UNIVERSIDADE NOVA DE LISBOA

MScCBBi

MASTER IN
COMPUTATIONAL BIOLOGY
& BIOINFORMATICS

FILIPA COELHO TAVARES
BSc in Applied Biology

Characterization of HIV-1 reverse transcriptase mutations in Brazil

Set, 2023

NOVA SCHOOL OF SCIENCE & TECHNOLOGY
NOVA MEDICAL SCHOOL
NOVA IMS Information Management School
DESDE 1902 INSTITUTO DE HIGIENE E MEDICINA TROPICAL UNIVERSIDADE NOVA DE LISBOA

CHARACTERIZATION OF HIV-1 REVERSE TRANSCRIPTASE MUTATIONS IN BRAZIL

FILIPA COELHO TAVARES

BSc in Applied Biology

Adviser: Nuno Miguel Sampaio Osório
Assistant Researcher, ICVS/3B's

Co-adviser: Ana Barroso Abecasis
Assistant Professor, NOVA University Lisbon - IHMT

Examination Committee:

Chair: Paula Maria Theriaga Mendes Bernardo Gonçalves
Associate Professor, FCT-NOVA

Rapporteurs: Agostinho Antunes Pereira
Associate Professor, FCUP

Members: Nuno Miguel Sampaio Osório
Assistant Researcher, ICVS/3B's

MASTER IN COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

NOVA University Lisbon
September, 2023

Characterization of HIV-1 reverse transcriptase mutations in Brazil

ACKNOWLEDGEMENTS

First and foremost, I would like to express a special thank you to my adviser, Nuno Osório, and co-adviser, Ana Abecasis, for both allowing me to do research on a topic that I'm incredibly passionate about, and also for helping and supporting me throughout this year-long journey.

To Maria Isabel Veiga and everyone at the PEvoGen research group at ICVS-3B's at University of Minho, thank you for letting me be a part of your group and by welcoming me with open arms.

To my parents and sisters, thank you for always supporting me and my decisions, and by listening to me explain what I was researching about and appearing interested, despite knowing that you weren't understanding a thing. And to my grandparents, for letting me spend my summer in your kitchen table writing this dissertation.

Finally, I'll always be tremendously thankful for every friend that I came across on this academic journey, both my friends from BA in Braga, and my new friends from MCBBi in Lisbon, for helping me and not letting me give up when I most wanted to.

ABSTRACT

Forty years after its discovery, HIV-1 remains a global challenge, with thousands of new infections and deaths annually even in upper-middle-income nations like Brazil. While antiretroviral therapy (ART) has revolutionized treatment, HIV-1's ability to evolve enables the virus to develop drug resistance mutations (DRMs). The combination of DRMs in complex patterns is a great concern, as this can render commonly-used ART regimens ineffective for many patients. Therefore, proactive resistance surveillance using innovative data analysis methods is essential.

With this context in mind, the overarching objective of this dissertation was to harness an array of bioinformatics techniques, including dimensionality reduction and clustering, phylogenetic analysis, and protein structure prediction, to curate, compile, and analyse HIV-1 reverse transcriptase genomic sequences and clinical data from HIV-1-infected individuals across diverse clinical settings in Brazil. The aim was to characterize prevalent combinations of HIV-1 genetic variants and its possible implications for ART resistance, in granular detail.

Our results highlight the K103N+L100I NNRTI-associated double mutant in HIV-1 reverse transcriptase as particularly concerning in Brazil, as it is linked to higher viral load, lower CD4+ cell count, and moderate-to-severe immunodeficiency, while increasing in frequency over the years. One possible reason for this increase is the relation between this double mutant and the NRTI-associated K65R mutation, which was likely favoured by a change in Brazilian ART regimens and increased use of TDF after 2014. In silico analysis of the structural implications of the K103N+L100I mutant points to an impact at the RNase H subunit, which could be beneficial to the virus in this context.

These findings suggest that RNase H inhibitors that bind to wild-type and K103N+L100I mutated reverse transcriptase have the potential to be a highly valuable addition to the current ART in clinical settings alike Brazil.

Keywords: Human Immunodeficiency Virus (HIV-1); Drug Resistance Mutations (DRMs); Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTIs); Clustering; Phylogeny; Protein Structure Prediction

RESUMO

Quarenta anos após descoberta, o VIH-1 permanece um desafio global, com milhares de novas infeções e mortes anualmente, mesmo em países de rendimento médio-alto como o Brasil. Enquanto a terapia antirretroviral (ARV) tem revolucionado o tratamento, a evolução do VIH-1 facilita o desenvolvimento de mutações de resistência. A combinação de mutações em padrões complexos é preocupante, visto que pode levar à falência terapêutica. Portanto, vigilância proativa de resistência usando métodos de análise de dados inovadores é essencial.

Assim, o objetivo abrangente desta dissertação foi juntar técnicas bioinformáticas, incluindo redução de dimensionalidade e clustering, análise filogenética e previsão de estrutura proteica, para organizar, compilar e analisar sequências genómicas da transcriptase reversa do VIH-1 e dados clínicos de indivíduos infetados com VIH-1 em diferentes condições clínicas no Brasil. O objetivo foi caracterizar detalhadamente combinações de variantes genéticas do VIH-1 predominantes e as suas possíveis implicações funcionais.

Os nossos resultados apontam o duplo mutante K103N+L100I associado a NNRTIs na transcriptase reversa do VIH-1 como preocupante no Brasil, dado que associou a maiores cargas virais, menor quantidade de células CD4+ e imunodeficiência moderada e severa, enquanto aumenta em frequência ao longo dos anos. Uma potencial razão para este aumento é a relação entre este duplo mutante e a mutação K65R associada a NRTIs, que foi provavelmente favorecida por uma mudança de regimes de ARV no Brasil e o aumento do uso de TDF depois de 2014. Análise in silico das implicações estruturais do mutante K103N+L100I aponta para um impacto da subunidade RNase H, o que poderá beneficiar os vírus neste contexto.

Estes achados sugerem que inibidores de RNase H que se ligam à transcriptase reversa wild-type e mutada com K103N+L100I têm o potencial de serem mais-valias para a ARV atual no Brasil e em condições clínicas semelhantes à do Brasil.

Palavras-Chave: Vírus da Imunodeficiência Humana (VIH-1); Mutações de Resistência; Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTIs); Clustering; Filogenia; Previsão de Estrutura Proteíca

INDEX

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

3

3D- Three-Dimensional

3TC – Lamivudine

A

ABC – Abacavir

ACR – Ancestral Character Reconstruction

ADR – Acquired Drug Resistance

AIDS – Acquired Immunodeficiency Syndrome

ART – Antiretroviral Therapy

ASE – Accuracy Self-Estimate

ATP – Adenosine Triphosphate

ATV – Atazanavir

AZT/ZDV – Zidovudine

B

BA5 – (or XZ462) methyl 4-azanyl-1-oxidanyl-2-oxidanylidene-1,8-naphthyvidine-3-carboxylate

BDT – Binding-Site Test

BI – Bayesian Inference

C

CAMEO-LB – Continuous Automated Model Evaluation – Ligand Binding

CASP – Critical Assessement of Techniques for Protein Structure Prediction

CRF – Circulating Recombinant Form

D

d4T – staduvine

DAPY – Diraylpyrimidine

DC – 2'-deoxycytidine-5'-monophosphate

ddI – didanosine

DEDD – D443, E478, D498, D549 RNase H active site motif

DNA – Deoxyribonucleic Acid

dNTP – Deoxynucleoside Triphosphate

DP – Domain Prediction

DR – Disorder Prediction

DRM – Drug Resistance Mutation

DRV – Darunavir

E

EFV – Efavirenz

EMA – Estimates of 3D Model Accuracy

ENF – Enfuvirtide

ETR – Etravirine

F

FN – Binding Site Prediction

FPV – Fosamprenavir

FTC – Emtricitabine

G

GDT-TS – Global Distance Test Total Score

H

HIV – Human Immunodeficiency Virus

HMM – Hidden Markov Model

I

IDU – Intravenous Drug Use

INI/INSTI – Integrase Inhibitor

IntFOLD – Integrated Protein Structure and Function Prediction Server

J

JC – Junkes-Cantor model

K

K2P – Kimura Two-Parameter model

L

LDA – Linear Discriminant Analysis

LMIC – Low- and Middle-Income Countries

LPV/RTV – Ritonavir

M

MCC – Matthews Correlation Coefficient

MCMC – Markov Chain Monte Carlo

MD, MC – Molecular Dynamics and Monte Carlo

MDS – Multidimensional Scaling

MEGA – Molecular Evolutionary Genetics Analysis

ML – Maximum Likelihood

MP – Maximum Parsimony

MPPA – Marginal Posterior Probabilities Approximation

MQA – Model Quality Assessment

MSA – Multiple Sequence Alignment

MSM – Men who have Sex with Men

MVC – Maraviroc

N

NAM – Nucleoside Analogue Mutation

NGS – Next-Generation Sequencing

NJ – Neighbour-Joining

NMR – Nuclear Magnetic Resonance

NNI – Nearest Neighbour Interchange

NNIBP – NNRTI Binding Pocket

NNRTI – Non-Nucleotide Reverse Transcriptase Inhibitor

NRTI – Nucleoside/Nucleotide Reverse Transcriptase Inhibitor

NVP – Nevirapine

O

ON1 - 2-(3,4-dichlorobenzyl)-5,6-dihydroxy-pyrimidine-4-carboxylic acid

P

PCA – Principal Component Analysis

PDB – Protein Data Bank

PI – Protease Inhibitor

PrEP – Pre-Exposure Prophylaxis

Q

QA – Model Quality Prediction

QID - 3-hydroxy-6-(phenylsulfonyl)-2,4(1H,3H)-dione

R

RAL – Raltegravir

RMSD – Root Mean Square Deviation

RNA – Ribonucleic Acid

RNase H – Ribonuclease H

RNHI – RNase H Inhibitor

RPV – Rilpivirine

RRE – Rev Response Element

RT – Reverse Transcriptase

S

SARS – Severe Acute Respiratory Syndrome

SIV- Simian Immunodeficiency Virus

T

TAM – Thymidine Analogue Mutation

TAR – Transactivation Response Element

TBM – Template-Based Models

TDF – Tenofovir

TDR – Transmitted Drug Resistance

TFM – Template-Free Models

TPV – Tipranavir

TS – Tertiary Structure Prediction

t-SNE – t-distributed Stochastic Neighbour Embedding

U

UMAP – Uniform Manifold Approximation and Projection

UPGMA – Unweighted Pair-Group Method using an Arithmetic Average

URF – Unique Recombinant Form

W

WHO – World Health Organization

WT – Wild-Type

Y

Y55 - 7-(furan-2-yl)-2-hydroxyisoquinoline-1,3(2H,4H)-dione

1 INTRODUCTION

1.1 MOTIVATION

There have been many pandemics in the world, since the ancient times, in the Roman Empire, with the Antonine Plague, to our current times, with the COVID-19 pandemic. A significant number of these pandemics, including the one caused by the human immunodeficiency virus (HIV), are attributed to diseases that are caused by viruses.

HIV is linked to one of the most notorious and lethal epidemics in history – the epidemic of Acquired Immunodeficiency Syndrome (AIDS). Since its start in 1981 [1], investigators and scientists have worked tirelessly in order to understand how to treat the disease and also how to stop HIV transmission. Nowadays, individuals infected with HIV typically undergo a regimen of antiretroviral therapy. This involves a combination of drugs, often referred to as a 'cocktail', that largely inhibits the virus's ability to replicate, thereby preventing the progression to AIDS in the infected person.

Despite the fact that antiretroviral therapy has largely benefited its users by helping them live a fairly comfortable life, the usage of these drugs still leads to more problems than what is desired, specifically due to the emergence of drug resistance mutations and consequent therapy failure associated with them.

In 2021, Brazil was reported to have the highest number of people living with HIV in Latin America, with approximately 960,000 patients [2], in spite of it being the first middle-income country to give free access to antiretroviral therapy to all that were infected with the virus, in 1996 [3]. Brazil's vast size, the substantial number of infections over the years, and the multicultural nature of its population, could be factors influencing the Drug Resistance problem in the country [4].

This way, a careful characterization and study of drug resistance mutations (DRMs) is important to prevent their distribution, by knowing how and why they do so and in what frequency, so that antiretroviral therapy (ART) doesn't fail and actually helps the infected individual who takes it. Moreover, it's also crucial to know which treatment options are causing the emergence of said DRMs, especially in a country with a population size like Brazil.

In light of this, we employed methods from bioinformatics and computational biology, such as protein structure prediction, cluster analysis, and phylogenetic analysis. Our aim was to discern the most prevalent mutation patterns and understand their evolution and impact across Brazil.

1.2 GOALS

The aim of this dissertation was curate, compile, and analyse HIV-1 genomic sequences and clinical information from HIV-1 infected individuals from different clinical settings, in Brazil, whilst characterizing in detail the prevalence and possible transmission of genetic variants with impact on ART resistance, specifically in the virus's reverse transcriptase. The more detailed goals were as follows:

(1) Characterize and analyse HIV-1 reverse transcriptase (RT) genomic sequences obtained from HIV-1 infected individuals from different clinical settings in Brazil,

using an array of bioinformatics tools and protocols in order to characterize the prevalence of genetic variants with impact on ART resistance.

(2) Determine factors that associate with the emergence of HIV-1 genetic variants that might risk the success of ART in Brazil.

1.3 STRUCTURE OF THE DOCUMENT

This dissertation is organized in the following way:

Chapter 2 – Human Immunodeficiency Virus

Characterization of the human immunodeficiency virus, its therapy/treatment options and consequent drug resistance mutations, and an in more detail view of the HIV reverse transcriptase and its binding site, NNIBP

Chapter 3 – State-of-the-Art Computational Techniques and Methods

Enumeration and description of state-of-the-art studies, methods and tools associated with the computational study of the HIV and its associated mutations performed throughout this dissertation, and also the possible problems and challenges linked to each of these methodologies

Chapter 4 – Methods

Description of the tools and methodologies used along this dissertation, where the main addressed topics are clustering analysis, phylogenetic analysis and protein structure prediction tools

Chapter 5 – Results

The main results generated in this dissertation are presented in this section, which included data statistics, sequence logos, clustering information, reverse transcriptase protein structure prediction, and phylogenetic trees representing the transmission of HIV-1 associated drug resistance mutations in Brazil

Chapter 6 – Discussion

Discussion of some of the major results presented in this dissertation. Emphasis was given to the discussion of the RT NNIBP-associated K103N+L100I double mutant and its characteristics and possible alternative treatment options

Chapter 7 – Conclusions and Future Perspectives

In this section, the principal conclusions of this dissertation are presented, as well as future research perspectives and improvements

2 HUMAN IMMUNODEFICIENCY VIRUS

2.1 EPIDEMIOLOGY

AIDS, through the zoonotic infection by HIV, is one of the most serious, prominent and deadly transmissible diseases in the 20th and 21st centuries [5,6], with a fatality rate of close to 100% [1]. Today, it's still a major global public health problem, even in well developed countries. The AIDS epidemic started in 1981, once the increase in homosexual men with rare opportunistic infections, such as pneumonia caused by *Pneumocystis crinii*, in the United States of America, was noted [1]. The subsequent discovery and isolation of the causing virus, HIV-1, happened in 1983 [1,6,7,8]. According to data by the World Health Organization (WHO), 40.1 million people have died so far globally from the disease, with 84.2 million people having been infected with HIV. It was estimated that, by the end of 2021, 38.4 million people were living with HIV, two thirds of whom (25.6 million) were in the WHO African Region.

AIDS, which occurs at a late stage in HIV infection, can be caused through infection by one of two different HIV types: HIV-1 and HIV-2 [9]. HIV-2 was discovered three years after HIV-1, in 1986 [6]. Despite having slightly different genome structures [1], these two HIV viruses are similar in many ways, including having the same modes of transmission and the same intracellular replication pathways [9]. However, their main difference lies on their clinical progression to AIDS: HIV-2 progresses more slowly than HIV-1 and has a relatively lower transmissibility [9]. Thus, HIV-1 is the main perpetrator of the global AIDS epidemic.

In spite of their differences, both HIV-1 and HIV-2 most likely originated from cross-species transmission of a particular type of lentivirus from African primates, called simian immunodeficiency virus (SIV) [6, 8], which despite its name, does not induce immunodeficiency in its natural hosts [6]. These cross-species transmissions are believed to have happened mainly due to the hunting, slaughtering, preparation and selling of bush meat in African markets [1, 6, 8], and also the capture, trade and keeping of monkeys as pets [1, 6], through the human exposure to blood and other bodily fluids from these primates [6].

However, HIV-1 and HIV-2 are related to different SIVs and thus have different evolutionary origins [1]. HIV-1 is associated with SIVcpz, which is found in the *Pan troglodytes troglodytes* (SIVcpz*Ptt*) and *Pan troglodytes schweinfurthii* (SIVcpz*Pts*) chimpanzee sub-species [1, 5, 6]. These sub-species are different due to their mitochondrial DNA (mtDNA) sequences [5] and, therefore, are geographically separated in the African continent [6]: *P. t. troglodytes* are located in equatorial Western Africa [1, 6], specifically in southern Cameroon, Gabon and the Republic of Congo [5], and *P. t. schweinfurthii* are in East Central Africa [6], specially the Democratic Republic of Congo and countries to the east [5]. In turn, HIV-2 is related to SIVsmm, found in sooty mangabey monkeys of the *Cercocebus atys* species [1, 6, 10], which are located in West Africa [1, 6], from Senegal to Ivory Coast [6].

These events of cross-species transmission happened independently several times over time, which eventually led to different groups of both HIV-1 and HIV-2 [5, 6, 8] – groups M, N, O and P for HIV-1 [8, 11] and groups A through I for HIV-2 [10, 11]. HIV-1's group M (Major) is the one that caused the global AIDS pandemic, group O (Outlier) caused a few tens of thousands of infections in West-Central Africa, group N (non-M, non-O) was found in a handful of people in Cameroon [1, 8, 11, 12] and group P, until now, was only found in two Cameroonian individuals, in 2009 [13]. When it comes to HIV-2, only groups A and B play a part in the HIV-2 epidemic [6,10], with the other groups only appearing in one single infected individual [6, 10].

HIV-1 group M is further divided into nine different subtypes (A, B, C, D, F, G, H, J and K) [1, 6, 11, 12], given HIV-1's high genetic diversity [12], which is the result of several isolated epidemic events from strains present in a certain source population [1], each representing different lineages of the M group [12]. Further sub-subtypes exist, like A1-A4, F1 and F2, which are not genetically distant enough to justify a new subtype designation [12]. There can also be recombinants between different subtypes, which are denominated circulating recombinant forms

(CRFs) if they have a significant epidemic spread, meaning if they are fully sequenced and found in three or more unlinked individuals [8, 12], or unique recombinant forms (URFs) if they don't meet these terms [8]. In certain regions, recombinants account for at least 25% of all HIV infections [1].

Given the fact that the HIV-1 epidemic started in the African continent, knowing that all the different subtypes, sub-subtypes and recombinants are found there [6] is not surprising. However, in all the other continents, even from country to country or within countries, the geographical distribution of these subtypes is quite heterogeneous, and this may be due to multifactorial, including founder, effects, population growth and urbanisation, cultural and sexual factors, and transport links and migration [6, 8].

As observed in Figure 2.1, which shows the global distribution of the major HIV-1 subtypes and recombinants, subtype A predominates the Eastern Africa region, the Soviet Union and other former Soviet Union countries [11, 14, 15]; subtype B is the most common in North America, Europe, Australia and South America, this last one reporting increasing levels of BF recombinants [11, 14, 15]; and subtype C is the most widely spread strain in Southern Africa and India [11, 14, 15]. In some other regions, it's the CRFs that dominate the reports of infection, particularly West and West Central Africa (CRF02_AG) and East and Southeast Asia areas (CRF01_AE) [1, 14, 15]. Globally, the C subtype is the most common, accounting for 46.6% of HIV-1 cases, followed by the B subtype (12.1%) and the A subtype (10.2%) [14].

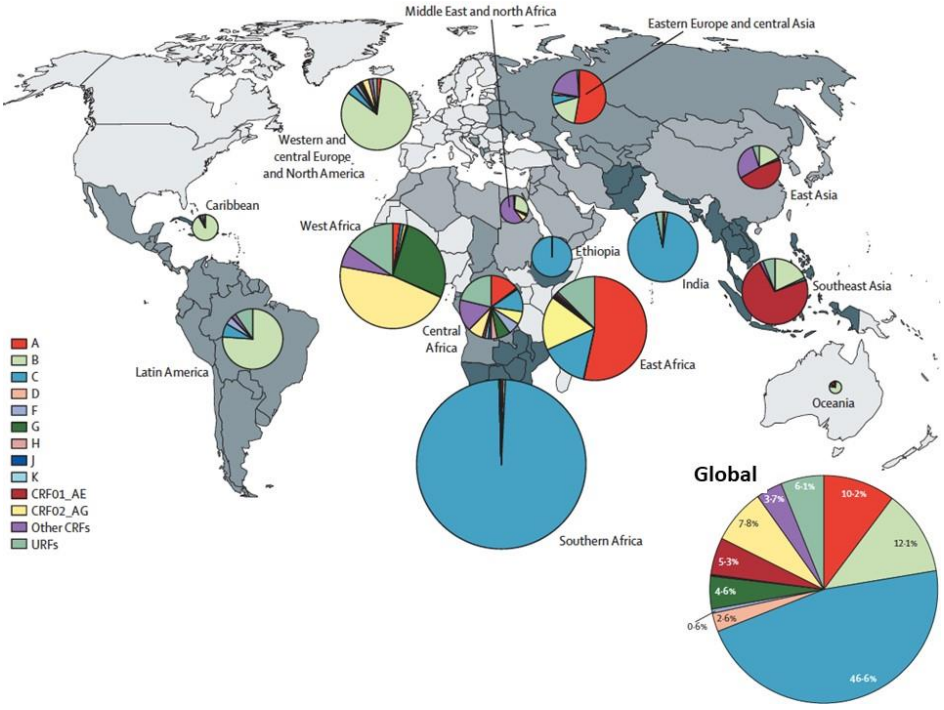


Figura 2.1 - Global distribution of the major HIV-1 subtypes, by pie charts, with different colours representing different subtypes, CRFs and URFs, with the size of the pie chart representing the comparative percentage of HIV-1 cases in each region. On the bottom right corner, there is a pie chart with the percentage of subtypes and recombinants globally [14]

2.2 PATHOPHYSIOLOGY AND PATHOGENESIS OF HIV-1

HIV-1 and -2 are both retroviruses [6, 16] and belong to the largest Retroviridae subfamily, Orthovirinae, which is comprised of six different genera – alpha, beta, gamma, delta, epsilon and lentivirus [17], this last one being the one that HIV is a member of [16, 18]. Being retroviruses entails that they have two identical copies of single-stranded RNA molecules as their genome [16, 18, 19] and, therefore, have to have a reverse transcriptase mechanism in order to

produce DNA out of their RNA, so that the virus can replicate and, then, integrate into the host genome [18, 19, 20].

The spherical HIV-1 viral particle has a diameter of 100nm [16] and is composed of a capsid and a sheath [20] surrounded by a lipidic membrane [16], as seen in Figure 2.2, derived from the membrane host cell [17, 18]. The capsid consists of p24 proteins surrounding the two copies of unspliced RNA [16, 18, 20], which is stabilized as a ribonucleoprotein complex [18], and the reverse transcriptase, integrase and protease enzymes [16, 18], while the sheath encircles a matrix (p17) located in the centre of the viral particle containing glycoproteins that establish the structure of the virus [18, 20]. On the other hand, the exterior of the viral particle is scattered with trimeric envelope glycoprotein spikes that are comprised of an ectodomain (gp120) and a transmembrane domain (gp41) [16, 17].

All these structures are encoded by nine genes assembled in one chromosome [21, 22], which have different functions in the viral replication cycle: three structural genes (*gag*, *env* and *pol*) and six regulatory/accessory genes (*tat* and *rev*, and *vif*, *vpr*, *vpu*, and *nef*, respectively) [16, 21, 23]. When it comes to the structural genes, the *gag* gene, just like in other retroviruses, encodes for the structural proteins of the capsid (p24, p7 and p6) and the matrix (p17) [16], which together with the proteins/enzymes encoded by the *pol* gene – reverse transcriptase (converts viral RNA into DNA), integrase (incorporates the viral DNA into the host's chromosomal DNA) and protease (cleaves large Gag and Pol protein precursors into their respective components) [16] - form the nucleus of the maturing HIV viral particle [16]. The *env* gene, in turn, encodes for the gp120 and gp41 glycoproteins [16], which promote receptor-mediated fusion of viral and cellular membranes to initiate the infection process [17] – specifically, gp41 allows the virus to enter the cell, whilst gp120 allows the virus to attach to DNA [20]. Regarding the regulatory genes, the *tar* gene encodes for the Tar protein, which binds to the TAR site (Transactivation Response Element) in the 5' end of all HIV-1 RNA transcripts, that promotes the transcription and the synthesis of complete and longer virus mRNAs [16, 23]. The *rev* gene encodes for the Rev protein, which, by its direct binding to the RRE target sequence (Rev Response Element) in the *env* coding sequence [23], facilitates the expression of structural and enzymatic genes [16]. Lastly, the accessory genes and their respective proteins have different functions: the Vpr protein is involved in the arrest of the cell cycle and enables the reverse transcribed DNA to gain access to the nucleus in non-dividing cells, such as macrophages; the Vpu protein performs the correct release of the virus particle; the Vif protein enhances the infectiveness of progeny viral particles; and the Nef protein allows virus budding in the late stages of the virus replication cycle [16].

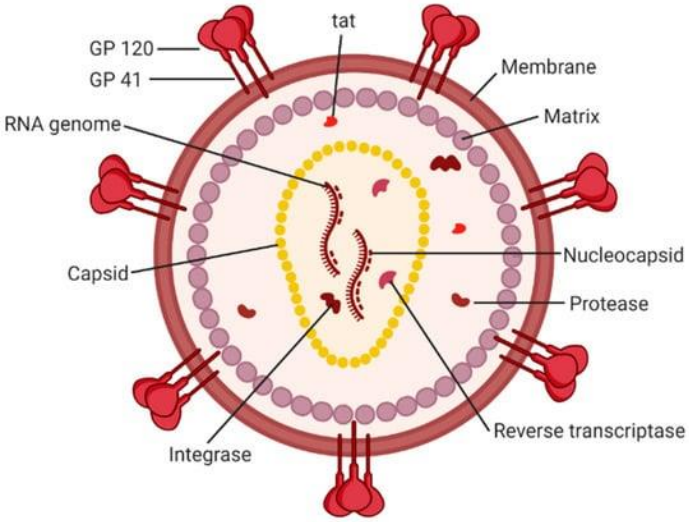


Figure 2.2 - Schematic representation of the HIV-1 viral particle [24]

HIV-1 transmission happens through direct exposure to infected blood or secretions of skin damage, such as needles or sharp tools, or abrasions in mucosal tissues during sexual

intercourse – in fact, infection acquired during heterosexual intercourse is the most common infection method worldwide [16]. This is required as HIV-1 doesn't have the ability to survive outside the bloodstream or lymphatic tissue [16]. After the virus infects the cells lining the vaginal or anorectal mucosae, it infects cells of the immune system, mainly CD4+ cells, which are T-lymphocytes that have monomeric glycoproteins, called CD4, on their surface [16, 18, 23], which will then lead to the spread of the virus to regional lymph nodes and subsequently into the bloodstream [16].

