distances from 3.16 Å to 7.01 Å, involving residues such as Gly48 and Asp29.

Subtype C had fewer alkyl interactions observed, with a focus on key residues such as Val82 and Ile84. The distances for alkyl interactions in subtype C were more consistent and slightly shorter, indicating potentially stronger and more localized hydrophobic interactions compared to subtype B. Subtype C had three hydrogen bonds with slightly longer distances, indicating a variation in bonding strength. Subtype C contained nine conventional hydrogen bonds with distances from 2.98 Å to 6.73 Å. The distances were generally shorter compared to subtype B, suggesting potentially stronger binding interactions in certain regions. Within subtype C, unique to this dataset were two unfavourable donor-donor interactions, which were absent in the subtype B map. This could indicate potential steric or electrostatic clashes within the complex.

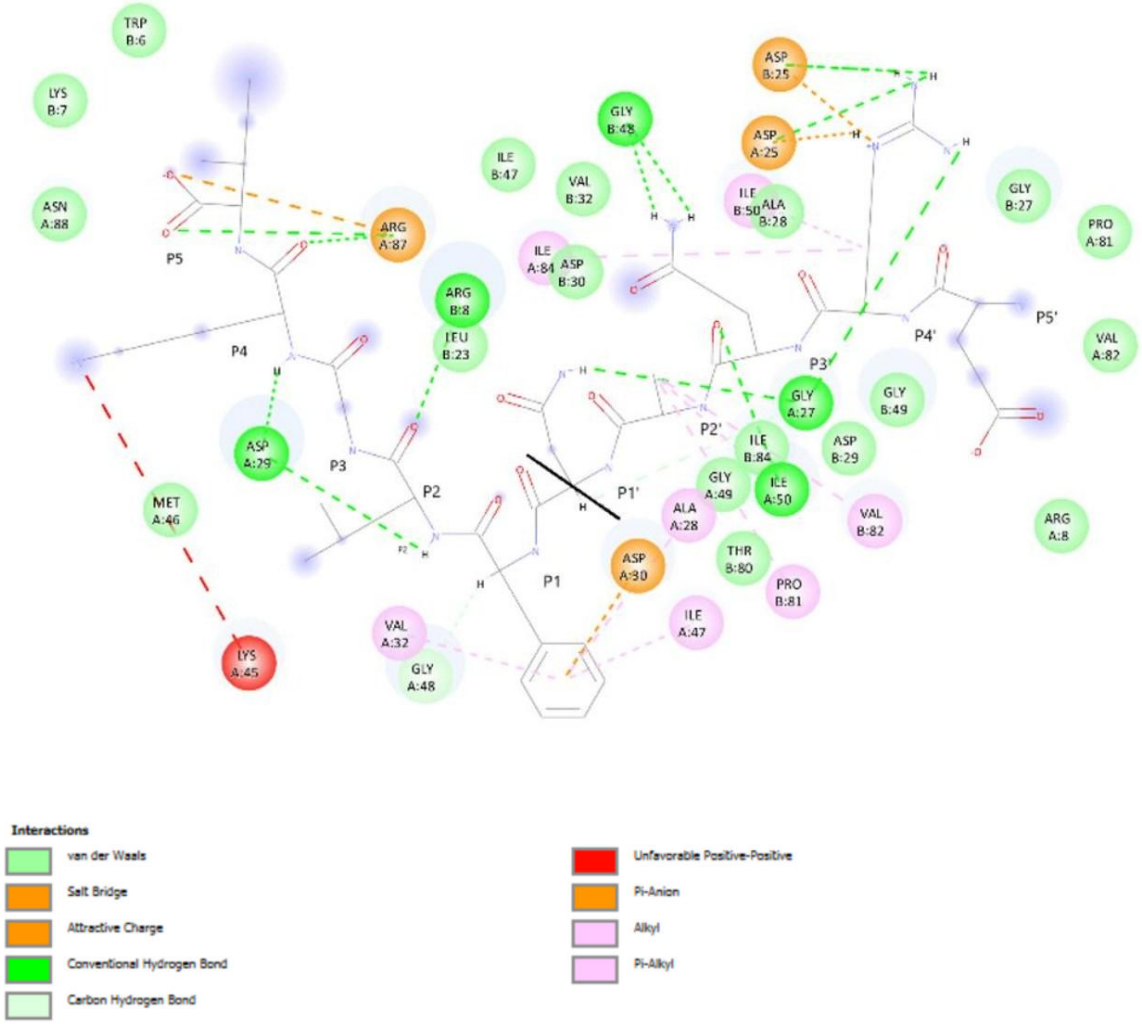


Figure 36: AA Interaction map of the ERQAN/FLGKI CS in subtype B complexed with WT PR

3.3.5.4 NC/P1 interactions with HIV-1 PR in Subtype B

In subtype B (ERQAN/FLGKI-PR) (Figure 36), has a binding affinity of -7.8kcal/mol. One of the primary interaction types observed is the salt bridge, particularly involving D25. The distance between D25 and position P3' was 4.73 Å, indicating a strong attractive charge interaction. Conversely, lysine residue Lys45, with a distance of 7.71 Å to position P4, demonstrated an unfavourable positive-positive repulsion, which may influence binding stability negatively. Additionally, subtype B exhibited significant π-based interactions. The π-anion interaction between Ala28 and position P1, with a distance of 4.75 Å, contributed positively to the peptide's stability.

Similarly, π-alkyl interactions which involved Ile47 and Val32 with P1 at distances of 4.91 Å and 7.14 Å respectively, were crucial for maintaining the hydrophobic core of the ligand.
Hydrogen bonding was another key feature of the ligand's interaction profile. Conventional hydrogen bonds involving Asp29 with positions P2 and P4, Gly48 with positions P1 and P2, and Arg8 with positions P1 and P2, contributed significantly to the structural stability of the ligand. Notably, the interaction between catalytic Gly27 and positions P4' is observed with distances of 4.74 Å and 4.40 Å, underscoring its role in maintaining the ligand's conformation. Furthermore, the specific interactions of catalytic residues were critical. D25 in both Chain A and Chain B showed interaction distances of 4.73 Å and 5.16 Å to position P3', which highlighted its important role across different chains. Gly27 interacted with position P4' which could further support the ligand's functional conformation.

3.3.5.4.1 Comparing NC/P1-PR in Subtypes B vs C

Both subtype B and subtype C of the NC/P1 ligand demonstrated a range of interactions that contributed to their stability and function. However, there were notable differences in the nature and distribution of these interactions. Subtype C showed π-alkyl interactions involving residues Ile50, Ile47, and Ala28 with position P1, at distances of 6.83 Å, 5.34 Å, and 4.82 Å, respectively. These interactions were relatively similar to those observed in subtype B, where π-alkyl interactions were also noted with residues such as Ile47 and Val32 with P1, though the distances (Ile47: 4.91 Å and Val32: 7.14 Å) differ slightly. This indicated that both subtypes maintained a hydrophobic core around P1, albeit with slight variations in interaction strength and distances.

An unfavourable donor-donor interaction was observed in subtype C between Ile50 (Chain B) and P2' with a distance of 4.11 Å. This type of interaction was not prominently noted in the subtype B data set, which suggested that subtype C may have some destabilizing forces that could affect ligand stability in this region. Both subtypes show a significant number of conventional hydrogen bonds, but with differences in their distribution and interacting partners. Subtype C had six hydrogen bonds involving residues such as Asp29 (interacting with P5, P3, and P2), Arg8 (with P3), TrP6 (with P4), and Asp30 (with P3'). The distances of these bonds range from 2.51 Å to 6.58 Å, with Asp30 forming a particularly short hydrogen bond at 2.51 Å with P3'. In contrast, Subtype B also demonstrated a high number of hydrogen bonds, with residues such as Asp29, Gly48, and Arg8 interacting with positions P2, P4, P1 and P2 at distances ranging from 3.22 Å to 6.98 Å. Subtype B showed a broader distribution of these bonds across different positions, highlighting its strong hydrogen-bonding network.

Both subtypes emphasized the role of catalytic residues in their interaction profiles. In Subtype B, catalytic residues such as D25 and G27 were frequently involved in interactions, such as hydrogen bonding and attractive charge interactions, which were crucial for maintaining the peptide's functional integrity. These residues were less prominently mentioned in subtype C, indicating that while they may have played a role, other interactions, such as those involving Asp29 and Asp30, may be more dominant in this subtype.

The differences in interaction types and distances between the two subtypes suggested potential variations in their structural stability and functional conformations. Subtype B, with its more diverse set of hydrogen bonds and well-distributed interactions across different positions, might possess a more stable and rigid conformation. Meanwhile, Subtype C's presence of unfavourable donor-donor interactions and shorter hydrogen bonds, particularly with Asp30, could suggest a different folding pattern or local flexibility, potentially affecting its binding affinity and interaction with other molecules.

While both subtype B and subtype C shared common types of interactions, such as π-alkyl interactions and hydrogen bonds, the specifics of the interaction distances, involved residues and the presence of unfavourable interactions varied. Subtype B appeared to have a more robust network of stabilizing interactions, while Subtype C exhibited some unique characteristics, such as the unfavourable donor-donor interaction and a distinct pattern of hydrogen bonding, which might have impacted its structural dynamics differently.

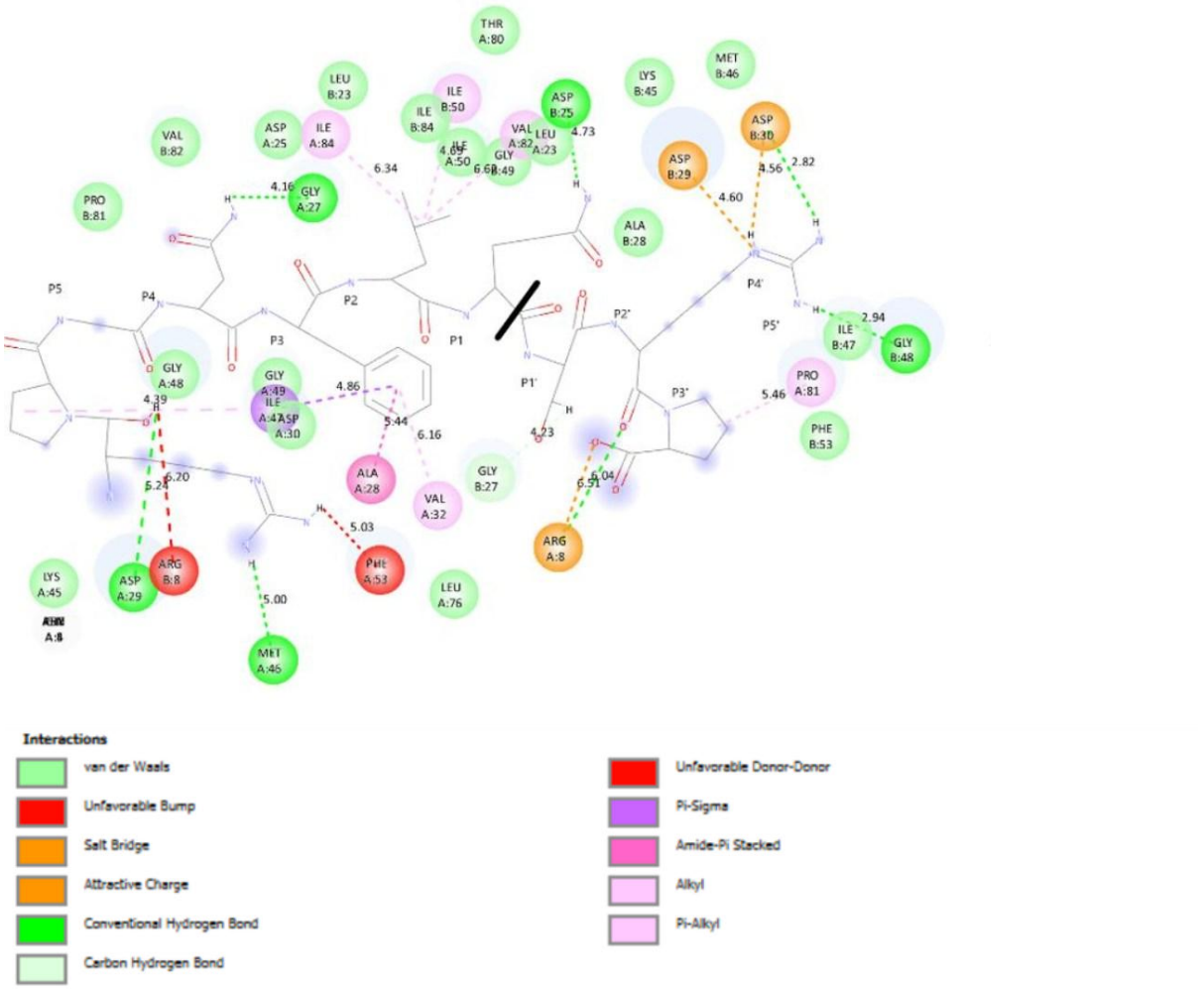


Figure 37: AA Interaction map of RPGNF/LQSRP CS in subtype B complexed with WT PR

3.3.5.5 P1/P6 interactions with HIV-1 PR in Subtype B

In subtype B (RPGNF/LQSRP) (Figure 37), had a binding affinity of -5.0kcal/mol. The interaction analysis revealed interactions that contributed to its structural stability and functional integrity. A notable feature was the presence of salt bridges, with Asp29 and Asp30 forming strong interactions with position P4' at distances of 4.60 Å and 4.56 Å, respectively. These salt bridges played a crucial role in maintaining the ligands' overall charge balance and structural conformation. Additionally, an attractive charge interaction is observed between an Arg8 residue and position P3' at a distance of 6.51 Å, further stabilizing the peptide through electrostatic interactions. The peptide also exhibited a significant number of conventional hydrogen bonds, with nine identified interactions. Key residues included Asp29 (interacting with P3 at 5.24 Å), Met46 (P5 at 5.00 Å), and Arg8 (P3' at 6.04 Å). Gly48 formed multiple hydrogen bonds, notably with P5' (2.94 Å) and P5 (3.19 Å) in Chain A,

underscoring its importance in ligand stabilization. The catalytic residues, such as D25 and G27, were also actively involved, with G27 forming hydrogen bonds with P1 (4.16 Å) and P2' (4.23 Å), and D25 interacting with P2' at 4.73 Å. Additionally, Asp30 formed a particularly strong hydrogen bond with P5' at a distance of 2.82 Å.

Other significant interactions included three hydrogen bonds involving residues such as Gly49 (P3 at 3.54 Å) and Ala28 (P1' at 3.91 Å), along with one amide-π stacked interaction between Ala28 and P1 at a distance of 5.44 Å. Hydrophobic interactions were also prominent, where alkyl and π-alkyl interactions contributed to the ligand's stability. Residues such as Ile47 (Chain A) and Ile84 (Chain A) interacted with P1 at distances of 4.86 Å and 3.84 Å, respectively, while Val82 interacted at 6.62 Å. π-Alkyl interactions were noted with Ile47 and Val32 at distances ranging from 4.39 Å to 6.16 Å, indicating a strong hydrophobic core that is critical for maintaining the peptide's structure.

3.3.5.5.1 Comparing P1/P6-PR IN Subtypes B vs C

π-Alkyl interactions were present in both subtypes, indicating the importance of hydrophobic interactions in maintaining binding stability. In subtype C, π-alkyl interactions were observed between Ala28 and Ile47 with P1' at distances of 4.45 Å and 6.26 Å, respectively. This contrasted with subtype B, where similar interactions were also found involving Ile47 (Chain A) with P1 at 4.86 Å and Val32 with P2 at 6.16 Å. While both subtypes featured π-alkyl interactions, the interacting residues and distances differed, suggesting subtle variations in the hydrophobic core organization between subtypes B and C.

Subtype C showed an unfavourable donor-donor interaction between Ile50 and P1 at a short distance of 3.63 Å and an unfavourable acceptor-acceptor interaction involving the G27 with P3. These types of destabilizing interactions were minimal or absent in subtype B, where no significant unfavourable donor-donor or acceptor-acceptor interactions were observed. This indicated that subtype C might contain repulsion effects that were not prominent in subtype B, potentially influencing its conformational flexibility and stability.

Hydrogen Bonds were a significant stabilizing force in both subtypes, though their distribution and interacting residues varied. Subtype C had four key hydrogen bonds, including Gly49 with P1 at 4.01 Å, D25 with P1' at 6.46 Å and Ala28 with P3 at distances of 3.58 Å and 4.08 Å. Subtype B, in comparison, demonstrated a more extensive hydrogen-bonding network, with nine conventional hydrogen bonds involving residues such as Gly48, Arg8, Asp29, and catalytic G27. The distances of these hydrogen bonds ranged from 2.82 Å to 6.04 Å, indicating strong and varied hydrogen-bonding interactions. This robust hydrogen-bonding network in subtype B suggested a more stabilized and

rigid peptide conformation compared to subtype C, which displayed fewer hydrogen bonds with more localized distribution.

Furthermore, conventional hydrogen bonds in subtype C were formed by residues such as Arg8, Asp29, Gly29, Asp30, Gly48, Leu5, and Pro81, contributing to the ligand's structural stability across various positions (distances ranging from 3.99 Å to 6.60 Å). In contrast, subtype B featured conventional hydrogen bonds involving D25 and G27, as well as others such as Gly48, which played crucial roles in maintaining binding integrity. The broader range of distances and the involvement of multiple catalytic residues in subtype B suggested a more diversified hydrogen-bonding network that could contribute to stronger structural stabilization.

In conclusion, the overall comparative analysis revealed that while both subtypes B and C utilized similar interaction types, including π-alkyl interactions and hydrogen bonds, their specific residues, distances, and interaction patterns differed. Subtype B showed a broader and more varied interaction network, particularly with its extensive hydrogen-bonding patterns and fewer unfavourable interactions, likely resulted in a more stable conformation. Subtype C, on the other hand, exhibited a distinct interaction profile with several unfavourable donor-donor and acceptor-acceptor interactions and a narrower hydrogen-bonding distribution, potentially leading to different folding dynamics and local structural flexibility.
Subtype B's robust and well-distributed interactions imply a stable, rigid conformation, while subtype C's interactions indicate potential flexibility or local instability due to unfavourable interactions, providing insights into the distinct structural characteristics of each subtype.

Tables 4 and 5 (see appendix) show a detailed analysis of the molecular interactions between HIV-1 PR and its various CS sequences during polyprotein processing. Each CS was characterized by its unique peptide sequence and the types of molecular bonds formed with protease residues. These include hydrogen bonds (both conventional and unconventional), van der Waals forces, salt bridges, alkyl and π-alkyl interactions, as well as less favourable interactions such as donor-donor or positive-positive clashes. The data also outlines the number of bonds, key interacting PR residues, the distances between interacting atoms and binding affinities in kcal/mol where available. The ligand positions relative to the CS (P1, P1', P2', etc.) help map where each interaction occurs. A key finding was that the MA/CA CS demonstrated the strongest binding affinity, supported by numerous interactions including nine conventional hydrogen bonds and six van der Waals interactions. This site involved catalytic residues such as D25 and G27, along with structural residues such as Asp29, Gly48, and Arg8, indicating a highly stabilized and tightly bound substrate. On the other hand, the CA/P2 site, despite having a higher number of interactions (including eleven alkyl and fifteen

hydrogen bonds), showed a significantly weaker binding affinity . This suggests that the nature and positioning of interactions, rather than their sheer number, contribute more critically to binding strength.

Across all CSs, catalytic residues D25 and G27 consistently participate in key interactions, underscoring their central role in proteolytic activity. Similarly, residues like Asp29, Asp30 and Gly48 frequently appear in hydrogen bonding networks, highlighting their importance in substrate recognition and stabilization. Hydrophobic interactions such as alkyl and π-alkyl contacts were prominent at the P1 and P2' positions of many CS, contributing to the overall binding affinity through structural complementarity. Interestingly, some sites also exhibited unfavourable interactions (e.g., donor-donor or acceptor-acceptor), which may introduce local instability or reduce binding strength. In conclusion, while all CS interact with the protease through a mix of favourable and unfavourable contacts, the MA/CA site stands out as the most energetically favourable, likely due to optimal alignment with the catalytic core and a well-distributed network of stabilizing interactions.

3.3.6 The substrate Envelope

Figures 38-43 depicts the binding poses and interaction profiles of the superimposed CSs within the PR of HIV-1 subtype C, highlighting the substrate envelope. Figures (38-42) A shows the overall structure of the HIV-1 subtype C PR with the CSs bound within the active site. The PR is represented as a ribbon diagram, highlighting its secondary structural elements, with the bound substrate shown as a yellow or grey surface representation. This view provides a macroscopic perspective, emphasizing how the CS fits within the enzyme's active site. In Figures (38-42) B, the yellow surface represents the substrate envelope, while the residues are shown in stick format. Figures (38-42) C presents a superimposed view of each CS, allowing for a comparative analysis. The highlighted yellow structure represents the most common sequence.

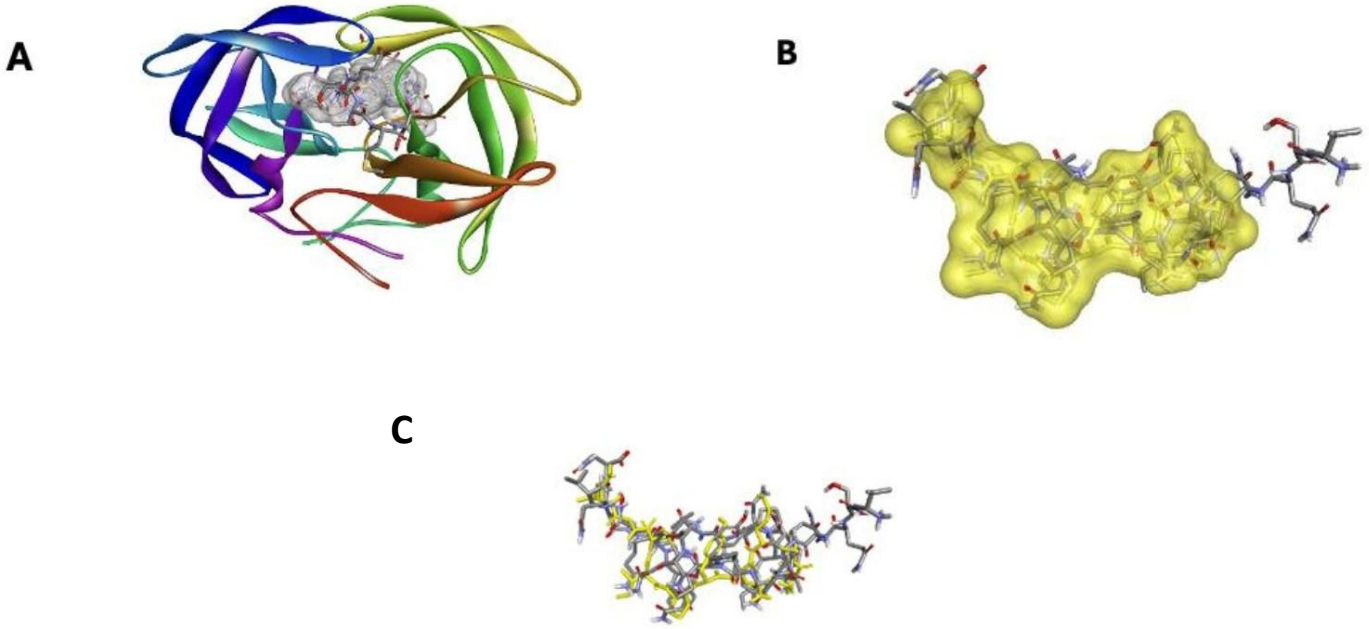


Figure 38: Superimposed subtype C MA/CA substrate envelope comparison

A) All four variants superimposed and bound to PR. B) Four variants superimposed and their substrate envelope shown in yellow. C) Four variants superimposed with the most common sequence highlighted in yellow.