After it infects the host cell, HIV-1 goes through its replication cycle, which can be divided into six different stages: (1) binding and entry; (2) uncoating; (3) reverse transcription; (4) provirus integration; (5) virus protein synthesis and assembly; and (6) budding [16], as schematically represented in Figure 2.3.

The HIV-1 particle binds, through the gp120 glycoprotein in the envelope spikes, to the CD4+ cells [16]. However, in order for the viral particle to fuse to the host cell's membrane, lentiviruses, as is the case for HIV-1, need additional cell surface proteins [18] – these proteins are called chemokine receptors, and the ones used by HIV-1 are, among others, CXCR4 and CCR5 [16, 18] -, that, mediated by the gp41 envelope component, allow the viral particle to enter the host CD4+ cell [23].

After the core of the viral particle is uncoated into the cytoplasm of the host cell, freeing its viral RNA, creating a large enzymatic complex ("preintegration complex") [16, 17], which is then reverse transcribed into a double stranded DNA particle called the provirus [17], the viral DNA is integrated into the host's chromosomes in the cell nucleus [17]. After transcription, unspliced and spliced mRNA transcripts are transported out of the cell nucleus and into the cell cytoplasm to suffer translation [16, 18] and, consequently, synthesize all the needed structural, regulatory and enzymatic products [16, 23] required to maintain infection and propagate new virions [17]. The last step is budding, which allows the new infectious viral particles to acquire a new envelope while budding through the host cell membrane, effectively leaving the host cell [16].

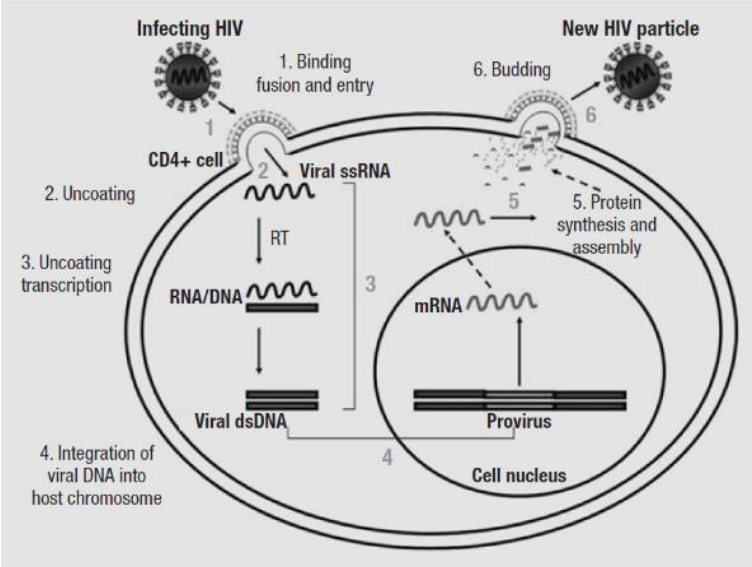


Figure 2.3 - Schematic representation of the HIV-1 replication cycle [16]

This infection leads to a progressive and near complete loss of all CD4+ cells in the host [23, 25], due to the virus replication and the chronic activation of immune system cells [16, 23]. This progress can be divided into three different clinical stages, based on, and not only, the rate of decline in CD4+ cell count, increase in plasma virus load and the HIV-1 associated opportunistic infections, as well as host factors, like genetic factors, innate immunity and HIV-specific immune susceptibility [16, 25, 26]: an acute stage, which corresponds to the primary infection; a chronic stage characterized by a clinical latency period; and, lastly, a crisis stage, where the host has developed AIDS [16, 23].

The acute stage, which is characterized by high levels of viral particles in the plasma and infection of a considerable percentage of lymphocytes [23], usually lasts for 16 weeks after the initial infection, despite symptoms only occurring after 2 to 4 weeks [26]. However, neutralising antibodies against HIV-1 only appear at the 12 weeks' mark and they target a very small range of epitopes, which leads to a rapid viral escape through different variations [8]. In this phase, infected hosts present flu-like symptoms, such as fever, fatigue, headaches, rashes, oral ulcers, myalgia, weight loss and chills [16, 20, 23, 26]. The more severe these symptoms are, the faster the progression to the AIDS stage is [16].

Following the primary infection, the chronic stage frequently lasts for 7 to 11 years [23], which is a clinical latency and asymptomatic period where there is a slow progressive loss of CD4+ lymphocytes and impairment of the immune system, while the HIV-associated pathogenic effects persist [16].

Lastly, during the AIDS stage, the percentage of infected lymphocytes increases (can exceed to 1 in 40 infected CD4+ cells) and, consequently, the number of CD4+ lymphocytes keeps decreasing (less than 200 cells/µl) [16, 23]. These levels increase the risk of opportunistic infections in the host, namely *Microcystis carinii*, *Candida albicans*, *Cytomegalovirus*, *Herpes zoster* or *Pneumocystis jiroveci* [16, 26], which can lead to possibly deadly diseases. This stage is characterized by lymph node swelling, diarrhea, extreme weight loss, respiratory and gastro-intestinal problems and fever [16, 20].

HIV-1, as a lentivirus, has a remarkable genetic variability [8, 20, 23] and capacity for viral gene expression [23]. This is due to (1) the reverse transcriptase mechanism that characterizes it, which has no proof-reading system [8, 16, 27, 28], leading to a high number of mutations, (2) the rapid viral replication [16, 27, 28], which means that more than 1000 virions are generated each day in the infected individual [16], and (3) the genetic recombination processes [8, 16, 28] between different viral particles within the same infected individual [16]. All these provide HIV-1 with a variability that is responsible for the virus being able to resist through the host's immunity and the effects of both antiretroviral and prophylactic therapy [16, 28], which has made the rapid development of an effective and safe vaccine very difficult [23].

2.3 REVERSE TRANSCRIPTASE

HIV-1 reverse transcriptase (RT) plays a crucial role in the virus's replication cycle, as it is responsible for the synthesis of double-stranded DNA using single-stranded RNA as the template [29]. It is an asymmetric multifunctional heterodimer comprised of two subunits with the same sequence but different three-dimensional structures [19, 29, 30, 31], due to the cleavage of the Gag-Pol glycoprotein [19, 22, 32] – p66 (with 66 kDa of molecular mass and 560 amino acid residues) and p51 (with 51 kDa of molecular mass and 440 amino acid residues, which correspond to the first 440 amino acid residues of the p66 subunit [33], given that it's missing the RNase H domain [30, 34]) [19, 22, 31, 32].

RT has two main enzymatic active sites conferred by the p66 subunit [31]: polymerase active site which acts on DNA polymerization using DNA or RNA as the template, and is composed of three catalytic carboxylates (D110, D185 and D186) that bind two divalent Mg^{2+} ions [32], and the RNase H domain that is responsible for the cleavage of the DNA-RNA hybrid, by the hydrolysis of RNA - both work to complete the process of converting single-stranded viral RNA into double-stranded DNA [19, 22, 32, 33] in the cytoplasm of the infected cell [32]. On the other hand, the p51 unit plays a structural role [32].

The p66 subunit resembles the shape of a right fist [19, 22, 34], hence the names of the different sections of this subunit: fingers, palm (where the polymerase active site is present [19]), thumb, connection (tether between the polymerase and RNase H sections [34], which are, according to crystallography studies, approximately 60 Å apart [32]) and RNase H domain [19, 22, 30, 32], as observed in Figure 2.4.

Based on different crystal RT structures, it was determined that this enzyme has the tendency to alter its conformation so to be able to perform its functions [19]. Thus, when the

nucleic acid template binds to RT, as shown in Figure 2.4, there is a conformational change in the p66 thumb and finger's position, where it goes from a "closed" position, where the p66 thumb subdomain folds down onto the fingers subdomain [30], to an "open" position [19, 32].

Due to its important function in the replication cycle of HIV-1, more than half of the available antiretroviral drugs target the RT [29], which are divided in two different types – nucleotide/nucleoside reverse transcriptase inhibitors (NRTIs) and non-nucleotide reverse transcriptase inhibitors (NNRTIs).

NRTIs, which are the most commonly used components of HIV-1 ART treatment schemes [35], inhibit replication by blocking the polymerase reaction of the RT enzyme [19, 29]. They achieve this due to competition with naturally occurring deoxynucleoside triphosphates (dNTPs) for their incorporation as chain terminators of the growing DNA chain, due to a lack of a 3'-OH group [36], consequently causing chain termination [29, 37]. This way, given that NRTIs were developed as structurally different analogs of said dNTPs [19], NRTIs have first to be converted into their respective triphosphate or diphosphate form by the host's cellular enzymes before they can be incorporated into the DNA chain [37].

NNRTIs work by (1) allosterically inhibiting the translocation of the template-primer substrate following nucleotide binding during DNA polymerization, preventing elongation of DNA strands [22, 33, 38] or by (2) affecting the binding of the catalytic carboxylate residues of the polymerase active site to the Mg^{2+} cations, causing a shift in the position of the primer grip, which helps position the primer strand in the polymerase active site, so to create the template-primer substrate [32], and a decrease of the thumb mobility in the p66 subunit [29, 38]. When present, NNRTIs cause RT to adopt an "open" conformation, where the p66 thumb domain is ~30 Å away from the fingers subdomain [30, 49].

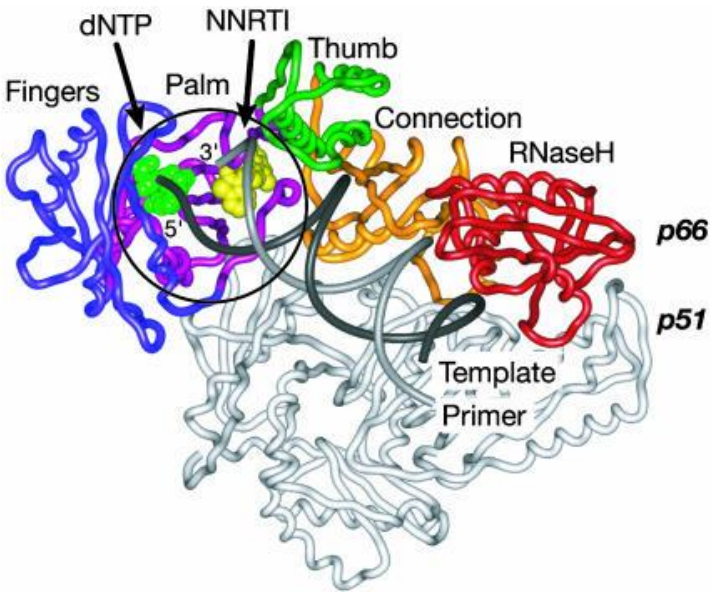


Figure 2.4 - Ribbon representation of HIV-1 RT with a nucleic acid and NNRTI. The fingers, palm, thumb, connection and RNase H subdomains of the p66 subunit are shown in blue, purple, green, orange and red, respectively. The p51 subunit is represented in light grey. The template and primer DNA strands are in dark grey and grey, respectively. The NNRTI is represented in yellow and the incorporating dNTP in fluorescent green [31]

Approximately 10 Å away from the polymerase active site [29, 32, 39], but not contiguous with, specifically between the β6-β10-β9 and β12-β13-β14 sheets in the palm subdomain of the p66 subunit [29], is the hydrophobic elastic pocket where the NNRTIs bind to, called NNRTI binding pocket (NNIBP) [33, 38], represented in Figure 2.5. It consists of sixteen different residues: L100, K101, K103, V106, T107, V108, V179, Y181, Y188, V189, G190, F227, W229, L234 and Y318 of the p66 subunit, and E138 of the p51 subunit [32], which by different types of interactions and forces with the NNRTIs, help them stabilize in the NNIBP [29].

The NNIBP only exists in the RT structure when NNRTIs are present/bound to it [22, 31, 32, 40]. Normally, this cavity is mostly filled by the aromatic-ring containing Y181 and Y188

residues' side chains [30], which after NNRTIs bind, suffers structural rearrangements to get the hydrophobic core from a "closed" to an "open" form [40], with the p66 thumb and fingers subdomains in their hyperextended conformations [32], by displacing the Y181 and Y188 residues, and by rotating the β12-β13-β14 palm subdomain sheet, resulting in an extended conformation of the primer grip [32, 40].

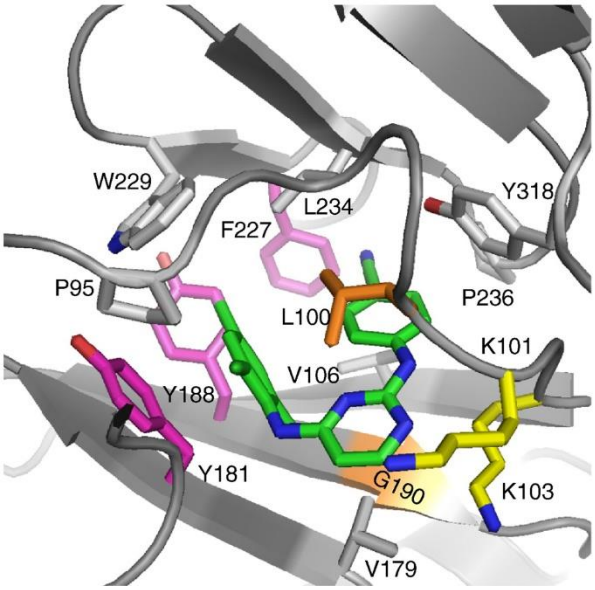


Figure 2.5 - Ribbon representation of the NNRTI-binding pocket (NNIBP) , showing the residues where NNRTI-resistance mutations occur. (To note that the T107, V108, V189 and E138 residues of the NNIBP are not shown in the representation) [32]

2.4 ANTIRETROVIRAL THERAPY

As mentioned previously, developing a safe and effective vaccine for HIV-1 infection is a particularly difficult task. However, despite there being no form of cure or treatment with permanent results for HIV-1 infection [20], antiretroviral therapy (ART) that targets certain proteins/enzymes in the HIV-1 replication cycle has been produced, thus delaying the progression of the disease and decreasing its incidence and mortality/morbidity [18, 41, 42, 43]. This way, HIV-1 infection can be a more controlled chronic disease for whoever has it, leading to a far more comfortable life. These antiretroviral drugs act by blocking enzyme's active sites or interrupting protein interactions [21], which, ideally, leads to a decrease in viral load and transmission rates [28, 44].

There are currently six different classes of antiretroviral drugs in the market, each of them with different antiretroviral potency, meaning that they decrease the plasma HIV-1 RNA levels in different extents [43]: nucleoside/nucleotide reverse transcriptase inhibitors (NRTIs); non-nucleoside reverse transcriptase inhibitors (NNRTIs); protease inhibitors (PIs); integrase inhibitors (INIs or INSTIs); CCR5 antagonists; and fusion inhibitors. In addition to the use of antiretroviral medication by HIV-positive individuals, the WHO recommends pre-exposure prophylaxys (PrEP) for individuals that are at substantial risk of HIV infection, such as sex workers and their clients, serodiscordant couples, people who inject drugs, among others [42].

ART is intended to be administered in regimens of three or more different drugs and drug classes so to achieve the best possible and more effective results [41], generally in fixed-dose formulation and once a day [43], which, according to WHO, should be started as soon as the diagnosis is given to the patient (Treat All). The most common regimens include a combination of 2 NRTIs and a potent third agent from another drug class [43]. However, the regimen should be individually chosen for each patient given a set of conditions, like resistance testing results, predicted virologic efficacy, toxicity, tolerability, pill burden, dosing frequency, drug-drug interactions and, when feasible, patient and practitioner preference [43].

NRTIs are one of the two classes of antiretroviral drugs that target the virus's reverse transcriptase. They act by blocking the polymerase reaction of the reverse transcriptase enzyme, inhibiting replication [19, 29]. They do this by competing with naturally occurring dNTPs for its incorporation into the growing DNA chain, therefore causing chain termination [29, 37]. There are currently 5 different NRTIs used in significant frequency: abacavir (ABC), emtricitabine (FTC), lamivudine (3TC), tenofovir (TDF) and zidovudine (AZT/ZDV) [45].

The other class that targets the reverse transcriptase is NNRTIs. As opposed to NRTIs, NNRTIs inhibit polymerase activity not by competing with dNTPs, but by indirectly interfering with their incorporation to the growing DNA chain [19, 29, 46]. They do this allosterically, by binding to a hydrophobic pocket close to, but not contiguous with, the RT active site [19, 29, 37], displacing the catalytic aspartate residues relative to the polymerase-binding site [29, 38], leading to the non-binding of the incoming dNTP during polymerization [47]. Four NNRTIs are regularly used: efavirenz (EFV), etravirine (ETR), nevirapine (NVP) and rilpivirine (RPV) [45].

On the other hand, there are drugs that don't act on the reverse transcriptase, like PIs, which in turn act on the protease's activity, inhibiting maturation of virions, resulting in the production of non-infectious particles [41]. They do this by preventing protease from processing the Gag and Gag/Pol polyprotein precursors, which are necessary for viral maturation [37]. Atazanavir (ATV), darunavir (DRV), ritonavir (RTV), fosamprenavir (FPV) and tipranavir (TPV) are the five most used PIs [45].

INIs inhibit integrase by binding to its active site and disrupting the correct positioning of the viral DNA relative to the active site, and also by disrupting the integrase interaction with two essential magnesium ions [37]. The only approved INI is raltegravir (RAL) [45].

CCR5 antagonists act by allosterically inhibiting the binding of the gp120 glycoprotein to the host CCR5 coreceptor [37]. However, this process does not work in occasions where there are undetected minority variant CXCR4 tropic viruses instead [37]. Maraviroc (MVC) is currently the only CCR5 antagonist in use [45].

Lastly, fusion inhibitors disrupt the interaction of the gp41 glycoprotein hairpin formation (where two complementary parts of gp41 fold onto one another, shortening the protein, thus bringing the host cell and viral cell membranes together) [37]. As well as the last two ART drugs, there is also only one approved fusion inhibitor called enfuvirtide (ENF), which has a limited use due to its subcutaneous administration, which elicits painful injection site reactions [37].

Despite all of these different drugs, several others have been previously on the market and administered to patients, but because of either drug side-effects, tendency to accumulate resistance mutations (DRMs) or high rates of toxicity, among other factors, or simply that more effective drugs have been produced, they have been discontinued or rarely/not used in a significant frequency.

Incidentally, drug resistance is one of, if not the, primary causes for ART failure [41], given the fact that it results in increased viral load and decreased CD4+ counts [44], ultimately posing a potential public-health threat that complicates the path towards AIDS elimination [41, 42]. DRMs appear due to pharmacological selection pressure from antiretroviral drugs [29], and also due to several patient limitations when it comes to the taking of the drugs, such as the inconvenience of taking a large number of drugs, the complexity of regimens and drug-drug interactions within these regimens, the side effects linked to the drugs, insufficient patient adherence/compliance and high cost of drugs [41, 44].

DRMs can be primary, accessory or compensatory according to their role in the drug resistance acquisition: primary mutations are the ones that lead directly to drug resistance, whereas the accessory ones appear alongside the primary DRMs to increase their drug resistance level; compensatory DRMs reduce the fitness cost associated with the drug resistance conferred by other DRMs [37, 42, 48]. The number of DRMs needed and the effect of each of them to the viral fitness contribute to the treatment's genetic barrier to resistance [37, 42].

Patients can obtain DRMs either through acquisition or transmission - acquired drug resistance (ADR), through drug selection pressure, or transmitted drug resistance (TDR), from

the transmission from person to person [37]. ADR is developed in more than 70-80% of patients experiencing virological failure [42], and the lack of any ADR in these conditions usually means there wasn't a correct therapy adherence [42]. ADR is also likely to occur in patients that are using PrEP, even if in low resistance levels, however likely mitigated through frequent laboratory testing so to initiate ART as soon as infection is identified [42]. TDR, in particular, is very worrisome, given that, when it occurs in treatment-naïve patients, it decreases by a lot the possible initial treatment options available, leading thus to rapid drug failure, and severe outcomes, if the mutations aren't identified first, which is why the analysis of DRMs is critical [28, 37]. Given the high and increasing prevalence of transmitted drug resistance (TDR), particularly affecting non-nucleoside reverse transcriptase inhibitors (NNRTIs), in low- and middle-income countries (LMIC) – now up to 25% [42, 49] - it is critically important to ensure not only correct adherence to therapy, but also the appropriate selection of treatment. This selection should take into account factors such as resistance testing results at the time of diagnosis and at the initiation of therapy. Despite this problem, some transmitted DRMs commonly revert back to wild type gradually in a span of a few years [37]. Moreover, even though uncommon, naturally occurring drug-resistant viruses can also be generated, despite most DRMs being largely non-polymorphic [37].

When it comes to cross-resistance, it is highly unlikely between drug classes, meaning that viruses that are resistant to one drug class are susceptible to a different drug class, even between NRTIs and NNRTIs, that both target the reverse transcriptase activity [37, 42]. However, there is a high cross-resistance between drugs from the same class, signifying that one DRM can typically decrease susceptibility to more than one drug from the same class [37].

In this manner, it is an absolute humanitarian necessity that patients in LMIC have access to ART that is properly selected and correctly administered, to avoid high rates of mortality and infection [41]. Be that as it may, there are several limitations in these countries to make this happen correctly, for instance, the cost, restrictive licensing policies, access to diagnostic testing, drugs stock outs, and suboptimal retention in care [41, 42], which results in the estimated 80% of infected persons with HIV-1 in LMIC not having access to antiretroviral treatment [41].

2.5 NNIBP DRUG RESISTANCE MUTATIONS

The NNRTI-binding site in the NNIBP, when ligand-free, has high plasticity, which gives way to the conformational changes that occur when NNRTIs bind to it, consequently reducing the activity of the reverse transcriptase, due to the decrease in this plasticity, whether by influencing the catalytic residues directly, or by influencing the enzyme-substrate or template-primer interactions [29, 30, 39, 50]. The conformational changes to the NNIBP depend on the inhibitor in question, regardless of whether or not there are drug resistance mutations at play [30].

As the NNRTI binding takes place, the Y181 and Y188's side-chains from the hydrophobic pocket are displaced [29, 50], which, thus, contribute to the stabilization of the drug-enzyme complex [29]. Three-dimensional representations of the interaction between each of the following NNRTIs with the binding site is present in Figure 2.6.

When it comes to NNRTIs, the first-generation drug was nevirapine (NVP), which binds to the NNIBP by adopting hydrophobic, stacking interactions between the aromatic side chains of the Y181 and Y188 residues and its pyridine groups [29, 50]. Additional electrostatic interactions with the K103 residue allow for the total stabilization of this NVP-RT complex [50].

With the goal of decreasing the contribution of the stacking interactions in the stabilization of the NVP-RT complex in the NNIBP, second-generation NNRTIs were developed, specifically efavirenz (EFV) [29], which is one of the most used NNRTI, due to its high tolerability and potency [33]. It performs by making direct or water-mediated hydrogen bonds and Van der Waals interactions with the protein backbone of the K101 and K103 residues,

crucially the backbone carbonyl oxygen of K101 [29], strengthening the stability and affinity of EFV to the binding site, more than other NNRTIs [29].

For the purpose of avoiding cross-resistance with NVP and EFV, and increase the genetic barrier, third-generation NNRTIs appeared, such as etravirine (ETR), which is a diraylpyrimidine (DAPY), characterized by a strong torsional flexibility [29, 32], which allows it to adapt to and have compensatory interactions with the altered binding site by DRMs, restricting the accommodation of more DRMs [29, 32]. It is stabilized in the binding site by a large number of non-bonding contacts, specifically with the p51's E138 residue, interaction not existing in the first- and second-generation NNRTIs [29].

Rilpivirine (RPV) was designed to optimize DAPY compounds, by interacting with the conserved amino acids in the NNIBP, particularly the primer grip's W229 [29], which can help in stabilization of inhibitor binding, even in the presence of DRMs [40]. It shares the same flexibility and its functions as EFV [29].

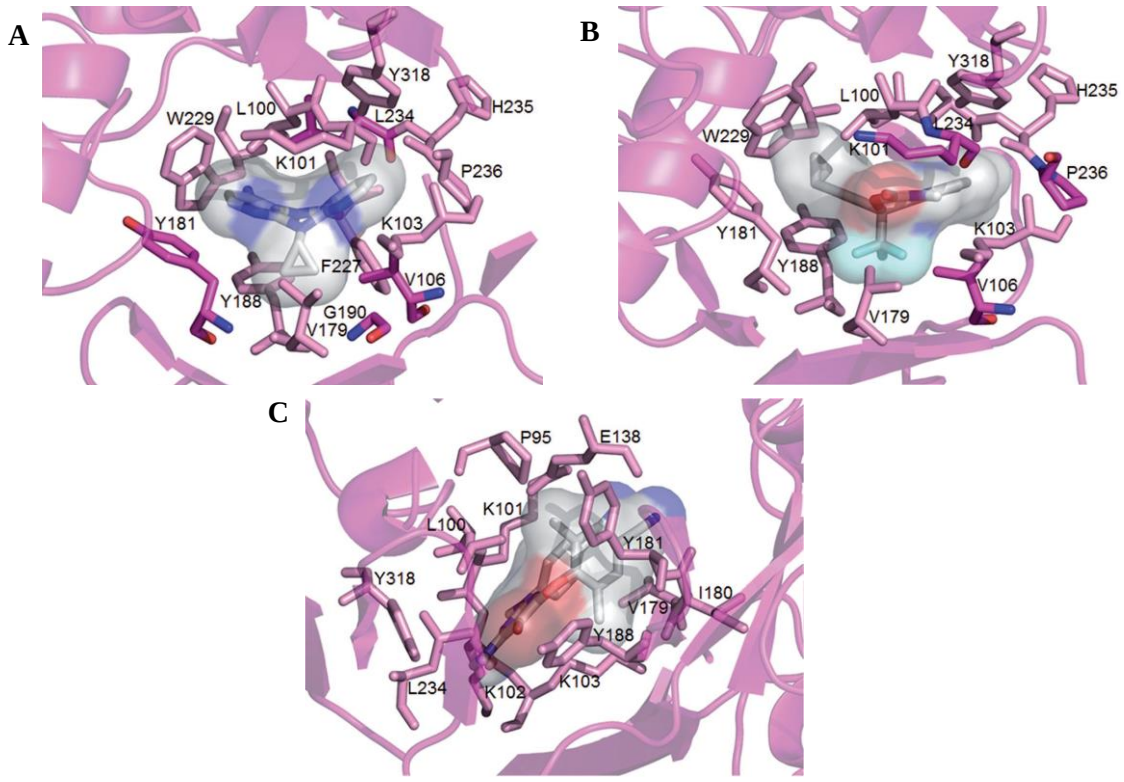


Figure 2.6 - Three-dimensional representation of NNRTI interactions in the NNIBP, showing the best posed for NVP (A), EFV (B) and ETR (C) bound to wild-type RT. The NNRTI in question is shown as grey carbon sticks with its correspondent surface area. The NNIBP residues involved in these interactions are represented in pink; and amino acids which are able to establish additional contacts or hydrogen bonds are given in magenta [29]

As mentioned in chapter 2.4, DRMs are one of the reasons why treatment failure exists, and the ones associated with NNRTIs are mostly present in and around the NNIBP [32]. They hinder the normal behaviour of NNRTIs by either directly impeding the binding of the drug to the NNIBP by altering its size, shape or polarity, or by indirectly preventing access to the binding pocket, despite some mutations also influencing RT functions other than the most common DNA polymerization [29].