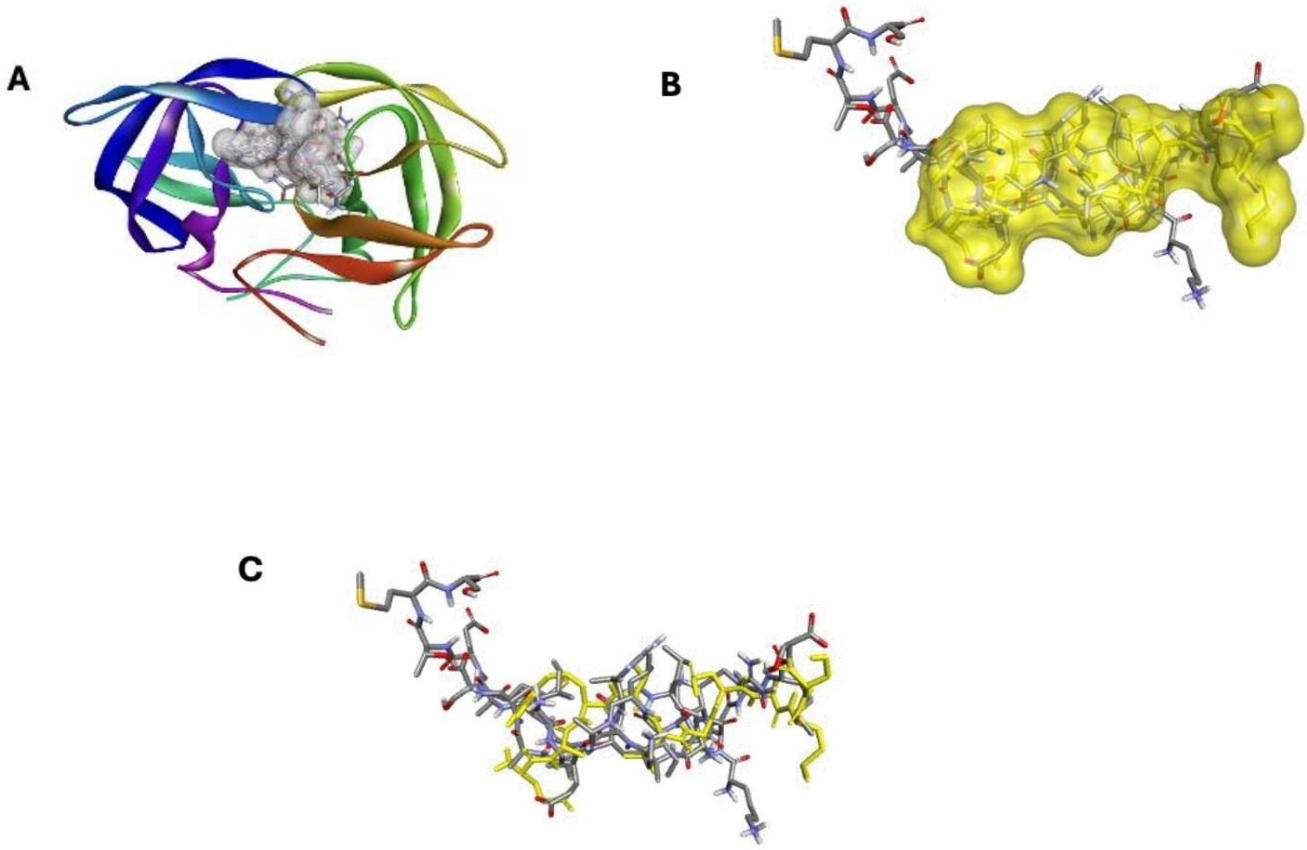


Figure 39: Superimposed subtype C CA/P2 substrate envelope comparison

A) All four variants superimposed and bound to PR. B) Four variants superimposed and their substrate envelope shown in yellow. C) Four variants superimposed with the most common sequence highlighted in yellow.

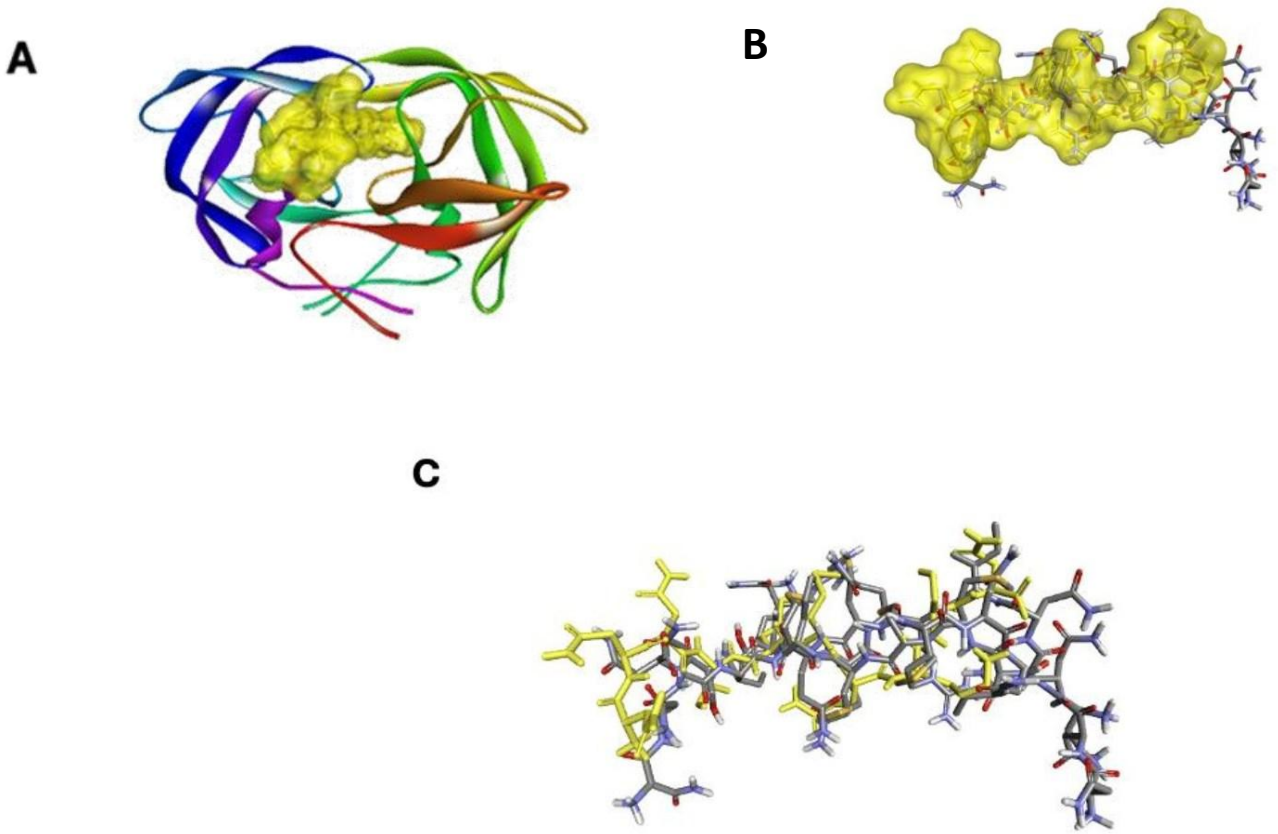
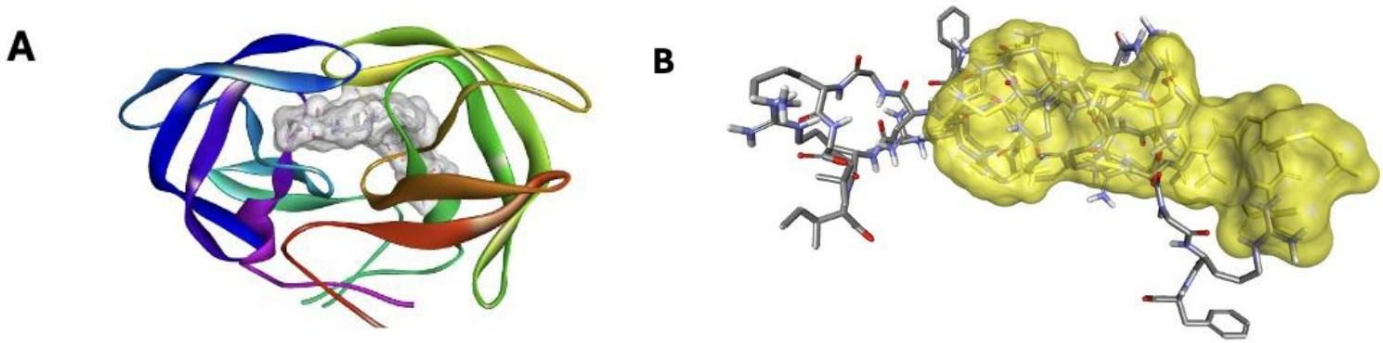


Figure 40: : Superimposed subtype C P2/NC substrate envelope comparison

A) All four variants superimposed and bound to PR. B) Four variants superimposed and their substrate envelope shown in yellow. C) Four variants superimposed with the most common sequence highlighted in yellow.



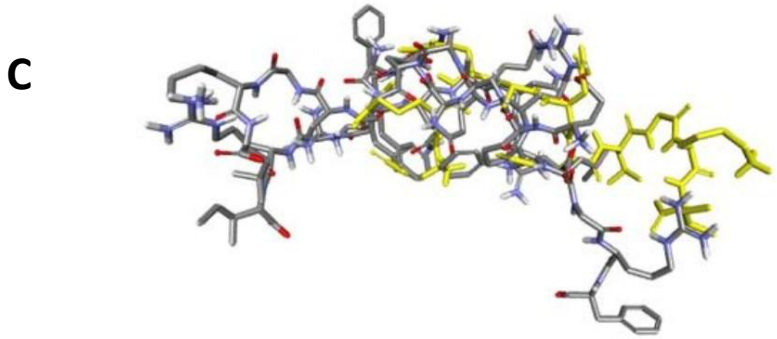


Figure 41: Superimposed subtype C NC/P1 substrate envelope comparison

A) All four variants superimposed and bound to PR. B.) Four variants superimposed and their substrate envelope shown in yellow. C) Four variants superimposed with the most common sequence highlighted in yellow.

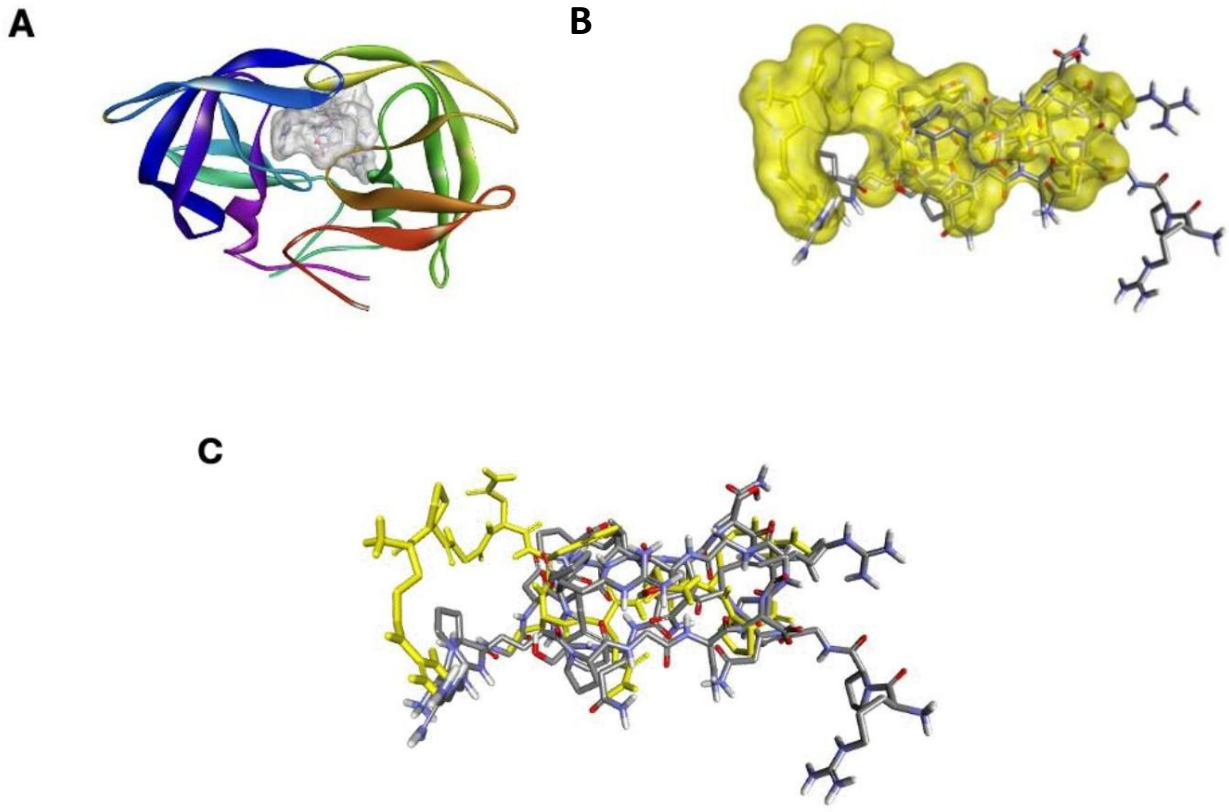


Figure 42: Superimposed subtype C P1/P6 substrate envelope comparison

A) All four variants superimposed and bound to PR. B.) Four variants superimposed and their substrate envelope shown in yellow. C) Four variants superimposed with the most common sequence highlighted in yellow.

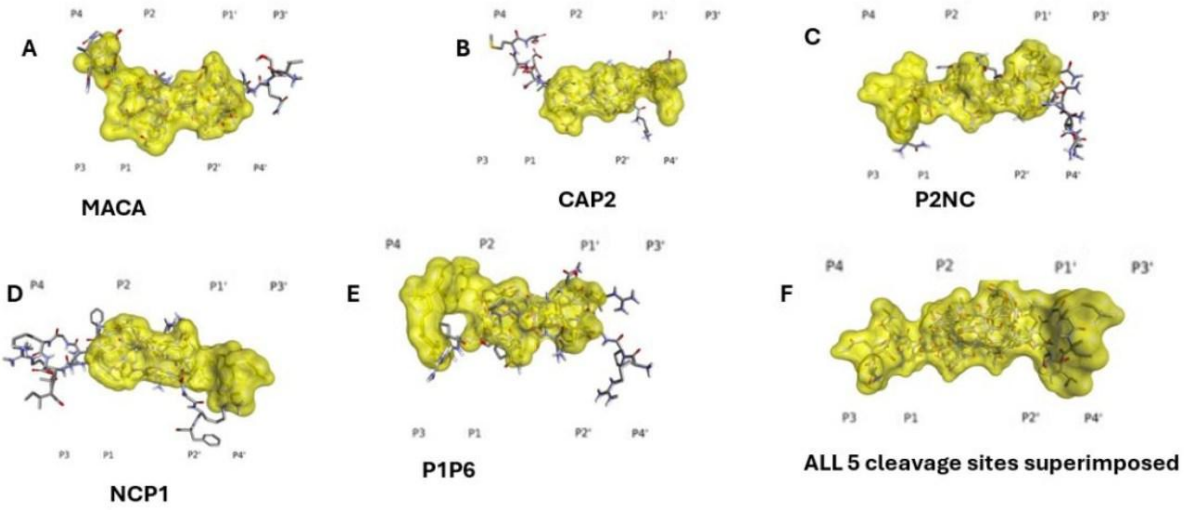


Figure 43: Comparison of the five-*gag* subtype C CS substrate envelopes and the substrate envelope for all CS superimposed

The MA/CA CS (Figure 43A) presented a detailed view of the binding interactions within the substrate envelope. This site showcased a strong fit, particularly at the P4 to P2' positions, where the potential interaction density is highest. The structural alignment of the substrate within the PR active site implies that MA/CA may have a higher binding affinity, potentially making it a more effective site for PR inhibition.

The CA/P2 CS (Figure 43B) demonstrated a similar binding profile to MA/CA but in the opposite direction, with notable differences in the interaction regions. The binding pose indicated strong interactions at the P2 and P5' positions, while the P3 to P5 regions showed more variability.

The P2/NC CS (Figure 43C) had a distinctive binding pose with significant interactions across the substrate envelope. The binding density is particularly high at the P5 to P2' positions, suggesting a strong and stable interaction in these regions. However, P3' to P5 positions show less interaction density, indicating potential areas of flexibility or weaker binding. This profile suggested that P2/NC could have a varied binding affinity, with certain regions contributing more to stability than others.

The NC/P1 CS (Figure 43D) had an extended interaction profile across the substrate envelope, with significant contacts at P2 to P5' positions. This site showed a stable binding pose that spans multiple interaction points, indicating a potentially high binding affinity. The interaction density across these regions suggested that NC/P1 maintains strong and consistent contact with the PR active site,

contributing to its overall stability. This extensive binding profile highlights the importance of multiple interaction points in maintaining a stable substrate envelope, which is important for effective PR inhibition.

The P1/P6 CS (Figure 43E) had robust interactions across a wide range of positions from P5 to P1'. The binding pose indicated a well-fitted and stable envelope, with significant interaction density at the P1 and P2' positions. This implies that P1/P6 may have a high binding affinity and stability, making it a potential target for effective PR inhibition. The uniformity of the interaction profile indicates that the substrate is well-accommodated within the PR active site, contributing to a stable and effective binding pose.

The superimposed structure of all five CSs (Figure 43F) provided a comprehensive view of the common interaction hotspots and areas of variability within the substrate envelope. The overlapping regions indicate strong consensus points, particularly at the P2 to P2’ positions, where interactions are consistently strong across all sites. The variability in other regions, such as P3 and P3', highlights the flexibility and differences in binding poses among the CSs. This view underscores the importance of targeting these consensus interaction points for designing effective PIs, while also considering the unique interaction profiles of each CS to enhance binding affinity and stability.

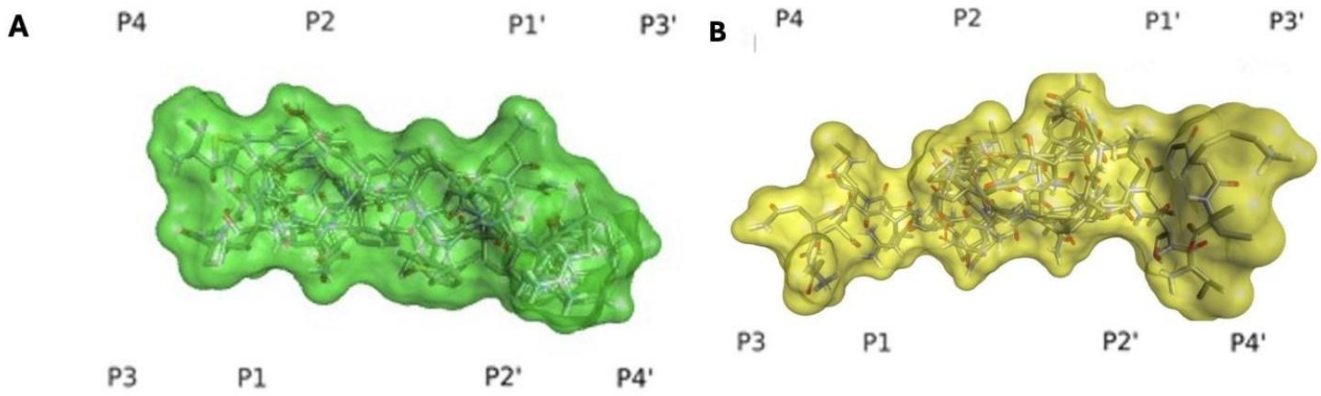


Figure 44: Comparison of subtype B (A) vs subtype C (B) substrate Envelopes

In Figure 44, both the subtype B(A) and subtype C(B) substrate envelopes exhibited similar elongated shapes, indicative of the extended binding regions within the HIV-1 PR active site. This elongation spans multiple interaction points, essential for stable binding and effective inhibition. However, subtle differences in the contours and surface representations suggest variations in how each substrate fits within the active site.

3.4 Discussion

This study described the natural polymorphisms at *gag* CSs and their potential impact on the substrate envelope structure of HIV-1 subtype C. The CS sequences were obtained from treatment-naïve samples and the study identified that these natural polymorphisms could lead to alterations in the shape of the substrate envelope, potentially influencing how the viral PR interacts with its natural substrates. Such changes may impact the efficiency of polyprotein cleavage, which is crucial for proper viral maturation and replication (Konvalinka et al., 2015).

MD simulations were employed to evaluate the binding affinity of *gag* CS complexes. MD simulations are necessary for understanding the biological function of protein systems, as they provide insights into the dynamic behaviour and interactions of these proteins (Hollingsworth and Dror, 2018; Karplus and McCammon, 2002). Multiple studies have utilized this approach to examine the dynamic behaviour of the HIV-1 PR and *gag* (Badaya and Sasidhar, 2020; Banerjee et al., 2024; Chatfield and Brooks, 1995; Monje-Galvan and Voth, 2020; Salsbury, 2010; Samsudin et al., 2021). In this context, models for HIV-1 subtype C *gag*-PR complexes were predicted based on the WT PR, which allowed for a detailed analysis of how mutations at the *gag* CS impacted binding and function. When experimental data is unavailable, homology modelling becomes an essential tool for studying molecular structural dynamics, enabling researchers to predict the behaviour of proteins based on known structures of related molecules (Hillisch et al., 2004; Schmidt et al., 2014; Vyas et al., 2012).