In particular, NVP has a very low genetic barrier, as a single point mutation near the NNIBP is enough for the development of high-level resistance and, thus, turning the drug ineffective [19, 29, 51]. This may have something to do with the fact that it has been shown that there are very weak, or no positive correlations or interactions between NVP resistance mutations [29]. Mutations at the Y181 and Y188 positions, which are located in the hydrophobic core of the

NNIBP [32], like Y181C, which is one of the most common NNRTI resistance mutation [19, 39], and Y188L, are responsible for this high-level resistance [29, 31].

These mutations, through the loss of their aromatic ring stacking interactions with the drug [29, 32], directly affect NVP binding by destabilizing protein-inhibitor hydrophobic interactions, therefore decreasing its stability and affinity to the binding pocket [19, 29, 40, 39, 50]. This way, both the association and dissociation rates decrease, meaning the rate of entry and exit of the inhibitor into the binding pocket, respectively [50]. Both Y181C and Y118L confer >50-fold reduced susceptibility to NVP, according to the HIV Drug Resistance Database by the Stanford University [52, 53].

Second-generation NNRTIs, like EFV, on the other hand, have a higher genetic barrier than NVP, as it has been supported that high-level EFV resistance is only acquired through the accumulation of multiple mutations [29]. The most common EFV resistance mutation is K103N [19, 33, 39], as well as most other NNRTIs, as it confers cross-resistance to all of them [39, 54], causing 20- to 55-fold reduced susceptibility to all licensed NNRTIs [31].

K103N, unlike most other NNRTI DRMs, does not interfere directly with NNRTI binding by provoking conformational changes in the NNBIP – instead, it stabilizes the closed conformation of the binding pocket, not allowing the NNRTI to bind to it, or doing it more slowly [31, 32, 39, 40], given that the K103 residue is located at the rim of the entrance to the binding pocket, with its side chains pointing out [19, 32]. Therefore, the association rate is reduced [39, 50] and, thus, the energy barrier is increased [29]. It does this by forming a hydrogen bond between the N103 flexible side chain amide and the Y188 phenoxyl oxygen at the pocket entrance [29, 39, 40, 50].

Although K103N is often the first NNRTI resistance mutation that appears in patients who are failing EFV treatment, it is usually followed by accessory mutations, like L100I or V108I, to form double- or even triple-mutants with higher levels of drug resistance than K103N alone [29, 54]. In particular, in NVP and EFV failing therapies, L100I with K103N increases the 20- to 55-fold reduced susceptibility up to 100-fold [51], whereas V108I together with K103N only increases about 2- to 5-fold [51], making the K103N+L100I double mutant very detrimental [39].

Mutations in the L100 residue, which is located in the central region of the binding pocket [32], in specific L100I, causes resistance by changing the conformation of the NNIBP, through steric interference between the β-branched isoleucine and the bound NNRTI [19, 29, 40], turning it from β- to γ-branched [32]. This is the result of a decrease in side chain length, as it is responsible for forming a bottleneck at the entrance of the binding site together with the V179 residue [39], and distortion of the NNIBP, bringing about the loss of protein-inhibitor interactions [29]. So, both these mutations together lead to decreased association and dissociation rates for the NNRTI in question, meaning there's a low energy barrier both for the entry and exit of the inhibitor out of the NNRTI binding pocket [39].

Regarding the third-generation NNRTI ETR, it needs more than one mutation to cause high-level resistance (>10-fold reduced susceptibility), as it can resist the appearance of more DRMs once only one is present, due to its flexibility [29, 32]. In particular, it is able to inhibit the K103N mutant by interacting with the side chain of the N103 residue, something first- and second-generation NNRTIs couldn't do [32], making the mutated enzyme conformationally more opened than normal [29]. Despite this, mutations at the 181 position, and to a lesser extent, the 190 position, may provide the base for high-level ETR resistance [51]. This is particularly concerning, as patients developing NVP therapy failure are at higher risk of also developing ETR treatment failure, as mutations in the Y181 residue occur in about 45% of NVP receiving patients [38].

3 STATE-OF-THE-ART COMPUTATIONAL METHODS AND TECHNIQUES

3.1 CLUSTERING

Given the wide variety and increasing volumes of data being produced each day, it is crucial to be able to deal with it – one way to do it is by classifying or grouping them into categories or clusters according to their similarities and dissimilarities. This can be achieved by using machine learning methods.

The three main categories of machine learning are supervised learning, unsupervised learning and reinforcement learning. Supervised learning entails the training of the model on labelled data, or a set of input variables and their corresponding output variable [55], to generate predictions on unlabelled data, meaning the "correct answer" is already know in the training process [55, 56]. On the other hand, unsupervised learning uses unlabelled data, or input variables without their respective output, for the training of the model at hand [55], where it automatically and independently learns from the input variables by extracting and selecting features and discovering hidden and interesting patterns from it based on the data's similarities and differences [55, 56], trying to find the "correct answer", or the correspondent output variable. In reinforcement learning, there is an agent that, by learning from errors committed during the training process, enables it to find the optimum solution and accomplish a goal in a certain situation in their respective environment [56].

In many cases, obtaining labelled data is very difficult, costly and time consuming, whereas acquiring unlabelled data is a much easier process [55]. Thus, unsupervised learning is most widely adopted, even though the process of building a model from unlabelled data is more arduous [57]. There are four types of unsupervised learning methods: clustering, principal component analysis, anomaly detection and autoencoders [56], and clustering is the one of the most important.

Clustering is the process in which data is grouped into clusters based on some measure of similarity [58, 59, 60], usually distance metrics [55, 58], like Euclidean Distance, Manhattan Distance and Minkowski Distance [55, 58]. Therefore, a cluster is a collection of organized objects which are "similar" (have less distance) between each other and "dissimilar" (have more distance) to objects in different clusters [56, 59, 60].

Clustering can be further divided into four types: partitioning, or exclusive clustering, where data is grouped so that each data point can only belong to one cluster, and the number of clusters in which to divide the data is pre-defined [60] – one example is K-means; overlapping, where fuzzy sets are used to cluster the data, in which each point may belong to two or more clusters at a time; probabilistic, which uses probability distribution to create the clusters; and hierarchical, where every data point is a cluster, and the iterative unions that happen between the two nearest clusters gradually reduce the number of clusters [56].

Hierarchical clustering, given the already mentioned large-scale data sets that are increasingly appearing, has allowed for an improved analysis and detailed description of these same data sets [60]. It works by organizing the data into a hierarchical structure according to a distance matrix [60], where the process starts with each object belonging to a separate cluster that sequentially start merging according to their shortest distance [56]. After merging into a new cluster, the previously merged individual clusters are removed from the distance matrix, and again the distance between this new merged cluster and all the other clusters is calculated and the distance matrix updated [56] – it's worth noting that, after each merge operation happens, the

proximity between the two similar clusters to merge next decreases [57]. This process is repeated until all individual objects are merged into a single cluster [56].

This results in a single, all-inclusive cluster that gradually divides into smaller singleton clusters of individual objects [59], where each intermediate level is viewed as combining two clusters of a lower level or splitting clusters from a higher level [59], given the type of hierarchical clustering at hand, in line with the pairwise distance/proximity between the objects. This results in a tree-like structure called dendrogram [57, 59, 60], in which the root node represents the entire data set, each leaf node is one of the clusters constructed, the intermediate nodes describe the cluster relations illustrating the merge or split operation performed, and its height symbolizes the distance between each pair of clusters [57, 60]. After the construction of the dendrogram, the user may manually select the number of clusters, or *k*, by examining it and choosing an appropriate cut-off point [57, 60].

In conformity with the approach used in hierarchical decomposition, hierarchical clustering can be agglomerative or divisive. Agglomerative clustering applies a bottom-up approach [57, 58, 59], where it begins with all the individual clusters and the most similar pair of them is merged [55, 57] until the single cluster will all of the data is reached, whereas the divisive clustering alternative uses a top-down strategy [57, 58, 59], in which it begins with the single cluster will all of the data objects and a chosen cluster is split into two sub-clusters [57] until there are only the individual clusters.

Despite its advantages, the use of hierarchical clustering has its limitations. First, the inflexibility of the agglomerative merge or divisive split step in the hierarchical decomposition, as once they're done, they cannot be undone [59, 60], which leads to possible misclassification and low-quality clusters that can't be corrected after [59, 60]. The second disadvantage, and the most common critic, is that hierarchical clustering algorithms tend to lack robustness and are, thus, sensitive to noise and outliers [60]. Another limitation is the application high process and computational cost, especially in large-scale data sets, as all objects have to compared to one another before every clustering step [58, 60]. The final disadvantage is that hierarchical clustering methods are inclined to form spherical shapes and reversal phenomenon, which induces the distortion of the normal hierarchical structure [60].

For this dissertation, hierarchical clustering was utilized, specifically the Louvain method, developed by Blondell et al. in 2008 [63]. The Louvain method is an iterative, greedy heuristic-based approach with the aim of partitioning a graph network into communities that best optimize a quality function, in this case modularity [61, 62, 63, 64], which is illustrated by a scalar value between -1 and 1 that measures the density of links inside communities as opposed to between communities [63, 64].

The Louvain algorithm can be divided in two different phases: the first phase is the local moving phase, where each node in the graph is moved to the neighbouring community that yields the biggest increase in the modularity quality function (if the modularity value is negative or does not improve from the original community where the node was, it stays in its current community [61, 64]) [61, 62, 63, 64]. This phase is complete when no individual node move from a community to one of its neighbouring communities improves the modularity [62, 63]. The second phase, which can be called aggregation of the network phase, entails the creation of a new graph network, where each node will now be the communities constructed in the first phase [62, 63], where the weights of the links between the new nodes are given by the sum of the weight of the links between nodes in the corresponding two communities [63]. The edges binding the nodes are all the external/outgoing connections from nodes of the different communities created in the first phase, whose weight is the sum of the weight of all the edges of the nodes of the original first phase community merged together [64]. After this newly weighted network is produced, the first phase is reapplied to it and iterated [63]. Iterations keep occurring until there are no more changes and a local maxima of the modularity is attained [63].

This algorithm has several favourable qualities, like its accuracy, so much that when tested on ad hoc modular networks, showed better results than other much slower community detection methods [63]. It is also faster than other methods, given its linear complexity on typical and sparse data [63], as most of the running time is concentrated on the first iterations of the two phases where the number of communities gradually decreases [63]. Nonetheless, it still has its negative traits, like the fact that it might find arbitrarily badly or weakly connected communities [62], since one node moved from one community to the neighbouring one may have acted as a bridge/link between nodes in its old community, disconnecting it [62]. Another disadvantage of the Louvain method is the resolution limit of the modularity quality metric, which may cause smaller communities to be clustered into larger communities, therefore "hiding" them [62], one problem that the original author of the method claims has been circumvented [63] thanks to its the intrinsic multi-level nature [63]. Storage of the networks in main memory may also become an issue [63].

Even though clustering is the method to use when visualization and classification of structure in large data sets is wanted/needed, most of the times these data sets are so high-dimensional that clustering becomes very difficult, hence the introduction of dimensionality reduction algorithms.

Dimensionality reduction is a pre-processing step in machine learning where the dimensionality, meaning the number of features in the dataset [65], is reduced to remove redundant features, noisy and/or irrelevant data [65, 66], mitigating the "curse of dimensionality", which is the group of phenomena that occurs in high-dimensional data sets that otherwise wouldn't happen in lower-dimensional ones [65, 67]. This allows for the improvement of the learning feature accuracy and the reduction of training time in the machine learning process [65]. However, this dimensionality reduction cannot be done in any type of way: the reduced representation of the data should account for the intrinsic dimensionality of the data, or the minimum number of parameters needed to account for the observed properties of the data [67].

This pre-processing step can be achieved by using two different methods, either in combination or isolation, to improve the calculated accuracy and precision of machine learning methods [65]: feature selection, where essential features are selected from the initial data set, and/or feature extraction, in which there is creation of new features from the existing ones in the data set [65]. Features are extracted as either linear (when the data forms a straight line when plotted) or nonlinear (when the data forms, in some way, a curved line when plotted) functions of the original set of features [68]. This selection/extraction is done by classifying the features as suitable, unnecessary or repeated [65].

In spite of improving the performance and accuracy of the clustering or machine learning algorithms [65], decreasing the data storage space and computational time and cost as the dimensions are reduced [60, 65] and allowing for a better visualization and examination of patterns in high-dimensional data sets [65], the process can cause the loss of information, damaging the interpretability of the results and possibly distorting the real clusters [60], especially when using feature selection, as some features are excluded from the data set [65].

Dimensionality reduction falls into two categories: linear techniques, with algorithms such as PCA (principal component analysis), LDA (linear discriminant analysis) and factor analysis [67], which assume that the data lies on or near a linear subspace of lower dimensionality [67], using linear functions for this [69, 70], and nonlinear (or manifold learning) techniques, like t-SNE (t-distributed stochastic neighbour embedding), MDS (multidimensional scaling) and Isomap [66], which do not rely on the linearity assumption, as more complex embeddings, or the low-dimensional space in which high-dimensional data can be translated, of the data in the high-dimensional space can be identified [67], using nonlinear functions for their purpose [69, 70].

One example of nonlinear dimensionality reduction algorithm is the novel UMAP, or Uniform Manifold Approximation and Projection, designed by McInnes et al. [66]. From a computational perspective, UMAP can be described as a weighted graph [66] where the Nearest-

Neighbour-Descent algorithm is used to identify a pre-set number of nearest neighbours for each individual object in the data set, and the distances to each of the neighbours is represented in the graph by the weights, where smaller the distance between neighbours, the heavier the weight is, and vice versa [66, 71]. On the other hand, t-SNE uses Gaussian probability functions for this [67]. The goal of UMAP is thus to discover a low dimensional manifold in which the high dimensional data neighbours can be uniformly distributed and represented, in this way maintaining the topology, or shape, of the high dimensional data set [66, 71]. UMAP is based on the principle of accurately representing local distance rather than preserve the global, large-distance structure of the data set [66, 71]. This allows for UMAP to "stretch-out" data-dense regions in the representation, which can help in reducing overcrowding of the low-dimensional representation, for better visualization, despite potentially turning interpretation of the calculated neighbour distances more challenging [71].

The UMAP algorithm implementation can be divided into two phases: the graphic construction, where a fuzzy weighted k-neighbour graph is constructed, in which the weight of the edges represents the local distance between each neighbour in the nodes, and the graphic layout, in which a low dimensional layout of the previously constructed graph is computed, by the definition of an objective function that preserves desired characteristics of the k-neighbour graph [66].

This novel manifold learning technique is competitive against t-SNE, which is the current state-of-the-art for dimensionality reduction and visualization quality [66]. It does this by preserving more of the global structure, despite it not being its main goal, due to the Laplacian Eigenmaps initialization, whereas t-SNE, for the most part, uses random initialization [66, 72], having a faster run time performance, scaling to significantly larger data sets, and having no computational restrictions on embedding dimension, which makes it viable as a dimensionality reduction technique for machine learning [66]. On the other hand, UMAP lacks, similarly to other non-linear dimensionality reduction techniques, the strong interpretability of, for example, PCA, as the dimensions of the embedding have no particular meaning, whereas in PCA the dimensions represent the directions of greatest variance in the original data set [66]. Moreover, by the assumption that UMAP makes that there is manifold structure in the data set, it can tend to find it in noisy data – however, as more data is sampled, the amount of structure that the algorithm might find in the noise decreases, consequently becoming more robust [66]. Therefore, caution should be taken about this particular issue when working with small data sets [66]. Lastly, as UMAP does not penalize uniform stretching that happens in data-dense regions, its representations might contain arbitrarily small distances between points, which might, or might not be, an actual faithful representation of the original data [71].

3.2 PHYLOGENY

The large quantities of data generated by, for example, NGS (next-generation sequencing), not only give way to the implementation of clustering algorithms, as explained in chapter 3.1, but also to the further development of the phylogenomics and phylogenetics areas [73, 74, 75], with the goal of extracting all the possible information from these large data sets [75].

Phylogenetic inference is a basic bioinformatics problem [75] that aims to estimate the evolutionary history, through ancestor-descendant relationships, of a taxon, species, gene or even whole genome, based on a tree representation [76, 77, 78]. This is particularly important in understanding the spread of contagious diseases, such as the transmission of HIV-1 and the evolution of the SARS (severe acute respiratory syndrome) - associated coronavirus [79].

A phylogenetic tree is, in computational and mathematical terms, an acyclic connected graph (V, E), where V represents the nodes and E represents the edges connecting these nodes [74, 80]. The external or terminal nodes, or leaves, represent the extant sequences, and are

connected only through a single edge [74]. The internal nodes are the common ancestors of all the sequences in a specific subtree, thus are connected by more than one edge [74]. The edges are called branches and represent the hypothetical ancestor-descendant relationships between the nodes [74]. Their length denotes the evolutionary time that occurred from one ancestor to its descendant [74], as units of time, number of generations or amount of change [77].

The tree can be constructed using a large plethora of characters, such as DNA or protein sequences, base composition and G+C content, phenotypical traits or even biogeographic areas [81]. The commonly used character is DNA or protein sequences, the first usually in the form of multiple sequence alignment (MSA) that contain insertions and deletions (indels) [75, 79, 82]. They are denoted as matrices of letters, where the rows represent the sequences and the columns are the positions of said genomic or genetic sequences [82], which will give the state of a certain character at a certain position [74]. Tree construction consists on the grouping of taxa based on a decreasing number of homologies, that is, similar characters that reflect the state of the common ancestor, to form progressively wider clades with older ancestry [76].

Trees can be built under the assumption that the branch lengths are proportional to time, meaning that the mutation rate across these branches is equal – they would follow the molecular clock hypothesis, which states that the rate of evolution, and therefore the mutation rate, is approximately constant over time [74, 76, 81]. This would ultimately allow researchers to infer the point in time when two lineages diverged [76]. However, there has been some research that has found that there is variation among lineages, implying that the molecular clock hypothesis does not work in many cases [81]. Today's phylogenetic inference techniques consider heterogeneity among lineages [76], such as the fact that transitions, meaning the substitution within purines (A and G) or within pyrimidines (C and T), are twice more likely to occur than transversions, which are the substitutions of a purine by a pyrimidine or vice versa [76].

A phylogenetic tree can also be rooted, when the location of the common ancestor of all the taxonomic units in the tree is known, or unrooted, when there is no common ancestor identified [74, 79, 83]. The root provides directionality and path of evolution within the tree [83]. There are two main methods of rooting trees: the most used is the outgroup method, which, by assuming that there are one or more taxa that are divergent from the remaining taxa, making it the ingroup, branches this ingroup with an outgroup, turning this link into the starting point of the tree [83]; and the other method is midpoint rooting, which places the root halfway between the two most divergent taxa [83] – however, this only works in trees that follow the molecular clock hypothesis, as this method is solely dependent on the branch length of the tree [83].

Phylogenetic inference, or phylogenetic reconstruction, methods can be divided into distance- or character-based methods.

Distance-based methods are semiparametric methods that convert the original data matrix into a matrix of pairwise distances between the taxa, based on models on nucleotide substitution [78], which are then used to build the phylogeny – methodology akin to hierarchical clustering is then used to build the tree, using dendrograms [74, 79, 82]. The two most used distance-based phylogenetic inference methods are Neighbour-Joining and UPGMA (unweighted pair-group method using an arithmetic average).

Neighbour-Joining (NJ), an algorithm developed by Saitou and Nei in 1987 [84], produces a unique final unrooted tree under the principle of minimum evolution, which tries to minimize branch lengths by minimizing the distance between them, despite it not guaranteeing a minimum evolution tree, meaning that the tree with the smallest sum of branch length estimates might not be the true tree, making this algorithm a heuristic one [84]. It does this by finding the pairs of taxa connected through a single interior node in an unrooted, bifurcating tree, or neighbours, as the name of the algorithm entails, that minimizes the total branch length at each stage of clustering of the taxa, starting with a star-like tree [84]. The first step is to create the pairwise distance matrix that gives name to the type of method NJ is. Then, the smallest distance between two taxa in the distance matrix must be chosen, which represents the two closest

neighbours/the least distant pair of nodes, and these are joined to create a new node, which is connected to the central ancestral node. The distance between this new node and the remaining taxa is then calculated. The new node is now fixed and considered as a single taxon, and subsequent joining steps, given this, will then be repeated until the fully resolved tree is achieved, making the NJ algorithm agglomerative [85].

UPGMA, introduced in 1958 by Sokal and Michener, is an average linkage clustering algorithm that works in a bottom-up approach, meaning it's also agglomerative, just like the NJ algorithm, by constructing a rooted tree that is as close as possible to the input distance matrix in the least squares sense, by also minimizing the minimum evolution score [86]. After the pairwise distance matrix is constructed, the two taxa with the smallest distance between one another are found and then clustered together, replacing the two original taxa in the data set, where the distance between the cluster and each of its leaves is half the distance between said leaves. The process is repeated, whilst the distance matrix shrinks and the phylogenetic tree grows, until there is a fully constructed tree. The distance between the newly formed cluster and subsequent formed clusters in the data set is the arithmetic average, hence the algorithm name, of the distances between the members, or leaves, of each one of the clusters. The UPGMA algorithm can only work in ultrametric distances, that is, distances which follow the molecular clock hypothesis - thus, all the leaves will have the same distance to the tree root and also the distance between the newly formed cluster and the all the other taxa [86].

When comparing the NJ and UPGMA algorithms, the NJ algorithm has the advantages of being faster, making it appropriate for large data sets, and also allows for lineages with largely different branch lengths, as it works with additive distances, which UPGMA cannot do, as it works with a molecular clock hypothesis. However, it only allows for the construction of one tree and it's strongly dependent on the evolution method utilized. By working with additive distances, it can potentially assign negative branch lengths, which can usually be interpreted as an estimate for the substitutions, but with negative ones, there end up being difficulties in interpreting the results [87]. On the other hand, UPGMA is a much simpler algorithm and easy to compute by hand, and as rooted trees are generated, they are easier to interpret and analyse.

Character-based methods, on the other hand, use probabilistic approaches to test for phylogenetic hypotheses and for the statistical inference of phylogenies [79, 81], by comparing all taxa simultaneously considering one character/site at a time [79], rather than relying on the pairwise distances between them. At the end, the goal is to find the tree with the best score/highest probability of being the correct tree, meaning the tree that presents the smallest number of accumulated changes given the initial sequence alignment, which is the starting point of this type of methods [78, 79]. Unlike distance-based methods, which can only produce one tree at a time, character-based methods produce many optimal and sub-optimal trees that can be compared to the hypothesis, making the distance-based methods less computationally intense than the character-based methods [78, 79]. The three most used algorithms in this branch of phylogenetic inference are maximum-parsimony, maximum likelihood and Bayesian inference.

Maximum parsimony (MP) seeks to find the tree that implies the fewest mutations/substitutions along its branches to explain the data at hand [74, 81]. The length of the tree, which is the sum of the length of all branches for a specific character in the data matrix, and then summing all characters in the same matrix, represents the amount of accumulated change [74, 81]. Despite the fact that it implies that character changes across the branches occur independently, a weight can be added to allow for differential weighting of the characters, like, for example, applying a lower weight cost to transitions than to transversions, or applying different weight costs to synonymous vs nonsynonymous mutations [74]. In regular conditions, the cost function of each character change from a parent to a child along an edge, which is classified as a substitution cost, is weighted equally, i.e. if the character state changes on two sides of a branch, the cost function increases the tree score by 1, and if it doesn't, the score doesn't

increase, regardless of the state the character is [74, 88]. This way, the most parsimonious tree is the one that minimizes the sum of weights on the tree edges, meaning the overall tree score [88].

There are two different maximum parsimony algorithms, the Fitch algorithm and the Sankoff algorithm, which are both dynamic programming algorithms, as they reduce the overall problem of computing the best scoring tree into smaller subproblems, which makes the process much more easy and efficient [74].

The Fitch algorithm takes as input data such as multiple sequence alignment and starts by labelling the leaves with its correspondent observed character state. Then, a depth-first traversal of the tree is carried out, as each internal node is labelled given its two immediate descendants' character states, in which, if the intersection of these two nodes is nonempty, the node is labelled by the intersection, meaning that the character states of the two descendants is the same; otherwise, the intersection is empty, the character states of the two descendants are different and the node is labelled by the union of said states, which increases the length of the tree by 1 [74]. Once all trees have been evaluated, the shortest tree is chosen as the most parsimonious tree. Given that this algorithm implies a model of symmetrical mutations, such that the cost matrix's diagonal elements are all 0 and the off-diagonal elements are all 1 [88], it can only derive unrooted trees. However, since changes in either direction are equally weighted, a tree can be arbitrarily rooted at the most convenient location [74].

Sankoff's algorithm works by assigning a particular function to each node of the tree, which computes the minimum costs of each possible state if a subtree was rooted in that node [74]. If a certain tree leave has the correct state for a specific character, then the function is equalled to 0, otherwise it will be equalled to ∞ [74]. The algorithm traverses the nodes of the tree in a post-order fashion, and then the best possible tree, that is, the one with the minimum length that corresponds to the minimum value of the previously assigned function, is found by backtracking from the root to the leaves by traversing the tree nodes in a pre-order manner [74, 88].

Maximum likelihood (ML) is one of the most accurate and most used algorithms for phylogenetic tree reconstruction [76, 80]. This is a probabilistic algorithm that takes into account the phylogenetic tree, which represents the hypothesis of the evolutionary history (topology) and amount of accumulated change (branch lengths), as well as a sequence evolution model, which is generally a Markov model [75, 78, 81].

These Markov substitution models describe the evolution of certain characters along the tree, given that the probability of a certain nucleotide changing to another one does not depend on the ancestral states, but rather on the actual nucleotide and the time over which the change has occurred [74, 81]. However, these Markov models do not take into account insertions and deletions (indel), either excluding sites that contain these indels, or ignoring the branches that lead to nodes with indels [74, 75]. The two most common Markov substitution models are the Jukes-Cantor (JC) model, which assumes that, regardless of the nucleotide, all rates of nucleotide change are the same, as all nucleotides are considered to be equally frequent [74]; and the Kimura two-parameter (K2P) model, which gives a different rate of change for transitions and transversions, with a bigger ate given to transitions over transversions [74].

With this in mind, the ML algorithm seeks the tree that maximizes the likelihood/probability of observing some taxa at the leaves of said tree given a hypothesis on how the data was generated and the evolutionary model utilized [80]. Once the evolutionary model is chosen, the likelihood is calculated by summing the probabilities of all possible scenarios by which the taxa at the leaves of the tree could have evolved, depending on the tree topology as well as the branch lengths [74]. A dynamic programming approach called pruning was established by Felsenstein in 1982, which allows the likelihood calculation to be completed without having to iterate through all the possible scenarios [74].

Lastly, Bayesian inference (BI), unlike the classical, frequentist MP and ML algorithms, which assume a null hypothesis by which alternative ones can only be rejected, assigns

probabilities to all the tested hypotheses, assuming, thus, that they are true [81]. Bayesian inference is based on a posterior probability of the tree, basically the probability that the tree is correct, given the data, meaning the prior distribution [78]. This posterior probability is computed by considering both the prior distribution associated with the data and a likelihood function of the observed data, given a particular model of evolution [74], giving way to the Bayes Theorem [74, 81], shown in the following equation:

$$P(Tree|Data) = \frac{P(Data|Tree) \times P(Tree)}{P(Data)}$$

where P (Data | Tree) is the likelihood function and P (Tree) is the prior distribution [74]. Selecting appropriate prior distribution for model parameters is essential [81]. When it comes to the posterior distribution/probability, P (Data), it cannot be easily computed [74], which means that BI relies on Markov Chain Monte Carlo (MCMC) methods to approximate this posterior distribution and account for phylogenetic uncertainty [74, 81]. Once the posterior distribution is estimated, the estimate of the phylogeny is considered to be the one with the largest posterior probability [74].