The comparative analysis across MA/CA, CA/P2, P2/NC, NC/P1, and P1/P6 CSs revealed several consistent patterns. For example, higher binding affinity was generally associated with extensive hydrogen bonding and π-alkyl interactions. Conserved residues played a critical role in maintaining strong interactions, while sequence variability did affect the binding strength. Unfavourable interactions, such as steric clashes and repulsive forces, contributed to lower binding affinities. All five ligands (MA/CA, CA/P2, P2/NC, NC/P1, and P1/P6) featured a balance of hydrophobic and polar residues that contributed to their structural stability and functional efficacy.

Based on our data and analysis, we showed that PR identifies a combination of specific residues and specific forces that are crucial for substrate recognition and cleavage. This is consistent with reports from Prabu-Jeyabalan *et al.* (2002) who stated that the PR senses how deeply and stably each substrate binds within its active site. The PR does not merely recognize a single residue but rather a 3D arrangement of residues that interact with it through a variety of forces (Prabu-Jeyabalan et al., 2000, 2002). This interpretation aligns with both the 3D substrate envelopes and the 2D interaction maps that are discussed below.

3.4.1 Hydrophobic and Hydrogen Bond Interactions:

Across all ligands, hydrophobic residues were central to maintaining structural stability. Valine, leucine, isoleucine, phenylalanine, and methionine frequently appeared in the CSs, forming a hydrophobic centre. MA/CA had valine at P5 and isoleucine at P2', while CA/P2 had valine and leucine at similar positions. P2/NC and NC/P1 also had isoleucine, leucine, and methionine residues, emphasizing the importance of hydrophobic residues in strengthening the ligand's internal structure (Scheraga, 1998; Sun, 2022). Hydrophobic residues on protein surfaces tended to cluster in binding regions, while charged residues were usually found elsewhere, reflecting general protein-ligand interaction principles (Du et al., 2016; Williams, 2013). Kathuria *et al.* (2016) showed that side chains of isoleucine, leucine, and valine tend to form large hydrophobic clusters that shield underlying hydrogen bond networks from water, improving resistance to solvent exchange. Other studies have shown that the hydrophobicity of residues plays a more significant role in guiding the folding process, with a smaller contribution to protein stabilization (Priyakumar, 2012; Van den Burg et al., 1994).

3.4.2 Polar and Ionic Interactions:

Hydrogen bonding and electrostatic interactions are critical for ligand specificity and binding affinity (Chen et al., 2016). All ligands contained polar residues such as asparagine, glutamine, and serine that participated in hydrogen bonding. MA/CA and P2/NC were more reliant on polar residues (asparagine, glutamine) for hydrogen bonding rather than charged interactions. Additionally, CA/P2 and NC/P1 included glutamic acid, with a lower pKa (4.3), indicating involvement in more acidic ionic interactions compared to the other ligands.

Additionally, basic residues such as arginine (present in CA/P2, NC/P1 and P1/P6) facilitated electrostatic interactions. These charged residues may played a role in other stages of the binding process and the long-range nature of electrostatic interactions was likely a key factor in the recognition phase (Sheinerman et al., 2000). Additionally, favourable electrostatic interactions may have kept the two proteins close to each other for a sufficient duration, allowing the structural rearrangements needed for binding to occur (Zhou and Pang, 2018).

3.4.3 Structural Rigidity and Flexibility:

Proline residues, present in both MA/CA and P1/P6, provided rigidity to the ligands by constraining conformational flexibility (Levitt, 1981). MA/CA had proline at the P1' cleavage site which could account for why it was a more conserved CS with the highest binding affinity. According to Potempa

et al. (2018) the P1' proline at the MA/CA site is key for capsid changes after cleavage, while HIV-1 protease also targets substrates with large hydrophobic residues, indicating two substrate classes. Spielvogel *et al.* (2023) found that subtle modifications in the P1'-equivalent positioning of the inhibitor influenced the preferential selection of one resistance pathway over another. Alterations in the P2'-equivalent positioning significantly impacted the inhibitor's efficacy, with these distinctions persisting in resistant viral strains, thereby shaping the development of mutations (Spielvogel et al., 2023).

P1/P6 had proline in the P4 position which accounts for its low conservation rate and low binding affinity. Proline is the most inflexible residue and can be incorporated to stiffen flexible regions, improving the thermostability of proteins (Yu et al., 2015). Proline is unique among amino acids in that it contains a secondary amine, technically an imine, rather than the typical primary amine. This structural feature classifies it as an imino acid. Its three-carbon side chain forms a covalent bond with the α-nitrogen, creating a rigid, cyclic structure that limits rotational freedom. As a result, proline residues introduce kinks or constraints within polypeptide chains, often disrupting regular folding patterns. (National Centre for Biotechnology Information (2025). PubChem Compound Summary for CID 145742, Proline. Retrieved March 31, 2025 from https://pubchem.ncbi.nlm.nih.gov/compound/Proline.). On the other hand, glycine, found in MA/CA, NC/P1, and P1/P6, contributed to flexibility, allowing the ligands to adjust conformation for optimal binding.

3.4.4 Importance of Hydrogen Bond Networks:

Hydrogen bonding was a fundamental stabilizing force in nearly all analysed CSs. This is consistent with previous reports by Prabu-Jeyabalan *et al.* (2002), Tie *et al*. (2005) and Wlodawer & Vondrasek (1998). The data consistently showed that sequences with higher binding affinities tended to have extensive networks of conventional hydrogen bonds, particularly involving residues such as Asp29, Asp30, Gly48, and catalytic G27. These residues were frequently located at strategic positions in the PR, where they engaged in multiple hydrogen bonds with different peptide substrates. For instance, the sequence VSQNY/PIVQN formed nine conventional hydrogen bonds with residues such as Asp29, Asp30, and Arg8, which significantly contributed to its binding affinity of -8.3 kcal/mol. Similar patterns were observed in other sequences, such as EARVL/AEAMS (-8.2 kcal/mol), where robust hydrogen bonding networks with residues like Gly49, Asp30, and G27 may have contributed to binding stability.

However, variability in hydrogen bond distances, often ranged from 3.00 Å to over 6.00 Å, suggesting that not all hydrogen bonds contributed equally to binding stabilization. Shorter bond distances

(approximately 3.00-3.50 Å) generally indicated stronger interactions and were more frequently associated with higher binding affinities. This highlights the need to not only quantify hydrogen bonds but also consider their quality and proximity in evaluating their contribution to the overall binding affinity (Chen et al., 2016).

3.4.5 Role of Alkyl and π-Alkyl Interactions in Stabilization

Alkyl and π-alkyl interactions played significant roles in the binding of CSs to HIV-1 PR, with certain residues such as Val82, Ile50, Ile84, and Pro81 which appeared frequently in these interactions. These hydrophobic contacts often occurred within a distance range of approximately 4.00 Å to 6.00 Å, where they provided stabilization by optimizing hydrophobic packing. For example, sequences such as RPGNF/LQNRP (-7.9 kcal/mol) and RPGNF/VQSRP (-8.1 kcal/mol) featured strong π-alkyl interactions with Val82, Pro81, and other residues, which could have had a role in maintaining their high binding affinities (Tie et al., 2004).

A common trend in this study was that sequences with higher binding affinities tended to have shorter and more effective alkyl and π-alkyl interaction distances, often below 5.00 Å. For instance, the EARVL/AEAMS sequence exhibited several key alkyl interactions with Val82 and Ile84 at unfavourable distances of approximately 6.13 Å and 6.09 Å, which, while relatively long, were still within the effective range for significant hydrophobic interactions. Conversely, sequences with slightly weaker affinities, such as ASQNY/PIVQN (-7.5 kcal/mol), often displayed shorter alkyl interaction distances, ranging from 4.20 Å to 6.61 Å, indicating weaker hydrophobic stabilization (Tie et al., 2004; Wallace et al., 1995) .

3.4.6 Contributions of Catalytic Residues and Aromatic Interactions:

Aromatics such as phenylalanine, tryptophan, and tyrosine, as well as catalytic residues such as D25 and G27, were frequently involved in critical π-cation interactions. These residues played a central role in stabilizing the PR-substrate complex by engaging in aromatic stacking or cation-π interactions, which were typically stronger than van der Waals interactions (Gallivan and Dougherty, 1999). For instance, the sequence VSQNY/PIVQN featured both π-alkyl interactions with Val82 and π-cation interactions with G27, reinforcing its high binding affinity of -8.3 kcal/mol. The distances of these interactions (4.71 Å and 4.16 Å, respectively) further underscored their significance in binding stabilization. Catalytic residues such as D25 and G27 also frequently participated in stabilizing hydrogen bonds and other interactions. For instance, sequences such as RPGNF/VQSRP and KARIL/AEAMS exhibited hydrogen bonds with these catalytic residues, which may be necessary for

binding stability and specificity (Prabu-Jeyabalan et al., 2002; Tie et al., 2012; Wlodawer and Erickson, 1993).

3.4.7 Impact of Unfavourable Interactions on Binding Affinity:

Despite the predominance of stabilizing interactions, the presence of unfavourable donor-donor or acceptor-acceptor interactions can significantly impact binding affinity. These interactions often involved steric clashes or electrostatic repulsion, which reduced binding efficiency (Spassov, 2024). For instance, the CS RPGNF/LQSRP, with a lower binding affinity of -5.0 kcal/mol, presented an unfavourable clash with Gly22 at 4.65 Å, which likely contributed to its reduced binding stability as glycine does provide flexibility (Yan and Sun, 1997). Lack thereof would therefore likely lead to decreased binding affinity. Given glycine's established role in conferring conformational flexibility (Yan and Sun, 1997), such a clash likely restricts local mobility, thereby compromising the stability of the ligand–protease complex.

Moreover, the sequences ASQNY/PIVQN (–7.5 kcal/mol) and ISQNY/PIVQN (–7.3 kcal/mol), despite their more favourable overall binding energies, also exhibited unfavourable donor–donor interactions with Arg8, with interatomic distances extending up to 6.01 Å. Although these distances exceeded conventional hydrogen bonding thresholds, the persistence of overlapping electrostatic fields or subtle steric strain can result in long-range repulsive effects (Kojić-Prodić and Molcanov, 2008; Kukić and Nielsen, 2010) . These effects may offset some of the stabilizing contributions elsewhere in the interaction interface, ultimately reducing the net binding affinity (Spassov, 2024; Warshel et al., 2006). Interestingly, some sequences, such as VSQNF/PIVQN (-7.4 kcal/mol), manage to compensate for these unfavourable interactions through other stabilizing forces, such as robust hydrogen bonding networks or effective π-alkyl interactions. This suggests that while unfavourable interactions are detrimental, their impact can be mitigated by the presence of other strong stabilizing interactions (Xiao et al., 2020).

3.4.8 Effect of variation at the CS on substrate binding and cleavage

The binding poses of HIV-1 PR with the different sequence variants revealed differences in binding affinities and interaction profiles. Overall, the binding affinities of the various CSs ranged from -5.0kcal/mol to -8.3 kcal/mol. High-affinity sequences typically exhibited a combination of strong hydrogen bonds, effective hydrophobic interactions, and minimal unfavourable contacts. In contrast, sequences with lower affinities often showed a mix of weaker hydrogen bonds, less optimal hydrophobic interactions, and the presence of repulsive forces.

More specifically, sequences such as VSQNY/PIVQN (MA/CA) and EARVL/AEAMS (CA/P2) exhibited high binding affinities. The high binding affinity suggested that these sequences fit very well within the substrate envelope, maintaining extensive and strong contact points with the PR residues (King et al., 2004; Prabu-Jeyabalan et al., 2002; Weber et al., 1989). The higher binding affinity suggested that these are more efficiently cleaved (Könnyű et al., 2013). This tight fitting is often facilitated by a combination of multiple hydrogen bonds, van der Waals interactions, and other non-covalent forces, which collectively stabilize the substrate within the active site (Chaudhury and Gray, 2009). Consequently, the substrate envelope for these sequences was well-defined and stable.

On the other end of the spectrum, sequences such as KARVL/AEAMS (CA/P2) and NNNIM/MQRNN (P2/NC) showed lower binding affinities. These sequences likely fit less optimally within the PR active site, resulting in a smaller substrate envelope. The higher Gbind values suggest weaker interactions and possibly fewer hydrogen bonds or less effective van der Waals interactions. As a result, these sequences might experience potential distortions in the substrate envelope, which could limit the overall stability and effectiveness of binding. This less optimal fitting can lead to increased flexibility or movement within the active site (Nalam et al., 2010; Nalam and Schiffer, 2008; Özen et al., 2011).

The sequential cleavage of the HIV-1 Gag polyprotein is highly regulated (Pettit et al., 2005). The early cleavage of the P2/NC site in the HIV-1 Gag precursor is likely due to its structural positioning, making it more accessible to viral PR and it has been shown that factors such as spatial arrangement and structural flexibility influence the order of processing (Louis et al., 1999a; Louis et al., 1999b; Pettit et al., 2005). This is consistent with the data from this study where P2/NC exhibited the highest sequence variability. Spielvogel *et al*. (2023) suggested that the high variability at this site is associated with selective pressure-driven adaptability, allowing modulation of cleavage rates to influence downstream processing. In addition, the lower binding affinity of P2/NC seen in this study indicated a weaker fit within the PR active site. It has been suggested that the greater flexibility and adaptability at the P2/NC CS allows it to act as a maturation checkpoint, regulating the overall cleavage process (Freed, 2015; Kleinpeter and Freed, 2020). Since cleavage at P2/NC triggers downstream conformational changes in Gag, its weaker binding and increased variability may allow fine-tuning of processing rates (Pettit et al., 1994). Thus, the cleavage hierarchy is not only dictated by PR accessibility but also by the inherent binding properties and stability of substrate interactions within the active site (Deshmukh et al., 2017).

In contrast, CA/P2 cleavage has been shown to occur last, reflecting its role in stabilizing the capsid structure until final core condensation (Gross et al., 2000; Pettit et al., 2005). The conservation of CA/P2 in this study highlighted its functional role in preventing premature core collapse and ensuring

proper virion infectivity (Aiken and Rousso, 2021). In addition, the high binding affinity of CA/P2 suggested that it fit well within the PR's substrate envelope, facilitating efficient cleavage due to strong non-covalent interactions. This aligns with CA/P2's role as a structurally constrained site, ensuring proper virion maturation before cleavage occurs (Adamson and Freed, 2010; Freed, 2015; Pettit et al., 2005).

3.4.9 The Substrate Envelope:

Despite the variability in the sequences of the substrates, the substrate envelope captured the essential binding region that remained consistent across different substrates, ensuring effective PR function (Prabu-Jeyabalan et al., 2002). This concept is particularly important in drug design, where inhibitors are crafted to fit within this conserved space to mimic the natural substrates' binding (Ghosh et al., 2016; Nalam and Schiffer, 2008). By designing inhibitors that align with the substrate envelope, they are less likely to be evaded by viral mutations, as such mutations would likely disrupt the PR's ability to process its natural substrates, thereby reducing the chance of resistance (Marie and Gordon, 2019; Ragland et al., 2014; Shen et al., 2013). Consequently, the substrate envelope serves as a critical blueprint for developing drugs that are both effective and resilient against the emergence of drug-resistant viral strains.

Analysis of the substrate envelopes for the variants seen at the five CS in *gag* (MA/CA, CA/P2, P2/NC, NC/P1, and P1/P6) suggested that different AA side chains can be accommodated at the CS and that they exhibited slightly different conformations depending on the substrate. For example, the MA/CA substrate envelope showed a distinct protrusion at the P1' position, while CA/P2 had a more extended substrate envelope at P2 and P2'. Potempa *et al.* (2018) found that a P1' proline at the MA/CA cleavage site is key for capsid changes post-cleavage, with P2 and P2' side chains switching their interactions between polar and hydrophobic subsites depending on P1' presence.
These findings suggested that HIV-1 PR's bi-functional subsites could be leveraged in inhibitor design (Potempa et al., 2018). These differences could influence substrate recognition and cleavage rate. Samant *et al*. (2022) reported that positions adjacent to the scissile bond demonstrated consistent yet complex physical properties affecting cleavage across different cut-sites, highlighting their crucial role in substrate recognition. Long-range sequence interdependencies significantly influenced cut-site efficiency, particularly for positions further from the scissile bond. Notably, aromatic AA and negative charges at specific positions can enhance cutting function, potentially altering the substrate envelope structure (Özen et al., 2011; Samant et al., 2022). In this study, the effects of long-range sequence variation (up to 15 AA away from the scissile bond) on the structure of the substrate envelope was not investigated and requires further study.

The structures showed that PR consistently interacted with the central core regions of the substrates, specifically around the P1 and P1' positions. This is consistent with previous studies which found that PR recognized a conserved core shape at P1 and P1' . These positions were seen as highly conserved and dense in the substrate envelope across all CSs. The PR "sees" these positions as the primary anchors that must align precisely for effective binding and cleavage (Özen et al., 2011; Özen et al., 2014; Prabu et al., 2006).

From the analysis of the generated subtype B structures, the consensus substrate envelope appeared to have a continuous and well-defined volume, suggesting a tight and specific fit for its corresponding substrate and is similar to the substrate envelope reported by (Prabu-Jeyabalan et al., 2002). The compact shape of the substrate envelope suggested that it might have strong hydrogen bonding and van der Waals interactions, with P1, P2, P1' and P2' forming most of these interactions. A tighter and more specific fit may imply higher substrate specificity and stronger binding affinity. Prabhu-Jeyabalan *et al*. (2002) reported that the hydrogen bonding network between backbone atoms in the substrate was essential for recognition by PR.

The subtype C substrate envelope had a slightly broader and less defined volume compared to the subtype B substrate envelope. This could indicate more flexibility in substrate binding, with less stringent requirements for specific interactions. The subtype C substrate envelope demonstrated well-defined binding interactions primarily around the P2 and P1' positions. These regions exhibited high interaction density, suggesting robust contacts that are likely critical for the stability and binding affinity of the substrate. The regions with high interaction density (strong hydrogen bonds or hydrophobic contacts) were particularly important because they provided the stability required for the substrate to be held in the correct orientation for the catalytic reaction (Ferenczy and Kellermayer, 2022).

This is similar to the subtype B substrate envelope reported by Prabhu-Jeyabalan *et al.* (2002) which showed significant interactions at the P1 and P2' positions, along with notable contact points extending to P3 and P3'. The subtype C substrate envelope showed less interaction density at the P3 and P3' positions, indicating potential areas of flexibility or weaker binding. This suggested that while the core interactions at P2 and P1' are strong, there was some degree of flexibility that could influence the overall stability of the substrate envelope. Key residues P1, P2, P1', and P2' were still important but allowed for more variability. The broader envelope also suggested a potential for accommodating a wider range of substrates or ligand molecules. According to Prabhu-Jeyabalan *et al.* (2002), PR preserves its structural stability despite frequent mutations, thanks to a network of backbone and water-mediated hydrogen bonds, which allows it to effectively recognize multiple substrate sites

within the Gag and Pol polyproteins. However, this flexibility might reduce the binding affinity and specificity compared to the subtype B envelope.

3.5 General Discussion and Conclusion

This study examined the natural variability of HIV-1 *gag* CS, focusing on subtype C compared to the more widely studied subtype B and how this influenced the shape of the substrate envelope and affected interactions with the viral protease.

In this study, the analysis of sequence variability at the CSs included AA up to 15 AA on either side of the scissile bond. Previous studies had analysed up to 5 AA on either side of the scissile bond (Oliveira et al., 2003). Subtype C showed greater variability particularly at the P2/NC and P1/P6 sites, with P2/NC showing 100% polymorphism. Variability increased with longer sequence windows, especially where regions overlapped. Overall, subtype C appeared to have a more diverse and flexible CS profile compared to subtype B, which may affect protease-substrate interactions (Rhee et al., 2003).

Furthermore, MD simulations provided insights into the binding affinities and interaction profiles between the PR and its substrates. These simulations showed that viral PRs utilized a combination of hydrophobic and hydrogen bonding interactions to stabilize substrate binding, ensuring proper positioning for catalysis. Strong binding was also driven by hydrogen bonds and π-alkyl interactions with conserved residues, while steric clashes reduced binding.

The substrate envelope concept provides a framework for understanding how variations in viral sequences can still result in effective substrate binding and cleavage by the PR (Nalam et al., 2013; Özen et al., 2014; Prabu-Jeyabalan et al., 2003, 2004). Our findings highlighted that despite the presence of polymorphisms, the substrate envelope remained largely conserved in key regions essential for the recognition and cleavage of natural substrates by the viral PR. Specifically, residues at the P1 and P1' positions played significant roles in stabilizing the binding of substrates within PR's active site (Spielvogel et al., 2023).

Mutations in subtype C have been associated with enhanced viral assembly efficiency and budding, suggesting that the variations observed in this study can affect not only PR function but also the overall viral replication process (Gartner et al., 2020). This interplay between sequence variability and structural integrity underscores the adaptability of HIV-1 and its capacity to evade immune pressures while maintaining critical functions (Saito and Yamashita, 2021).

Understanding the effect of mutations on the substrate envelope could also benefit drug development. By focusing on the conserved substrate envelope, researchers can design inhibitors that mimic natural substrates, which are less likely to be affected by mutations. This approach could potentially reduce the risk of drug resistance, as viral mutations that disrupt essential binding interactions would compromise PR function (Nalam et al., 2013). Increasing the sequence length around the scissile bond generally raised polymorphism frequency, notably at the CA/P2 site (3.5% to 96%) and MA/CA site (8% to 76%), while P2/NC remained highly variable (~95%). Spacer peptides caused overlapping recognition sequences, suggesting flexibility that allows tolerance of sequence variation without losing cleavage efficiency. When spacers were at the amino end, variability at 15 AA matched that at 5 AA, implying the amino-terminal region may limit tolerated variation and needs further study. Subtype C showed greater *gag* cleavage site variability than subtype B, suggesting increased adaptability and resistance potential.

In conclusion, this study highlighted the delicate balance between sequence variability and functional conservation in HIV-1 subtype C, compared to subtype B. Understanding these dynamics will offer an essential understanding of the processes behind viral replication and the development of effective therapeutic strategies to combat HIV-1.

3.6 Limitations

Studying natural polymorphisms at Gag CS and their impact on the substrate envelope structure presented several limitations. Firstly, the inherent variability in HIV-1 sequences across different isolates complicated the identification of consistent patterns and impacts of these polymorphisms. This genetic diversity necessitated extensive sampling and sequencing to capture a representative dataset, which was resource intensive. Additionally, computational models used to predict changes in binding affinities and substrate envelope alterations may have limited accuracy due to assumptions and approximations inherent in the algorithms. These models often required high-quality structural data, which was not available for all polymorphic variants. Furthermore, the dynamic nature of protein interactions and conformational flexibility of the PR and substrates added layers of complexity that were difficult to capture in static models. Finally, translating these molecular-level findings to an understanding of their impact on viral fitness and drug resistance mechanisms involved multiple layers of biological interactions, making it challenging to draw direct correlations.

One notable limitation of this study lies in the mismatch between some of the peptide sequences used for molecular modelling and the final, corrected CS variants identified through updated statistical analysis. The modelling work was conducted prior to the recalculation of sequence prevalence, which led to the selection of some sequences that were later found not to be the most statistically representative. Although this affected the direct correspondence between Chapters 2 and 3, the

modelled structures still provided valuable insights into protease–substrate binding interactions and supported broader structural interpretations. This discrepancy underscored the importance of iterative data validation and highlighted the dynamic nature of computational and sequence-based research.

4 References

Adamson, C.S., Freed, E.O., 2010. Novel approaches to inhibiting HIV-1 replication. Antiviral Res 85, 119-141.

Adcock, S.A., McCammon, J.A., 2006. Molecular dynamics: survey of methods for simulating the activity of proteins. Chem Rev 106, 1589-1615.

Agu, P.C., Afiukwa, C.A., Orji, O.U., Ezeh, E.M., Ofoke, I.H., Ogbu, C.O., Ugwuja, E.I., Aja, P.M., 2023. Molecular docking as a tool for the discovery of molecular targets of nutraceuticals in diseases management. Scientific Reports 13, 13398.

Aiken, C., Rousso, I., 2021. The HIV-1 capsid and reverse transcription. Retrovirology 18, 29.

Alberts B, J.A., Lewis J, et al., 2002. The Shape and Structure of Proteins. Molecular Biology of the Cell. 4th edition. .

Alencar, C., Sabino, E., Diaz, R., Mendrone, A., Nishiya, A., 2024. Genetic diversity in the partial sequence of the HIV-1 gag gene among people living with multidrug-resistant HIV-1 infection. Revista do Instituto de Medicina Tropical de São Paulo 66.

Alfadhli, A., Huseby, D., Kapit, E., Colman, D., Barklis, E., 2007. Human immunodeficiency virus type 1 matrix protein assembles on membranes as a hexamer. J Virol 81, 1472-1478.

Anfinsen, C.B., 1972. The formation and stabilization of protein structure. Biochem J 128, 737-749.

Anfinsen, C.B., 1973. Principles that govern the folding of protein chains. Science 181, 223-230.

Arts, E.J., Hazuda, D.J., 2012. HIV-1 antiretroviral drug therapy. Cold Spring Harb Perspect Med 2, a007161.

Badaya, A., Sasidhar, Y.U., 2020. Inhibition of the activity of HIV-1 protease through antibody binding and mutations probed by molecular dynamics simulations. Scientific Reports 10, 5501.

Bagaria, A., Jaravine, V., Huang, Y.J., Montelione, G.T., Güntert, P., 2012. Protein structure validation by generalized linear model root-mean-square deviation prediction. Protein Sci 21, 229-238.

Bailes, E., Gao, F., Bibollet-Ruche, F., Courgnaud, V., Peeters, M., Marx, P.A., Hahn, B.H., Sharp, P.M., 2003. Hybrid origin of SIV in chimpanzees. Science 300, 1713.

Balasubramaniam, M., Freed, E.O., 2011. New insights into HIV assembly and trafficking. Physiology (Bethesda) 26, 236-251.

Bally, F., Martinez, R., Peters, S., Sudre, P., Telenti, A., 2000. Polymorphism of HIV type 1 gag p7/p1 and p1/p6 cleavage sites: clinical significance and implications for resistance to protease inhibitors. AIDS Res Hum Retroviruses 16, 1209-1213.

Banerjee, P., Qu, K., Briggs, J.A.G., Voth, G.A., 2024. Molecular dynamics simulations of HIV-1 matrix-membrane interactions at different stages of viral maturation. Biophysical Journal 123, 389-406.

Barré-Sinoussi, F., Chermann, J.C., Rey, F., Nugeyre, M.T., Chamaret, S., Gruest, J., Dauguet, C., Axler-Blin, C., Vézinet-Brun, F., Rouzioux, C., Rozenbaum, W., Montagnier, L., 1983. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). Science 220, 868-871.

Barrie, K.A., Perez, E.E., Lamers, S.L., Farmerie, W.G., Dunn, B.M., Sleasman, J.W., Goodenow, M.M., 1996. Natural variation in HIV-1 protease, Gag p7 and p6, and protease cleavage sites within gag/pol polyproteins: amino acid substitutions in the absence of protease inhibitors in mothers and children infected by human immunodeficiency virus type 1. Virology 219, 407-416.

Beck, Z.Q., Morris, G.M., Elder, J.H., 2002. Defining HIV-1 Protease Substrate Selectivity. Current Drug Targets - Infectious Disorders 2, 37-50.
Bell, N.M., Lever, A.M.L., 2013. HIV Gag polyprotein: processing and early viral particle assembly. Trends in Microbiology 21, 136-144.
Berkhout, B., 1999. HIV-1 evolution under pressure of protease inhibitors: climbing the stairs of viral fitness. J Biomed Sci 6, 298-305.
Bieniasz, P.D., 2006. Late budding domains and host proteins in enveloped virus release. Virology 344, 55-63.
Borhani, D.W., Shaw, D.E., 2012. The future of molecular dynamics simulations in drug discovery. J Comput Aided Mol Des 26, 15-26.
Branden, C.I., Tooze, J., 2012. Introduction to protein structure. Garland Science.
Briggs, J.A.G., Kräusslich, H.-G., 2011. The Molecular Architecture of HIV. Journal of Molecular Biology 410, 491-500.
Brik, A., Wong, C.H., 2003. HIV-1 protease: mechanism and drug discovery. Org Biomol Chem 1, 5-14.
Bryngelson, J.D., Wolynes, P.G., 1987. Spin glasses and the statistical mechanics of protein folding. Proceedings of the National Academy of Sciences 84, 7524-7528.
Buonaguro, L., Del Guadio, E., Monaco, M., Greco, D., Corti, P., Beth-Giraldo, E., Buonaguro, F.M., Giraldo, G., 1995. Heteroduplex mobility assay and phylogenetic analysis of V3 region sequences of human immunodeficiency virus type 1 isolates from Gulu, northern Uganda. The Italian-Ugandan Cooperation AIDS Program. J Virol 69, 7971-7981.
Buttler, C.A., Pezeshkian, N., Fernandez, M.V., Aaron, J., Norman, S., Freed, E.O., van Engelenburg, S.B., 2018. Single molecule fate of HIV-1 envelope reveals late-stage viral lattice incorporation. Nature Communications 9, 1861.
Cao, A., 2020. The Last Secret of Protein Folding: The Real Relationship Between Long-Range Interactions and Local Structures. The Protein Journal 39, 422-433.
Carlson, J.M., Listgarten, J., Pfeifer, N., Tan, V., Kadie, C., Walker, B.D., Ndung'u, T., Shapiro, R., Frater, J., Brumme, Z.L., Goulder, P.J., Heckerman, D., 2012. Widespread impact of HLA restriction on immune control and escape pathways of HIV-1. J Virol 86, 5230-5243.
Carter, R., Luchini, A., Liotta, L., Haymond, A., 2019. Next Generation Techniques for Determination of Protein-Protein Interactions: Beyond the Crystal Structure. Curr Pathobiol Rep 7, 61-71.
Case, D., Darden, T., Cheatham, T., Simmerling, C., Wang, J., 2012. AMBER 12 San Francisco: University of California.[Google Scholar].
Case, D.A., Aktulga, H.M., Belfon, K., Cerutti, D.S., Cisneros, G.A., Cruzeiro, V.W.D., Forouzesh, N., Giese, T.J., Götz, A.W., Gohlke, H., Izadi, S., Kasavajhala, K., Kaymak, M.C., King, E., Kurtzman, T., Lee, T.-S., Li, P., Liu, J., Luchko, T., Luo, R., Manathunga, M., Machado, M.R., Nguyen, H.M., O'Hearn, K.A., Onufriev, A.V., Pan, F., Pantano, S., Qi, R., Rahnamoun, A., Risheh, A., Schott-Verdugo, S., Shajan, A., Swails, J., Wang, J., Wei, H., Wu, X., Wu, Y., Zhang, S., Zhao, S., Zhu, Q., Cheatham, T.E., III, Roe, D.R., Roitberg, A., Simmerling, C., York, D.M., Nagan, M.C., Merz, K.M., Jr., 2023. AmberTools. Journal of Chemical Information and Modeling 63, 6183-6191.
Case, D.A., Cheatham III, T.E., Darden, T., Gohlke, H., Luo, R., Merz Jr., K.M., Onufriev, A., Simmerling, C., Wang, B., Woods, R.J., 2005. The Amber biomolecular simulation programs. Journal of Computational Chemistry 26, 1668-1688.
Case, D.A., Duke, R.E., Walker, R.C., Skrynnikov, N.R., Cheatham Iii, T.E., Mikhailovskii, O., Simmerling, C., Xue, Y., Roitberg, A., Izmailov, S.A., Merz, K.M., Kasavajhala, K., Luo, R., Belfon, K., Li, P., Shen, J., Darden, T., Harris, R., Sagui, C., Onufriev, A., Pan, F., Izadi, S., Wang, J., Xiong, Y., Roe, D.R., Xiongwu, W., LeGrand, S., Gohlke, H., Swails, J., Schott-

Verdugo, S., Götz, A.W., Qi, R., Smith, J., Wei, H., Cerutti, D., Zhao, S., Lee, T., King, E., Berryman, J., Et, A., 2022. AMBER 22 Reference Manual. University of California.
Case, D.A., Rutgers, T.S.U.o.N.J.R.D.o.C.a.C.B.
Castley, A., Sawleshwarkar, S., Varma, R., Herring, B., Thapa, K., Dwyer, D., Chibo, D., Nguyen, N., Hawke, K., Ratcliff, R., Garsia, R., Kelleher, A., Nolan, D., 2017. A national study of the molecular epidemiology of HIV-1 in Australia 2005-2012. PLoS One 12, e0170601.
Cavasotto, C.N., Adler, N.S., Aucar, M.G., 2018. Quantum Chemical Approaches in Structure-Based Virtual Screening and Lead Optimization. Front Chem 6, 188.
Chamanian, M., Purzycka, K.J., Wille, P.T., Ha, J.S., McDonald, D., Gao, Y., Le Grice, S.F., Arts, E.J., 2013. A cis-acting element in retroviral genomic RNA links Gag-Pol ribosomal frameshifting to selective viral RNA encapsidation. Cell Host Microbe 13, 181-192.
Chameettachal, A., Mustafa, F., Rizvi, T.A., 2023. Understanding Retroviral Life Cycle and its Genomic RNA Packaging. J Mol Biol 435, 167924.
Charneau, P., Borman, A.M., Quillent, C., Guétard, D., Chamaret, S., Cohen, J., Rémy, G., Montagnier, L., Clavel, F., 1994. Isolation and Envelope Sequence of a Highly Divergent HIV-1 Isolate: Definition of a New HIV-1 Group. Virology 205, 247-253.
Chatfield, D.C., Brooks, B.R., 1995. HIV-1 Protease Cleavage Mechanism Elucidated with Molecular Dynamics Simulation. Journal of the American Chemical Society 117, 5561-5572.
Chaudhury, S., Gray, J.J., 2009. Identification of structural mechanisms of HIV-1 protease specificity using computational peptide docking: implications for drug resistance. Structure 17, 1636-1648.
Checkley, M.A., Luttge, B.G., Freed, E.O., 2011. HIV-1 envelope glycoprotein biosynthesis, trafficking, and incorporation. J Mol Biol 410, 582-608.
Chellappan, S., Kairys, V., Fernandes, M.X., Schiffer, C., Gilson, M.K., 2007. Evaluation of the substrate envelope hypothesis for inhibitors of HIV-1 protease. Proteins 68, 561-567.
Chen, D., Oezguen, N., Urvil, P., Ferguson, C., Dann, S.M., Savidge, T.C., 2016. Regulation of protein-ligand binding affinity by hydrogen bond pairing. Sci Adv 2, e1501240.
Chen, L., Li, Q., Nasif, K.F.A., Xie, Y., Deng, B., Niu, S., Pouriyeh, S., Dai, Z., Chen, J., Xie, C.Y., 2024. AI-Driven Deep Learning Techniques in Protein Structure Prediction. Int J Mol Sci 25.
Chen, X., Wang, X., 2024. The HIV-1 gag p6: a promising target for therapeutic intervention. Retrovirology 21, 1.
Chien, A.-I., Liao, W.-H., Yang, D.-M., Wang, C.-T., 2006. A domain directly C-terminal to the major homology region of human immunodeficiency type 1 capsid protein plays a crucial role in directing both virus assembly and incorporation of Gag–Pol. Virology 348, 84-95.
Childers, M.C., Daggett, V., 2017. Insights from molecular dynamics simulations for computational protein design. Mol Syst Des Eng 2, 9-33.
Chothia, C., Lesk, A.M., 1986. The relation between the divergence of sequence and structure in proteins. Embo j 5, 823-826.
Chou, K.-C., 1996. Prediction of Human Immunodeficiency Virus Protease Cleavage Sites in Proteins. Analytical Biochemistry 233, 1-14.
Chukkapalli, V., Oh, S.J., Ono, A., 2010. Opposing mechanisms involving RNA and lipids regulate HIV-1 Gag membrane binding through the highly basic region of the matrix domain. Proceedings of the National Academy of Sciences 107, 1600-1605.
Chung, H.S., Eaton, W.A., 2018. Protein folding transition path times from single molecule FRET. Current Opinion in Structural Biology 48, 30-39.
Clavel, F., Mammano, F., 2010. Role of Gag in HIV Resistance to Protease Inhibitors. Viruses 2, 1411-1426.

Coffin J, H.A., Levy JA, Montagnier L, Oroszlan S, Teich N, Temin H, Toyoshima K, Varmus H, Vogt P and Weiss R. 1986. , 1986. What to call the AIDS virus? Nature, 321, 10.
Collins, J.R., Burt, S.K., Erickson, J.W., 1995. Flap opening in HIV-1 protease simulated by 'activated' molecular dynamics. Nat Struct Biol 2, 334-338.
Coren, L.V., Thomas, J.A., Chertova, E., Sowder, R.C., 2nd, Gagliardi, T.D., Gorelick, R.J., Ott, D.E., 2007. Mutational analysis of the C-terminal gag cleavage sites in human immunodeficiency virus type 1. J Virol 81, 10047-10054.
Côté, H.C., Brumme, Z.L., Harrigan, P.R., 2001. Human immunodeficiency virus type 1 protease cleavage site mutations associated with protease inhibitor cross-resistance selected by indinavir, ritonavir, and/or saquinavir. J Virol 75, 589-594.
Cournia, Z., Allen, B., Sherman, W., 2017. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. Journal of Chemical Information and Modeling 57, 2911-2937.
Creighton, T.E., 1993. Proteins: structures and molecular properties. Macmillan.
Crick, F.H., 1958. On protein synthesis. Symp Soc Exp Biol 12, 138-163.
Datta, S.A., Curtis, J.E., Ratcliff, W., Clark, P.K., Crist, R.M., Lebowitz, J., Krueger, S., Rein, A., 2007. Conformation of the HIV-1 Gag protein in solution. J Mol Biol 365, 812-824.
Davies, D.R., 1990. The structure and function of the aspartic proteinases. Annu Rev Biophys Biophys Chem 19, 189-215.
De Vivo, M., Masetti, M., Bottegoni, G., Cavalli, A., 2016. Role of Molecular Dynamics and Related Methods in Drug Discovery. Journal of Medicinal Chemistry 59, 4035-4061.
Debrunner, P., Tsibris, J., Münck, E., 1969. Mössbauer Spectroscopy in Biological Systems: Proceedings of a Meeting Held at Allerton House, March 17 and 18, 1969, Monticello, Ill., Organized by the University of Illinois at Urbana-Champaign, Departments of Chemistry and Physics. University of Illinois.
Deng, Y., Roux, B., 2009. Computations of standard binding free energies with molecular dynamics simulations. J Phys Chem B 113, 2234-2246.
Desantis, F., Miotto, M., Di Rienzo, L., Milanetti, E., Ruocco, G., 2022. Spatial organization of hydrophobic and charged residues affects protein thermal stability and binding affinity. Sci Rep 12, 12087.
Deshmukh, L., Ghirlando, R., Clore, G.M., 2015. Conformation and dynamics of the Gag polyprotein of the human immunodeficiency virus 1 studied by NMR spectroscopy. Proceedings of the National Academy of Sciences 112, 3374-3379.
Deshmukh, L., Tugarinov, V., Louis, J.M., Clore, G.M., 2017. Binding kinetics and substrate selectivity in HIV-1 protease−Gag interactions probed at atomic resolution by chemical exchange NMR. Proceedings of the National Academy of Sciences 114, E9855-E9862.
Dill, K.A., 1990. Dominant forces in protein folding. Biochemistry 29, 7133-7155.
Dill, K.A., Chan, H.S., 1997. From Levinthal to pathways to funnels. Nature Structural Biology 4, 10-19.
Dill, K.A., MacCallum, J.L., 2012. The protein-folding problem, 50 years on. Science 338, 1042-1046.
Doyon, L., Croteau, G., Thibeault, D., Poulin, F., Pilote, L., Lamarre, D., 1996. Second locus involved in human immunodeficiency virus type 1 resistance to protease inhibitors. J Virol 70, 3763-3769.
Du, X., Li, Y., Xia, Y.L., Ai, S.M., Liang, J., Sang, P., Ji, X.L., Liu, S.Q., 2016. Insights into Protein-Ligand Interactions: Mechanisms, Models, and Methods. Int J Mol Sci 17.
Duan, Y., Wu, C., Chowdhury, S., Lee, M.C., Xiong, G., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T., Caldwell, J., Wang, J., Kollman, P., 2003. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. Journal of Computational Chemistry 24, 1999-2012.

Dubois, N., Marquet, R., Paillart, J.C., Bernacchi, S., 2018. Retroviral RNA Dimerization: From Structure to Functions. Front Microbiol 9, 527.
Engelman, A., Cherepanov, P., 2012. The structural biology of HIV-1: mechanistic and therapeutic insights. Nature Reviews Microbiology 10, 279-290.
Eswar, N., John, B., Mirkovic, N., Fiser, A., Ilyin, V.A., Pieper, U., Stuart, A.C., Marti-Renom, M.A., Madhusudhan, M.S., Yerkovich, B., Sali, A., 2003. Tools for comparative protein structure modeling and analysis. Nucleic Acids Research 31, 3375-3380.
Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M.Y., Pieper, U., Sali, A., 2006. Comparative protein structure modeling using Modeller. Curr Protoc Bioinformatics Chapter 5, Unit-5.6.
Fadahunsi, A.A., Uzoeto, H.O., Okoro, N.O., Cosmas, S., Durojaye, O.A., Odiba, A.S., 2024. Revolutionizing drug discovery: an AI-powered transformation of molecular docking. Medicinal Chemistry Research 33, 2187-2203.
Fehér, A., Weber, I.T., Bagossi, P., Boross, P., Mahalingam, B., Louis, J.M., Copeland, T.D., Torshin, I.Y., Harrison, R.W., Tözsér, J., 2002. Effect of sequence polymorphism and drug resistance on two HIV-1 Gag processing sites. European Journal of Biochemistry 269, 4114-4120.
Ferenczy, G.G., Kellermayer, M., 2022. Contribution of hydrophobic interactions to protein mechanical stability. Comput Struct Biotechnol J 20, 1946-1956.
Ferreira, L.G., Dos Santos, R.N., Oliva, G., Andricopulo, A.D., 2015. Molecular docking and structure-based drug design strategies. Molecules 20, 13384-13421.
Fischer, E., 1894. Einfluss der Configuration auf die Wirkung der Enzyme. Berichte der deutschen chemischen Gesellschaft 27, 2985-2993.
Fischer, E., Fischer, E., 1909. Einfluß der Konfiguration auf die Wirkung der Enzyme. I. Untersuchungen Über Kohlenhydrate und Fermente (1884–1908), 836-844.
Fiser, A., 2010. Template-based protein structure modeling. Methods Mol Biol 673, 73-94.
Forli, S., Huey, R., Pique, M.E., Sanner, M.F., Goodsell, D.S., Olson, A.J., 2016. Computational protein-ligand docking and virtual drug screening with the AutoDock suite. Nat Protoc 11, 905-919.
Forrest, L.R., Tang, C.L., Honig, B., 2006. On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins. Biophys J 91, 508-517.
Fossen, T., Wray, V., Bruns, K., Rachmat, J., Henklein, P., Tessmer, U., Maczurek, A., Klinger, P., Schubert, U., 2005. Solution structure of the human immunodeficiency virus type 1 p6 protein. J Biol Chem 280, 42515-42527.
Freed, E.O., 2015. HIV-1 assembly, release and maturation. Nat Rev Microbiol 13, 484-496.
Freire, E., 2008. Do enthalpy and entropy distinguish first in class from best in class? Drug Discovery Today 13, 869-874.
Fuller, J.C., Burgoyne, N.J., Jackson, R.M., 2009. Predicting druggable binding sites at the protein-protein interface. Drug Discov Today 14, 155-161.
Fun, A., van Maarseveen, N.M., Pokorná, J., Maas, R.E., Schipper, P.J., Konvalinka, J., Nijhuis, M., 2011. HIV-1 protease inhibitor mutations affect the development of HIV-1 resistance to the maturation inhibitor bevirimat. Retrovirology 8, 70.
Fun, A., Wensing, A.M.J., Verheyen, J., Nijhuis, M., 2012. Human Immunodeficiency Virus gag and protease: partners in resistance. Retrovirology 9, 63.
Gallivan, J.P., Dougherty, D.A., 1999. Cation-pi interactions in structural biology. Proc Natl Acad Sci U S A 96, 9459-9464.
Gallo, R.C., Salahuddin, S.Z., Popovic, M., Shearer, G.M., Kaplan, M., Haynes, B.F., Palker, T.J., Redfield, R., Oleske, J., Safai, B., et al., 1984. Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. Science 224, 500-503.