Maximum parsimony can perform poorly in situations with long-branch attraction, meaning in instances where the branch lengths are much longer relative to others, as it ignores them, i.e. it gives equal probability of change on all branches [74, 79, 81]. In contrast, maximum likelihood makes uses of the branch length information [81]. Thus, maximum parsimony becomes less consistent and reliant than maximum likelihood [74]. However, if these long branches are subdivided, inconsistent estimation and, consequently, possibly incorrect trees, maximum parsimony becomes faster than maximum likelihood [74]. Moreover, maximum parsimony assumes a relatively slow and symmetric rate of trait evolution, unlike maximum likelihood, which can estimate asymmetric rates of trait evolution [81].

Moreover, whereas maximum likelihood methods are not feasible for more than 50 taxa data sets, as Bayesian inference methods use a faster search strategy through the use of MCMC, it can be used on data sets of several hundred taxa in order to find the most probable tree, its branch lengths and support [78].

Optimality criteria such as the MP and ML algorithms produce a wide number of possible trees, but this vast group of phylogenies still needs to be efficiently searched in order to find the most probable tree. There are two types of methods to solve this problem: heuristic and stochastic search methods.

Heuristic search methods encompass methods that divide the general problem into smaller problems to make the search process much easier and more efficient.

The branch-and-bound method works by sequentially adding branches to a starting three-branch tree and determining a certain bound value (which would be an upper bound if working with MP trees or lower bound if working with ML trees) – considering we're working with a MP tree, if the value of the parsimony criterion exceeds the bound value, then all the trees that are obtained from adding branches to the initial tree are not considered; otherwise, then each possible tree after adding branches is considered and their parsimony criterion is compared to the bound value [74]. By doing the tree search this way, large portions of the data set are never considered, since it's known they won't contain the optimal tree [74]. Despite the fact that this method guarantees a phylogenetic tree that optimizes the criterion of interest, it only works in small data sets (up to 15 taxa), making it not ideal otherwise [74].

The stepwise addition and branch swapping method begins by constructing an initial tree and then sequentially adding branches to the current working tree to the branch that has the best value of optimality criterion – in the end, a complete tree containing all the taxa in the data set is achieved [74]. However, a final tree is not necessarily optimal, as different orders of taxa addition to the working tree give different final trees. Thus, the branch swapping technique is used to

search for the optimal tree, by rearranging portions of the final tree using a specific strategy and then evaluating the optimality criterion of this new tree [74].

There are three main rearranging methodologies that differ in how localized their rearrangements of the phylogeny are. The first one is nearest neighbour interchange (NNI), which is the most localized and least rigorous method, starts by choosing an internal branch of the tree and removing it along with the total of additional four branches that were connecting this internal branch to the nodes, in total of five branches removed. This allows for four subtrees, with three new alternative topologies that emerge from the reconnection of the nodes, with two that are different from the original tree topology [74]. This way, from an unrooted tree, which has n - 3 internal branches, we reach a possible 2(n - 3) NNI rearrangements [74]. The second one is called subtree pruning and regrafting, which works by removing either an internal or external branch, and its associated subtree from the tree, and then adding it again to one of the remaining tree branches [74]. This equals to a total of 2n – 8 possible rearrangements with selection of one of n – 3 internal branches, or 2n – 6 possible rearrangements if an external branch from a selection of n branches was chosen – thus, there are a total of 4(n – 2)(n – 3) possible rearrangements, although not all of them are distinct – nevertheless, subtree pruning and regrafting still leads to a wider search of tree space than NNI [74]. The third and final method is tree bisection and reconnection, which is the most global strategy [74]. It entails the removal of an internal branch and the resulting two subtrees are then reconnected in various ways – thus, if the resulting subtrees are of n_1 and n_2 taxa, there are a total for $(2n_1 – 3)(2n_2 – 3)$ possible tree reconnection rearrangements [74].

The central problem of heuristic search methods is that they do not guarantee to find the global optimum, usually only reaching a local optimum, especially in data sets of larger magnitude [74]. This way, stochastic search methods present an alternative to this problem.

One example of a stochastic search method is the simulated annealing method, which functions by making perturbations to an initial tree, possibly using one of the branch swapping methods explained above, and the value of the optimality criterion of interest is calculated for the perturbed tree – if the criterion value of the perturbed tree is higher than of the initial tree, then it becomes the current tree and the next one to suffer perturbations; otherwise, the decision of whether or not the perturbed tree will replace the initial tree is made in a probabilistic fashion, which depends on the actual criterion value and the number of iterations already performed. As the algorithm proceeds, the probability of accepting trees with a worse criterion value decreases according to a cooling schedule. That being so, if the algorithm is performed long enough and the cooling schedule satisfies certain conditions, it will converge onto the global solution [74].

Despite the advantage of theoretically being able to reach a global optimum, stochastic search methods have the disadvantage of setting several parameters that determine how the search will proceed – in the case of simulated annealing, it's the cooling schedule – which turns the systematic study of the properties of the algorithm more difficult, as well as the fact that, in practice, the algorithm's properties do not work, as the run time for any given problem cannot be infinite [74].

After the optimum tree is chosen by one of the many methods explain above, the estimated tree must have a desirable measure of confidence/internal support [74, 78]. That's where probabilistic/resampling methods come into play. They allow for the testing of appropriateness of the different trait evolutionary models and different hypotheses concerning the mode of evolution of the trait itself [81].

One of these methods is called bootstrapping, and, after it was introduced by Efron, it was proposed to be used for phylogenetic purposes by Felsenstein in 1983 [89]. It works by assessing repeatability, meaning the probability that another such taxa shares the same clade with the original taxa in the tree [82]. It can be divided into nonparametric and parametric bootstrapping.

Nonparametric bootstrapping requires, for each bootstrap simulation step, the columns, or the characters, of the data matrix to be chosen at random with replacement until the same number of characters as the original data matrix is achieved, and thus a new, replicate data matrix

is produced [74, 82, 89]. After the new data matrix is set, a phylogenetic search is applied to it, including both the optimality criterion and the search method, with an estimated phylogeny being obtained, with which a consensus tree of all the replicate data sets' phylogenies being constructed [74, 89]. The number of replications falls between 100 and 1,000 according to the size of the data set, the specifications of the analysis, with 100 being used for more detailed searches, and 1,000 for "faster" methods that conduct little to no branch swapping per replicate [74, 78]. Thus, the result of bootstrapping can only be as good as the data and the tree inference algorithm [89].

Parametric bootstrapping is pretty similar to its nonparametric counterpart, but instead of resampling the original data, it uses a model of sequence evolution to simulate new, independent data sets from the original one [74, 89].

The result of bootstrapping is given by what is called bootstrap value, which is a measure of phylogenetic accuracy for a certain node or clade [78], as it is the proportion/percentage of replicates that show a certain clade/node as appearing in the tree [74, 89]. What is considered to be the bootstrap value for strong tree support is not yet agreed among all researchers: Felsenstein himself considered the value of 95% or greater as representing statistically significant support [89], whereas other researchers consider the value of 70% or greater as an indicator of strong support [76, 78]. However, it should be noted that bootstrapping does not measure the "truth", but only a confidence interval in which it contains not the "true" tree, but instead an estimated one [89].

Nevertheless, bootstrapping only provides confidence to single nodes or clades rather than a joint confidence statement of the entire tree, which means that, even if the single confidence values are all high, the joint confidence value might not be [89]. Moreover, bootstrapping may present a "multiple test" problem if interpreted in a strict statistical manner, as, by chance, one in twenty clades might show a 95% bootstrap value [89]. Furthermore, bootstrap values may be affected by the total number of characters in the data set, as its support for a certain clade might decrease with the addition of invariant characters, that are compatible, but not informative to the clade in question, which may be selected for the replicate instead of an actual relevant character [89].

Another way for the measure of confidence in a tree is done within the Bayesian framework through the use of the Bayes Factor, which is defined as the ratio between marginal likelihoods, or the ratio between the posterior probabilities and the prior probabilities, which in turn determines how much the prior probabilities are changed by the data when computing the posterior probabilities [81]. This will give support values for each node in the tree, which are the frequency it appears in the posterior samples [74], that won't be considered reliable or significant if they are less than 90% or even 95% [76].

3.3 PROTEIN STRUCTURE PREDICTION

Methods like NGS produce large-scale data, which is causing a gap between the number of sequences and the respective protein structures. There are approximately 200 million unique protein sequences in UniProt, which is the universal protein knowledgebase, but only 150.000 solved protein structures in the Protein Data Bank (PDB) [90, 91, 92]. Despite the determination of protein structure moving slower than their sequencing, there is a need to bridge this protein sequence-structure gap, and this is currently being worked on by computational tools [90, 91, 92].

However, determining the structure of proteins is not only important for the decrease of the sequence-structure gap, but also to help understand the protein's function [90, 93, 94] and because they have a crucial role in potential drug design, thus allowing for the understanding of diseases and to find potential new treatments [91, 94]. This is being achieved mainly by the use of computational methods due to the difficulty of experimental techniques using crystallization [94]. These methods elucidate the protein's structure by using their respective sequence [91, 94], given that early computational studies showed that proteins tend to conserve their structure, rather

than their sequence, more efficiently – thus, sequence similarity suggests structural similarity, enabling the prediction of protein's structure by using their sequence [94].

Proteins have the ability to fold and bind their amino acids into different structures and conformations, hence their characterization by their primary, secondary, tertiary and quaternary structures [91], as shown in Figure 3.1. The primary structure, which is the lowest energized one [91], represents the bound together amino acid residues that form long peptide chains [91]. The folding of these long peptide chains results in hydrogen bonds being formed between the atoms of the amino acids, which causes them to fold into repeating, regular patterns called coils or loops, such as alpha helices, beta sheets and beta turns - these form the secondary structure of the protein [91]. Due to side chain interactions, proteins can fold into a large number of conformations, thus resulting in the protein's overall three-dimensional (3D) structure, which is the protein's tertiary structure [91]. Finally, the protein's quaternary structure represents a complex protein structure consisting of various peptide chains [91].

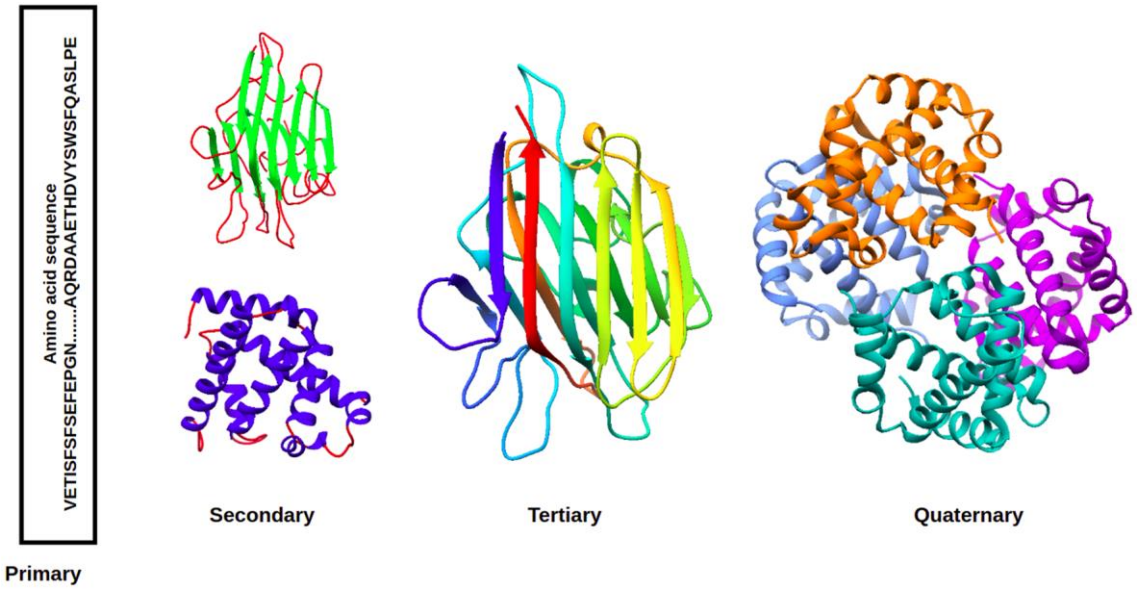


Figure 3.1 - Levels of protein structure, respectively primary, secondary, tertiary and quaternary structures [91]

To aid in the protein structure prediction method's progress, a biannual event named CASP (Critical Assessment of Techniques for Protein Structure Prediction) started in 1994 [90, 91, 93], where the participants have to build models, within a stipulated time, that give the structure of target sequences for which their structure was previously unknown; after, the predicted structures are compared to their experimentally determined counterparts [91]. In the last CASP experiments, there have been great advancements, particularly in template-based modelling (TBM), template-free modelling (TFM) and estimates of 3D model accuracy (EMA) [92].

Hereby, protein structure prediction's methods, which work by the use of heuristic algorithms [91], can be divided in template-based and template-free methods, based on the availability of a template, which is one protein with a known structure, in protein databases, for the determination of the structure of the target protein sequence, for which the structure is unknown [91, 94].

Template-based methods use a template structure, which is homologous with the target protein or has similar folds, to build a 3D model of the target sequence, for which they use homology/comparative modelling and threading/fold recognition techniques [91, 94]. When neither a significant evolutionary relationship or sequence similarity is found, threading

techniques are applied, as even proteins without any evolutionary relationship show structural similarity [91].

If there is conservation of the 3D structure between a template and a target protein sequences, the prediction of the structure of the target sequence given the structural features of the template sequence is possible, whose process is known as comparative modelling [94]. This process can be divided in four different steps.

The first step is fold assignment and template selection, where, by using tools like BLAST and FASTA, the target sequence undergoes a pairwise comparison to a set of potential template protein sequences with known structure in databases like PDB to find to most suitable template [91, 94]. Profile-based algorithms, which apply an alignment process that includes information from related proteins, work better than pairwise alignment-based algorithms, specifically ones that use hidden Markov models (HMMs), like SAM, HMMER and HHpred [94]. If more than one suitable template is found, in order to increase the probability of a higher quality prediction, the one with the highest sequence identity to the target sequence is chosen [94].

Once a template is chosen, it needs to be aligned with the target sequence [91, 94], to correctly correspond the residues of the target with the template's one [94]. First, standard sequence alignment methods, like Needleman-Wunsch or Smith-Waterman, are used, which calculate an alignment score through substitution scoring matrices like BLOSUM or PAM [94] – this should give a relatively high sequence identity between the sequences. However, if this identity is less than 40%, methods like multiple sequence alignment need to be implemented, as they add more information about structurally conserved regions and thus can detect more distantly related template sequences [91, 94]. From the target-template sequence alignment, one can distinguish the well aligned regions with the template and regions with no alignments – from these aligned regions, the 3D model can be build [91], which is the next step.

Model building is used to explicitly predict the target structure atomic coordinates through the equivalent residues obtained in the sequence alignment [94], and its methods are based on three different approaches: rigid body assembly, where small rigid bodies obtained from structurally aligned proteins are used, in which, firstly, the atomic coordinates of the conserved regions are used to build the main core of the protein, then loops and folds are built in, given peptide fragments that fit the conformation of the protein's core after a database search, and finally, the side chains, which are the branches that are added to the backbone of a protein to complete its structure [91], are predicted [94]; segment matching, which is based on the fact that approximately 100 hexamers can account for 76% of conformational space, and thus, can be used to search for similar fragments in databases and use those, as well as, similarly to the rigid body assembly approach, the atomic coordinates of the Cα atoms from their conserved residues, to construct both the core of the protein, as well as the side-chains [94, 95]; and lastly spatial restraints satisfaction, that assumes that structural features of conserved regions between the target and template sequences, like distances and angles between them, are similar, with which homology-based restraints supplemented with stereo chemical restraints are generated – the number of these restraints' violations is then minimized to search for the global low energy protein model [94]. Spatial restraints satisfaction modelling methods are the most flexible, as many different types of restraints can be incorporated, which can be derived from various template conformations or experimental data sources, like nuclear magnetic resonance (NMR) and fluorescence spectroscopy [94]. After building the model, it needs refinement, which aims to improve the quality of the predicted structure [91].

The final step in comparative modelling is to evaluate the structure predicted in the protein model, which can focus on assessing both the target-template alignment and the geometry and stereochemistry of the model on the model itself or on specific segments [94]. There are four main techniques with which the estimation of a predicted protein's structure quality can be performed. Root Mean Square Deviation (RMSD) calculates the average distance between equivalent atom pairs in the superposition of the target and template structures, and usually only

backbone residues are considered for its calculation [91]. A lower RMSD means that the atoms under consideration are close together or belong in the same cluster, and if the RMSD is significant, it is more sensitive to local structure deviation rather than global topology [91]. Global Distance Test Total Score (GDT_TS), which is the most widely used scoring method for the overall structure evaluation, gives the count of amino acid pairs, after both target and template structures are superposed, that fall within a certain distance cut-off range of their position, which gives a score average of results at 1, 2, 4 and 8 Å, which typically range from 0 to 100, with a higher score representing a perfectly constructed conformation of the target's core [91]. As it is insensitive to outlier regions, which RMSD is, GDT_TS is presented as more accurate than RMSD [91]. TM-scores also find the similarity between protein structures, where in the interval (0,1) of possible scores, if it is below 0.17, it indicates two unrelated proteins, and when the score is above 0.5, it means the target and template proteins have better similarity between each other [91]. Unlike the previous two methods, TM-scores uses a distance scale variable in its calculations, D_0, that normalizes distances and thus makes TM-scores not depend on protein size [91]. Lastly, Z-score is used to calculate, unlike all the other methods, the difference between the native fold of a protein and a group of misfolded structures, meaning it compares the target-template pair score with the distribution of scores for all the other possible target-template pairs, to measure the ability of various techniques to correctly identify native structure from all the incorrect ones – a high Z-score will indicate a high quality model [91, 94].

Comparative modelling is used when sequence identity is above 25% - otherwise, threading techniques are used, given that, in this case, as homologous sequences templates would be hard to find, more sensitive structural relationships need to be found instead, particularly through the use of fold recognition methods, because protein sequences with both low sequence identities and similar fold are rarely observed [91]. Hence, threading, or fold-recognition, searches for a template by performing several sequence-structure comparisons with various protein template alternatives to select the one with the most similar fold, which is represented as physiochemical properties, secondary structure and solvent accessibility, to the one that was previously set as a target [84]. Then, the protein that presented the similar fold to the target fold will be selected as the template protein structure [91].

On the other hand, template-free methods are used when TBM fail and don't find any suitable templates in any database [91], as they don't need any template structure, which gives them broad applicability [94]. Instead, they can either use physiochemical principles, being classified as ab initio methods, or knowledge from fragments of known proteins or restraints from threading results, called de novo or fragment-based methods [94] - this allows for the possible conformation space to be reduced [91, 94]. The physics-based approach uses the principles of thermodynamics to produce the native protein structures, as well as give insight into their folding mechanisms, through the application of MD, MC (molecular dynamics and Monte Carlo) simulations, genetic algorithms and simulated annealing techniques [91]. On the other hand, knowledge-based techniques extract fragments and statistical information from known PDB 3D protein structures in order to create a model for the target protein [91].

Generally, TBM are more highly accurate, faster, more reliable and can perform predictions that show structures very similar to the native one, whereas TFM have often lower quality and is often more computationally expensive [94]. However, TBM application is limited to the presence of an available template structure of a homologous protein, and template-free physics-based methods are the only ones that allow for a view into the actual folding mechanisms of the protein [91, 94].

One relatively novel TBM method is IntFOLD-TS (Integrated Protein Structure and Function Prediction Server), which was initially developed by McGuffin and Roche in 2011, with the main goal being generating high quality 3D models of proteins, as well as annotating them with accurately predicted per-residue errors [96], specifically through accurate Model Quality Assessment (MQA), given that improvements in Estimates of Model Accuracy (EMA) provide

higher quality protein models [97]. The current IntFOLD server provides researchers who use it with six different component methods: IntFOLD-TS for tertiary structure prediction, ModFOLD for Accuracy Self-Estimate (ASE) scoring, ReFOLD for 3D model refinement, DISOclust for disorder prediction, DomFOLD for structural domain prediction and FunFOLD for ligand binding site prediction [92]. These tools have been extremely helpful in several cases, such as the modelling of novel *Drosophila melanogaster* genome proteins, structurally and functionally annotating the proteome of barley powdery mildew (*Blumeria graminis* f. sp. *hordei*), understanding the effect of the missense mutations associated with dermatosparaxis, revealing new interactions and mechanisms for the regulation of mammalian GCKIII kinases and explaining the evolutionary resurrection of flagellar motility in *Pseudomonas fluorescens* [92].

The server only requires the target protein sequence as input, with the following pipeline: the modelling stages start with 14 different target-template alignment methods, which include SP3, SPARKS2, HHsearch, COMA, SPARKSC, CNFsearch and eight other threading methods that are integrated in the LOMETS package (PPA, dPPA, dPPA2, sPPA, MUSTER, wPPA, wdPPA, and wMUSTER), which results in 10 alignments each, thus 140 total target-template alignments. These methods are applied to MODELLER to build the models, which does it by generating a set of restraints for the target sequence based on the template's aligned regions, restraints that are represented by a molecular probability density function – this probability function must be optimised in order to get the best model with the least violation of restraints [91]. These alignments, plus ones contributed by the HHpred and the template-free I-TASSER light methods, are then assessed using ASE scoring through the ModFOLD7_rank method, in order to minimise local errors in the final generated models. At last, the final models are pooled and then scored and ranked with the help of ModFOLD7_rank, with the best five models presented in descending order of global model quality [92, 97]. The ASE scores allow researchers to view the local model quality score as a temperature gradient that can be mapped onto the 3D structures, where blue indicates residues in the model that are predicted to be close to the native structure and red represents residues in the model that are far from the native structure [92].

Not only is the structure of a protein important to understand its cellular role and its functions, but also the location of its binding site and the respective ligand binding residues [98]. Thus, the IntFOLD server uses FunFOLD for this same goal.

FunFOLD uses a 3D model of the target protein and a list of templates used for the model building's PDB IDs as input, under the assumption that ligand containing PDB templates with the same folds as the target's 3D model must also have similar binding sites [98]. The FunFOLD algorithm works by superposing each of the relevant template structures onto the target model using the TM-align method, where if the TM-score is >= 0.4 (TM-scores between 0.4 and 0.6 were shown to demonstrate significantly related folds), then the superposition is saved. These superposition files are then used to correctly orientate the original PDB with bound ligands structures to the model. Ligands are then used to indicate binding pockets by assigning them to clusters via agglomerative hierarchical clustering methods that identify continuous masses of contacting ligands. These ligands are considered to be part of a cluster if as at least one of their atoms is in contact with the continuous mass given the criteria <= the Van der Waals radius of an atom plus 0.5 Å distance for contacts between ligands. This way, the cluster with the largest number of ligands will be considered the location for the most likely binding pocket. After this, in order to determine which residues are to be in contact with the predicted ligand, the distances between all atoms within the mass of ligands and all atoms in the 3D protein model are calculated, again given the <= the Van der Waals radius of an atom plus 0.5 Å distance criteria. Once the residues that are in contact with the ligand are determined, the residues that effectively bind to the predicted ligand need to be selected, which is done by a "residue voting", that says that a residue is to be included in the prediction if it has, at least, one contact with 2 or more ligands and, at least, 25% of the ligands in the cluster. Lastly, only the coordinates of the model and relevant ligands are saved and presented to the researcher [98].

Afterwards, FunFOLD uses a quality assessment tool to score the quality of the protein-ligand binding site residue predictions – FunFOLDQA [99]. It uses protein feature analysis to predict ligand binding residues prior to the experimental solution of the protein structures and their ligand interactions, by utilizing neural networks [99]. Upon the IntFOLD results, there are presented six different scores: BDTalign Score, which determines the distance between equivalent residues of the model binding site and each template binding site, meaning how well the model and the template structures fit together; Identity Score, which decides the binding residues that are "equivalent" in the model and template structures' binding sites and scores them according to their amino acid identity, hence the name; Rescaled BLOSUM62 Score, which scores residues that are equivalent in 3D space according to the BLOSUM62 scoring matrix; Equivalent Residue Ligand Distance Score, which scores the equivalent residues in relation to their distance from the bound ligand; Model Quality Score, calculated using ModFOLDclust2; and the Predicted BDT (Binding-Site Test) and MCC (Matthews Correlation Coefficient) scores, which both produce continuous scores relating to the distance between the predicted and observed residues [100] – BDT and MCC scores are strongly correlated to each other, but BDT score is less susceptible to the subjectivity of defining binding residues, as MCC score does not take into account the actual 3D placement of the predicted residues from the observed binding site [100]. For simplicity sake, only the Model Quality Score was used to determine how good the FunFOLD tool worked on predicting the binding residues, their locations and their respective ligands.

4 METHODS

4.1 SEQUENCE LOGOS

Before any sort of analysis was performed, all sequences were aligned using the AB873942 GenBank sequence, defined as "Human immunodeficiency virus 1 pol gene for pol protein, partial cds, isolate: KI-560" [101], as a reference for the pol gene, where the RT coding region is in. With this, they were inserted into the MAFFT program in the Galaxy-EU web platform.

MAFFT is a multiple sequence alignment (MSA) tool for unix-like operating systems, which is classified as a similarity-based method rather than an evolutionary-based method – despite this, evolutionary data is still useful, as the to be aligned sequences are generated from a common ancestor in the course of evolution [102]. It offers various MSA methods, such as L-INS-i and FFT-NS-2. The one used was FFT-NS-2, as it is fast and recommended for less than 30.000 sequences, whereas L-INS-i, however more accurate, only works for less than 200 sequences, and there were ~20.000 sequences to align. The default value of 1.53 was used as the gap opening penalty at group-to-group alignment. When it came to matrix selection, BLOSUM-62 was the chosen one. In the end, a FASTA file with the MSA was obtained.

The Galaxy web platform was started in 2005 and offers a combined, user-friendly interface that allows users without programming skills to run tools and workflows, making computation accessible. Additionally, Galaxy ensures reproducibility by automatically capturing information for each analysis step, so that any user can repeat and understand a complete analysis. Furthermore, Galaxy ensures transparency by enabling researchers to share and publish their Galaxy objects [103]. There are three main servers of Galaxy: the United States one (www.usegalaxy.org), the Australian one (www.usegalaxy.org.au) and the European one (www.usegalaxy.eu), this last one being the one used for these analyses.

After all sequences were aligned, sequence logos were constructed to determine how conserved the NNIBP residues' nucleotides were among the 22,838 sequences. Sequence logos are a graphical method that are presented as a display of patterns in a set of aligned sequences, where the sequence characters are stacked on top of each other for each position in the aligned sequences – the height of the letters representing each of the four nucleotides is proportional to its frequency in a particular sequence position, with the most common one being on top. With this, both the consensus sequence and the relative frequency of bases and the information content (that can be measured either in bits or probability) at every position are determined [104].