Ganser-Pornillos, B.K., Yeager, M., Sundquist, W.I., 2008. The structural biology of HIV assembly. Current Opinion in Structural Biology 18, 203-217.
Gartner, M.J., Roche, M., Churchill, M.J., Gorry, P.R., Flynn, J.K., 2020. Understanding the mechanisms driving the spread of subtype C HIV-1. EBioMedicine 53, 102682.
Gatanaga, H., Suzuki, Y., Tsang, H., Yoshimura, K., Kavlick, M.F., Nagashima, K., Gorelick, R.J., Mardy, S., Tang, C., Summers, M.F., Mitsuya, H., 2002. Amino acid substitutions in Gag protein at non-cleavage sites are indispensable for the development of a high multitude of HIV-1 resistance against protease inhibitors. J Biol Chem 277, 5952-5961.
Gelderblom, H.R., Hausmann, E.H., Ozel, M., Pauli, G., Koch, M.A., 1987. Fine structure of human immunodeficiency virus (HIV) and immunolocalization of structural proteins. Virology 156, 171-176.
Genheden, S., and Ryde, U., 2015. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. Expert Opinion on Drug Discovery 10, 449-461.
Gershenson, A., Gierasch, L.M., 2011. Protein folding in the cell: challenges and progress. Curr Opin Struct Biol 21, 32-41.
Ghosh, A.K., Osswald, H.L., Prato, G., 2016. Recent Progress in the Development of HIV-1 Protease Inhibitors for the Treatment of HIV/AIDS. J Med Chem 59, 5172-5208.
Ghosn, J., Carosi, G., Moreno, S., Pokrovsky, V., Lazzarin, A., Pialoux, G., Sanz-Moreno, J., Balogh, A., Vandeloise, E., Biguenet, S., 2010. Unboosted atazanavir-based therapy maintains control of HIV type-1 replication as effectively as a ritonavir-boosted regimen. Antivir Ther 15, 993-1002.
Gilson, M.K., Sharp, K.A., Honig, B.H., 1988. Calculating the electrostatic potential of molecules in solution: Method and error assessment. Journal of Computational Chemistry 9, 327-335.
Gilson, M.K., Zhou, H.X., 2007. Calculation of protein-ligand binding affinities. Annu Rev Biophys Biomol Struct 36, 21-42.
Gladnikoff, M., Shimoni, E., Gov, N.S., Rousso, I., 2009. Retroviral assembly and budding occur through an actin-driven mechanism. Biophys J 97, 2419-2428.
Goddard, T.D., Huang, C.C., Meng, E.C., Pettersen, E.F., Couch, G.S., Morris, J.H., Ferrin, T.E., 2018. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. Protein Science 27, 14-25.
Goedert, J.J., Gallo, R.C., 1985. Epidemiological evidence that HTLV-III is the AIDS agent. Eur J Epidemiol 1, 155-159.
Goodsell, D.S., Morris, G.M., Olson, A.J., 1996. Automated docking of flexible ligands: Applications of autodock. Journal of Molecular Recognition 9, 1-5.
Göttlinger, H.G., Dorfman, T., Sodroski, J.G., Haseltine, W.A., 1991. Effect of mutations affecting the p6 gag protein on human immunodeficiency virus particle release. Proceedings of the National Academy of Sciences 88, 3195-3199.
Gray, T.M., Matthews, B.W., 1984. Intrahelical hydrogen bonding of serine, threonine and cysteine residues within alpha-helices and its relevance to membrane-bound proteins. J Mol Biol 175, 75-81.
Griffiths, J.T., Phylip, L.H., Konvalinka, J., Strop, P., Gustchina, A., Wlodawer, A., Davenport, R.J., Briggs, R., Dunn, B.M., Kay, J., 1992. Different requirements for productive interaction between the active site of HIV-1 proteinase and substrates containing -hydrophobic-hydrophobic- or -aromatic-Pro- cleavage sites. Biochemistry 31, 5193-5200.
Grinter, S.Z., Zou, X., 2014. Challenges, applications, and recent advances of protein-ligand docking in structure-based drug design. Molecules 19, 10150-10176.
Gromiha, M.M., Selvaraj, S., 1999. Importance of long-range interactions in protein folding. Biophys Chem 77, 49-68.

Gross, I., Hohenberg, H., Wilk, T., Wiegers, K., Grättinger, M., Müller, B., Fuller, S., Kräusslich, H.G., 2000. A conformational switch controlling HIV-1 morphogenesis. Embo j 19, 103-113.
Groves, M.R., Dhanaraj, V., Badasso, M., Nugent, P., Pitts, J.E., Hoover, D.J., Blundell, T.L., 1998. A 2.3 A resolution structure of chymosin complexed with a reduced bond inhibitor shows that the active site beta-hairpin flap is rearranged when compared with the native crystal structure. Protein Engineering, Design and Selection 11, 833-840.
Guo, J., Wu, T., Anderson, J., Kane, B.F., Johnson, D.G., Gorelick, R.J., Henderson, L.E., Levin, J.G., 2000. Zinc finger structures in the human immunodeficiency virus type 1 nucleocapsid protein facilitate efficient minus- and plus-strand transfer. J Virol 74, 8980-8988.
Hageman, J.H., 1977. Biochemistry. (Lehninger, Albert L.). ACS Publications.
Hahn, B.H., Shaw, G.M., De Cock, K.M., Sharp, P.M., 2000. AIDS as a zoonosis: scientific and public health implications. Science 287, 607-614.
Hall, T., 1999. "BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl. Acids., 95-98.
Hill, M.K., Shehu-Xhilaga, M., Crowe, S.M., Mak, J., 2002. Proline residues within spacer peptide p1 are important for human immunodeficiency virus type 1 infectivity, protein processing, and genomic RNA dimer stability. J Virol 76, 11245-11253.
Hillisch, A., Pineda, L.F., Hilgenfeld, R., 2004. Utility of homology models in the drug discovery process. Drug Discov Today 9, 659-669.
Hollingsworth, S.A., Dror, R.O., 2018. Molecular Dynamics Simulation for All. Neuron 99, 1129-1143.
Homeyer, N., Gohlke, H., 2012. Free Energy Calculations by the Molecular Mechanics Poisson−Boltzmann Surface Area Method. Molecular Informatics 31, 114-122.
Houzet, L., Morichaud, Z., Didierlaurent, L., Muriaux, D., Darlix, J.L., Mougel, M., 2008. Nucleocapsid mutations turn HIV-1 into a DNA-containing virus. Nucleic Acids Res 36, 2311-2319.
Huang, L., Chen, C., 2013. Understanding HIV-1 protease autoprocessing for novel therapeutic development. Future Med Chem 5, 1215-1229.
Huang, S.-Y., Zou, X., 2010. Advances and Challenges in Protein-Ligand Docking. International Journal of Molecular Sciences 11, 3016-3034.
Huey, R., Morris, G.M., Olson, A.J., Goodsell, D.S., 2007. A semiempirical free energy force field with charge-based desolvation. J Comput Chem 28, 1145-1152.
Hurley, J.H., Boura, E., Carlson, L.A., Różycki, B., 2010. Membrane budding. Cell 143, 875-887.
Isaacs, D., Mikasi, S.G., Obasa, A.E., Ikomey, G.M., Shityakov, S., Cloete, R., Jacobs, G.B., 2020. Structural Comparison of Diverse HIV-1 Subtypes using Molecular Modelling and Docking Analyses of Integrase Inhibitors. Viruses 12.
Jiang, J., Ablan, S.D., Derebail, S., Hercík, K., Soheilian, F., Thomas, J.A., Tang, S., Hewlett, I., Nagashima, K., Gorelick, R.J., Freed, E.O., Levin, J.G., 2011. The interdomain linker region of HIV-1 capsid protein is a critical determinant of proper core assembly and stability. Virology 421, 253-265.
Joshi, A., Garg, H., Nagashima, K., Bonifacino, J.S., Freed, E.O., 2008. GGA and Arf proteins modulate retrovirus assembly and release. Mol Cell 30, 227-238.
Jouvenet, N., Neil, S.J.D., Bess, C., Johnson, M.C., Virgen, C.A., Simon, S.M., Bieniasz, P.D., 2006. Plasma Membrane Is the Site of Productive HIV-1 Particle Assembly. PLOS Biology 4, e435.
Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl,

S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583-589.
Karplus, M., McCammon, J.A., 2002. Molecular dynamics simulations of biomolecules. Nature Structural Biology 9, 646-652.
Karplus, M., Petsko, G.A., 1990. Molecular dynamics simulations in biology. Nature 347, 631-639.
Kauzmann, W., 1959. Some Factors in the Interpretation of Protein Denaturation11The preparation of this article has been assisted by a grant from the National Science Foundation, in: Anfinsen, C.B., Anson, M.L., Bailey, K., Edsall, J.T. (Eds.), Advances in Protein Chemistry. Academic Press, pp. 1-63.
Kelly, B.N., Howard, B.R., Wang, H., Robinson, H., Sundquist, W.I., Hill, C.P., 2006. Implications for Viral Capsid Assembly from Crystal Structures of HIV-1 Gag1-278 and CAN133-278. Biochemistry 45, 11257-11266.
Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., Wyckoff, H., Phillips, D.C., 1958. A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. Nature 181, 662-666.
Kieken, F., Arnoult, E., Barbault, F., Paquet, F., Huynh-Dinh, T., Paoletti, J., Genest, D., Lancelot, G., 2002. HIV-1(Lai) genomic RNA: combined used of NMR and molecular dynamics simulation for studying the structure and internal dynamics of a mutated SL1 hairpin. Eur Biophys J 31, 521-531.
King, N.M., Prabu-Jeyabalan, M., Nalivaika, E.A., Wigerinck, P., de Béthune, M.P., Schiffer, C.A., 2004. Structural and thermodynamic basis for the binding of TMC114, a next-generation human immunodeficiency virus type 1 protease inhibitor. J Virol 78, 12012-12021.
Kirchhoff, F., 2013. HIV Life Cycle: Overview, Encyclopedia of AIDS, pp. 1-9.
Kitchen, D.B., Decornez, H., Furr, J.R., Bajorath, J., 2004. Docking and scoring in virtual screening for drug discovery: methods and applications. Nature Reviews Drug Discovery 3, 935-949.
Kleinpeter, A.B., Freed, E.O., 2020. HIV-1 Maturation: Lessons Learned from Inhibitors. Viruses 12.
Kojić-Prodić, B., Molcanov, K., 2008. The Nature of Hydrogen Bond: New Iinsights Into Old Theories. Acta Chimica Slovenica 55, 692-708.
Kolli, M., Stawiski, E., Chappey, C., Schiffer Celia, A., 2009. Human Immunodeficiency Virus Type 1 Protease-Correlated Cleavage Site Mutations Enhance Inhibitor Resistance. Journal of Virology 83, 11027-11042.
Könnyű, B., Sadiq, S.K., Turányi, T., Hírmondó, R., Müller, B., Kräusslich, H.-G., Coveney, P.V., Müller, V., 2013. Gag-Pol Processing during HIV-1 Virion Maturation: A Systems Biology Approach. PLOS Computational Biology 9, e1003103.
Konvalinka, J., Kräusslich, H.-G., Müller, B., 2015. Retroviral proteases and their roles in virion maturation. Virology 479-480, 403-417.
Krieger, E., Nabuurs, S.B., Vriend, G., 2003. Homology modeling. Structural bioinformatics 44, 509-523.
Krishnan, L., Engelman, A., 2012. Retroviral integrase proteins and HIV-1 DNA integration. J Biol Chem 287, 40858-40866.
Kukić, P., Nielsen, J.E., 2010. Electrostatics in proteins and protein-ligand complexes. Future Med Chem 2, 647-666.

Kumar, S., Bansal, M., 1996. Structural and sequence characteristics of long alpha helices in globular proteins. Biophys J 71, 1574-1586.
Kurt Yilmaz, N., Swanstrom, R., Schiffer, C.A., 2016. Improving Viral Protease Inhibitors to Counter Drug Resistance. Trends in Microbiology 24, 547-557.
Kuzembayeva, M., Dilley, K., Sardo, L., Hu, W.S., 2014. Life of psi: how full-length HIV-1 RNAs become packaged genomes in the viral particles. Virology 454-455, 362-370.
Lehninger, A.L., Nelson, D.L., Cox, M.M., 2005. Lehninger principles of biochemistry. Macmillan.
Leidner, F., Kurt Yilmaz, N., Schiffer, C.A., 2021. Deciphering Complex Mechanisms of Resistance and Loss of Potency through Coupled Molecular Dynamics and Machine Learning. J Chem Theory Comput 17, 2054-2064.
Lengauer, T., Rarey, M., 1996. Computational methods for biomolecular docking. Current Opinion in Structural Biology 6, 402-406.
Lesk, A.M., Chothia, C., 1980. How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins. Journal of Molecular Biology 136, 225-270.
Lever, R.A., Lever, A.M., 2011. Intracellular defenses against HIV, viral evasion and novel therapeutic approaches. J Formos Med Assoc 110, 350-362.
Levin, J.G., Mitra, M., Mascarenhas, A., Musier-Forsyth, K., 2010. Role of HIV-1 nucleocapsid protein in HIV-1 reverse transcription. RNA Biology 7, 754-774.
Levinthal, C., 1969. Mossbauer spectroscopy in biological systems, Proceedings of a meeting held at Allerton House. P. Debrunner, JCM Tsibris, and E. Munck, editors. University of Illinois Press, Urbana, IL.
Levitt, M., 1981. Effect of proline residues on protein folding. J Mol Biol 145, 251-263.
Li, G., Verheyen, J., Rhee, S.-Y., Voet, A., Vandamme, A.-M., Theys, K., 2013. Functional conservation of HIV-1 Gag: implications for rational drug design. Retrovirology 10, 126.
Liang, F., Sun, M., Xie, L., Zhao, X., Liu, D., Zhao, K., Zhang, G., 2024. Recent advances and challenges in protein complex model accuracy estimation. Computational and Structural Biotechnology Journal 23, 1824-1832.
Liu, Y., Rao, U., McClure, J., Konopa, P., Manocheewa, S., Kim, M., Chen, L., Troyer, R.M., Tebit, D.M., Holte, S., Arts, E.J., Mullins, J.I., 2014. Impact of mutations in highly conserved amino acids of the HIV-1 Gag-p24 and Env-gp120 proteins on viral replication in different genetic backgrounds. PLoS One 9, e94240.
Llewellyn, G.N., Hogue, I.B., Grover, J.R., Ono, A., 2010. Nucleocapsid Promotes Localization of HIV-1 Gag to Uropods That Participate in Virological Synapses between T Cells. PLOS Pathogens 6, e1001167.
Louis, J.M., Clore, G.M., Gronenborn, A.M., 1999a. Autoprocessing of HIV-1 protease is tightly coupled to protein folding. Nat Struct Biol 6, 868-875.
Louis, J.M., Wondrak, E.M., Kimmel, A.R., Wingfield, P.T., Nashed, N.T., 1999b. Proteolytic Processing of HIV-1 Protease Precursor, Kinetics and Mechanism*. Journal of Biological Chemistry 274, 23437-23442.
Madhusudhan, M.S., Marti-Renom, M.A., Eswar, N., John, B., Pieper, U., Karchin, R., Shen, M.-Y., Sali, A., 2005. Comparative Protein Structure Modeling, in: Walker, J.M. (Ed.), The Proteomics Protocols Handbook. Humana Press, Totowa, NJ, pp. 831-860.
Maguire Michael, F., Guinea, R., Griffin, P., Macmanus, S., Elston Robert, C., Wolfram, J., Richards, N., Hanlon Mary, H., Porter David, J.T., Wrin, T., Parkin, N., Tisdale, M., Furfine, E., Petropoulos, C., Snowden, B.W., Kleim, J.-P., 2002. Changes in Human Immunodeficiency Virus Type 1 Gag at Positions L449 and P453 Are Linked to I50V Protease Mutants In Vivo and Cause Reduction of Sensitivity to Amprenavir and Improved Viral Fitness In Vitro. Journal of Virology 76, 7398-7406.

Malet, I., Roquebert, B., Dalban, C., Wirden, M., Amellal, B., Agher, R., Simon, A., Katlama, C., Costagliola, D., Calvez, V., Marcelin, A.G., 2007. Association of Gag cleavage sites to protease mutations and to virological response in HIV-1 treated patients. J Infect 54, 367-374.
Mariani, C., Desdouits, M., Favard, C., Benaroch, P., Muriaux, D.M., 2014. Role of Gag and lipids during HIV-1 assembly in CD4(+) T cells and macrophages. Front Microbiol 5, 312.
Marie, V., Gordon, M., 2019. Gag-protease coevolution shapes the outcome of lopinavir-inclusive treatment regimens in chronically infected HIV-1 subtype C patients. Bioinformatics 35, 3219-3223.
Marie, V., Gordon, M.L., 2022. The HIV-1 Gag Protein Displays Extensive Functional and Structural Roles in Virus Replication and Infectivity. Int J Mol Sci 23.
Martin-Serrano, J., Neil, S.J., 2011. Host factors involved in retroviral budding and release. Nat Rev Microbiol 9, 519-531.
Massiah, M.A., Starich, M.R., Paschall, C., Summers, M.F., Christensen, A.M., Sundquist, W.I., 1994. Three-dimensional structure of the human immunodeficiency virus type 1 matrix protein. Journal of molecular biology 244, 198-223.
Massiah, M.A., Worthylake, D., Christensen, A.M., Sundquist, W.I., Hill, C.P., Summers, M.F., 1996. Comparison of the NMR and X-ray structures of the HIV-1 matrix protein: Evidence for conformational changes during viral assembly. Protein Science 5, 2391-2398.
Massova, I., Kollman, P.A., 2000. Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding. Perspectives in Drug Discovery and Design 18, 113-135.
Matthew, A.N., Leidner, F., Lockbaum, G.J., Henes, M., Zephyr, J., Hou, S., Rao, D.N., Timm, J., Rusere, L.N., Ragland, D.A., Paulsen, J.L., Prachanronarong, K., Soumana, D.I., Nalivaika, E.A., Kurt Yilmaz, N., Ali, A., Schiffer, C.A., 2021. Drug Design Strategies to Avoid Resistance in Direct-Acting Antivirals and Beyond. Chem Rev 121, 3238-3270.
Mayer, M.P., 2006. Protein Folding, Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1495-1503.
McCutchan, F.E., Hegerich, P.A., Brennan, T.P., Phanuphak, P., Singharaj, P., Jugsudee, A., Berman, P.W., Gray, A.M., Fowler, A.K., Burke, D.S., 1992. Genetic variants of HIV-1 in Thailand. AIDS Res Hum Retroviruses 8, 1887-1895.
Meng, E.C., Pettersen, E.F., Couch, G.S., Huang, C.C., Ferrin, T.E., 2006. Tools for integrated sequence-structure analysis with UCSF Chimera. BMC Bioinformatics 7, 339.
Meng, X.Y., Zhang, H.X., Mezei, M., Cui, M., 2011. Molecular docking: a powerful approach for structure-based drug discovery. Curr Comput Aided Drug Des 7, 146-157.
Miller, M., Schneider, J., Sathyanarayana, B., Toth, M., Marshall, G., Clawson, L., Selk, L., Kent, S., Wlodawer, A., 1989. Structure of complex of synthetic HIV-1 protease with a substrate-based inhibitor at 2.3 A resolution. Science 246, 1149-1152.
Miteva, M., Robert, C., Maréchal, J.-D., Perahia, D., 2011. Receptor Flexibility in Ligand Docking and Virtual Screening, pp. 99-117.
Miyazaki, Y., Miyake, A., Nomaguchi, M., Adachi, A., 2011. Structural Dynamics of Retroviral Genome and the Packaging. Frontiers in Microbiology 2.
Monje-Galvan, V., Voth, G.A., 2020. Binding mechanism of the matrix domain of HIV-1 gag on lipid membranes. eLife 9, e58621.
Morita, E., Sundquist, W.I., 2004. Retrovirus budding. Annu Rev Cell Dev Biol 20, 395-425.
Morris, G.M., Goodsell, D.S., Halliday, R.S., Huey, R., Hart, W.E., Belew, R.K., Olson, A.J., 1998. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. Journal of Computational Chemistry 19, 1639-1662.
Muriaux, D., Darlix, J.L., 2010. Properties and functions of the nucleocapsid protein in virus assembly. RNA Biol 7, 744-753.

Muzammil, S., Ross, P., Freire, E., 2003. A major role for a set of non-active site mutations in the development of HIV-1 protease drug resistance. Biochemistry 42, 631-638.
Myers, G., 1994. Tenth anniversary perspectives on AIDS. HIV: between past and future. AIDS Res Hum Retroviruses 10, 1317-1324.
Myint, L., Matsuda, M., Matsuda, Z., Yokomaku, Y., Chiba, T., Okano, A., Yamada, K., Sugiura, W., 2004. Gag non-cleavage site mutations contribute to full recovery of viral fitness in protease inhibitor-resistant human immunodeficiency virus type 1. Antimicrob Agents Chemother 48, 444-452.
Nalam, M.N., Ali, A., Altman, M.D., Reddy, G.S., Chellappan, S., Kairys, V., Ozen, A., Cao, H., Gilson, M.K., Tidor, B., Rana, T.M., Schiffer, C.A., 2010. Evaluating the substrate-envelope hypothesis: structural analysis of novel HIV-1 protease inhibitors designed to be robust against drug resistance. J Virol 84, 5368-5378.
Nalam, M.N., Ali, A., Reddy, G.S., Cao, H., Anjum, S.G., Altman, M.D., Yilmaz, N.K., Tidor, B., Rana, T.M., Schiffer, C.A., 2013. Substrate envelope-designed potent HIV-1 protease inhibitors to avoid drug resistance. Chem Biol 20, 1116-1124.
Nalam, M.N., Schiffer, C.A., 2008. New approaches to HIV protease inhibitor drug design II: testing the substrate envelope hypothesis to avoid drug resistance and discover robust inhibitors. Curr Opin HIV AIDS 3, 642-646.
Newberry, R.W., Raines, R.T., 2019. Secondary Forces in Protein Folding. ACS Chem Biol 14, 1677-1686.
Ni, N., Nikolaitchik, O.A., Dilley, K.A., Chen, J., Galli, A., Fu, W., Prasad, V.V.S.P., Ptak, R.G., Pathak, V.K., Hu, W.-S., 2011. Mechanisms of Human Immunodeficiency Virus Type 2 RNA Packaging: Efficient trans Packaging and Selection of RNA Copackaging Partners. Journal of Virology 85, 7603-7612.
Nijhuis, M., van Maarseveen, N.M., Lastere, S., Schipper, P., Coakley, E., Glass, B., Rovenska, M., de Jong, D., Chappey, C., Goedegebuure, I.W., Heilek-Snyder, G., Dulude, D., Cammack, N., Brakier-Gingras, L., Konvalinka, J., Parkin, N., Kräusslich, H.G., Brun-Vezinet, F., Boucher, C.A., 2007. A novel substrate-based HIV-1 protease inhibitor drug resistance mechanism. PLoS Med 4, e36.
Nisole, S., Saïb, A., 2004. Early steps of retrovirus replicative cycle. Retrovirology 1, 9.
Nussinov, R., Zhang, M., Liu, Y., Jang, H., 2022. AlphaFold, Artificial Intelligence (AI), and Allostery. J Phys Chem B 126, 6372-6383.
Ocwieja, K.E., Sherrill-Mix, S., Mukherjee, R., Custers-Allen, R., David, P., Brown, M., Wang, S., Link, D.R., Olson, J., Travers, K., Schadt, E., Bushman, F.D., 2012. Dynamic regulation of HIV-1 mRNA populations analyzed by single-molecule enrichment and long-read sequencing. Nucleic Acids Res 40, 10345-10355.
Olesen, J.S., Jespersen, S., da Silva, Z.J., Rodrigues, A., Erikstrup, C., Aaby, P., Wejse, C., Hønge, B.L., 2018. HIV-2 continues to decrease, whereas HIV-1 is stabilizing in Guinea-Bissau. Aids 32, 1193-1198.
Oliveira, d., Engelbrecht, S., Rensburg, E.J.v., Gordon, M., Bishop, K., Megede, J.z., Barnett, S.W., Cassol, S., 2003. Variability at Human Immunodeficiency Virus Type 1 Subtype C Protease Cleavage Sites: an Indication of Viral Fitness? Journal of Virology 77, 9422-9430.
Ott, D.E., Coren, L.V., Shatzer, T., 2009. The nucleocapsid region of human immunodeficiency virus type 1 Gag assists in the coordination of assembly and Gag processing: role for RNA-Gag binding in the early stages of assembly. J Virol 83, 7718-7727.
Özen, A., Haliloğlu, T., Schiffer, C.A., 2011. Dynamics of Preferential Substrate Recognition in HIV-1 Protease: Redefining the Substrate Envelope. Journal of Molecular Biology 410, 726-744.

Özen, A., Lin, K.H., Kurt Yilmaz, N., Schiffer, C.A., 2014. Structural basis and distal effects of Gag substrate coevolution in drug resistance to HIV-1 protease. Proc Natl Acad Sci U S A 111, 15993-15998.
Pace, C.N., Fu, H., Lee Fryar, K., Landua, J., Trevino, S.R., Schell, D., Thurlkill, R.L., Imura, S., Scholtz, J.M., Gajiwala, K., 2014. Contribution of hydrogen bonds to protein stability. Protein Science 23, 652-661.
Pauling, L., Corey, R.B., 1951. Atomic Coordinates and Structure Factors for Two Helical Configurations of Polypeptide Chains. Proceedings of the National Academy of Sciences 37, 235-240.
Pauling, L., Corey, R.B., Branson, H.R., 1951. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. Proceedings of the National Academy of Sciences 37, 205-211.
Peramo, A., 2016. Solvated and generalised Born calculations differences using GPU CUDA and multi-CPU simulations of an antifreeze protein with AMBER. Molecular Simulation 42, 1263-1273.
Perutz, M.F., Rossmann, M.G., Cullis, A.F., Muirhead, H., Will, G., North, A.C.T., 1960. Structure of Hæmoglobin: A Three-Dimensional Fourier Synthesis at 5.5-Å. Resolution, Obtained by X-Ray Analysis. Nature 185, 416-422.
Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., Ferrin, T.E., 2004. UCSF Chimera—A visualization system for exploratory research and analysis. Journal of Computational Chemistry 25, 1605-1612.
Pettit, S.C., Henderson, G.J., Schiffer, C.A., Swanstrom, R., 2002. Replacement of the P1 amino acid of human immunodeficiency virus type 1 Gag processing sites can inhibit or enhance the rate of cleavage by the viral protease. J Virol 76, 10226-10233.
Pettit, S.C., Lindquist, J.N., Kaplan, A.H., Swanstrom, R., 2005. Processing sites in the human immunodeficiency virus type 1 (HIV-1) Gag-Pro-Pol precursor are cleaved by the viral protease at different rates. Retrovirology 2, 66.
Pettit, S.C., Moody, M.D., Wehbie, R.S., Kaplan, A.H., Nantermet, P.V., Klein, C.A., Swanstrom, R., 1994. The p2 domain of human immunodeficiency virus type 1 Gag regulates sequential proteolytic processing and is required to produce fully infectious virions. J Virol 68, 8017-8027.
Pettit, S.C., Simsic, J., Loeb, D.D., Everitt, L., Hutchison, C.A., Swanstrom, R., 1991. Analysis of retroviral protease cleavage sites reveals two types of cleavage sites and the structural requirements of the P1 amino acid. Journal of Biological Chemistry 266, 14539-14547.
Poon, A.F.Y., Ndashimye, E., Avino, M., Gibson, R., Kityo, C., Kyeyune, F., Nankya, I., Quiñones-Mateu, M.E., Arts, E.J., Paton, N.I., Walker, S., Hoppe, A., Rinke de Wit, T.F., Sigaloff, K.C.E., Hamers, R., Boender, T.S., Boerma, R.S., Ondoa, P., Nijboer, M., Kroeze, S., Inzaule, S., Kityo Mutuluuza, C., Akanmu, A.S., The Ugandan Drug, Resistance Study T., 2019. First-line HIV treatment failures in non-B subtypes and recombinants: a cross-sectional analysis of multiple populations in Uganda. AIDS Research and Therapy 16, 3.
Pornillos, O., Ganser-Pornillos, B.K., Yeager, M., 2011. Atomic-level modelling of the HIV capsid. Nature 469, 424-427.
Potempa, M., Lee, S.-K., Kurt Yilmaz, N., Nalivaika, E.A., Rogers, A., Spielvogel, E., Carter, C.W., Schiffer, C.A., Swanstrom, R., 2018. HIV-1 Protease Uses Bi-Specific S2/S2′ Subsites to Optimize Cleavage of Two Classes of Target Sites. Journal of Molecular Biology 430, 5182-5195.
Prabu-Jeyabalan, M., Nalivaika, E., Schiffer, C.A., 2000. How does a symmetric dimer recognize an asymmetric substrate? a substrate complex of HIV-1 protease11Edited by I. Wilson. Journal of Molecular Biology 301, 1207-1220.

Prabu-Jeyabalan, M., Nalivaika, E., Schiffer, C.A., 2002. Substrate Shape Determines Specificity of Recognition for HIV-1 Protease: Analysis of Crystal Structures of Six Substrate Complexes. Structure 10, 369-381.
Prabu-Jeyabalan, M., Nalivaika, E.A., King, N.M., Schiffer, C.A., 2003. Viability of a Drug-Resistant Human Immunodeficiency Virus Type 1 Protease Variant: Structural Insights for Better Antiviral Therapy. Journal of Virology 77, 1306-1315.
Prabu-Jeyabalan, M., Nalivaika, E.A., King, N.M., Schiffer, C.A., 2004. Structural Basis for Coevolution of a Human Immunodeficiency Virus Type 1 Nucleocapsid-p1 Cleavage Site with a V82A Drug-Resistant Mutation in Viral Protease. Journal of Virology 78, 12446-12454.
Prabu, M., Nalivaika, E., Romano, K., Schiffer, C., 2006. Mechanism of Substrate Recognition by Drug-Resistant Human Immunodeficiency Virus Type 1 Protease Variants Revealed by a Novel Structural Intermediate. Journal of virology 80, 3607-3616.
Prado, J.G., Wrin, T., Beauchaine, J., Ruiz, L., Petropoulos, C.J., Frost, S.D., Clotet, B., D'Aquila, R.T., Martinez-Picado, J., 2002. Amprenavir-resistant HIV-1 exhibits lopinavir cross-resistance and reduced replication capacity. Aids 16, 1009-1017.
Priyakumar, U.D., 2012. Role of hydrophobic core on the thermal stability of proteins - molecular dynamics simulations on a single point mutant of Sso7d abstract. J Biomol Struct Dyn 29, 961-971.
Qiu, X., Li, H., Ver Steeg, G., Godzik, A., 2024. Advances in AI for Protein Structure Prediction: Implications for Cancer Drug Discovery and Development. Biomolecules 14.
Ragland, D.A., Nalivaika, E.A., Nalam, M.N., Prachanronarong, K.L., Cao, H., Bandaranayake, R.M., Cai, Y., Kurt-Yilmaz, N., Schiffer, C.A., 2014. Drug resistance conferred by mutations outside the active site through alterations in the dynamic and structural ensemble of HIV-1 protease. J Am Chem Soc 136, 11956-11963.
Ragland, D.A., Whitfield, T.W., Lee, S.K., Swanstrom, R., Zeldovich, K.B., Kurt-Yilmaz, N., Schiffer, C.A., 2017. Elucidating the Interdependence of Drug Resistance from Combinations of Mutations. J Chem Theory Comput 13, 5671-5682.
Raiborg, C., Stenmark, H., 2009. The ESCRT machinery in endosomal sorting of ubiquitylated membrane proteins. Nature 458, 445-452.
Rauwerdink, A., Kazlauskas, R.J., 2015. How the Same Core Catalytic Machinery Catalyzes 17 Different Reactions: the Serine-Histidine-Aspartate Catalytic Triad of α/β-Hydrolase Fold Enzymes. ACS Catal 5, 6153-6176.
Ravikumar, A., Ramakrishnan, C., Srinivasan, N., 2019. Stereochemical Assessment of (φ,ψ) Outliers in Protein Structures Using Bond Geometry-Specific Ramachandran Steric-Maps. Structure 27.
Rhee, S.Y., Gonzales, M.J., Kantor, R., Betts, B.J., Ravela, J., Shafer, R.W., 2003. Human immunodeficiency virus reverse transcriptase and protease sequence database. Nucleic Acids Res 31, 298-303.
Richards, F.M., 1974. The interpretation of protein structures: total volume, group volume distributions and packing density. J Mol Biol 82, 1-14.
Richardson, J.S., 1981. The anatomy and taxonomy of protein structure. Adv Protein Chem 34, 167-339.
Rihn, S.J., Wilson, S.J., Loman, N.J., Alim, M., Bakker, S.E., Bhella, D., Gifford, R.J., Rixon, F.J., Bieniasz, P.D., 2013. Extreme Genetic Fragility of the HIV-1 Capsid. PLOS Pathogens 9, e1003461.
Rossi, E., Meuser, M.E., Cunanan, C.J., Cocklin, S., 2021. Structure, function, and interactions of the HIV-1 capsid protein. Life 11, 100.
Saito, A., Yamashita, M., 2021. HIV-1 capsid variability: viral exploitation and evasion of capsid-binding molecules. Retrovirology 18, 32.

Sali, A., 1989-2022. UCSF MODELLER, University of California San Francisco, San Francisco, CA 94143, USA.
Sali, A., Blundell, T.L., 1993. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234, 779-815.
Salomon-Ferrer, R., Case, D.A., Walker, R.C., 2013. An overview of the Amber biomolecular simulation package. WIREs Computational Molecular Science 3, 198-210.
Salsbury, F.R., Jr., 2010. Molecular dynamics simulations of protein dynamics and their relevance to drug discovery. Curr Opin Pharmacol 10, 738-744.
Samant, N., Nachum, G., Tsepal, T., Bolon, D., 2022. Sequence dependencies and biophysical features both govern cleavage of diverse cut-sites by HIV protease. Protein Science 31.
Samsudin, F., Gan, S.K.-E., Bond, P.J., 2021. The impact of Gag non-cleavage site mutations on HIV-1 viral fitness from integrative modelling and simulations. Computational and Structural Biotechnology Journal 19, 330-342.
Sanches, M., Krauchenco, S., Martins, N.H., Gustchina, A., Wlodawer, A., Polikarpov, I., 2007. Structural characterization of B and non-B subtypes of HIV-protease: insights into the natural susceptibility to drug resistance development. J Mol Biol 369, 1029-1040.
Sanger, F., 1952. The Arrangement of Amino Acids in Proteins, in: Anson, M.L., Bailey, K., Edsall, J.T. (Eds.), Advances in Protein Chemistry. Academic Press, pp. 1-67.
Sanger, F., Tuppy, H., 1951. The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. Biochem J 49, 481-490.
Sankaran, S.V., Krishnan, S.R., Sayed, Y., Gromiha, M.M., 2024. Mechanism of drug resistance in HIV-1 protease subtype C in the presence of Atazanavir. Current Research in Structural Biology 7, 100132.
Santos, L.H.S., Ferreira, R.S., Caffarena, E.R., 2019. Integrating Molecular Docking and Molecular Dynamics Simulations. Methods Mol Biol 2053, 13-34.
Saravanan, K.M., Selvaraj, S., 2017. Dihedral angle preferences of amino acid residues forming various non-local interactions in proteins. J Biol Phys 43, 265-278.
Sarkar, A., Concilio, S., Sessa, L., Marrafino, F., Piotto, S., 2024. Advancements and novel approaches in modified AutoDock Vina algorithms for enhanced molecular docking. Results in Chemistry 7, 101319.
Sasmal, D.K., Pulido, L.E., Kasal, S., Huang, J., 2016. Single-molecule fluorescence resonance energy transfer in molecular biology. Nanoscale 8, 19928-19944.
Scheraga, H.A., 1998. Theory of hydrophobic interactions. J Biomol Struct Dyn 16, 447-460.
Schmidt, T., Bergner, A., Schwede, T., 2014. Modelling three-dimensional protein structures for applications in drug design. Drug Discovery Today 19, 890-897.
Schweighardt, B., Wrin, T., Meiklejohn, D.A., Spotts, G., Petropoulos, C.J., Nixon, D.F., Hecht, F.M., 2010. Immune escape mutations detected within HIV-1 epitopes associated with viral control during treatment interruption. J Acquir Immune Defic Syndr 53, 36-46.
Scott, W.R.P., Schiffer, C.A., 2000. Curling of Flap Tips in HIV-1 Protease as a Mechanism for Substrate Entry and Tolerance of Drug Resistance. Structure 8, 1259-1265.
Seabra, G.d.M., Walker, R.C., Elstner, M., Case, D.A., Roitberg, A.E., 2007. Implementation of the SCC-DFTB Method for Hybrid QM/MM Simulations within the Amber Molecular Dynamics Package. The Journal of Physical Chemistry A 111, 5655-5664.
Senn, H.M., Thiel, W., 2009. QM/MM Methods for Biomolecular Systems. Angewandte Chemie International Edition 48, 1198-1229.
Sharp, P.M., Hahn, B.H., 2011. Origins of HIV and the AIDS Pandemic. Cold Spring Harbor Perspectives in Medicine 1.
Shehu-Xhilaga, M., Kraeusslich, H.G., Pettit, S., Swanstrom, R., Lee, J.Y., Marshall, J.A., Crowe, S.M., Mak, J., 2001. Proteolytic processing of the p2/nucleocapsid cleavage site is

critical for human immunodeficiency virus type 1 RNA dimer maturation. J Virol 75, 9156-9164.
Sheinerman, F.B., Norel, R., Honig, B., 2000. Electrostatic aspects of protein-protein interactions. Curr Opin Struct Biol 10, 153-159.
Shen, Y., Altman, M.D., Ali, A., Nalam, M.N., Cao, H., Rana, T.M., Schiffer, C.A., Tidor, B., 2013. Testing the substrate-envelope hypothesis with designed pairs of compounds. ACS Chem Biol 8, 2433-2441.
Sitkoff, D., Sharp, K.A., Honig, B., 1994. Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models. The Journal of Physical Chemistry 98, 1978-1988.
Son, A., Kim, W., Park, J., Lee, W., Lee, Y., Choi, S., Kim, H., 2024. Utilizing Molecular Dynamics Simulations, Machine Learning, Cryo-EM, and NMR Spectroscopy to Predict and Validate Protein Dynamics. Int J Mol Sci 25.
Spassov, D.S., 2024. Binding Affinity Determination in Drug Design: Insights from Lock and Key, Induced Fit, Conformational Selection, and Inhibitor Trapping Models. Int J Mol Sci 25.
Spielvogel, E., Lee, S.K., Zhou, S., Lockbaum, G.J., Henes, M., Sondgeroth, A., Kosovrasti, K., Nalivaika, E.A., Ali, A., Yilmaz, N.K., Schiffer, C.A., Swanstrom, R., 2023. Selection of HIV-1 for resistance to fifth-generation protease inhibitors reveals two independent pathways to high-level resistance. Elife 12.
Srinivasan, J., Miller, J., Kollman, P.A., Case, D.A., 1998. Continuum Solvent Studies of the Stability of RNA Hairpin Loops and Helices. Journal of Biomolecular Structure and Dynamics 16, 671-682.
Sriramulu, D.K., Lee, S.-G., 2021. Effect of molecular properties of the protein-ligand complex on the prediction accuracy of AutoDock. Journal of Molecular Graphics and Modelling 106, 107921.
Stern A, A.R., 2016. It Is All About Mutations. Viral Evolution: Viral Pathogenesis. :. , 233–240.
Strack, B., Calistri, A., Craig, S., Popova, E., Göttlinger, H.G., 2003. AIP1/ALIX Is a Binding Partner for HIV-1 p6 and EIAV p9 Functioning in Virus Budding. Cell 114, 689-699.
Stremlau, M., Owens, C.M., Perron, M.J., Kiessling, M., Autissier, P., Sodroski, J., 2004. The cytoplasmic body component TRIM5α restricts HIV-1 infection in Old World monkeys. Nature 427, 848-853.
Su, C.T., Kwoh, C.K., Verma, C.S., Gan, S.K., 2018. Modeling the full length HIV-1 Gag polyprotein reveals the role of its p6 subunit in viral maturation and the effect of non-cleavage site mutations in protease drug resistance. J Biomol Struct Dyn 36, 4366-4377.
Suguna, K., Padlan, E.A., Smith, C.W., Carlson, W.D., Davies, D.R., 1987. Binding of a reduced peptide inhibitor to the aspartic proteinase from Rhizopus chinensis: implications for a mechanism of action. Proc Natl Acad Sci U S A 84, 7009-7013.
Sun, Q., 2022. The Hydrophobic Effects: Our Current Understanding. Molecules 27.
Sundquist, W.I., Krausslich, H.G., 2012. HIV-1 assembly, budding, and maturation. Cold Spring Harb Perspect Med 2, a006924.
Sutherland, K.A., Collier, D.A., Claiborne, D.T., Prince, J.L., Deymier, M.J., Goldstein, R.A., Hunter, E., Gupta, R.K., 2016. Wide variation in susceptibility of transmitted/founder HIV-1 subtype C Isolates to protease inhibitors and association with in vitro replication efficiency. Scientific Reports 6, 38153.
Sutherland, K.A., Goodall, R.L., McCormick, A., Kapaata, A., Lyagoba, F., Kaleebu, P., Thiltgen, G., Gilks, C.F., Spyer, M., Kityo, C., Pillay, D., Dunn, D., Gupta, R.K., 2015. Gag-Protease Sequence Evolution Following Protease Inhibitor Monotherapy Treatment Failure in HIV-1 Viruses Circulating in East Africa. AIDS Res Hum Retroviruses 31, 1032-1037.
Systèmes, D., 2021. Biovia Discovery Studio 2021. Dassault Systèmes, San Diego, CA, USA.

Tanford, C., 1962. Contribution of Hydrophobic Interactions to the Stability of the Globular Conformation of Proteins. Journal of the American Chemical Society 84, 4240-4247.
Teto, G., Tagny, C.T., Mbanya, D., Fonsah, J.Y., Fokam, J., Nchindap, E., Kenmogne, L., Njamnshi, A.K., Kanmogne, G.D., 2017. Gag P2/NC and pol genetic diversity, polymorphism, and drug resistance mutations in HIV-1 CRF02_AG- and non-CRF02_AG-infected patients in Yaoundé, Cameroon. Scientific Reports 7, 14136.
Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22, 4673-4680.
Tie, Y., Boross, P.I., Wang, Y.F., Gaddis, L., Hussain, A.K., Leshchenko, S., Ghosh, A.K., Louis, J.M., Harrison, R.W., Weber, I.T., 2004. High resolution crystal structures of HIV-1 protease with a potent non-peptide inhibitor (UIC-94017) active against multi-drug-resistant clinical strains. J Mol Biol 338, 341-352.
Tie, Y., Wang, Y.F., Boross, P.I., Chiu, T.Y., Ghosh, A.K., Tozser, J., Louis, J.M., Harrison, R.W., Weber, I.T., 2012. Critical differences in HIV-1 and HIV-2 protease specificity for clinical inhibitors. Protein Sci 21, 339-350.
Todd, M.J., Semo, N., Freire, E., 1998. The structural stability of the HIV-1 protease. J Mol Biol 283, 475-488.
Torrecilla, E., Llácer Delicado, T., Holguín, Á., 2014. New findings in cleavage sites variability across groups, subtypes and recombinants of human immunodeficiency virus type 1. PLoS One 9, e88099.
Torres, P.H., Sodero, A.C., Jofily, P., Silva-Jr, F.P., 2019. Key topics in molecular docking for drug design. International journal of molecular sciences 20, 4574.
Trott, O., Olson, A.J., 2010. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 31, 455-461.
Usami, Y., Popov, S., Popova, E., Inoue, M., Weissenhorn, W., H, G.G., 2009. The ESCRT pathway and HIV-1 budding. Biochem Soc Trans 37, 181-184.
Vajda, S., Guarnieri, F., 2006. Characterization of protein-ligand interaction sites using experimental and computational methods. Current Opinion in Drug Discovery and Development 9, 354.
Van den Burg, B., Dijkstra, B.W., Vriend, G., Van der Vinne, B., Venema, G., Eijsink, V.G., 1994. Protein stabilization by hydrophobic interactions at the surface. Eur J Biochem 220, 981-985.
Velazquez-Campoy, A., Muzammil, S., Ohtaka, H., Schön, A., Vega, S., Freire, E., 2003. Structural and thermodynamic basis of resistance to HIV-1 protease inhibition: implications for inhibitor design. Curr Drug Targets Infect Disord 3, 311-328.
Velazquez-Campoy, A., Todd, M.J., Vega, S., Freire, E., 2001. Catalytic efficiency and vitality of HIV-1 proteases from African viral subtypes. Proceedings of the National Academy of Sciences 98, 6062-6067.
Venkatachalam, S., Murlidharan, N., Krishnan, S.R., Ramakrishnan, C., Setshedi, M., Pandian, R., Barh, D., Tiwari, S., Azevedo, V., Sayed, Y., Gromiha, M.M., 2023. Understanding Drug Resistance of Wild-Type and L38HL Insertion Mutant of HIV-1 C Protease to Saquinavir. Genes (Basel) 14.
Verheyen, J., Verhofstede, C., Knops, E., Vandekerckhove, L., Fun, A., Brunen, D., Dauwe, K., Wensing, A.M., Pfister, H., Kaiser, R., Nijhuis, M., 2010. High prevalence of bevirimat resistance mutations in protease inhibitor-resistant HIV isolates. AIDS 24, 669-673.
Visseaux, B., Damond, F., Matheron, S., Descamps, D., Charpentier, C., 2016. Hiv-2 molecular epidemiology. Infect Genet Evol 46, 233-240.