The construction of the sequence logos was done using R Statistical Software (v.4.3.1) [105], with the ggplot2 and ggseqlogo packages. The ggseqlogo package is built on the ggplot2 package and it offers native illustration of publication-ready DNA, RNA and protein sequence logos in a highly customizable fashion with features including multi-logo plots, qualitative and quantitative colour schemes, annotation of logos and integration with other plots [106]. The ggplot2 package on which the ggeqlogo is built on produces statistical, or data, graphics, whilst using a deep underlying grammar based on the Grammar of Graphics, unlike other graphics packages, which is composed of a set of independent components that can be composed in many different ways. This makes ggplot2 very powerful, because one is not limited to a set of pre-specified graphics, thus new graphics that are precisely tailored for a specific problem can be constructed [107].

The code used in this part of the dissertation is available in the following GitHub link: https://github.com/filipactavares/msc_dissertation/tree/main/Chapter_4.1.

4.2 CLUSTERING AND TEMPORAL DISTRIBUTION ANALYSES

For the purpose of determining DRMs of each sequence in the data set, the HIVdb algorithm by the HIV Drug Resistance Database of Stanford University was employed [108]. For simplicity sake, if the nucleotides of the sequence couldn't give certainty when it came to which mutation was in that position (for example, RAA could either be a lysine (K) or glutamic acid (E)), if the study was to check whether that sequence had the K101E mutation, for instance, then it was considered that it didn't have the K101E mutation. Afterwards, a manual verification of all mutations was performed.

In order to determine whether there were any patterns, partitions or groups that could be formed in the data set, a clustering analysis using the Louvain method, preceded by dimensionality reduction by the UMAP technique, was employed. This was done using the SCANPY package in Python (v.3.9.7) in Spyder 5.1.5. Spyder is a free and open-source scientific environment for Python, combining advanced analysis, debugging, editing, and profiling with data exploration.

SCANPY is a scalable toolkit for analysing single-cell gene expression data and other types of omics data, and it includes methods for pre-processing, visualization, clustering, among others [109].

SCANPY relies on the use of the ANNDATA class, which stores a data matrix efficiently with the most general annotations possible. All statistics and machine-learning tools extract information from a data matrix, which can be added to an ANNDATA object while leaving its structure unaffected [109].

Using Seaborn (v.3.9.7) [110], a Python data visualization library, we generated a heat map to reveal the mutations that distinguished the clusters of sequences.

As CD4+ cell count and viral load in the bloodstream are two of the most important factors that can determine the stage of progress of the AIDS infection, violin plots were constructed using these factors in order to observe the distribution of CD4+ cell count and viral load in the data set sequences in each cluster, comparing them. A violin plot, with the use of density curves, shows the distribution of numeric data for one or more groups, where the width of the density curve corresponds to the approximate frequency of data points in that region. This was also done with the help of the Seaborn package.

The code used in this part of the dissertation is available in the following GitHub link: https://github.com/filipactavares/msc_dissertation/tree/main/Chapter_4.2, where all parameters are explained.

In the interest of tracking the distribution of the viral strains that characterize each one of the clusters determined before over time, as well as the distribution of the most common antiretroviral treatment schemes over the years, bump charts, which are similar to linear plots but focused on observing changes in rank over time, were constructed using RAWGraphs, which is an open source data visualization framework.

4.3 PROTEIN STRUCTURE PREDICTION

In an effort to understand how the mutations in clusters 1, 3, 4, and 5 affected the RT conformation and the NNRTI binding, we predicted the RT structure and the ligand binding to the NNIBP, given only the amino acid sequence. This was performed using the IntFOLD (v. 7) server, that includes the FunFOLD program to predict the bound ligands [92, 111], which only needs the amino acid sequence of the target protein to be able to predict its structure. This analysis resulted in four different pdb (program database) files for each of the clusters analysed, as well as 5 txt files for each of the analysed clusters' mutated conformations with information regarding Tertiary Structure Prediction (TS), Disorder Prediction (DR), Domain Prediction (DP), Binding Site Prediction (FN) and Model Quality Prediction (QA).

However, before any protein structure prediction was to be carried out, the nucleotide sequence of the RT needed to be translated into its respective amino acid sequence. This was done using the MEGA11 (Molecular Evolutionary Genetics Analysis) (v. 11) software, which contains a vast collection of computational methods and tools for molecular evolution, to discover organismal and genome evolutionary patterns and processes [112].

After the RT structures were predicted, USCF ChimeraX (v. 1.6.1) was used for visualization of these protein structures using the pdb files obtained from the IntFOLD analysis. USCF ChimeraX is a next-generation molecular visualization program from the Resource for Biocomputing, Visualization, and Informatics (RBVI) [113], which uses Matchmaker as its superposition tool. Matchmaker superposes related structures by superposing proteins or nucleic acids through a pairwise sequence alignment and then matching the sequence-aligned residues in 3D, without having to worry about numbering or missing residues [113]. This tool then gives both the sequence alignment score, as well as the RMSD score between the pruned atoms, by running an iterative pruning process to remove residue pairs (which are far apart from each other structurally and conformationally) using a default cut-off value of 2 Å, and the overall RMSD score. The Matchmaker parameters used were the default ones by the software, which were as follows: chain pairing – bb; alignment algorithm – Needleman-Wunsch; similarity matrix – BLOSUM-62; ssfraction – 0.3; gap open (HH/SS/other) – 18/18/6; gap extend – 1; SS matrix – (H, O): -6 (S, S): 6 (H, H): 6 (O, S): -6 (O, O): 4 (H, S): -9; and iteration cut-off: 2.

Regarding the FunFOLD results, which were present in the FN file for each of the mutated conformations, it generated a PyMOL image of ligand binding residues prediction, and provided the putative binding site residues, as well as the most likely ligands at each site, with their respective PDB three-letter code. Furthermore, the FunFOLDQA scores were given, particularly the Model Quality Score. Lastly, the CAMEO-LB (Continuous Automated Model EvaluatiOn – Ligand Binding) program, which allows for a more fine-grained ligand binding prediction and a more detailed assessment, gives the putative ligands for each of the binding residues in each of the four different conformations at study. Particularly, its format is divided in three sections that, respectively, give the unique identifier for a residue or atom (‘r’ is the residue name and ‘n’ the residue number); contain the predicted p-values for each of the four possible ligand categories (‘I’ for ions, ‘O’ for organics, ‘N’ for polynucleotides and ‘P’ for peptides); and, finally, and optionally, the specification of ligands by their three-letter PDB code [114].

4.4 PHYLOGENETIC TREES

Phylogenetic trees were used to explore the evolution of the K103N mutation and its combinations with V108I, L100I and K65R mutations.

Six different trees were built – three with B subtype sequences and the other three with C subtype sequences. To reduce the computational complexity of building phylogenetic trees, which can be very challenging when dealing with large and diverse datasets, all trees were built from 200 sequences, 50 from each of the clusters 1, 3 and 5 and the remaining 50 from cluster 0, which was the one that had minimal resistance mutations. Each of these sequences was picked in a random manner from the total 22,838 sequences database pool.

The trees were built using the IQ-TREE software in the Galaxy-EU web platform. IQ-TREE is a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies [115] by feeding it the aligned sequences in a FASTA file. All parameters were left as the default IQ-TREE parameters, except for the ultrafast bootstrap replicates, which was set to 1000. This resulted, in this case, in six different files with the consensus tree in Newick format.

After this, to visualize the trees in a simplified and comprehensive way, the PastML method was utilized. PastML is a method that uses ancestral character reconstruction (ACR) to trace the evolution and origin of the characters of interest. This is done by using decision-theory concepts to associate each tree node to a set of likely states, where a unique state is predicted in the tree

regions with low uncertainty, whereas several states are predicted in uncertain regions, typically around the tree root. In the end, to visualize the results, the neighbouring nodes associated with the same states are clustered together through the use of graph visualization tools [116]. The algorithm used as inputs the tree in Newick format and a CSV file with the annotation table specifying tip states for the characters of interest. The ancestral state prediction method was maximum likelihood: MPPA (marginal posterior probabilities approximation) and the evolutionary model for state changes was Felsenstein 1981 (F81).

In this case, the characters of interest in the annotated table were MUTATION (with the K103N, K103N+V108I, K103N+L100I and NONE – if none of the above mutations were present in the sequence in question – states), K65R (with the K65R – if the K65R mutation is present in the sequence in question - and NOK65R – if the K65R is absent from the sequence at hand – states).

Also, it was verified if any of these sequences was a part of a possible transmission cluster, which would determine if any transmission of these mutations happened between individuals, particularly the K103N+L100I+K65R triple mutant, according to transmission clusters identified in [3], who used the same sequence data base for their analyses.

4.5 STATISTICAL ANALYSIS

Statistical analysis was performed using Epi InfoTM for Windows (v. 7.2.5 of 23 of November of 2021), provided by the Centers for Disease Control and Prevention (Atlanta, United States of America), available for download in the following link: https://www.cdc.gov/epiinfo/support/por/pt_downloads.html). Jamovi Stats. Open. Now for desktop (v. 2.3.28 solid), available for download in the following link: https://www.jamovi.org/download.html) was also used.

The analyses performed included simple data tabulations, comparison of proportions through contingency tables relying on Mantel-Haenszel chi-square test or the Fisher's exact test (depending on the quantity of occurrences being greater or smaller than 5, respectively), Pearson correlation matrices and logistic regression analyses.

5 RESULTS

5.1 CHARACTERIZATION OF THE STUDY POPULATION

The study population consisted of 22,838 samples of different HIV-1 infected individuals in Brazil, result of HIV-1 sequence genotyping done between 2008 and 2017. Out of these 22,838 individuals, 20,227 of them, or 88.57% of the total sample pool, were on antiretroviral therapy, meaning they are treatment Non-Naïve, whereas the remaining 2611 individuals, or 11.43% of the sample pool, still weren't on any sort of antiretroviral therapy, making them treatment-Naïve. The main characteristics of this sample pool are illustrated in Table 5.1.

More than half of the population at study are male, counting 55.07% of it (n = 12,577), whereas 44.88% (n = 10,249) account for the female individuals.

When it comes to the patients' age at the time of diagnosis, 1.07% (n = 244) were younger than 1 year old, 2.64% (n = 603) were between 2 and 9 years old, 4.92% (n = 1123) were between 10 and 17 years old, 14.45% (n = 3300) were between 18 and 30 years old, 57.52% (n = 13,136), which accounted for most of the infected individuals, were between 31 and 49 years old, 19.31% (n = 4409) were between 50 and 70 years old, and a mere 0.08% (n = 18) were older than 80 years old. The diagnosis was performed at different federative units in Brazil, namely São Paulo, with 22.41% (n = 5119), Rio Grande do Sul accounting for 11.75% (n = 2683), followed by Rio de Janeiro explaining 10.04% (n = 2292), then Minas Gerais with 9.05% (n = 2066), and Paraná at 7.49% (n = 1711), while other federative units report 39.26% of individuals (n = 8967).

Unfortunately, almost 50% (49.94%; n = 11,405) of individuals didn't report their route of transmission of the virus. Within the existing data, the most common routes are, in order: heterosexual contact (33.56%; n = 7664), men who have sex with men (MSM) (7.52%; n = 1717), perinatal (transmission from mother to child) (4.03%; n = 921), bisexual contact (2.88%; n = 658), intravenous drug use (IDU) (1.94%; n = 444), blood transfusion (0.12%; n = 28) and by accident (0.004%; n = 1).

Regarding the different treatment schemes of the Non-Naïve patients, there are many, but the five most frequent ART regimens are, in order: 3TC, EFV, TDF (21.38%, n = 4324); 3TC, AZT, EFV (19.80%, n = 4004); 3TC, AZT, LPV (11.16%, n = 2258); 3TC, LVP, TDF (8.24%, n = 1666) and 3TC, ATV, RTV, TDF (6,70%, n = 1355), whilst 32.73% (n = 6620) of Non-Naïve individuals were administered other ART regimens.

The B subtype is the most prevalent in this Brazilian population, amounting 65.47% (n = 14,953), followed by the C subtype with 14.64% (n = 3344) and the F1 subtype with 9.56% (n = 2184). Other subtypes account for 10.32% (n = 2357) of cases.

Table 5.1 - Characterization of the study population, namely by presence/absence of treatment, gender, age at diagnosis (yrs), federative unit of diagnosis, route of transmission, treatment scheme and subtype

	n	%		n	%
Treatment			**Route of Transmission**		
NonNaive	20227	88.57	Accident	1	0.004
Naive	2611	11.47	Heterosexual	7664	33.56
Gender			Bisexual	658	2.88
Male	12577	55.07	MSM	1717	7.52
Female	10249	44.88	IDU	444	1.94
NA	12	0.05	Transfusion	28	0.12
Age at Diagnosis (yrs)			Perinatal	921	4.03
<1	244	1.07	NA	11405	49.94
2-9	603	2.64	**Treatment Scheme**		
10-17	1123	4.92	3TC,EFV,TDF	4324	21.38
18-30	3300	14.45	3TC,AZT,EFV	4004	19.80
31-49	13136	57.52	3TC,AZT,LPV	2258	11.16
50-79	4409	19.31	3TC,LVP,TDF	1666	8.24
>80	18	0.08	3TC,ATV,RTV,TDF	1355	6.70
NA	5	0.02	Others	6620	32.73
Federative Unit of Diagnosis			**Subtype**		
São Paulo	5119	22.41	B	14953	65.47
Rio Grande do Sul	2683	11.75	C	3344	14.64
Minas Gerais	2066	9.05	F1	2184	9.56
Rio de Janeiro	2292	10.04	Others	2357	10.32
Paraná	1711	7.49			
Others	8967	39.26			

n, number of HIV-1+ individuals; %, percentage of HIV-1+ individuals; yrs, years; NA, not applicable

5.2 K103 IS THE NNIBP RESIDUE WITH THE MOST VARIATION

In order to determine the genetic diversity of NNIBP residues, the encoding nucleotides among the 22,838 sequences were analysed. Sequence logos were constructed by aligning the sequences to a reference and grouping the sequences by presence/absence of treatment (Non-Naïve and Naïve, respectively) and by subtype (B and C). The results are shown in Table 5.2. It wasn't possible to construct a sequence logo for the Y318 residue due to the lack of sequences that had actual nucleotides present and not gaps, which was what happened after sequence alignment. The characters representing the sequence nucleotides are stacked on top of each other according to the percentage in which they appear in each position of the sequence, with the most common one being on top and having the biggest height [104].

The L100 residue sequence logo shows that the TTA codon, which translates to leucine (L), is present in more than 85% of cases, regardless of treatment and subtype. The K101 residue sequence logo also shows that the AAA codon, which translates to lysine (K), is the most common, accounting for about 85% of sequences. However, there is a difference in the order of the most common nucleotides in the K101 codon's first position. In Non-Naïve, subtype C sequences, it is A-G-C-T, whereas in all other cases, it is A-C-G-T. This indicates that, in this cohort, Non-Naïve, subtype C sequences have a tendency for a higher frequency of the K101N mutation (AAG) than other cases. The K103N mutation appears in about 50% of sequences in all cases, regardless of treatment and subtype. About 50% of sequences have AAA codon, which translates to lysine (K), while the other half have either AAC or AAT codons, which both translate to asparagine (N). V106 shows a curious case: in subtype B sequences, GTA codon, which translates to valine (V), is the most common. However, in subtype C sequences, another codon GTG, which also translates to valine (V), is more common. This means that a synonymous

mutation is present according to the sequence subtype in question. Regardless of which codon is being referenced, they both account for more than 75% of the sequences. T107 is one of the most conserved residues in the NNIBP, with the codon ACA, which translates to threonine (T), present in almost all sequences. V108 residue is also consistent across all conditions, with the most common codon being GTA, which translates to valine (V), accounting for approximately 90% of sequences. The next four residues, V179, Y181, Y188 and V189, are also largely unaffected by presence/absence of treatment of the subtype. The most common codon for each residue is GTT that stands for valine (V), TAT, which codes for tyrosine (Y), also TAT for the Y188 residue, and GTA, which also translates for valine (V), representing between 85% and close to 100% of sequences. G190 residue presents a similar situation to the K101 residue. The most common codon is GGA, which codes for glycine (G), accounting for close to 80% of sequences. However, there is a change in the order of the most common nucleotides in the codon's last position. In subtype B sequences, it is A-C-G-T, whereas in subtype C sequences, it is A-G-C-T. Nonetheless, both mutated codons GGC and GGG translate to glycine (G), meaning they are synonymous mutations. F227 shows a different most common codon for each of the subtypes accounted for, B and C. For subtype B, TTC codon translates to phenylalanine (F) and accounts for 75% of sequences. The remaining 25% have TTT codon that also translates to phenylalanine. For subtype C, on the other hand, the most common codon is TTT being shared among close to 100% of sequences. The W229 residue, like the T107 residue, is conserved in both conditions. The codon TGG, which codes for tryptophan (W), appears in close to 100% of sequences. The L234 residue is similar. The most common codon CTC, which codes for leucine (L), is consistent across all sequence types, accounting for about 95% of them. Lastly, the E138 residue, like F227, presents a different most common codon according to the subtype. For subtype B, the codon is GAG, accounting for 85% of sequences. The remaining 15% have the GAA codon. For subtype C, it is GAA that accounts for 90% of sequences, and the GAG codon accounts for the remaining 10%. Both of these codons translate to glutamic acid (E), which means they are synonymous mutations.

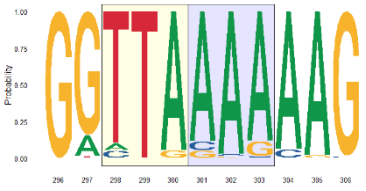
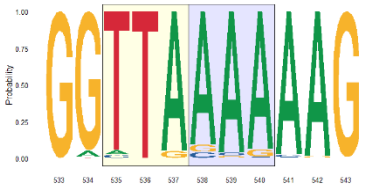
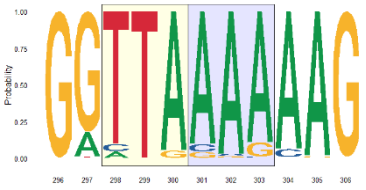
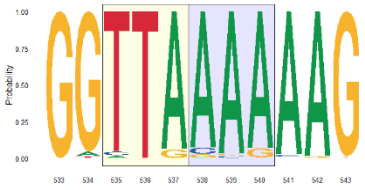
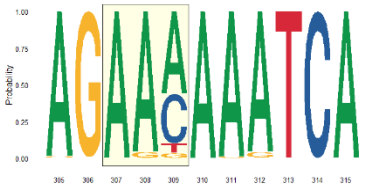
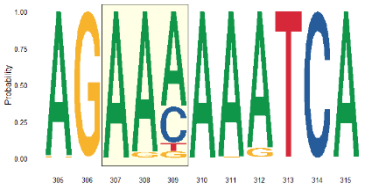
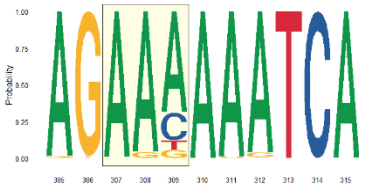
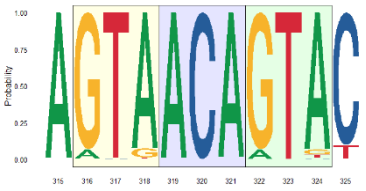
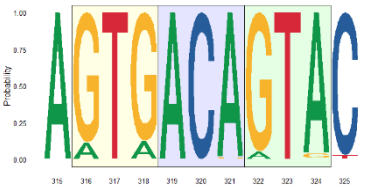
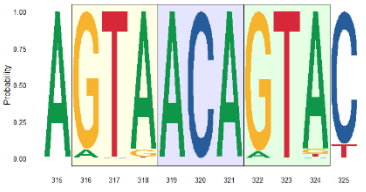
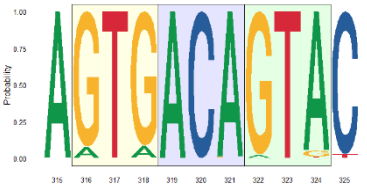
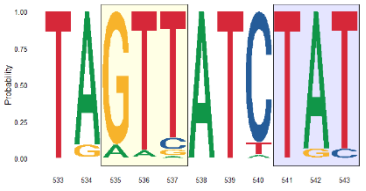
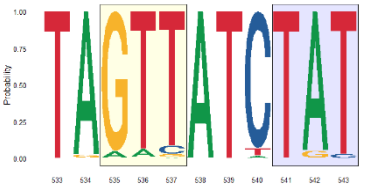
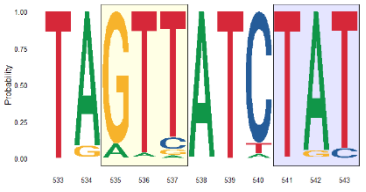
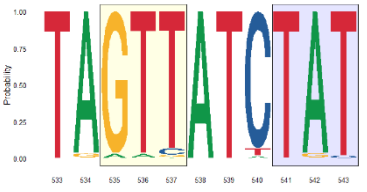
There was sequencing information available to analyse 15 of the 16 NNIBP residues. However, it wasn't possible to construct a sequence logo for the Y318 residue due to the lack of sequences that had actual nucleotides present and not gaps. Curiously, for most of the 15 residues in Table 5.2, there is a tendency for smaller substitution rate on the subtype C sequences when compared to the subtype B sequences, but of no particular significance.

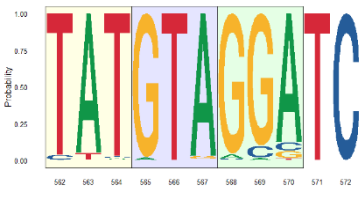
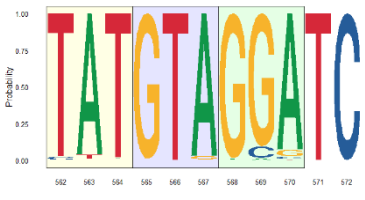
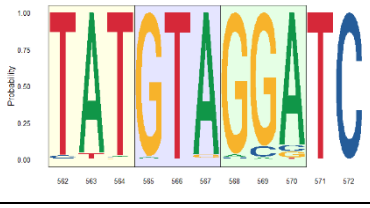
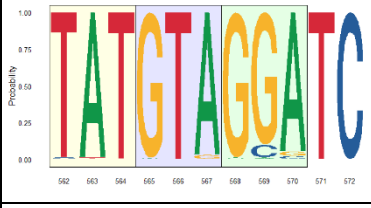
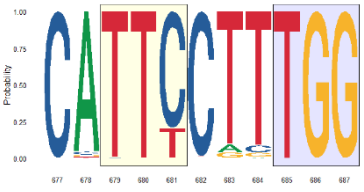
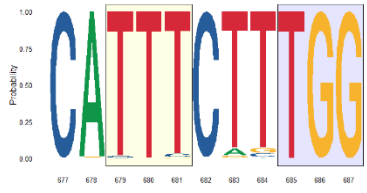
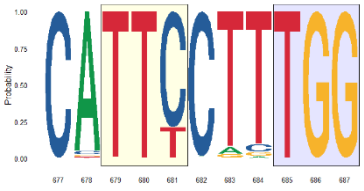
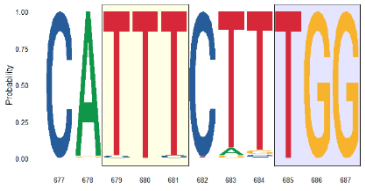
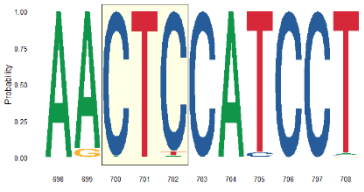
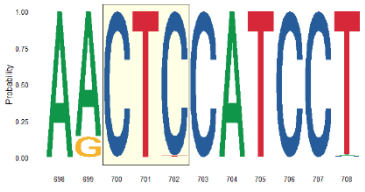
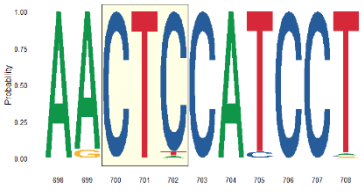
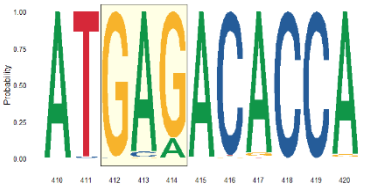
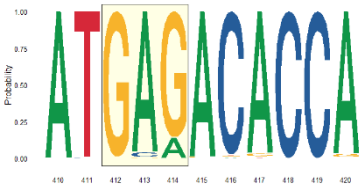
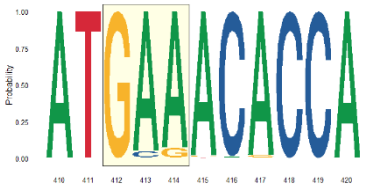
5.3 HIGHER VIRAL LOAD AND LOWER CD4+ CELL COUNT FOR K103N+L100I DOUBLE MUTANT

With a focus on determining if there are any patterns, partitions or groups that can be formed in the data set, given the different 132 mutations across the 16 NNIBP residues, a clustering analysis using the Louvain method, preceded by dimensionality reduction by the UMAP technique, was employed. This was done to explore the possibility that some mutations may be grouped together in the data set, and if yes, which ones.

Using the UMAP dimensionality reduction manifold for, in this case, Louvain, clustering, seven different clusters were able to be identified, as seen in Figure 5.1.

Table 5.2 - Sequence Logos of 15 of the 16 residues that comprise the NNIBP by presence/absence of treatment and subtype. The colours highlighting each residue name indicate their respective highlight colour in the sequence logo

		Subtype B	Subtype C
L100, K101	Non-Naïve		
	Naïve		
K103	Non-Naïve		
	Naïve		
V106, T107, V108	Non-Naïve		
	Naïve		
V179, Y181	Non-Naïve		
	Naïve		

Y188, V189, G190	Non-Naïve		
	Naïve		
F227, W229	Non-Naïve		
	Naïve		
L234	Non-Naïve		
	Naïve		
E138	Non-Naïve		
	Naïve		

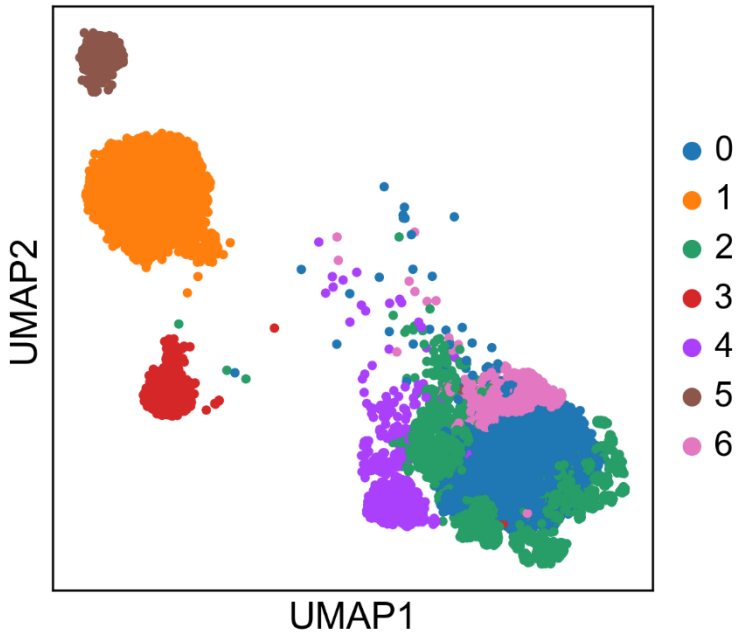


Figure 5.1 - UMAP visualization of all sequences based on the 132 mutations of the 16 NNIBP residue, showing the clustering by the Louvain method of 22,838 HIV-1 infected individuals in Brazil from 2008 to 2017 into seven different clusters

To check which mutations were the ones that were the cause of clustering of the sequences of the data set in the form shown in Figure 5.1, a heat map, where each column shows a different mutation and each row is one of the sequences of the data set, was constructed, where the yellow colour represents the presence of the mutation in question in a particular sequence and purple as the corresponding absence. This heat map is represented in Figure 5.2.

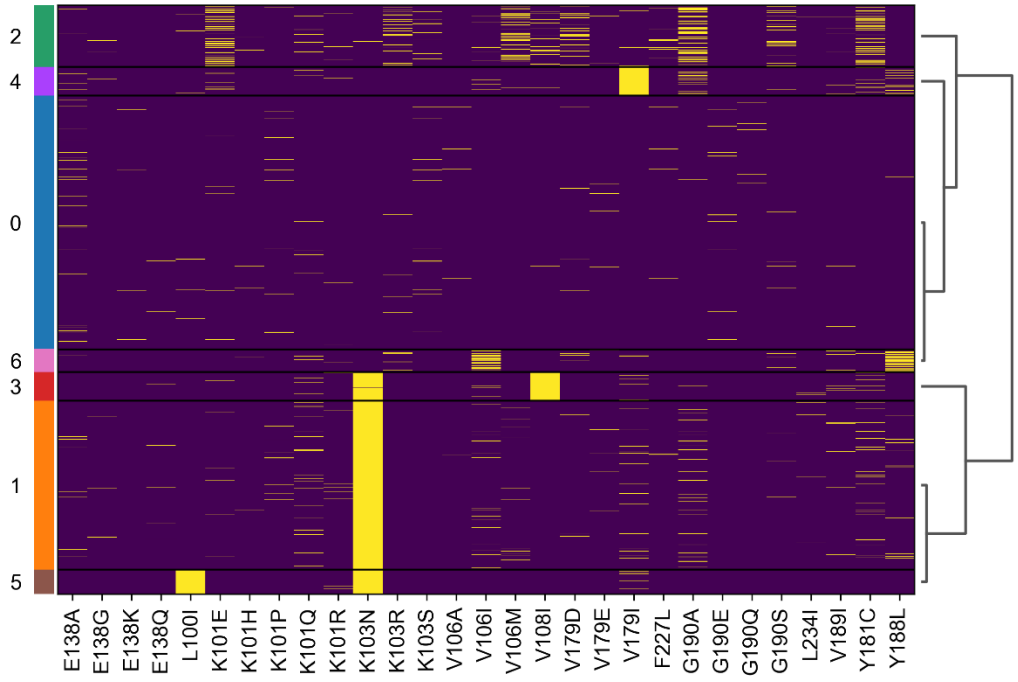


Figure 5.2 - Heat map visualization of the most prevalent mutations associated with each of the seven clusters, highlighting the representative mutations in clusters 1, 3, 4, and 5 (K103N, K103N+V108I, V179I, and K103N+L100I, respectively). Yellow means presence of mutation, whereas purple means absence

The heat map shows which mutations represent each of the different seven clusters – clusters 1, 3, 4 and 5 exhibited specific representative mutations (K103N, K103N+V108I, V179I, and K103N+L100I, respectively). Moreover, cluster 0 was not dominated in frequency by any mutation, while cluster 2 displayed a high diversity of different high frequency resistance mutations, including the following: K101E, K103R, V106M, V179D, G190A and Y181C.

Two of the factors that can determine the stage of progress of the AIDS infection are the CD4+ cell count and the viral load in the bloodstream, as stated in Chapter 2.2. With this in mind, violin plots were constructed using these factors to observe the distribution of CD4+ cell count and viral load in the data set sequences in each cluster, comparing them, as shown in Figure 5.3 and 5.4, respectively.

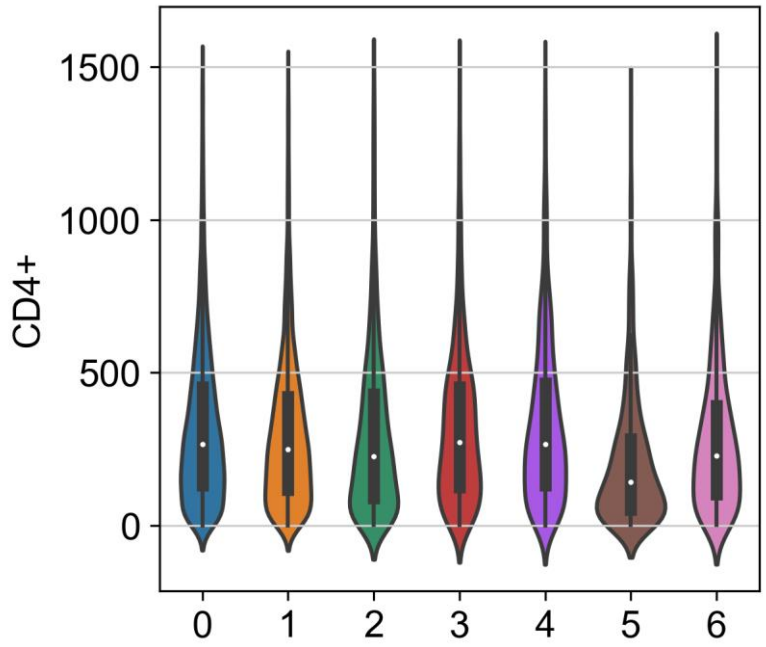


Figure 5.3 - Violin plot visualization of the CD4+ cell count for each of the seven clusters, revealing a significant difference between cluster 5 and the other clusters

In the violin plots, a significant difference between cluster 5, characterized by the K103N+L100I double mutant, and the other clusters is revealed, both in the CD4+ cell count and viral load, with cluster 5 showing a lower CD4+ cell count and a higher viral load than the other clusters. Regarding the CD4+ cell count, every patient in the data set was described as either having, or not, immunodeficiency, and whether this immunodeficiency was moderate or severe. Given this information, by statistical analysis, it showed that cluster 5 was one of the independent variables that influenced immunodeficiency ($p < 0.001$), by increasing the probability of it appearing by 89%. Moreover, the increase in immunodeficiency associated with cluster 5 was not innocent, meaning that infection by HIV-1 derived from sequences in cluster 5 influenced the appearance of moderated immunodeficiency ($p < 0.05$) [$p = 0.016$], increasing its probability by 38%, and also the appearance of severe immunodeficiency ($p < 0.001$), increasing the probability of it being the case by 139%. When it comes to viral load, statistical analysis also confirms that cluster 5 was one of the independent variables with higher influence in the appearance of higher levels of viral load ($p < 0.001$), increasing the probability of higher viral load by 53%.

To determine whether any other laboratory or personal data about each infected patient could also partition/group the data into different clusters, the UMAP embedding visualization was used and coloured differently according to these various variables, particularly sex, age, HIV-1 subtype, presence/absence of treatment and type of ART used. As seen in Appendix A.1 through A.7, there is high heterogeneity among all these variables, meaning there is no particular cluster

formed when using these variables as the condition to which to group the data with. This was in agreement with the lack of association of the clusters with other variables other than viral load and CD4+ counts in the statistical analysis.

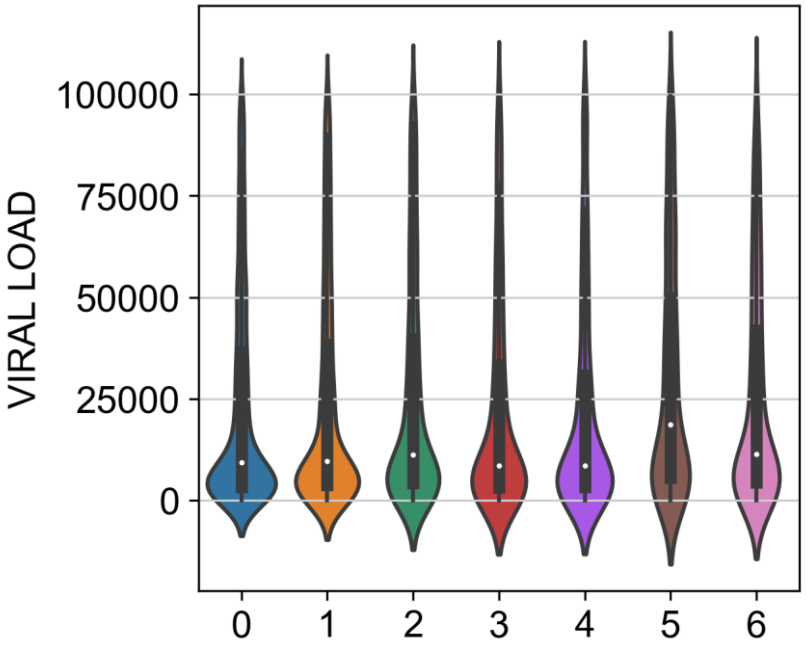


Figure 5.4 - Violin plot visualization of the viral load for each of the seven clusters, revealing a significant difference between cluster 5 and the other clusters

5.4 K103N+L100I DOUBLE MUTANT ASSOCIATES WITH DIFFERENCE IN PREDICTED BINDING RESIDUES AND LIGANDS

For the purpose of understanding the structural changes that resulted from the HIV-1 RT mutations most prevalent in each of the 4 clusters with specific determining mutations (clusters 1, 3, 4 and 5) and how they affect the binding of antiretroviral drugs to the NNIBP, the RT structure was predicted from amino acid sequences, and also the type the ligands bound to this protein and to which residues. The results of the protein structure prediction for each cluster mutations in relation to the wild-type (WT) RT conformation, are presented in Figure 5.5 and the results of ligand binding analysis are shown in Figure 5.6. The detailed results are present in the following GitHub link: https://github.com/filipactavares/msc_dissertation/tree/main/Chapter_4.3.

Given the MatchMaker results, which superposes proteins by first creating a pairwise sequence alignment, shows that, in the case of the K103N conformation in relation to the WT, the sequence alignment score was that of 2841.3, with a local RMSD score between 375 pruned atoms of 1.040 Å, whereas across all 560 pairs, it was 3.789 Å; as for the K103N+V108I conformation, the sequence alignment score was that of 2832.8, with a local RMSD score between 403 pruned atoms of 0.993 Å, whereas across all 560 pairs, it was 3.323 Å; the V179I conformation compared with the WT conformation had a sequence alignment score of 2836.3, with a local RMSD score between 394 pruned atoms of 1.059 Å, whereas across all 560 pairs, it was 3.680 Å; at last, the K103N+L100I conformation presented a sequence alignment score of 2839.9, with a local RMSD score between 347 pruned atoms of 1.000 Å, whereas across all 560 pairs, it was 3.089 Å. This is presented in Table 5.3.

Table 5.3 - MatchMaker results (particularly Sequence alignment score, Number of pruned atoms pairs, RMSD score between pruned atom pairs (Å) and RMSD score between all 560 atom pairs (Å)) for the four superposition 3D protein structures between the wild-type (WT) structure and the K103N, K103N+V108I, V179I and K103N+L100I structures, which are coloured in blue, green, orange and purple, respectively, as in Figure 5.5

		WT+K103N	WT+K103N+V108I	WT+V179I	WT+K103N+L100I
	Sequence alignment score	2841.3	2832.8	2836.3	2839.9
	Number of pruned atom pairs	375	403	394	347
RMSD	Between pruned atom pairs (Å)	1.040	0.993	1.059	1.000
	Between all 560 atom pairs (Å)	3.789	3.323	3.680	3.089

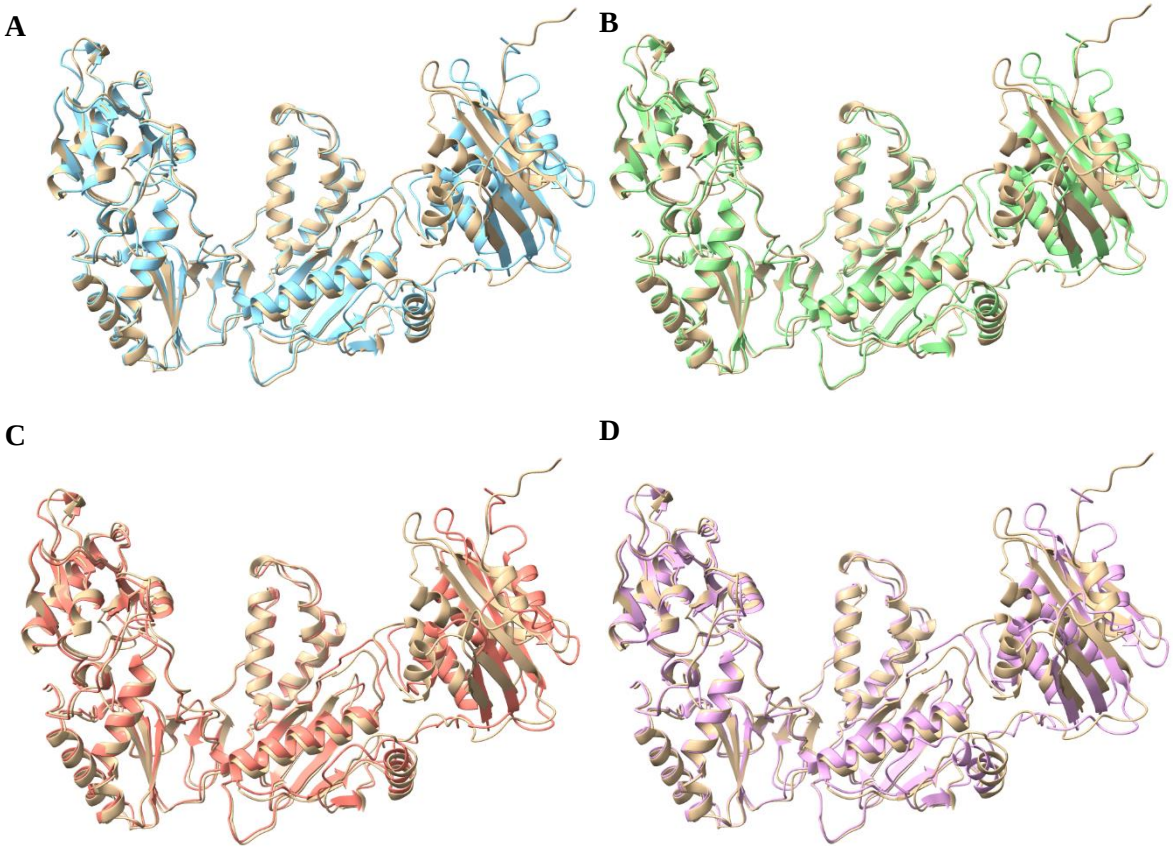


Figure 5.5 - 3D models of the predicted HIV-1 reverse transcriptase structure under different mutations: (A) K103N, (B) K103N+V108I, (C) V179I, and (D) K103N+L100I superposed with the wild type RT structure. The gold structure represents the WT model, whereas the blue, green, orange and purple structures represent the K103N, K103N+V108I, V179I and K103N+L100I mutated RT structures, respectively (3D short rotating videos of each of these superposed structures is present in the following GitHub link: https://github.com/filipactavares/msc_dissertation/tree/main/Chapter_5.4)

When it comes to the FunFOLD results, meaning the ligand binding residue prediction, the K103N conformation presented a model quality value of 0.75499755, with the following predicted ligand residues: 95, 100, 101, 103, 106, 108, 179, 181, 187, 188, 189, 190, 227, 228, 229, 230, 234, 235, 236 and 318, which, when compared with the WT conformation, it also has the 236 residue as a possible ligand binding site. The K103N conformation has as the most likely

ligands at each site DC (2'-deoxycytidine-5'-monophosphate) and T27 (which is a synonym for the NNRTI rilpivirine). For the K103N+V108I conformation, the model quality value is 0.7541711, and the predicted ligand residues are 95, 100, 101, 102, 103, 106, 108, 179, 181, 187, 188, 189, 190, 222, 227, 228, 229, 230, 234, 235, 236 and 318, that compared with the predicted WT binding residues, it additionally has the 102, 186, 222 and 236 residues. The K103N+V108I structure also has DC and T27 as putative binding ligands. Regarding the V179I conformation, it has a model quality value of 0.7582536, with the 95, 100, 101, 103, 106, 108, 179, 181, 186, 187, 188, 189, 190, 222, 227, 228, 229, 230, 234, 235, 236 and 318 as predicted ligand residues – this compared with the WT conformation also has the 186, 227 and 236 residues. The V179I conformation, as the previous two, too has DC and T27 as most likely ligands. At last, the K103N+L100I structure, with a model quality score of 0.7562798, has as possible binding residues 95, 100, 101, 103, 106, 108, 179, 181, 187, 188, 189, 190, 227, 228, 229, 230, 234, 235, 236, 318 and 443, which compared with the WT structure, additionally has the 236 and 443 residues. The K103N+L100I conformation has T27 and DC as putative ligands at each site.

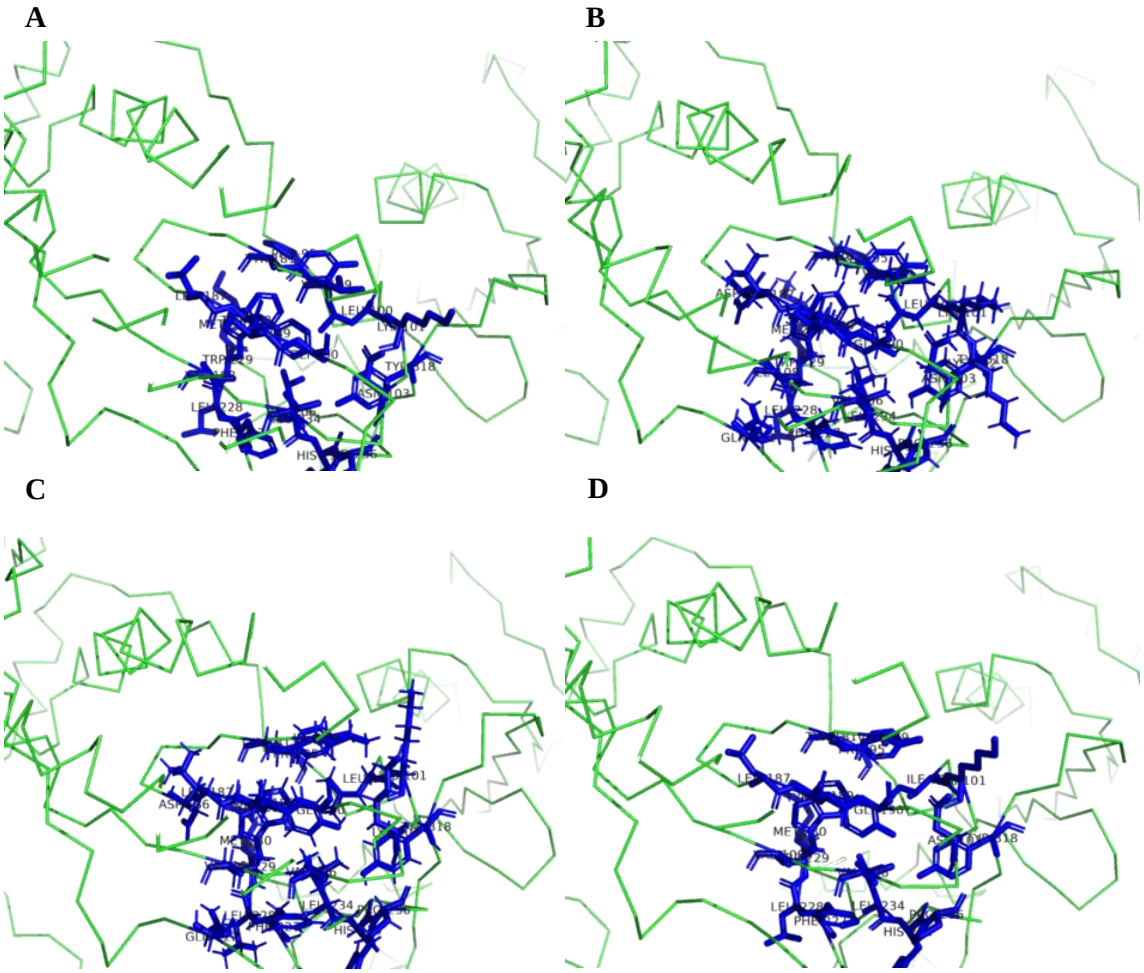


Figure 5.6 - 3D models of the predicted HIV-1 reverse transcriptase structure and its corresponding ligand binding residues under different mutations, generated using IntFOLD and FunFOLD, respectively. The models represent the impact of the following mutations: (A) K103N, (B) K103N+V108I, (C) V179I and (D) K103N+L100I. Predicted ligand binding residues are shown as blue sticks, and the predicted 3D structure of the RT enzyme is depicted in green. These models provide insight into the structural changes induced by specific mutations in the binding pocket of RT, which may have implications for the efficacy of antiretroviral drugs

Concerning all the ligands that bound to each of the residues in the four different 3D protein structures, when comparing with the WT results, which included the DDG, DG, DC, DA, RU,

R8E, T90, M9A, T27, 7N1, 5MA and NVP ligands, for a total of 12 possible ligands, the K103N conformation did not include the T90 and 7N1 ligands, totalling 10 ligands; the K103N+V108I conformation had additionally the MG and 1RY ligands, equalling 14 ligands; the V179I conformation also additionally included the MG and 1RY ligands, which sums up to 14 ligands; and, finally, the K103N+L100I conformation presented six additional ligands, which were MG, QID, BA5, Y55, MN and ON1, giving way to 18 putative ligands. More detailed information about which residue in which conformation had what ligand bound to it is present in Appendix B.1.

Moreover, as observed in Figure 5.6 (B) and (C), which represented the putative ligand binding residues for the K103N+V108I and V179I conformations, more of these residues' different atoms could bind to the different ligands than in the other conformations, particularly K103N and K103N+L100I.

5.5 TREATMENT SCHEME SHIFT ASSOCIATES WITH AN INCREASE IN THE PROPORTION OF K103N+L100I DOUBLE MUTANT INFECTED INDIVIDUALS

To track the distribution of the viral strains that characterize each one of the seven clusters over time, a bump chart was constructed, as observed in Figure 5.7.

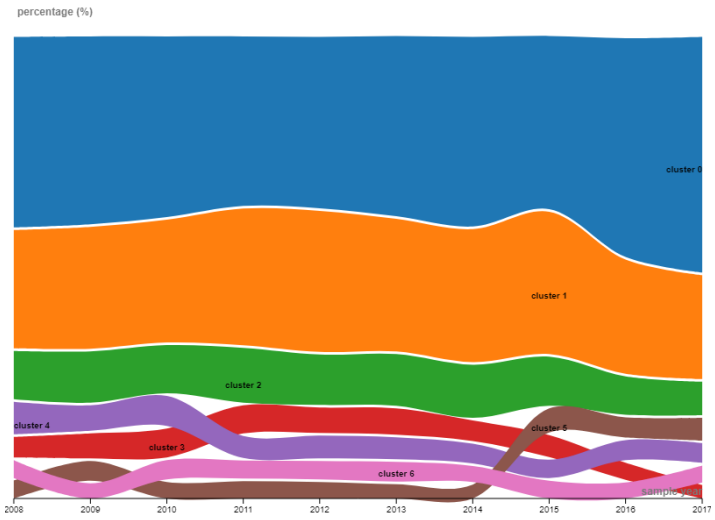


Figure 5.7 - Bump chart visualization of the temporal distribution of all sequences, grouped into seven clusters, from 2008 to 2017, providing an overview of the changes in the distribution of viral strains over time

The bump chart shows that the distribution of each cluster over time is relatively constant, with no significant increases or decreases over the years – that is, except cluster 5, characterized by the K103N and L100I mutations. There was an increase of 2.3% in the number of sequences associated with cluster 5 from 2014 to 2015.

Coincidently, as seen in Figure 5.8, there was a treatment scheme shift from 2014 to 2015, which is concurrent with the increase of the number of cluster 5 sequences. This particular treatment scheme shift was from 3TC+AZT+EFV to 3TC+EFV+TDF, meaning TDF, which is an NRTI, started being used instead of AZT, which is also an NRTI. The 3TC+EFV+TDF treatment scheme increased in 6.6% from 2014 to 2015, whereas the 3TC+AZT+EFV treatment scheme decreased in 5.0%. After that, they each kept increasing and decreasing, respectively.

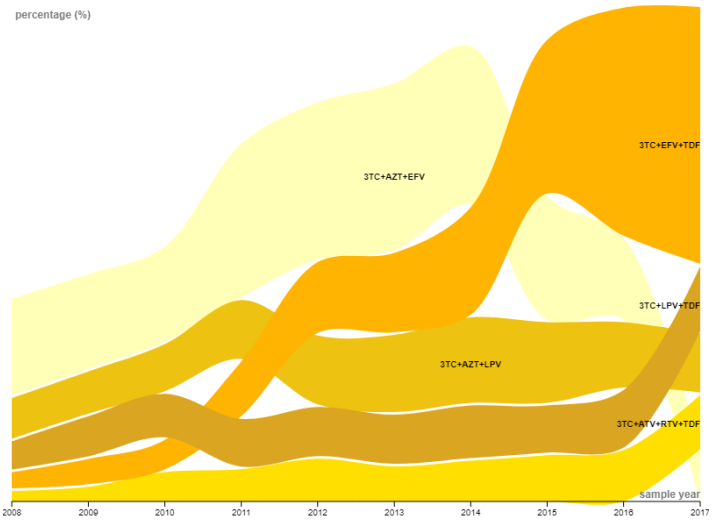


Figure 5.8 - Bump chart visualization of the temporal distribution of the 5 most common HIV treatment schemes used in Brazil from 2008 to 2017, providing insight into the prevalence of different treatment options over time

Also, as shown in [3], there was an increase in the number of sequences with the K65R mutation over the years, specifically from 2014 to 2015, when the treatment scheme shift occurred. In [3], they determined that the treatment scheme shift was the cause of the increase of the K65R mutation, which is an NRTI associated mutation, as it is often considered the TDF resistance mutation, despite also appearing when using different NRTIs [35]. This lead to the question of whether the increase of the K65R mutation, associated with the treatment scheme shift, might have been the cause of the increase in the number of cluster 5 sequences.

By statistical analyses, it was determined that, in fact, the increase in cluster 5 sequences was due to the increase in the number of sequences with the K65R mutation, by the correspondent shift in the treatment scheme ($p < 0.05$) [$p = 0.034$] – the prevalence of cluster 5 sequences was 34% greater after the treatment scheme shift rather than before. Moreover, TDF use increased the probability of the presence of cluster 5 sequences in the studied population 3.67 times, or 267%.

Furthermore, it was of importance to verify if any other variables could possibly also have a positive correlation when it came to the increase in the number of cluster 5 sequences. This demonstrated that only less treatment time had a directly proportional positive correlation with cluster 5, but only with sequences in cluster 5 that were associated with TDF use ($p = 0.05$), as its introduction to treatment schemes in this population is more recent than other drugs.

Moreover, despite there being an increase in the frequency of cluster 5 sequences over the years, which were previously explained in chapter 5.3 to be related to more aggressive viruses to the immunological system, the prevalence of patients with immunodeficiency and/or severe immunodeficiency decreased (5% and 31%, respectively).

5.6 K65R MUTATION COMMONLY APPEARS AFTER THE K103N+L100I DOUBLE MUTANT IS FORMED

With the intent of determining whether the K65R mutation appeared before or after the formation of the K103N+L100I double mutant, to verify the hypothesis that the K65R mutation was involved in the increase in numbers of the cluster 5 sequences, which was associated with this double mutant, phylogenetic trees were built. Six different trees were constructed, three with 200 subtype B sequences each and another three with 200 subtype C sequences each – one

example of each is shown in Figures 5.9 and 5.10, respectively. The other four trees are presented in Appendix C.8 through C.11.