Vlachakis, D., Bencurova, E., Papangelopoulos, N., Kossida, S., 2014. Chapter Seven - Current State-of-the-Art Molecular Dynamics Methods and Applications, in: Donev, R. (Ed.), Advances in Protein Chemistry and Structural Biology. Academic Press, pp. 269-313.
Voet, D., Voet, J.G., 2010. Biochemistry. John Wiley & Sons.
Votteler, J., Sundquist, W.I., 2013. Virus budding and the ESCRT pathway. Cell Host Microbe 14, 232-241.
Vyas, V.K., Ukawala, R.D., Ghate, M., Chintha, C., 2012. Homology modeling a fast tool for drug discovery: current perspectives. Indian J Pharm Sci 74, 1-17.
Wallace, A.C., Laskowski, R.A., Thornton, J.M., 1995. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. Protein Engineering 8, 127-134.
Wang, C., Greene, D., Xiao, L., Qi, R., Luo, R., 2017. Recent Developments and Applications of the MMPBSA Method. Front Mol Biosci 4, 87.
Wang, E., Sun, H., Wang, J., Wang, Z., Liu, H., Zhang, J.Z.H., Hou, T., 2019. End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design. Chemical Reviews 119, 9478-9508.
Wang, J., Wolf, R.M., Caldwell, J.W., Kollman, P.A., Case, D.A., 2004. Development and testing of a general amber force field. Journal of Computational Chemistry 25, 1157-1174.
Warshel, A., Levitt, M., 1976. Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. Journal of Molecular Biology 103, 227-249.
Warshel, A., Sharma, P.K., Kato, M., Xiang, Y., Liu, H., Olsson, M.H.M., 2006. Electrostatic Basis for Enzyme Catalysis. Chemical Reviews 106, 3210-3235.
Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L., Lepore, R., Schwede, T., 2018a. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res 46, W296-w303.
Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A P., Rempfer, C., Bordoli, L., Lepore, R., Schwede, T., 2018b. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Research 46, W296-W303.
Watson, G., Velasco-Berrelleza, V., Noy, A., 2022. Atomistic Molecular Dynamics Simulations of DNA in Complex 3D Arrangements for Comparison with Lower Resolution Structural Experiments, in: Leake, M.C. (Ed.), Chromosome Architecture: Methods and Protocols. Springer US, New York, NY, pp. 95-109.
Watts, J.M., Dang, K.K., Gorelick, R.J., Leonard, C.W., Bess, J.W., Jr., Swanstrom, R., Burch, C.L., Weeks, K.M., 2009. Architecture and secondary structure of an entire HIV-1 RNA genome. Nature 460, 711-716.
Webb, B., Sali, A., 2016. Comparative Protein Structure Modeling Using MODELLER. Curr Protoc Bioinformatics 54, 5.6.1-5.6.37.
Weber, I.T., Agniswamy, J., 2009. HIV-1 Protease: Structural Perspectives on Drug Resistance. Viruses 1, 1110-1136.
Weber, I.T., Miller, M., Jaskólski, M., Leis, J., Skalka, A.M., Wlodawer, A., 1989. Molecular Modeling of the HIV-1 Protease and Its Substrate Binding Site. Science 243, 928-931.
Wensing, A.M.J., Fun, A., Nijhuis, M., 2014. HIV Protease Inhibitor Resistance, in: Gotte, M., Berghuis, A., Matlashewski, G., Wainberg, M., Sheppard, D. (Eds.), Handbook of Antimicrobial Resistance. Springer New York, New York, NY, pp. 1-31.
Wensing, A.M.J., Fun, A., Nijhuis, M., 2017. HIV Protease Inhibitor Resistance, in: Berghuis, A., Matlashewski, G., Wainberg, M.A., Sheppard, D., Gotte, M. (Eds.), Handbook of Antimicrobial Resistance. Springer New York, New York, NY, pp. 567-602.

Wetlaufer, D.B., 1973. Nucleation, Rapid Folding, and Globular Intrachain Regions in Proteins. Proceedings of the National Academy of Sciences 70, 697-701.
Wibisono, A., Suhartanto, H., 2012. Cloud computing model and implementation of molecular dynamics simulation using Amber and Gromacs, 2012 International Conference on Advanced Computer Science and Information Systems (ICACSIS), pp. 31-36.
Wilen, C.B., Tilton, J.C., Doms, R.W., 2012. HIV: cell binding and entry. Cold Spring Harb Perspect Med 2.
Williams, K.C., Burdo, T.H., 2009. HIV and SIV infection: the role of cellular restriction and immune responses in viral replication and pathogenesis. Apmis 117, 400-412.
Williams, M.A., 2013. Protein–ligand interactions: Fundamentals. Protein-ligand interactions: Methods and applications, 3-34.
Wlodawer, A., Erickson, J.W., 1993. STRUCTURE-BASED INHIBITORS OF HIV-1 PROTEASE. Annual Review of Biochemistry 62, 543-585.
Word JM, L.S., Richardson JS, Richardson DC. , 1999. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. 285(4):1735-47. doi: 10.1006/jmbi.1998.2401. PMID: 9917408.

.
Wright, E.R., Schooler, J.B., Ding, H.J., Kieffer, C., Fillmore, C., Sundquist, W.I., Jensen, G.J., 2007. Electron cryotomography of immature HIV-1 virions reveals the structure of the CA and SP1 Gag shells. Embo j 26, 2218-2226.
Wu, K., Karapetyan, E., Schloss, J., Vadgama, J., Wu, Y., 2023. Advancements in small molecule drug design: A structural perspective. Drug Discov Today 28, 103730.
Xiao, F., Chen, Z., Wei, Z., Tian, L., 2020. Hydrophobic Interaction: A Promising Driving Force for the Biomedical Applications of Nucleic Acids. Adv Sci (Weinh) 7, 2001048.
Xu, L., Sun, H., Li, Y., Wang, J., Hou, T., 2013. Assessing the Performance of MM/PBSA and MM/GBSA Methods. 3. The Impact of Force Fields and Ligand Charge Models. The Journal of Physical Chemistry B 117, 8408-8421.
Xue, B., Mizianty, M., Kurgan, L., Uversky, V., 2011. Protein intrinsic disorder as a flexible armor and a weapon of HIV-1. Cellular and molecular life sciences : CMLS 69, 1211-1259.
Yan, B.X., Sun, Y.Q., 1997. Glycine residues provide flexibility for enzyme active sites. J Biol Chem 272, 3190-3194.
Yang, Z., Lasker, K., Schneidman-Duhovny, D., Webb, B., Huang, C.C., Pettersen, E.F., Goddard, T.D., Meng, E.C., Sali, A., Ferrin, T.E., 2012. UCSF Chimera, MODELLER, and IMP: An integrated modeling system. Journal of Structural Biology 179, 269-278.
Yu, H., Zhao, Y., Guo, C., Gan, Y., Huang, H., 2015. The role of proline substitutions within flexible regions on thermostability of luciferase. Biochim Biophys Acta 1854, 65-72.
Zephyr, J., Kurt Yilmaz, N., Schiffer, C.A., 2021. Viral proteases: Structure, mechanism and inhibition. Enzymes 50, 301-333.
Zhou, H.X., Pang, X., 2018. Electrostatic Interactions in Protein Structure, Folding, Binding, and Condensation. Chem Rev 118, 1691-1741.
Zhu, P., Chertova, E., Bess, J., Lifson, J.D., Arthur, L.O., Liu, J., Taylor, K.A., Roux, K.H., 2003. Electron tomography analysis of envelope glycoprotein trimers on HIV and simian immunodeficiency virus virions. Proceedings of the National Academy of Sciences 100, 15812-15817.

5 Appendix

Table 4: Summary of subtype B CS interactions

Cleavage Site	Sequence	Bond type	Number of bonds	Interacting residues	Distances (Angstroms)	Binding affinity(kcal/mol) must be Gbind	Position and ligand
MA/CA	VSQNY/PIVQN	Unfavourable donor-donor	1	Arg8	5.96		P3' valine
		van der waals	6	ASP29	3.90		P5' asparagine
				ALA28	3.95		P3'
				Asp30	5.28		P5' asparagine
				GLY48	4.57		P2'
				Catalytic Gly27	3.80		P2'
				Gly49	3.86		P1'
				Catalytic D25	3.21		P3
				Asp29	4.76		P3
		Conventional hydrogen bond	9	ASP30	4.08		P4'
				ASP29	3.51		P2'
					3.90		P3'
					5.09		P2'
				GLY48	4.57		P3'
					3.70		P3
					3.18	-7.8	P4
				Catalytic GLY27	4.14		P3
				GLY48	4.30		P2'
				ARG8	7.06		P2'
					6.43		P1
				THR80	4.47		P1
		Unfavourable acceptor-acceptor	1	GLY48	3.17		P1
		Alkyl	5	VAL32	6.17		P2'
				ILE50	6.07		P2'
				ILE84	6.35		P1'
				ILE47	5.92		P2'
				ILE50	6.07		P2'
				ALA28	4.04		P1
		π-Alkyl	5	ALA28	5.63		P1
				VAL32	6.62		P1
				ILE47	6.16		P1
				ILE50	4.08		P1
				ILE84	5.22		P1
CA/P2		Alkyl	11	ILE50	5.08	-4.2	P1'

	KARVL/AEAMS			ILE84	5.92		P1’
				VAL32	6.00		P1’
				ILE47	4.38		P1’
				ALA28	4.68		P1’
				ILE50	5.32		P1
				PRO81	4.06		P1
				ILE47	4.60		P1’
				LEU23	3.79		P4’
				VAL82	6.06		P1’
					3.97		P4’
				LEU23(Chain B)	6.73		P4’
		Hydrogen bond	7	ASP29	3.96		P2
				ASP30	5.43		P5’
				GLY49	3.70		P1
				GLY49	3.67		P2
				ASP30(Chain B)	2.88		P5
				ALA28	4.14		P3
				LEU23	3.79		P3
		Conventional hydrogen bond	15	GLY48	3.69		P5
					3.37		P5
				ASP29	3.36		P4
				GLY48(CHAIN A)	3.16		P1’
					4.17		P1’
				ILE50	5.32		P1
					3.70		P1
				MET46	5.42		P5’
					7.01		P5’
				PRO81	6.8		P4’
				ARG8	6.83		P2’
					7.02		P2’
				Catalytic GLY27	4.83		P3’
					4.10		P3’
				ASP29	3.72		P3’
		Attractive charge	2	ASP29	4.52		P3’
				ASP30	5.43		P4’
P2/NC	SATIM/MQRGN	Salt Bridge	1	ASP29	4.48		P2’
		Attractive Charge	3	ASP30	4.00		P2’
				ARG8	5.38		P1’
				ARG87	5.91		P4’
		Hydrogen Bond	4	ASP29	4.34		P1’
				GLY49	3.92		P2
				Catalytic GLY27	3.73		P1
				GLY48	4.50		P1
		Alkyl	8	ILE84(chain A)	5.89		P3
				ILE84(chain B)	5.85		P1
				Val82	6.51		P1

				ILE47 (Chain B)	4.44		P3
				ILE47 (chain A)	6.12		P1
				ALA28	3.80		P1
				VAL32	7.04		P1
				ILE50	7.05		P1
		Conventional hydrogen	12	GLY48	2.98		P4
				ARG8	6.05		P2
				Catalytic GLY27	3.92		P3’
				GLY48	3.27		P1
					3.42		P1
				ARG8	6.73		P4’
					3.38		P1’
				ARG87	6.34		P4’
				ASP30	3.67		P3’
				ASP29	4.85		P1’
					4.78		P2’
					3.93		P3’
NC/P1	ERQAN/FLGKI	Salt Bridge	1	Catalytic D25	7.55		P3’
		Attractive Charge	1	Catalytic D25	4.73		P3’
		Unfavourable positive-positive	1	LYS4	7.71		P4
		π-Anion	3	ALA28	4.75		P1
		Alkyl		ILE47	4.91		P1
		π-Alkyl		VAL32	7.14		P1
		Conventional Hydrogen Bond	12	ASP29	4.38		P2
					3.60		P4
				GLY48	3.22		P1
				ARG8	6.98		P1
					4.27		P2
				Catalytic GLY27	4.40		P4’
					4.02		P1
				ILE84	3.72		P3’
				GLY48(Chain B)	4.19		P2’
					4.41		P2’
				Catalytic D25 (chain A)	4.73		P3’
				Catalytic D25 (Chain B)	5.16		P3’
		Hydrogen Bond	2	Catalytic Gly27	4.74		P4’
				GLY48	3.31		P1

P1/P6	RPGNF/LQSRP	Salt Bridge	2	ASP29	4.60	-5.0	P4’
				ASP30	4.56		P4’
		Attractive Charge	1	ARG	6.51		P3’
		Conventional Hydrogen Bond	9	ASP29	5.24		P3
				MET46	5.00		P5
				ARG8	6.04		P3’
				GLY48	2.94		P5’
				GLY48(Chain A)	3.19		P5
				Catalytic GLY27	4.16		P1
				ILE50	3.84		P1
				Catalytic D25	4.73		P2’
				ASP30	2.82		P5’
		Hydrogen bond	3	GLY49	3.54		P3
				ALA28	3.91		P1’
				Catalytic GLY27	4.23		P2’
		Unfavourable Donor-Donor	0				
		Amide-Pi Stacked	1	ALA28	5.44		P1
		Alkyl	3	ILE47(Chain A)	4.86		P1
				Pro81	5.46		P4’
				ILE84(chain A)	3.84		P1
				VAL82	6.62		P1
		π-Alkyl	3	ILE47	4.86		P2
				VAL32	6.16		P2
				ILE47(chain B)	4.39		P2

Table 5: Summary of subtype C CS interactions

CS	Sequence	Bond type	Number of bonds	Interacting residues	Distances (Angstroms)	Binding affinity (kcal/mol)	Position
MA/CA	VSQNY/PIVQN	π-Alkyl	1	Benzene ring with Val82	4.71	-8.3	P1’
		π-cation	1	Benzene with catalytic Gly(chainA)27	4.16		P1’
		Unfavourable donor-donor	2	Arg8	2.22		P1’
				Ile47	2.48		P1’
		Hydrogen bond	2	Ala28,	3.10		P2’
				catalytic Gly (chain B 27),	2.81		P4
		Conventional hydrogen bond	9	Asp29 (Chain B)	2.55		P5
				Asp30 (chain B)	2.04		P4
				Asp30 (Chain B)	2.15		P5
				Arg8(Chain B)	2.85		P3
				Asp29(Chain A)	2.48		P4’
				Asp29(Chain A)	2.12		P2’
				Arg87	2.46		P1’
				Asp30(Chain A)	3.22		P4’
				Trp6	3.0		P5’
					2.91		P5’
	ASQNY/PIVQN	Alkyl	3	Ala28	5.22	-7.5	
				Ile50	5.95		
				Ile84	5.91		
		π-alkyl	3	Ile50	4.20		
				Ala28(chainB)	4.31		
				Ile47	6.61		
		π-anion	1	Benzene ring with Catalytic D25	6.01		
		Unfavourable acceptor-acceptor	1	Gly48	4.10		
		Unfavourable donor-donor		Arg8	6.01		
		Conventional hydrogen bond	8	Glu21	4.72		P5
				Arg8	3.4		P3
				Gly48(chainB)	4.31		P2
				Asp30	3.41		P3’
				Arg	6.26		P3’
				Pro81	3.20		P4’
				Pro81	2.31		P4’
				Gly	2.58		P5’
		Hydrogen bond	2	Pro81(chainA)	3.66		
				Gly49(chainA)	3.97		
	VSQNF/PIVQN	Alkyl	3	Val82	5.24	-7.4	
				Pro81	4.55		
				Ile50	5.19		

		π-alkyl	5	Val32	6.83		P1’
				Ile84	6.53		P1’
				Ala23	5.55		P1’
				Ile47	5.45		P1’
				Ile50	4.21		P1’
		Hydrogen bond	1	Gly49	3.79		P4’
		Conventional hydrogen bond	8	Val82(chain A)	3.32		
				Arg8 (chain A)	6.17		
				Asp30	4.35		
				Arg8 (chain B)	5.92		
				Ile50	4.20		
				Asp29	3.62		
				Asp30	3.95		
				Gly48	3.72		
					4.55		
	ISQNY/PIVQN	Alkyl	3	Phe53	5.25	-7.3	P5
				Pro81	4.86		P5
				Ile47	4.51		P2’
		π-alkyl	6	Ile84	4.47		P1’
				Val32(chain B)	5.52		P1’
				Ala28	3.76		P1’
				Ile47(chain B)	6.51		P1’
				Ile47(chain B)	4.51		P2’
				Ile50	6.36		P1
		Unfavourable donor-donor	1	Arg8	6.38		P4’
		Hydrogen bond	2	Gly49	3.97		P2
				Asp30(chain B)	3.93		P1’
		Conventional hydrogen bond	9	Arg8	7.21		P4
				Gly48	4.09		P4
				Gly48	4.02		P4
				Asp30	3.62		P3
				Asp29	4.86		P3
				Gly48(chain B)	3.11		P3’
				Arg8(chain B)	5.60		P3’
					6.41		P4’
					5.23		P5’
				Gly49 (chain B)	3.22		P4’
Capsid-P2	EARVL/AEAMS	Alkyl	3	Val82	6.13	-8.2	P1
				Ile84	6.09		P1
				Arg8	3.99		P4’
		Unfavourable acceptor-acceptor	1	Gly48	4.81		P3
		Hydrogen bond	2	Gly49	3.44		P3
				Ile50	4.85		P3
		Conventional hydrogen bond	13	Catalytic Gly27	4.71		P2
				Gly48	3.21		P5
					3.79		P3
				Asp29	4.14		P2’
				Ile50 (chain B)	3.88		P2
				Asp30	3.12		P2’

				Asp29	3.50		P2’
					4.17		P2’
				Arg8	5.44		P1’
					6.10		P3’
					3.60		P4’
				Arg87	4.60		P5’
					6.53		P5’
	KARIL/AEAMS	Alkyl	7	Val32	6.62	-7.5	P1
				Ile84	5.57		P1
				Ile50	4.50		P1
				Ala28	5.19		P1
				Val82	5.17		P5
					6.30		P1
				Leu23	6.51		P1
				Val82(chain A)	4.22		P4’
					4.87		P4’
		Unfavourable donor-donor	1	Arg8	5.44		P5’
		Hydrogen bond	2	Gly49	3.92		P1
				Asp30	5.64		P2
		Conventional hydrogen bond	11	Glu21	4.75		P5
				Pro81	3.36		P4
				Asp30	4.14		P2
				Arg8	6.92		P3
				Gly48	3.71		P2
					4.14		P3
				Asp30(chain B)	3.27		P2’
				Asp29(chain B)	4.33		P2’
				Pro81(chainA)	4.76		P5’
				Arg8	6.31		P5’
				Gly48(chain B)	3.16		P4’
	KAKVL/AEAMS	Alkyl	2	Ile47	6.30	-7.2	P1
				Ala28	3.93		P1
		Unfavourable donor-donor	1	Arg8	5.82		P4
		Hydrogen bond	2	Gly49	3.40		P1’
		Conventional hydrogen bond	12	Catalytic Gly27 (Chain A)	3.95		P4
					4.07		P4
					4.48		P4
					ILE50(chain A)		P2
					Ile50 (chain B)		P2
				Catalytic Gly27 (Chain B)	5.66		P2
				Gly48	3.81		P1
					3.52		P2
					4.40		P1
				Arg8	6.56		P2
					6.46		P3

				Arg87	4.41		P5
	KARVL/AEAMS	Alkyl	2	Val32	5.45	-4.2	P4
				Pro81	5.34		P4
		Hydrogen bond	2	Gly49	3.83		
				Catalytic Gly27(chain A)	5.20		
				Thr80	3.54		P5’
				Gly49(chain B)	3.89		P3’
		Conventional hydrogen bond	13	Catalytic D25 (chain A)	4.85		P5
				Catalytic Gly 27 (chain B)	3.32		P5
				Catalytic D25 (chain B)	5.06		P2
					5.36		P2
				Asp30	3.16		P3
				Gly48	3.45		P2
					3.63		P3
				Ile50	5.38		P2’
				Arg8	5.60		P4’
				Asp30(chain B)	3.18		P2’
					3.49		P2’
				Gly48(chain B)	3.83		P3’
P2NC	NNNIM/MQRSN	Alkyl	6	Val82	3.87	-7.1	P3
					5.19		P3
				Ile84	6.51		P3
					5.41		P3
				Ile50	6.73		P1
					4.88		P1
		Hydrogen bond	3	Ala28	3.51		P1’
				Gly49	3.41		P4’
					3.78		P4’
		Conventional hydrogen bond	9	Arg8	4.82		P4’
					6.31		P4’
				Pro81	4.73		P5’
				Glu21	4.42		P5’
				Gly48	3.20		P2’
				Asp30	3.89		P1’
				Arg8(chain B)	6.51		P5’
					6.70		P5’
				Gly48(chain A)	3.41		P4’
	NNNIM/MQRGN	Alkyl	5	Ile84	6.24	-7.1	P1’
				Val32	6.71		P1
				Ala28	6.21		P1’
				Ile47	3.77		P1’
					5.92		P2’
		Hydrogen bond	2	Gly84	4.03		P1
				Ala28	5.21		P1’
		Unfavourable donor-donor	2	Arg28	6.12		P2’
					3.42		P3’
			12	Arg8	5.83		P3

		Conventional hydrogen bond		Catalytic Gly27(chain B)	3.63		P5
				Gly41	4.33		P4
				Asp32	3.04		P2
					3.82		P2
				Ile50	3.32		P1
					4.04		P1’
				Asp29	4.30		P1’
					4.22		P3’
				Arg8	6.91		P2’
				Val82	3.47		P2
				Arg82	4.51		P3’
				Asp30	4.45		P5’
	NNNIM/MQRNN	Alkyl	8	Leu76	5.55	-6.2	P1’
				Val32	4.36		P2
				Ile47	4.57		P2
				Ile84	7.21		P2
				Val82(chain B)	6.22		P1’
				Pro81	6.25		P1’
				Ile50	4.50		P2
				Ile84(chain A)	7.45		P2
		Unfavourable donor-donor	1	Arg8	5.93		P3’
		Hydrogen bond	3	Gly49	3.03		P2
				Ala28	3.86		P3
				Asp29	3.80		P4
		Conventional hydrogen bond	9	Arg8	6.23		P4’
				Gly48	3.74		P5
					3.87		P4
					3.29		P3
				Arg87	5.28		P4’
				Arg8	6.21		P1’
				Asp29	4.44		P3’
				Asp30	4.14		P3’
	NNNIM/MQKSN	Alkyl	4	Val82	6.80	-6.1	P1’
				Ile50	3.85		P2’
				Ile84	5.98		P2’
				Pro81	6.49		P2’
		Unfavourable donor donor	1	Arg87	4.99		P4
		Hydrogen bond	1	Asn88	4.45		P4
		Conventional hydrogen bond	15	Trp6	6.47		P5
					5.22		P4
				Arg87	4.08		P4
					5.93		P4
				Leu5	5.69		P4
				Arg8	5.99		P1
				Gly48	3.69		P1
				Catalytic Gly27	4.83		P2’
				Asp30	2.86		P1’
				Catalytic D25	5.38		P2’

				Asp29	3.33		P3’
					4.66		P3’
				Asp30	4.13		P3’
					3.00		P5’
					4.27		P5’
NC-P1	ERQAN/FLGKV	π-alkyl	3	Ile50	6.83	-8.2	P1
				Ile47	5.34		P1
				Ala28	4.82		P1
		Unfavourable donor-donor	1	Ile50 (chain B)	4.11		P2’
		Conventional hydrogen bonds	6	Asp29	4.14		P5
					4.69		P3
					4.34		P2
				Arg8	4.03		P3
				Trp6	6.58		P4
				Asp30	2.51		P3’
	ERQAN/FLGKI	π-Alkyl	3	Val82	4.65	-7.8	P1’
				Ile84	6.25		P1’
				Ile50	5.65		P1’
		Alkyl	3	Pro81	6.23		P2’
				Phe53	5.93		P2’
				Arg8	4.73		P5’
		Hydrogen Bond	2	Catalytic D25	5.32		P3
				Ala23	3.53		P1’
		Conventional hydrogen bond	10	Asp29	3.55		P5
				Asp30(chain A)	3.10		P5
					3.90		P5
				Gly48	3.77		P3
				Catalytic Gly27	4.16		P1
				Asp32	3.24		P1’
				Gly41	4.16		P1’
					3.23		P2’
				Asp23	3.29		P1’
				Arg8	4.82		P5’
					5.61		P5’
	ERQAN/FLGRI	Alkyl	1	Ile47	4.88	-7.4	P2’
		Unfavourable acceptor-acceptor	1	Gly41	4.85		P4
		Unfavourable donor-donor	2	Catalytic Asp25	3.55		P5
				Arg82	5.80		P5’
		Hydrogen bond	6	Ala23	3.96		P4’
				Gly41(chain B)	4.02		P3
				Catalytic D25	6.26		P2
				Gly49	3.70		P1’
				Pro81	3.89		P1’
				Thr74	3.70		P4’
		Conventional hydrogen bond	13	Catalytic D25	3.77		P5
				Catalytic Gly27	4.28		P4
				Val32	4.25		P3
					3.13		P3

				Asp30	3.97		P3
					4.02		P5
				Ile50	5.07		P3
				Ile50(chain A)	4.15		P2
					4.27		P1’
				Arg8	4.73		P1
					4.29		P1’
					3.20		P1’
				Asp30	4.69		P4’
	ERQAN/FLGRL	π-Alkyl	5	Ala23	3.86	-6.5	P1
				Ile84	6.61		P1
				Ile50	6.53		P1
				Val32	6.71		P1
				Ile47	5.42		P1
		Unfavourable donor-donor	1	Asp30	2.73		P5
		Hydrogen bond	2	Gly41	3.27		P1
				Gly49	3.03		P1
		Conventional hydrogen bond	13	Asp30	2.73		P5
				Catalytic D25	4.74		P4
					3.90		P1
				Arg8	6.12		P3
				Catalytic Gly27	4.22		P1
				Gly43	4.11		P1’
				Arg8(chain B)	5.44		P1’
					6.33		P2’
					4.53		P3’
				Catalytic D25(chain B)	5.46		P4
P1-P6	RPGNF/VQSRP	π-Alkyl	2	Ala28	4.45	-8.1	P1’
				Ile47	6.26		P1’
		Unfavourable donor-donor	1	Ile50	3.63		P1
		Unfavourable acceptor-acceptor	1	Catalytic Gly27			P3
		Hydrogen bond	4	Gly49	4.01		P1
				Catalytic D25	6.46		P1’
				Ala28	3.58		P3
					4.08		P3
		Conventional hydrogen bond	9				
				Arg8	5.52		P3
				Asp29	5.11		P3
					3.99		P2
				Gly29	4.00		P3
				Asp30	4.65		P5
				Gly48	6.26		P2’
					4.45		P3’
				Leu5	5.43		P5’
				Pro81	6.60		P2’
	RPGNF/LQNRP	π-Alkyl	4	Val82	4.63	-7.9	P1
				Pro81	5.37		P1
				Val82(chain A)	5.95		P3’

				Pro81(chain A)	5.64		P3’
		Alkyl	1	Ile47	5.03		P1’
		Unfavourable acceptor-acceptor	1	Gly48	3.21		P1’
		Unfavourable donor-donor	2	Asp30	2.02		P5’
				Asp23	2.87		P5’
		Hydrogen bond	2	Ile50	3.97		P2’
				Ala28	3.06		P2’
		Conventional hydrogen bond	13	Trp6	5.01		P5
					6.15		P5
				Arg8	4.17		P3
				Asp30	3.01		P1’
					5.07		P1’
				Catalytic D25	5.05		P2’
				Catalytic D25 (Chain A)	4.21		P2’
				Ile50(chain B)	4.02		P3’
				Asp29	3.71		P1
					4.75		P1
				Gly48	3.65		P1’
				Gly43	4.93		P3’
				Asp29	3.42		P5’
	RPGNF/PQSRP	Amid-Pi Stacked	1	Gly48	4.93	-7.1	P1
		Alkyl	2	Val82	6.13		P4’
				Pro81	6.11		P4’
		π-Alkyl	2	Ile47	5.69		
				Ala28	4.99		
		Hydrogen bond	3	Ile50	3.85		P3
				Gly49	3.80		P3’
				Glu21	5.41		P4
		Conventional hydrogen bond	8	Glu21	2.20		P5
					3.06		P5
				Arg8	6.00		P3
				Val82	2.80		P2
				Gly48	4.54		P1
				Gly48(chain B)	3.88		P3’
				Catalytic	4.66		P5’
				D25(chain A)	4.40		P5’
				Catalytic D25(chain B)	4.37		P5’
	RPGNF/LQSRP	π-Alkyl	2	Val82	4.77	-5.0	P1
				Pro81	5.36		P1
		Alkyl	3	Val32	6.09		P3’
				Ile87	3.06		P3’
				Ile47	4.54		P1
		Unfavourable Bump	1	Gly22	4.65		P4
		Hydrogen bond	3	Pro81	3.35		P2’
				Catalytic Gly27	3.44		P1

				Ala28	5.71		P3’
		Conventional hydrogen bond	11	Arg8	6.41		P5
				Catalytic D25	3.74		P3
				Gly48	4.89		P4
					3.71		P4’
					5.71		P4’
				Asp29	3.57		P4’
				Asp30	4.07		P5’
				Asp29(chain B)	3.96		P1
				Asp30 (chain A)	3.54		P3
					4.67		P3
				Val32	3.62		P1’

Homology Modelling Scripts

1. **Build_profile.py**

```
from modeller import *
log.verbose()
env = Environ()
#-- Prepare the input files
#-- Read in the sequence database
sdb = SequenceDB(env)
sdb.read(seq_database_file='pdb_95.pir', seq_database_format='PIR',
        chains_list='ALL', minmax_db_seq_len=(30, 4000), clean_sequences=True)
#-- Write the sequence database in binary form
sdb.write(seq_database_file='pdb_95.bin', seq_database_format='BINARY',
         chains_list='ALL')
#-- Now, read in the binary database
sdb.read(seq_database_file='pdb_95.bin', seq_database_format='BINARY',
        chains_list='ALL')
#-- Read in the target sequence/alignment
aln = Alignment(env)
aln.append(file='K4.ali', alignment_format='PIR', align_codes='ALL')
#-- Convert the input sequence/alignment into
#   profile format
prf = aln.to_profile()
#-- Scan sequence database to pick up homologous sequences
prf.build(sdb, matrix_offset=-450, rr_file='${LIB}/blosum62.sim.mat',
```

```
        gap_penalties_1d=(-500, -50), n_prof_iterations=1,
        check_profile=False, max_aln_evalue=0.01)
#-- Write out the profile in text format
prf.write(file='build_profile.prf', profile_format='TEXT')
#-- Convert the profile back to alignment format
aln = prf.to_alignment()
#-- Write out the alignment file
aln.write(file='build_profile.ali', alignment_format='PIR')
```

2. **Alignment.py**

```
from modeller import *
env = Environ()
aln = Alignment(env)
mdl = Model(env, file='1f7a_1', model_segment=('FIRST:P','LAST:P'))
aln.append_model(mdl, align_codes='1f7a_1A', atom_files='1f7a_1.pdb')
aln.append(file='K4.ali', align_codes='K4')
aln.align2d(max_gap_length=50)
aln.write(file='K4-1f7a_1A.ali', alignment_format='PIR')
aln.write(file='K4-1f7a_1A.pap', alignment_format='PAP')
```

3. **Model.py**

```
from modeller import *
from modeller.automodel import *
#from modeller import soap_protein_od
env = Environ()
a = AutoModel(env, alnfile='K4-1f7a_1A.ali',
              knowns='1f7a_1A', sequence='K4',
              assess_methods=(assess.DOPE,
                              #soap_protein_od.Scorer(),
                              assess.GA341))
a.starting_model = 1
a.ending_model = 5
a.make()
```

4. Evaluate Model

```
from modeller import *
from modeller.scripts import complete_pdb
log.verbose()    # request verbose output
env = Environ()
env.libs.topology.read(file='$(LIB)/top_heav.lib') # read topology
env.libs.parameters.read(file='$(LIB)/par.lib') # read parameters
# read model file
mdl = complete_pdb(env, 'K4.B99990005.pdb')
# Assess with DOPE:
s = Selection(mdl)   # all atom selection
s.assess_dope(output='ENERGY_PROFILE NO_REPORT', file='K4.profile',
              normalize_profile=True, smoothing_window=15)
```

5. Evaluate template

```
from modeller import *
from modeller.scripts import complete_pdb
log.verbose()    # request verbose output
env = Environ()
env.libs.topology.read(file='$(LIB)/top_heav.lib') # read topology
env.libs.parameters.read(file='$(LIB)/par.lib') # read parameters
# directories for input atom files
env.io.atom_files_directory = './:../atom_files'
# read model file
mdl = complete_pdb(env, '1f7a_1.pdb', model_segment=('FIRST:P', 'LAST:P'))
s = Selection(mdl)
s.assess_dope(output='ENERGY_PROFILE NO_REPORT', file='1f7a_1A.profile',
              normalize_profile=True, smoothing_window=15)
```

6. Plot Profiles

```
import pylab
import modeller
def r_enumerate(seq):
   """Enumerate a sequence in reverse order"""
```

```
    # Note that we don't use reversed() since Python 2.3 doesn't have it
    num = len(seq) - 1
    while num >= 0:
        yield num, seq[num]
        num -= 1
def get_profile(profile_file, seq):
    """Read `profile_file` into a Python array, and add gaps corresponding to
       the alignment sequence `seq`."""
    # Read all non-comment and non-blank lines from the file:
    f = open(profile_file)
    vals = []
    for line in f:
        if not line.startswith('#') and len(line) > 10:
            spl = line.split()
            vals.append(float(spl[-1]))
    # Insert gaps into the profile corresponding to those in seq:
    for n, res in r_enumerate(seq.residues):
        for gap in range(res.get_leading_gaps()):
            vals.insert(n, None)
    # Add a gap at position '0', so that we effectively count from 1:
    vals.insert(0, None)
    return vals
e = modeller.Environ()
a = modeller.Alignment(e, file='K4-1f7a_1A.ali')
template = get_profile('1f7a_1A.profile', a['1f7a_1A'])
model = get_profile('K4.profile', a['K4'])
# Plot the template and model profiles in the same plot for comparison:
pylab.figure(1, figsize=(10,6))
pylab.xlabel('Alignment position')
pylab.ylabel('DOPE per-residue score')
pylab.plot(model, color='red', linewidth=2, label='Model')
pylab.plot(template, color='green', linewidth=2, label='Template')
pylab.legend()
pylab.savefig('dope_profile.png', dpi=65)
```

MD Scripts

1. **Minimization of waters**

```
&cntrl
 imin=1,
 maxcyc=5000, ncyc=3000,
 ntb=1, cut=10.0, ntpr=5,
 ntr=1, restraintmask=':1-
198',
 restraint_wt=2.0
&end
 /
```

2. **Minimization of the whole system**

```
&cntrl
 imin=1,
 maxcyc=10000, ncyc=7000,
 ntb=1, ntr=0, cut=10.0,
ntpr=5
&end
 /
```

3. **Heating of proteins from 0K to 300K**

```
&cntrl
 imin=0, irest=0, ntx=1,
ig=-1,
 nstlim=25000, dt=0.002,
 ntc=2, ntf=2,
 cut=8.0, ntb=1,
 ntpr=500, ntwx=500,
 ntt=3, gamma_ln=2.0,
 tempi=0.0, temp0=300.0,
```

```
 ntr=1, restraintmask=':1-
198',
 restraint_wt=2.0,
 nmropt=1, ioutfm=1
/
&wt TYPE='TEMP0', istep1=0,
istep2=25000,
 value1=0.1, value2=300.0, /
&wt TYPE='END' /
```

4. Equilibrate the density

```
&cntrl
 imin=0,
 irest=1,
 ntx=5,
 nstlim=25000, dt=0.002,
 ntc=2, ntf=2,
 cut=8.0, ntb=2,
 ntp=1, taup=1.0,
 ntpr=500, ntwx=500,
 ntt=3, gamma_ln=2.0,
 temp0=300.0,
 ntr=1, restraintmask=':1-
198',
 restraint_wt=2.0,
 ig=-1,
 ioutfm=1
&end
 /
```

5. Equilibrate

```
&cntrl
 imin=0, irest=1, ntx=5,
 nstlim=1000000, dt=0.002,
```

```
 ntc=2, ntf=2,
 cut=10.0, ntb=2,
 ntp=1, taup=2.0,
 ntpr=1000, ntwx=1000,
 ntt=3, gamma_ln=2.0,
 temp0=300.0,
 ig=-1, ioutfm=1
&end
 /
```

Ethics Certificate

UNIVERSITY OF KWAZULU-NATAL
INYUVESI YAKWAZULU-NATALI

Miss Laurinda Vuyolwethu Mqhaba (216056362)
School Of Lab Med & Medical Sc
Medical School

Dear Miss Laurinda Vuyolwethu Mqhaba,

Protocol reference number: 00014274
Project title: Natural polymorphisms at Gag cleavage sites and the impact they have on the substrate envelope structure of HIV-1 subtype C.

Exemption from Ethics Review

In response to your application received on , your school has indicated that the protocol has been granted **EXEMPTION FROM ETHICS REVIEW.**

Any alteration/s to the exempted research protocol, e.g., Title of the Project, Location of the Study, Research Approach and Methods must be reviewed and approved through an amendment/modification prior to its implementation. The original exemption number must be cited.

For any changes that could result in potential risk, an ethics application including the proposed amendments must be submitted to the relevant UKZN Research Ethics Committee. The original exemption number must be cited.

In case you have further queries, please quote the above reference number.

PLEASE NOTE:
Research data should be securely stored in the discipline/department for a period of 5 years.

I take this opportunity of wishing you everything of the best with your study.

Yours sincerely,

[redacted] 07 September 2021

Dr Brenda Zola De Gama
Academic Leader Research
School Of Lab Med & Medical Sc

L_V_Mqhaba_FINAL_dissertation_07TH_July25.docx

ORIGINALITY REPORT

16%	12%	11%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

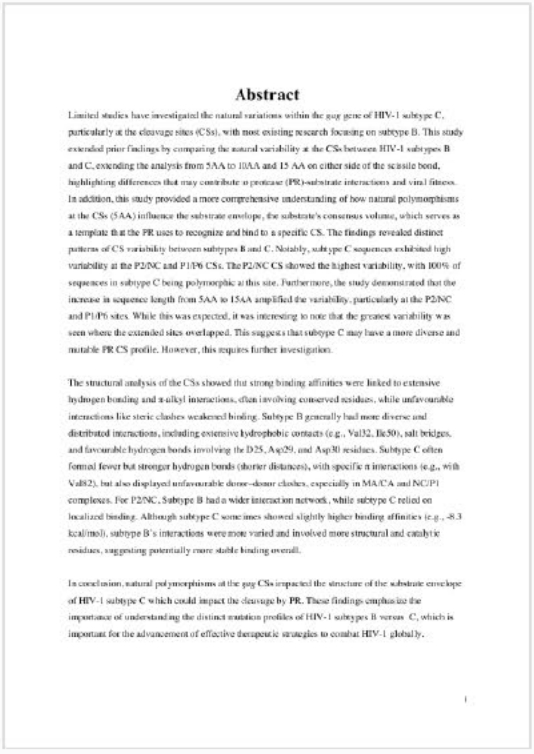
1	researchspace.ukzn.ac.za Internet Source	3%
2	hdl.handle.net Internet Source	1%
3	www.frontiersin.org Internet Source	1%
4	etheses.bham.ac.uk Internet Source	1%
5	discovery.ucl.ac.uk Internet Source	<1%
6	wiredspace.wits.ac.za Internet Source	<1%
7	escholarship.umassmed.edu Internet Source	<1%
8	www.mdpi.com Internet Source	<1%
9	"Applications of Computational Tools in Drug Design and Development", Springer Science	<1%

Digital Receipt

This receipt acknowledges that Turnitin received your paper. Below you will find the receipt information regarding your submission.

The first page of your submissions is displayed below.

Submission author:	Laurinda Mqhaba
Assignment title:	PostgradCheck
Submission title:	L_V_Mqhaba_FINAL_dissertation_07TH_July25.docx
File name:	L_V_Mqhaba_FINAL_dissertation_07TH_July25.docx
File size:	17.14M
Page count:	139
Word count:	40,915
Character count:	230,002
Submission date:	07-Jul-2025 01:15PM (UTC+0200)
Submission ID:	2637995148



Abstract

Limited studies have investigated the natural variations within the gag gene of HIV-1 subtype C, particularly at the cleavage sites (CSs), with most existing research focusing on subtype B. This study extended prior findings by comparing the natural variability at the CSs between HIV-1 subtypes B and C, extending the analysis from 5AA to 10AA and 15 AA on either side of the scissile bond, highlighting differences that may contribute to protease (PR)-substrate interactions and viral fitness. In addition, this study provided a more comprehensive understanding of how natural polymorphisms at the CSs (5AA) influence the substrate envelope, the substrate's consensus volume, which serves as a template that the PR uses to recognize and bind to a specific CS. The findings revealed distinct patterns of CS variability between subtypes B and C. Notably, subtype C sequences exhibited high variability at the P2/NC and P1/P6 CSs. The P2/NC CS showed the highest variability, with 100% of sequences in subtype C being polymorphic at this site. Furthermore, the study demonstrated that the increase in sequence length from 5AA to 15AA amplified the variability, particularly at the P2/NC and P1/P6 sites. While this was expected, it was interesting to note that the greatest variability was seen where the extended sites overlapped. This suggests that subtype C may have a more diverse and mutable PR CS profile. However, this requires further investigation.

The structural analysis of the CSs showed that strong binding affinities were linked to extensive hydrogen bonding and π-alkyl interactions, often involving conserved residues, while unfavourable interactions like steric clashes weakened binding. Subtype B generally had more diverse and distributed interactions, including extensive hydrophobic contacts (e.g., Val32, Ile50), salt bridges, and favourable hydrogen bonds involving the D25, Asp29, and Asp30 residues. Subtype C often formed fewer but stronger hydrogen bonds (shorter distances), with specific π interactions (e.g., with Val82), but also displayed unfavourable donor–donor clashes, especially in MA/CA and NC/P1 complexes. For P2/NC, Subtype B had a wider interaction network, while subtype C relied on localized binding. Although subtype C sometimes showed slightly higher binding affinities (e.g., -8.3 kcal/mol), subtype B's interactions were more varied and involved more structural and catalytic residues, suggesting potentially more stable binding overall.

In conclusion, natural polymorphisms at the gag CSs impacted the structure of the substrate envelope of HIV-1 subtype C which could impact the cleavage by PR. These findings emphasize the importance of understanding the distinct mutation profiles of HIV-1 subtypes B versus C, which is important for the advancement of effective therapeutic strategies to combat HIV-1 globally.