Overall, its noticeable that the K65R mutation typically appears after the formation of the K103N+L100I double mutant, although it was possible to find a case of the K65R appearing before the K103N+V108I double mutant in the subtype B tree. The cluster circles which present this K65R mutation are fairly small, between 1 and 11 sequences – in spite of that, the K65R occurring cluster circles are larger in the C subtype tree than in the B subtype tree.

Also, the possibility of these sequences belonging to a transmission cluster, according to [3], was studied, to determine if the transmission of viruses harbouring the K65R mutation could have been one of the causes, if not the cause, for the increase of the number of sequences with this mutation and, consequently, the sequences with the K103N+L100I double mutation – amongst the six trees built, 4 sequences belonged to different transmission clusters. Subtype B sequence with the ID 11PB150082, with the K103N+L100I double mutation, as well as the K65R mutation, was a part of the B9 transmission cluster. Regarding the C subtype sequences, 21PR110049 and 11SP139055 sequences, the first being a K103N+V108I double mutant, and without the K65R mutation, and the second a K103N+L100I double mutant with the K65R mutation, both belonged to the C1 transmission cluster. Subtype C K103N+L100I and K65R triple mutant 21PR130350 sequence, too, belonged to a transmission cluster, C3.

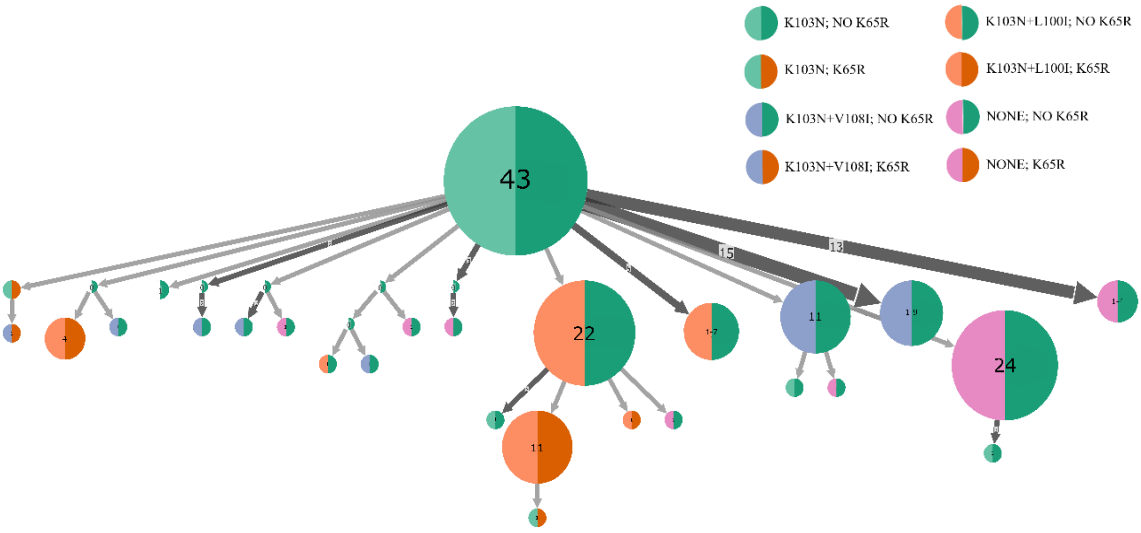


Figure 5.9 - Ancestral state reconstruction (ACR) given the cluster's representative mutation (K103N, K103N+V108I or K103N+L100I) and presence/absence of the K65R mutation visualized by PastML on a phylogenetic tree inferred from 200 subtype B sequences. Circles denote clusters of samples (the left half represents which mutation present in the formed clusters (K103N (light green), K103N+V108I (blue), K103N+L100I (dark orange) or NONE of the above (pink) and the right half denotes the absence/presence of the K65R mutation, where dark green represents absence and dark orange, presence). The sample sizes of clusters are indicated in the labels (for example, the circle '43' encompasses 43 sequences – thus, the larger the circle, the more sequences it represents). Clusters with a '0' and white colour indicate internal tree nodes for which two different cluster mutations had similar marginal probabilities. Arrows between two circles denote transmissions from the top to the bottom cluster. The size and number on top of the arrows indicate that the arrows represent multiple transmission events leading to clusters of similar sizes [48]

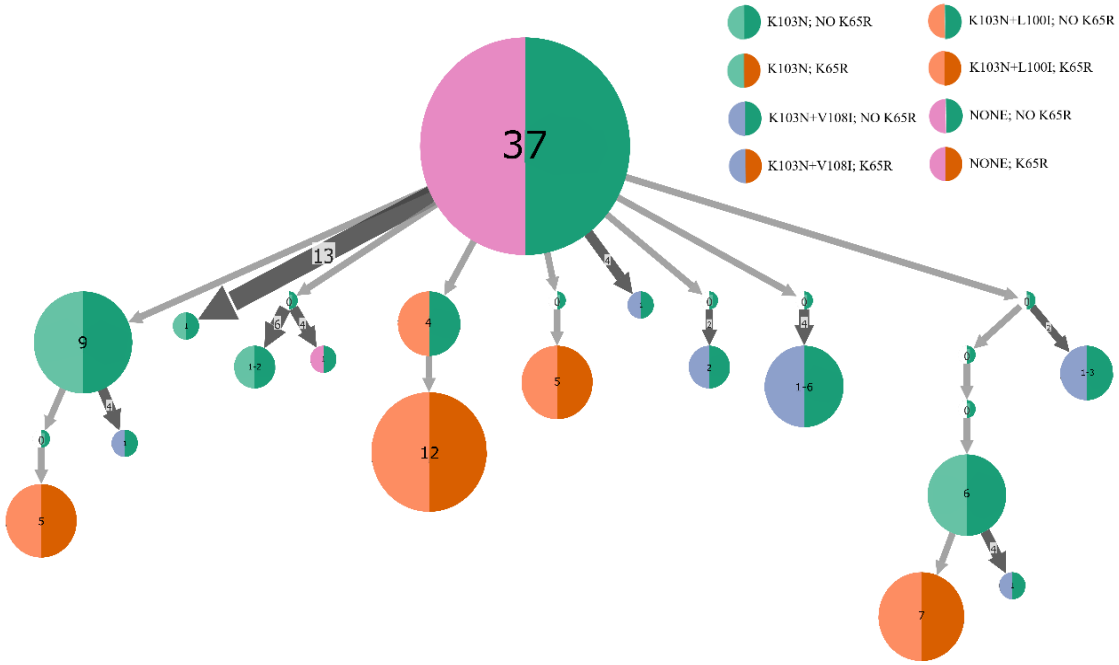


Figure 5.10 - Ancestral state reconstruction (ACR) given the cluster's representative mutation (K103N, K103N+V108I or K103N+L100I) and presence/absence of the K65R mutation visualized by PastML on a phylogenetic tree inferred from 200 subtype C sequences. Circles denote clusters of samples (the left half represents which mutation present in the formed clusters (K103N (light green), K103N+V108I (blue), K103N+L100I (dark orange) or NONE of the above (pink) and the right half denotes the absence/presence of the K65R mutation, where dark green represents absence and dark orange, presence). The sample sizes of clusters are indicated in the labels (for example, the circle '43' encompasses 43 sequences – thus, the larger the circle, the more sequences it represents). Clusters with a '0' and white colour indicate internal tree nodes for which two different cluster mutations had similar marginal probabilities. Arrows between two circles denote transmissions from the top to the bottom cluster. The size and number on top of the arrows indicate that the arrows represent multiple transmission events leading to clusters of similar sizes [48]

6 DISCUSSION

HIV-1 is a problematic virus and its study and continuous vigilance of its transmission patterns remain of the upmost importance, even today, almost 40 years after it was first identified. Specifically, Brazil, as the 1st country in South America to make anti-retroviral therapy (ART) available for the population [3], asks for the study of its resistance mutations to understand their evolution and transmission, and, possibly, prevent the most clinically dangerous ones from spreading any further. This becomes more important knowing that in 2010 alone, there was an increase of 17% of HIV-1 infections, with 48,000 reported infections and 14,000 AIDS-related deaths [3].

With this in mind, the characterization of these mutations in the Brazilian population is critical, with this dissertation focusing on the study of the mutations associated with the reverse transcriptase (RT) NNIBP residues. However, this decision was biased towards the NNIBP, despite the fact that all possible mutations should have been studied, which would have resulted in a more thorough and in-depth analysis of all residues in the main enzymes of the virus and their respective mutations. This would have putatively lead to a better understanding of the actual NNIBP-associated mutations studied, as there can be some sort of interaction between mutations that cause resistance to different classes of ARTs, given that most therapy treatment schemes don't include just one type of ART; also, some more rather troublesome mutations in other areas of the virus could have been discovered and studied. On the other hand, all of this would have ultimately lead to a plethora of information that would have been more difficult to characterize and group into useful results, and thus would not fit into the scope and main goals of this dissertation.

This way, mutations associated with the HIV-1 RT, specifically mutations in the NNIBP's residues where NNRTIs act, were picked as the object of study due to the fact that this enzyme plays an important role in the viral life cycle, and the fact that NNRTIs that act on it, despite having good pharmacological properties, show a low genetic barrier, meaning that a single mutation can cause resistance to several drugs in the same class of ART, which makes it crucial to study its associated mutations and all things related to them, to prevent future ART failure.

Either way, the K103 residue was the one with the highest percentage of mutation, regardless of treatment-naivety and virus subtype, particularly in the appearance of the K103N mutation, as seen in Table 5.2, which is the most frequent mutation in the virus's RT and is a rather complex and possibly dangerous one, as it not only confers resistance to one NNRTI, but to three different ones: efavirenz (EFV), nevirapine (NVP) and delavirdine (DLV) [117]. Moreover, it has the ability to be synergetic with accessory mutations [117], like L100I or V108I, as seen in the clustering-associated heatmap made in Figure 5.2.

In a general way, the clustering that was performed resulted in a much better understanding of the data pool that existed, which otherwise would have been too large to take from and form any sort of comprehensible conclusions. In spite of this, the clustering process was not able to completely pattern the sequences in the data base into separate, characterized by only a couple of mutations clusters. Case in point, cluster 2 displayed a high diversity of several different mutations, and as observed in the UMAP visualization in Figure 5.1, this cluster could have been further 'separated' into three or four smaller clusters that would have theoretically been characterized by only some of the several mutations in the original cluster 2. This was attempted by increasing the resolution parameter in the Louvain clustering method, but this only provided a further separation of the smaller clusters that were already pretty isolated from the rest, such as clusters 1, 3 and 5. This would have been immensely more helpful in the mutation analyses, as some of the cluster 2's mutations don't have any interaction between one another, and if separated into smaller clusters, they could have been analyzed in more detail.

Distinctly, the K103N+L100I double mutant, represented by cluster 5, was of great concern, considering that, not only does it increase the reduced susceptibility to both NVP and EFV by almost 50-fold when compared with the K103N mutation alone [51], but it also showed, as seen in the violin plots in Figures 5.3 and 5.4, a lower CD4+ cell count and a higher viral load, which are variables for which the infection progress can be measured with, respectively, than the other clusters. This entails that this cluster, represented by the K103N+L100I double mutant, has a singular clinical importance when compared to the other clusters, as it is potentially more critical to the person who has these mutations, since it hypothetically can lead faster to the AIDS stage. This is supported further by the fact that cluster 5 was described as being a factor that provoked the increase of severity of the virus' immunodeficiency, turning it more aggressive to the host's immunological system, meaning the increase in viral load and decrease in the count of CD4+ cells was not clinically innocent.

Also of concern, on account of the immunological aggressiveness of the K103N+L100I double mutant associated virus, was the increase in the number of sequences associated with cluster 5 from 2014 to 2015, such did not occur in the other clusters, as observed in the bump chart in Figure 5.7. This took place, amongst other factors, primarily due to the also increase of sequences with the K65R mutation within 2014 and 2015, which, in turn, happened thanks to the therapy treatment-scheme shift from 3TC+AZT+EFV to 3TC+EFV+TDF that too occurred from 2014 to 2015.

K65R is the NRTI-associated second most common discriminatory resistance mutation, meaning it enables RT to discriminate between dideoxy-NRTI chain terminators and the naturally produced dNTPs, so that NRTIs are prevented from being incorporated into a growing DNA chain [37]. It is primarily selected by tenofovir disoproxil fumarate (TDF), followed by abacavir (ABC), stavudine (d4T) and didanosine (ddI) [35, 36, 42], and confers low- to intermediate-level resistance to all approved NRTIs, except for zidovudine (AZT), where it causes an increase in susceptibility [35, 36, 37, 42].

It is characterized by having a reduced viral replication capacity which is the product of a decreased rate of incorporation of all natural substrates, like nucleotides, and also approved NRTI drugs, conferring its resistance [32, 35, 36]. Virus harbouring this mutation also show a decreased excision rate of NRTIs and also an increased fidelity [35], which translates to a low error rate and, therefore, a high degree of accuracy in the reverse transcription process.

On the other hand, the reduced NRTI excision rate is more prominent for AZT, resulting in a counteraction of the incorporation and excision resistance mechanisms that, in turn, restore AZT susceptibility [32, 36].

The earliest ART drugs used were AZT and d4T in combination with 3TC, and while they were of considerable importance at that time, they were soon replaced with drugs like TDF due to drug toxicities, clinical complications and the introduction of newer TAM-sparing NRTIs [35]. TAMs, or thymidine analogue mutations, are selected by AZT and d4T [118], causing resistance to these NRTIs [119], and they show a great negative association with K65R, given both have antagonistic mechanisms of resistance: TAMs facilitate the NRTI excision reaction by increasing the ATP binding, resulting in the removal of the chain-terminating NRTI and resumption of DNA polymerization [119], whereas K65R, as mentioned, reduces the NRTI excision activity of TAMs [32, 35, 36, 119]. This, consequently, makes K65R negate the resistance to AZT caused by TAMs, and vice versa, TAMs negate the resistance to TDF, ABC and other NRTIs conferred by K65R [119]. Thus, this elucidates why the percentage of the K65R mutation before 2014, when the treatment scheme shift occurred, where AZT was being used instead of TDF, was lower than after 2015.

TDF was, in turn, first introduced in 2001 and showed improved antiviral activity against harbouring TAMs and NAMs (nucleoside analogue mutations) [35], whose susceptibility is decreased by only 2-fold in the presence of the K65R mutation [42]. Unlike other NRTIs, which

need two phosphorylation steps, TDF only requires one, which results in high intracellular concentrations, allowing for antiviral activity in both resting and activated CD4+ cells [35].

All of these factors together return a low incidence of the K65R mutation, both in treatment-naïve and non-naïve patients [35].

Thus, as K65R can have a potential synergistic fitness effect with NNRTI-induced mutations, given that the use of TDF together with either EFV or NVP can cause the more rapid development of resistance through K65R [118], the emergence of this mutation, in this case, is possibly associated with the use of EFV, as the treatment-scheme change happened. This may explain why the K65R mutation was statistically shown to be one of the motives for the higher viral load and decreased CD4+ cell count associated with cluster 5, which in turn was characterized by the K103N+L100I double mutant. Essentially, the K65R mutation augmented the original resistance level that the K103N+L100I NNRTI-induced mutations conferred alone, which was already pretty high.

Nevertheless, the increase of sequences with the K103N+L100I double mutant that was observed cannot be explained by the putative synergistic interaction previously explained, as both K103N+L100I and K65R have a replication fitness defect, which would mean that this defect would be further increased, lowering further viral replication. This, however, can be elucidated by the uncommon transmission capacity of the K65R mutation in low- and middle-income countries (LMIC), which is higher than the usual, almost nonexistent transmission of K65R, as stated in [3], where out of the 396 sequences with the K103N+L100I+K65R triple mutant, 6 of them were a part of one of 5 different transmission clusters (B4, B9, B13, C1 and C3). Rightly, the shift from the use of AZT to TDF alone increased the number of sequences with the K65R, but not to the extent that it did, to just over 10% of all sequences in the cohort [3], when usually only 2 to 5% of sequences have K65R present [35]. Furthermore, given the phylogenetic trees in Figures 5.9 and 5.10, the fact that the K65R mutation only appears after the K103N+L100I is already formed seems to be of accordance with this theory, as if it was the other way around, there wouldn't be also an increase in number of the K103N+L100I double mutant-containing sequences. Additionally, for the most part, as observed in Figures 5.9 and 5.10, and Appendices C.8 through C.11, the nodes/circles that contain the K65R mutation are apparently larger than the internal nodes they originated from, most of them that already contained the K103N+L100I double mutant, which further appears to give credit to the theory.

However, as stated in chapter 5.5, despite the increase in frequency of patients with the K103N+L100I double mutant associated with cluster 5 over the years, the frequency of patients acquiring immunodeficiency or severe immunodeficiency decreased. This is curious, given that the new treatment scheme (3TC+EFV+TDF), even while including TDF and 3TC together, which seems to be a combination of choice for the management of the HIV-1 infection [35], also has TDF in combination with EFV, which has been shown to be highly associated with the emergence of the K65R mutation and also to the faster emergence of resistance and viral breakthrough [118]. Notwithstanding, there seems to be a better interaction in the treatment scheme when TDF is present, than when AZT was being used, or maybe some other factors that might have collaborated so that HIV-1-infected individuals can have less immunodeficiency.

Regarding the K103N+V108I double mutant, compared to the K103N+L100I double mutant, is more frequent (30-40% of resistant isolates to only 10%), but it is 25-30 fold less resistant than K103N+L100I [54]. As shown in [54], which states that, the lower than expected frequency of the K103N+L100I double mutant amongst clinical isolates could be the result of lower relative replication efficiency, despite the higher drug resistance, the almost absence of K103N+L100I-containing viruses in the data pool up until the treatment scheme shift of 2014 can be explained by this reduced replication fitness.

Concerning the V179I mutation, it is usually detected in subtype A or C virus-infected treatment-naïve individuals, and it increases in frequency in treatment-non-naïve patients who use ART containing ETR or RPV [120]. Contrarily to the K103N+V108I, or even the K103N+L100I,

double mutants, it does not confer drug resistance when alone – however, when present with Y181C or Y181V, it enhances drug resistance to some NNRTIs by 3- to 8-fold [120]. Given that in the studied data cohort the V179I mutation that characterizes cluster 4 has close to no interaction with the Y181C mutation (the Y181V mutation was not considered after the threshold for the heat map was set), it appears that this mutation does not have any particular significance in its further study, hence its absence from the constructed phylogenetic trees and the detailed analyses of only the K103N-containing mutation combinations. Regardless, as V179I was included in the protein structure prediction process, it should have also been included in the phylogeny analysis, as it could have possibly had some sort of interaction with the other mutations and their transmission and evolution through the years.

Overall, all the other clusters, including cluster 1 with K103N, cluster 3 with K103N+V108I and cluster 4 with V179I, showed all similar results amongst themselves, but different from the results of cluster 5 with K103N+L100I, with lower CD4+ cell count (all below 500 cell/mm^3, as would be expected for HIV-1 infected individuals) and higher levels of viral load.

Lastly, the protein structure for each of the studied conformations, as well as which ligands bound to which residues, were predicted, to verify if there were any differences, and if yes, which ones, and also whether the ligands that bound to each of them could possibly have any importance for the inhibition of the viruses containing any of these mutations. Protein structure prediction using computational methods has become more and more important over the years because of what it can achieve and in very short amounts of time – usually, protein crystallization is the method employed to see the protein's structure at an atomic level, to check interactions between ligand and ligand binding residue, and also to make future predictions in the world of drug discovery. However, protein crystallization is an arduous and time labouring job, making its cost particularly high, preventing the crystallization of more proteins, hence the discrepancy between the number of sequenced proteins and their respective 3D structures. Thus, protein structure prediction methods and programs like CASP are crucial to keep being developed, to narrow this discrepancy more and more.

In a general note, all the conformations had T27, or the NNRTI rilpivirine (RPV), as one of the most common ligands, which was not used in the data cohort. Both ETR and RPV, which are DAPY compounds, were first designed to be less bulky and more flexible, and have the ability to adapt better to resistance mutations in the NNIBP, allowing them to effectively inhibit both WT and resistant HIV-1 viruses [121]. Susceptibility to RPV, in particular, does not decrease with the presence of the K103N mutation, which is one of the most common DRMs, but it does select for the L100I, V108I and V179I mutations [122]. However, as both L100I and V108I are accessory mutations and rarely appear alone/not in the presence of K103N, even if, for example, EFV was previously used and effectively selected for K103N, RPV has the ability to work against the resistance provided by this mutation, which limits the further development of the K103N+L100I and K103N+V108I double mutants and, consequently, the evolution of this into the K103N+L100I+K65R triple mutant, which is already known to be particularly critical and dangerous.

This way, since T27, or RPV, is a ligand that binds in a positive manner to all the studied conformations, it might be advantageous for RPV to be a substitute for EFV, which in fact selects for two of the mutations present in the clinically critical K103N+L100I+K65R triple mutant.

Regarding the K103N+L100I double mutant, it had the N443 residue as an additional putative ligand binding residue, unlike the other conformations, which is a part of the DEDD motif in the RT RNase H active site. To this residue, there were four different binding ligands that didn't bind to any other residue in any other conformation: QID, BA5, Y55 and ON1.

Ribonuclease H (RNase H), which works together with the RT polymerase in the process of generating double-stranded DNA from the single-stranded RNA genome, degrades the RNA component of the RNA/DNA hybrid duplex formed during reverse transcription [123, 124]. Its tertiary structure is similar to all known RNase H enzymes, like the human RNase H1 enzyme,

and its active site contains four spatially conserved carboxylate residues (D443, E478, D498 and D549) located in a cavity that also includes the essential to catalysis and too highly conserved among diverse organisms H539 residue [123, 125, 126]. The DEDD motif forms two-metal binding pockets of Mg^{2+} cations that are required for the nucleotydil transfer reaction involved in the cleavage of the RNA phosphate backbone during the reverse transcription process [123, 124, 126].

Despite the serious efforts, the assays used to assess the RNase H activity are too time-consuming, thus there have only been a handful of "drug-like" molecules that have been described and can potentially inhibit RNase H activity, called RNase H inhibitors (RNHIs) – however, none of them have ever been approved [123, 124]. Yet, RNHI have great potential to prevent the replication of the HIV-1 virus, considering that, as none of the major mutations that are regularly associated with resistance to ART are found in the RNase H domain, RNHIs can specifically bind to or near the RNase H active site and still retain inhibition potency against clinically significant drug-resistant HIV viruses, like the ones that have been studied in this dissertation [123].

One of these potentially successful RNHIs is QID, or 3-hydroxy-6-(phenylsulfonyl)-2,4(1H,3H)-dione, which is a RNHI of the N-hydroxyquinazolinediones class and it inhibits the RNase H activity by chelating divalent metals in the RNase H active site [125]. It achieves this due to the fact that it has a pendant group that engages the conserved H539 residue via potential π-π interactions, increasing the activity against RNase H and reaching a significantly high level of inhibition potency, which also happens with ON1 [125]. Moreover, unlike other RNHIs, it does not have any activity against the RT polymerase active site, as it can't bind to it [125].

BA5, or XZ462, or methyl 4-azanyl-1-oxidanyl-2-oxidanylidene-1,8-naphthyvidine-3-carboxylate, inhibits RNase H activity by binding to its active site and reducing the amount of cleavage performed by this enzyme [124]. Overall, it shows a reduced inhibition potency [124]. Considering both NRTI- and NNRTI-associated mutations, while BA5 had reduced potency for NRTI-associated mutations K70R and M184V-containing viruses, which affected the susceptibility of viral replication to nucleoside analogs, the NNRTI-associated K103N+L100I double mutant did not, contrarily, cause a reduction in susceptibility of the virus to BA5, which, in turn, supports the idea that BA5 inhibits viral replication by binding to the RNase H active site, and not to the polymerase active site [124].

Y55, or 7-(furan-2-yl)-2-hydroxyisoquinoline-1,3(2H,4H)-dione, belongs to the RNHI N-hydroxyimide class and is based on inhibitors of the influenza virus endonuclease [123]. They can bind to the RH active site in multiple conformations, meaning they can inhibit RH with multiple mechanisms of action, which makes it easier to escape single-point mutations [127]. They inhibit RNase H given that the position and angle of the 3 oxygen atoms in its moiety are such that they mimic the RH active site metal ion interaction with the substrate during catalysis [123]. Moreover, they can inhibit both RH and polymerase activity, selectively inhibiting RH activity over polymerase activity, where Y55 is a noncompetitive inhibitor of RT polymerase for its nucleic acid substrate, and a competitive inhibitor of RT RH activity for the RNA/DNA substrate [127].

ON1, or 2-(3,4-dichlorobenzyl)-5,6-dihydroxy-pyrimidine-4-carboxylic acid, which has been explored as a hepatitis C (HCV) inhibitor [125], is a RNHI of the pyrimidinol carboxylic acid derivatives class that inhibits RNase H activity by the customization of dual metal chelation via modification of the acidity of the central hydroxyl group, where the metal ions are coordinated by its two phenolic oxygen atoms [128]. As QID, it doesn't bind to the polymerase active site, which means it doesn't have any inhibition activity towards the RT polymerase [125]. Generally, it showed a moderately high potency against RNase H activity – however, QID is more potent [125, 128]. Moreover, all pyrimidinol carboxylic acids exhibit minimal to no inhibitory activity against the human RNase H enzyme, consistent with good selectivity for the viral enzyme [128].

This way, there are potentially four different RNHI that can inhibit the K103N+L100I-containing virus' replication function, even if most of them, like BA5, QID and ON1, don't bind

to the polymerase active site where these residues are on. Especially, as BA5 has already been tested for viruses with the K103N+L100I double mutant, despite the fact that this double mutant does not decrease the low potency BA5 shows, it can be a potential RNHI candidate for viruses with K103N+L100I, which would help decrease the effect brought on by the transmission rates given by the K65R mutation, when together with the double mutant, and the high viral load and low CD4+ cell count it brings. However, the other three RNHIs ought to be tested in viruses with the K103N+L100I double mutation, as well as other mutation combinations, in order to determine whether they may also be good options.

7 CONCLUSIONS AND FUTURE PERSPECTIVES

Brazil, being a large and densely populated country, has a significant number of people living with HIV, who rely on antiretroviral therapy (ART). It has a long history of HIV treatment and its dominated by the two most relevant HIV-1 subtypes worldwide, B and C. This unique setting provides an excellent opportunity to study drug resistance mutations (DRMs) and the patterns of DRMs that arise upon treatment failure. Monitoring this setting is of the upmost relevance.

The K103N+L100I double mutant was highlighted in this work to be of significant clinical importance due to its association with high viral load and low CD4+ cell count, particularly after the treatment shift from AZT to TDF. This shift favoured and selected for the appearance of the K65R mutation, which, due to elusive factors, might have contributed to an increase in the frequency of the K103N+L100I double mutant over the years. This increase is a matter of great concern.

However, the frequency of people with severe immunodeficiency that was associated with the infection by this double mutant decreased between 2008 and 2017. Despite this, the possibility of the increase of the prevalence of K103N+L100I-containing viruses, with the also observed increase of the frequency of more aggressive viral loads, is concerning, since it can reduce the durability of the effect observed of the improvement of the clinical immunodeficiency condition.

In conclusion, the findings of this thesis highlight the importance of vigilance in the transmission of drug-resistant HIV-1 mutants, particularly the aggressive K103N+L100I double mutant. The study also emphasizes the need for regular testing for different DRMs upon HIV-1 diagnosis to administer the most effective treatment. Furthermore, the research suggests that RNase H inhibitors, such as BA5, ON1, QID, and Y55, may be a promising new generation of HIV-1 therapy, as they inhibit the RNase H activity in the reverse transcriptase of the virus, thereby preventing its replication and potentially blocking the increase in viral load. Additionally, the computational approach, using tools such as DiffDock, may be a cost- and time-efficient method to test the binding of various ligands to the RT structure and its mutations.

Moreover, this study underscores the importance of analysing and studying all DRMs associated with HIV-1 RT or other DRMs in general, as other critical mutations may impact resistance to different antiretroviral drugs, such as the K65R mutation. This comprehensive approach can ultimately lead to the development of more effective treatments and, hopefully, the elimination of resistance mutations.

In summary, this thesis provides valuable insights into the emergence and transmission of drug-resistant HIV-1 mutants and highlights the need for continued vigilance and research into new and innovative treatment approaches to address this pressing global health concern.

REFERENCES

[1] Rambaut A, Posada D, Crandall KA, Holmes EC. 2004. The causes and consequences of HIV evolution. *Nature Reviews Genetics* 5(1):52-61. (doi: 10.1038/nrg1246)

[2] UNAIDS DATA 2022. Geneva: Joint United Nations Programme on HIV/AIDS; 2022. Licence: CC BY-NC-SA 3.0 IGO.

[3] Santos-Pereira A, Triunfante V, Araújo PMM, Martins J, Soares H, Poveda E, Souto B, Osório NS. 2021. Nationwide Study of Drug Resistance Mutations in HIV-1 Infected Individuals under Antiretroviral Therapy in Brazil. *International Journal of Molecular Sciences* 22(10):5304. (doi: 10.3390/ijms22105304)

[4] Benzaken AS, Pereira GFM, Costa L, Tanuri A, Santos AF, Soares MA. 2019. Antiretroviral treatment, government policy and economy of HIV/AIDS in Brazil: is it time for HIV cure in the country?. *AIDS Research and Therapy* 16(1):19. (doi: 10.1186/s12981-019-0234-2)

[5] Keele BF, Heuverswyn FV, Li Y, Bailes E, Takehisa J, Santiago ML, Bibollet-Ruche F, Chen Y, Wain LV, Liegeois F, Loul S, Ngole EM, Bienvenue Y, Delaporte E, Brookfield JFY, Sharp PM, Shaw GM, Peeters M, Hahn BH. 2006. Chimpanzee Reservoirs of Pandemic and Nonpandemic HIV-1. *Science* 313(5786):523-526. (doi: 10.1126/science.1126531)

[6] Heuverswyn FV and Peeters M. 2007. The Origins of HIV and Implications for the Global Epidemic. *Current Infectious Disease Reports* 9(4):338-346. (doi: 10.1007/s11908-007-0052-x)

[7] Faria NR, Rambaut A, Suchard MA, Baele G, Bedford T, Ward MJ, Tatem AJ, Sousa JD, Arinaminpathy N, Pépin J, Posada D, Peeters M, Pybus OG, Lemey P. 2014. The early spread and epidemic ignition of HIV-1 in human populations. *Science* 346(6205):56-61. (doi: 10.1126/science.1256739)

[8] Hemellar J. The origin and diversity of the HIV-1 pandemic. 2012. *Trends in Molecular Medicine* 18(3):182-192. (doi: 10.1016/j.molmed.2011.12.001)

[9] Nyamweya S, Hegedus A, Jaye A, Rowland-Jones S, Flanagan KL, Macallan DC. 2013. Comparing HIV-1 and HIV-1 infection: Lessons for viral immunopathogenesis. *Reviews in Medical Virology* 23(4):221-240. (doi: 10.1002/rmv.1739)

[10] Visseaux B, Damond F, Matheron S, Descamps D, Charpentier C. 2016. Hiv-2 molecular epidemiology. *Infection, Genetics and Evolution* 16:233-240. (doi: 10.1016/j.meegid.2016.08.010)

[11] Bbosa N, Kaleebu P, Ssemwanga D. 2019. HIV subtype diversity worldwide. *Current Opinion in HIV and AIDS* 14(3):153-160. (doi: 10.1097/COH.0000000000000534)

[12] Abecasis A, Lemey P, Vidal N, Oliveira T, Peeters M, Camacho R, Shapiro B, Rambaut A, Vandamme A-M. 2007. Recombination Confound the Early Evolutionary History of Human Immunodeficiency Virus Type 1: Subtype G Is a Circulating Recombinant Form. *Journal of Virology* 81(16):8543-8551. (doi: 10.1128/JVI.00463-07)

[13] Plantier J-C, Leoz M, Dickerson JE, Oliveira FD, Cordonnier F, Lemée V, Damond F, Robertson DL, Simon F. 2009. A new human immunodeficiency virus derived from gorillas. *Nature Medicine* 15(8):871-872. (doi: 10.1038/nm.2016)

[14] Hemelaar J, Elangovan R, Yun J, Dickson-Tetteh L, Fleminger I, Kirtley S, Williams B, Gouws-Williams E, Ghys PD. 2019. Global and regional molecular epidemiology of HIV-1, 1990-2015: a systematic review, global survey, and trend analysis. *The Lancet Infectious Diseases* 19(2):143-155. (doi: 10.1016/S1473-3099(18)30647-9)

[15] Santos-Pereira A, Magalhães C, Araújo PMM, Osório NS. 2021. Evolutionary Genetics of Mycobacterium Tuberculosis and HIV-1: "The Tortoise and the Hare". *Microorganisms* 9(1):147. (doi: 10.3390/microorganisms9010147)

[16] Fanales-Belasio E, Raimondo M, Suligoi B, Buttò S. 2010. HIV virology and pathogenetic mechanisms of infection: a brief overview. *Annali dell'Istituto Superiore di Sanità* 46(1):5-14. (doi: 10.4415/ANN_10_01_02)

[17] Ganser-Pornilllos BK, Yeager M, Pornillos O. 2012. Assembly and Architecture of HIV. *Advances in Experimental Medicine and Biology* 726:441-465. (doi: 10.1007/978-1-4614-0980-9_20)

[18] Turner BG and Summer MF. 1999. Structural Biology of HIV. *Journal of Molecular Biology* 285(1):1-32. (doi: 10.1006/jmbi.1998.2354)

[19] Srivastava A, Birari V, Sinha S. 2020. Small Conformational Changes Underlie Evolution of Resistance to NNRTI in HIV Reverse Transcriptase. *Biophysical Journal* 118(10):2489-2501. (doi: 10.1016/j.bpj.2020.04.008)

[20] Das B, Kutsal M, Das R. 2022. Effective prediction of drug-target interaction on HIV using deep graph neural networks. *Chemometrics and Intelligent Laboratory Systems* 230:104676. (doi: 10.1016/j.chemolab.2022.104676)

[21] Li G, Piampongsant S, Faria NR, Voet A, Pineda-Peña A-C, Khouri R, Lemey P, Vandamme A-M, Theys K. 2015. An integrated map of HIV genome-wide variation from a population perspective. *Retrovirology* 12:18. (doi: 10.1186/s12977-015-0148-6)

[22] Kamil R, Debnath U, Verma S, Prabhakar YS. 2018. Identification of Adjacent NNRTI Binding Pocket in Multi-mutated HIV1-RT Enzyme Model: An *in silico* Study. *Current HIV Research* 16(2):121-129. (doi: 10.2174/1570162X16666180412165004)

[23] Feinberg MB and Greene WC. 1992. Molecular insights into human immunodeficiency virus type 1 pathogenesis. *Current Opinion in Immunology* 4(4):466-474. (doi: 10.1016/s0952-7915(06)80041-5)

[24] Rossi E, Meuser ME, Cunanan CJ, Cocklin S. 2021. Structure, Function and Interactions of the HIV-1 Capsid Protein. *Life* 11(2):100. (doi: 10.3390/life11020100)

[25] Paranjape RS. 2005. Immunopathogenesis of HIV infection. *Indian Journal of Medical Research* 121(4):240-255.

[26] Stekler J and Collier AC. 2004. Primary HIV Infection. *Current HIV/AIDS Reports* 1(2):68-73. (doi: 10.1007/s11904-004-0010-2)

[27] Lui Y, Li H, Wang X, Han J, Jia L, Li T, Li J, Li L. 2020. Natural presence of V179E and rising prevalence of E138G in HIV-1 reverse transcriptase in CRF55_01B viruses. *Infection, Genetics and Evolution* 77:104098. (doi: 10.1016/j.meegid.2019.104098)

[28] Steiner MC, Gibson KM, Crandall KA. 2020. Drug Resistance Prediction Using Deep Learning Techniques on HIV-1 Sequence Data. *Viruses* 12(5):560. (doi: 10.3390/v12050560)

[29] Alcaro S, Alteri C, Artese A, Ceccherini-Silberstein F, Costa G, Ortuso F, Bertoli A, Forbici F, Santoro MM, Parrotta L, Flandre P, Masquelier B, Descamps D, Calvez V, Marcelin A-G, Perno CF, Sing T, Svicher V. 2011. Docking Analysis and Resistance Evaluation of Clinically Relevant Mutations Associated with the HIV-1 Non-nucleoside Reverse Transcriptase Inhibitors Nevirapine, Efavirenz and Etravirine. *ChemMedChem* 6(12):2203-2213. (doi: 10.1002/cmdc.201100362)

[30] Sharaf NG, Ishima R, Gronenborn AM. 2016. Conformational plasticity of the NNRTI-binding pocket in HIV-1 reverse transcriptase – A fluorine NMR study. *Biochemistry* 55(28):3864-3873. (doi: 10.1021/acs.biochem.6b00113)

[31] Pata JD, Stirtan WG, Goldstein SW, Steitz TA. 2004. Structure of HIV-1 reverse transcriptase bound to an inhibitor active against mutant reverse transcriptases resistance to other nonnucleoside inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* 101(29):10548-10553. (doi: 10.1073/pnas.0404151101)

[32] Sarafianos SG, Marchand B, Das K, Himmel D, Parniak MA, Hughes SH, Arnold E. 2009. Structure and function of HIV-1 reverse transcriptase: molecular mechanisms of polimerization and inhibition. *Journal of Molecular Biology* 385(3):693-713. (doi: 10.1016/j.jmb.2008.10.071)

[33] Wang J, Dykes C, Domaoal RA, Koval CE, Bambara RA, Demeter LM. The HIV-1 reverse transcriptase mutants G190S and G190A, which confer resistance to non-nucleoside reverse transcriptase inhibitors, demonstrate reductions in RNase H activity and DNA synthesis from $tRNA^{Lys,3}$ that correlate with reductions in replication efficiency. *Virology* 348(2):462-474. (doi: 10.1016/j.virol.2006.01.014)

[34] Tachedjian G, Orlova M, Sarafianos SG, Arnold E, Goff SP. 2001. Nonnucleoside reverse transcriptase inhibitors are chemical enhancers of dimerization of the HIV type 1 reverse transcriptase. *Proceedings of the National Academy of Sciences of the United States of America* 98(13):7188-7193. (doi: 10.1073/pnas.121055998)

[35] Brenner BG and Coutsinos D. 2009. The K65R mutation in HIV-1 reverse transcriptase: genetic barriers, resistance profile and clinical implications. *HIV Therapy* 3(6):583-594. (doi: 10.2217/hiv.09.40)

[36] Das K, Bandwar RP, White KL, Feng JY, Sarafianos SG, Tuske S, Tu X, Clark Jr. AD, Boyer PL, Hou X, Gaffney BL, Jones RA, Miller MD, Hughes SH, Arnold E. 2009. Structural Basis for the Role of the K65R Mutation in HIV-1 Reverse Transcriptase Polymerization, Excision Antagonism, and Tenofovir Resistance. *The Journal of Biological Chemistry* 284(50):35092–35100. (doi: 10.1074/jbc.M109.022525)

[37] Tang MW and Shafer RW. 2012. HIV-1 Antiretroviral Resistance: scientific principles and clinical applications. *Drugs* 72(9):e1-e25. (doi: 10.2165/11633630-000000000-00000)

[38] Reuman EC, Rhee S-Y, Holmes SP, Shafer RW. 2010. Constrained patterns of covariation and clustering of HIV-1 non-nucleoside reverse transcriptase inhibitor resistance mutations. *Journal of Antimicrobial Chemotherapy* 65(7):1477-1485. (doi: 10.1093/jac/dkq140)

[39] Geitmann M, Unge T, Danielson UH. 2006. Interaction Kinetic Characterization of HIV-1 Reverse Transcriptase Non-Nucleoside Inhibitor Resistance. *Journal of Medicinal Chemistry* 49(8):2375-2387. (doi: 10.1021/jm0504050)

[40] Das K, Clark Jr AD, Lewi PJ, Heeres K, De Jonge MR, Koymans LMH, Vinkers HM, Daeyaert F, Ludovici DW, Kukla MJ, De Corte B, Kavash RW, Ho CY, Ye H, Lichtenstein MA, Andries K, Pauwels R, De Bethúne M-P, Boyer PL, Clark P, Hughes SH, Janssen PAJ, Arnold

E. 2004. Roles of conformational and positional adaptability in structure-based design of TMC125-R165335 (etravirine) and related non-nucleoside reverse transcriptase inhibitors that are highly potent and effective against wild-type and drug-resistant HIV-1 variants. *Journal of Medicinal Chemistry* 47(10):2550-2560. (doi: 10.1021/jm030558s)

[41] Simon V, Ho DD, Karim QA. 2006. HIV/AIDS epidemiology, pathogenesis, prevention, and treatment. *Lancet* 368(9534):489-504. (doi: 10.1016/S0140-6736(06)69157-5)

[42] Clutter DS, Jordan MR, Bertagnolio S, Shafer RW. 2016. HIV-1 Drug Resistance and Resistance Testing. *Infection, Genetics and Evolution* 46:292-307. (doi: 10.1016/j.meegid.2016.08.031)

[43] Thompson MA, Aberg JA, Cahn P, Montaner JSG, Rizzardini G, Telenti A, Gatell JM, Günthard HF, Hammer SM, Hirsch MS, Jacobsen DM, Reiss P, Richman DD, Volberding PA, Yeni P, Schooley RT. 2010. Antiretroviral Treatment of Adult HIV Infection: 2010 Recommendations of the International AIDS Society-USA Panel. *The Journal of the American Medical Association* 304(3):321-333. (doi: 10.1001/jama.2010.1004)

[44] Uhlmann EJ, Tebas P, Storch GA, Powderly WG, Lie YS, Whitcomb JM, Hellman NS, Arens MQ. 2004. Effects of the G190A substitution of HIV reverse transcriptase on phenotypic susceptibility of patient isolates to delavirdine. *Journal of Clinical Virology* 31(3):198-203. (doi: 10.1016/j.jcv.2004.03.012)

[45] Guidelines for the Use of Antiretroviral Agents in Adults and Adolescents with HIV. Department of Health and Human Services. 2023. Available at https://clinicalinfo.hiv.gov/en/guidelines/adult-and-adolescent-arv. Accessed 27/07/2023.

[46] Wang J, Zhang G, Bambara RA, Li D, Liang H, Wu H, Smith HM, Lowe NR, Demeter LM, Dykes C. 2011. Nonnucleoside Reverse Transcriptase Inhibitor-Resistant HIV Is Stimulated by Efavirenz during Early Stages of Infection. *Journal of Virology* 85(20):10861-10873. (doi: 10.1128/JVI.05116-11)

[47] Wang J, Liang H, Bacheler L, Wu H, Deriziotis K, Demeter LM, Dykes C. 2010. The non-nucleoside reverse transcriptase inhibitor efavirenz stimulates replication of human immunodeficiency virus type 1 harboring certain non-nucleoside resistance mutations. *Virology* 402(2):228-237. (doi: 10.1016/j.virol.2010.03.018)

[48] Blassel L, Zhukova A, Villabona-Arenas CJ, Atkins KE, Hué S, Gascuel O. 2021. Drug resistance mutations in HIV: new bioinformatics approaches and challenges. *Current Opinion in Virology* 51:56-64. (doi: 10.1016/j.coviro.2021.09.009)

[49] Charpentier C, Hingrat QL, Ferré VM, Damond F, Descamps D. 2023. Future of Antiretroviral Drugs and Evolution of HIV-1 Drug Resistance. *Viruses* 15(2):540. (doi: 10.3390/v15020540)

[50] Maga G, Amacker M, Ruel N, Hübscher U, Spadari S. 1997. Resistance to Nevirapine of HIV-1 Reverse Transcriptase Mutants: Loss of Stabilizing Interactions and Thermodynamic or Steric Barriers are Induced by Different Single Amino Acid Substitutions. *Journal of Molecular Biology* 274(5):738-747. (doi: 10.1006/jmbi.1997.1427)

[51] Shafer RW and Schapiro JM. 2008. HIV-1 Drug Resistance Mutations: and Updated Framework for the Second Decade of HAART. *AIDS Reviews* 10(2):67-84.

[52] Rhee S-Y, Gonzales MJ, Kantor R, Betts BJ, Ravela J, Shafer RW. 2003. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Research* 31(1):298-303. (doi: 10.1093/nar/gkg100)

[53] Shafer RW. 2006. Rationale and uses of a public HIV drug-resistance database. *The Journal of Infectious Diseases* 194(Suppl 1):S51-58. (doi: 10.1086/505356)

[54] Koval CE, Dykes C, Wang J, Demeter LM. 2006. Relative replication fitness of efavirenz-resistant mutants of HIV-1: Correlation with frequency during clinical therapy and evidence of compensation for the reduced fitness of K103N+L100I by the nucleoside resistance mutation L74V. *Virology* 353(1):184-192. (doi: 10.1016/j.virol.2006.05.021)

[55] Alashwal H, Halaby ME, Crouse JJ, Abdalla A, Moustafa AA. 2019. The Application of Unsupervised Clustering Methods to Alzheimer's Disease. *Frontiers in Computational Neuroscience* 13:31. (doi: 10.3389/fncom.2019.00031)

[56] Dridi, S. 2022. Unsupervised Learning - A Systematic Literature Review. OSF Preprints. April 4. (doi: 10.31219/osf.io/kpqr6)

[57] Greene, D, Cunningham, P, Mayer, R. 2008. Unsupervised Learning and Clustering. In: Cord, M, Cunningham, P. (eds) Machine Learning Techniques for Multimedia. Cognitive Technologies. Springer, Berlin, Heidelberg. (doi: 10.1007/978-3-540-75171-7_3)

[58] Gülağız, F and Şahin, S. 2017. Comparison of Hierarchical and Non-Hierarchical Clustering Algorithms. *International Journal of Computer Engineering and Information Technology* 9(1):6-14.

[59] Rani Y and Rohil H. 2013. A Study of Hierarchical Clustering Algorithm. *International Journal of Information and Computation Technology* 3(11):1225-1232.

[60] Xu R and Wunsch D. 2005. Survey of Clustering Algorithms. *IEEE Transactions on Neural Networks* 16(3):645-678. (doi: 10.1109/TNN.2005.845141)

[61] Lu H, Halappanavar M, Kalyanaraman A. 2014. Parallel Heuristics for Scalable Community Detection. *Parallel Computing* 47:19-37. (doi: 10.1016/j.parco.2015.03.003)

[62] Traag VA, Waltman L, Van Eck NJ. 2019. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* 9(1):5233. (doi: 10.1038/s41598-019-41695-z)

[63] Blondel VD, Guillaume J-P, Lambiotte R, Lefebvre E. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Methods: Theory and Experiment* (arXiv:0803.0476v2) (doi: 10.1088/1742-5468/2008/10/P10008)

[64] Motschnig N, Ramharter A, Schweiger O, Zabka P, Foerster K-T. 2021. On Comparing and Enhancing Two Common Approaches to Network Community Detection. *IEEE Global Communications Conference (GLOBECOM)* 1-6. (doi: 10.1109/GLOBECOM46510.2021.9685248)

[65] Velliangiri S, Alagumuthurkrishnan A, Joseph SIT. 2019. A Review of Dimensionality Reduction Techniques for Efficient Computation. *Procedia Computer Science* 165:104-111. (doi: 10.1016/j.procs.2020.01.079)

[66] McInnes L, Healy J, Melville J. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. (arXiv:1802.03426v3) (doi: 10.48550/arXiv.1802.03426)

[67] Van Der Maaten LJP, Postma EO, Van Der Herik HJ. 2009. Dimensionality Reduction: A Comparative Review. *Journal of Machine Learning Research* 10:1-41.

[68] De Backer S, Naud A, Scheunders P. 1998. Non-linear dimensionality reduction techniques for unsupervised feature extraction. *Pattern Recognition Letters* 19(8):711-720. (doi: 10.1016/S0167-8655(98)00049-X)

[69] Arunasakthi K and KamatchiPriya L. 2014. A REVIEW ON LINEAR AND NON-LINEAR DIMENSIONALITY REDUCTION TECHNIQUES. *Machine Learning and Applications: An International Journal (MLAIJ)* 1(1):65-76.

[70] Ayesha S, Hanif MK, Talib R. 2020. Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion* 59:44-58. (doi: 10.1016/j.inffus.2020.01.005)

[71] Diaz-Papkovich A, Anderson-Trocmé L, Gravel S. 2020. A review of UMAP in population genetics. *Journal of Human Genetics* 66(1):85-91. (doi: 10.1038/s10038-020-00851-4)

[72] Kobak D and Linderman GC. 2019. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nature Biotechnology* 39(2):156-157. (doi: 10.1038/s41587-020-00809-z)

[73] Chernomor O, Von Haeseler A, Minh BQ. 2016. Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Systematic Biology* 65(6):997-1008. (doi: 10-1093/sysbio/syw037)

[74] Kubatko LS. 2008. "Inference of Phylogenetic Trees" Lectures Notes in Mathematics, edited by Morel JM, Takens F, Teissier B. *Springer*. 1-38.

[75] Truszkowski J and Goldman N. 2016. Maximum Likelihood Phylogenetic Inference is Consistent on Multiple Sequence Alignments, with or without Gaps. *Systematic Biology* 65(2):328-333. (doi: 10.1093/sysbio/syv089)

[76] Masters JC and Pozzi L. 2017. Phylogenetic Inference. *The International Encyclopedia of Primatology*. (doi: 10.1002/9781119179313.wbprim0419)

[77] O'Meara BC. 2012. Evolutionary Inferences from Phylogenies: A Review of Methods. *Annual Review of Ecology, Evolution, and Systematics* 43:267-285. (doi: 10.1146/annurev-ecolsys-110411-160331)

[78] Soltis DE and Soltis PS. 2003. The Role of Phylogenetics in Comparative Genetics. *Plant Physiology* 132(4):1790-1800. (doi: 10.1104/pp.103.022509)

[79] Munjal G, Hanmandlu M, Srivastava S. Phylogenetic Algorithms and Applications. *Ambient Communications and Computer Systems* 904:187-194. (doi: 10.1007/978-981-13-5934-7_17)

[80] Jin G, Nakhleh L, Snir S, Tuller T. 2006. Maximum likelihood of phylogenetic networks. *Bioinformatics* 22(21):2604-2611. (doi: 10.1093/bioinformatics/btl452)

[81] Irisarri I and Zardoya R. 2013. Phylogenetic Hypothesis Testing. *Encyclopedia of Life Sciences (eLS)*. (doi: 10.1002/9780470015902.a0025163)

[82] Holmes S. 2003. Bootstrapping Phylogenetic Trees: Theory and Methods. *Statistical Science* 18(2):241-255.

[83] Kinene T, Wainaina J, Maina S, Boykin LM. 2016. Rooting Trees, Methods for. *Encyclopedia of Evolutionary Biology* 489-493. (doi: 10.1016/B978-0-12-800049-6.00215-8)

[84] Saitou N and Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4(4):406-25. (doi: 10.1093/oxfordjournals.molbev.a040454)

[85] Davidson R and Del Campo AM. 2020. Combinatorial and Computational Investigations of Neighbor-Joining Bias. *Frontiers in Genetics* 11. (doi: 10.3389/fgene.2020.584785)

[86] Moulton V, Spillner A, Wu T. 2018. UPGMA and the normalized equidistant minimum evolution problem. *Theoretical Computer Science* 721:1-15. (doi: 10.1016/j.tcs.2018.01.022)

[87] Kuhner MK and Felsenstein J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution* 11(3):459-468. (doi: 10.1093/oxfordjournals.molbev.a040126)

[88] Kannan L and Wheeler WC. 2012. Maximum Parsimony on Phylogenetic networks. *Algorithms for Molecular Biology* 7(9). (doi: 10.1186/1748-7188-7-9)

[89] Soltis PS and Soltis DE. 2003. Applying the Bootstrap in Phylogenetic Reconstruction. *Statistical Science*. 18(2):256-267. (doi: 10.1214/ss/1063994980)

[90] Ayoub R and Lee Y. 2020. Protein structure search to support the development of protein structure prediction methods. *Proteins* 89(6):648-658. (doi: 10.1002/prot.26048)

[91] Jisna VA and Jayaraj PB. 2021. Protein Structure Prediction: Conventional and Deep Learning Perspectives. *The Protein Journal* 40(4):522-544. (doi: 10.1007/s10930-021-10003-y)

[92] McGuffin LJ, Adiyaman R, Maghrabi AHA, Shuid AN, Brackenridge DA, Nealon JO, Philomina LS. 2019. IntFOLD: an integrated web resource for high performance protein structure and function prediction. *Nucleic Acids Research* 47(W1):W408-W413. (doi: 10.1093/nar/gkz322)

[93] Laurenzi A, Hung L-H, Samudrala R. 2013. Structure Prediction of Partial-Length Protein Sequences. *International Journal of Molecular Sciences* 14(7):14892-14907. (doi: 10.3390/ijms140714892)

[94] Liu T, Tang GW, Capriotti E. 2011. Comparative Modeling: The State of the Art and Protein Drug Target Structure Prediction. *Combinatorial Chemistry & High Throughput Screening* 14(6):532-547. (doi: 10.2174/138620711795767811)

[95] Levitt M. 1992. Accurate modeling of protein conformation by automatic segment matching. *Journal of Molecular Biology* 226(2):507-533. (doi: 10.1016/0022-2836(92)90964-l)

[96] McGuffin LJ and Roche DB. 2011. Automated tertiary structure prediction with accurate local model quality assessment using the IntFOLD-TS method. *Proteins* 79(10):137-146. (doi: 10.1002/prot.23120)

[97] McGuffin LJ, Shuid AN, Kempster R, Maghrabi AHA, Nealon JO, Salehe BR, Atkins JD, Roche DB. 2018. Accurate template-based modeling in CASP12 using the IntFOLD4-TS, ModFOLD6, and ReFOLD methods. *Proteins* 86(1):335-344. (doi: 10.1002/prot.25360)

[98] Roche DB, Tetchner SJ, McGuffin LJ. 2011. FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins. *BMC Bioinformatics* 12:160. (doi: 0.1186/1471-2105-12-160)

[99] Roche DB, Buenavista MT, McGuffin LJ. 2012. FunFOLDQA: A Quality Assessment Tool for Protein-Ligand Binding Site Residue Predictions. *PLoS ONE* 7(5):e38219. (doi: 10.1371/journal.pone.0038219)

[100] Roche DB, Tetchner SJ, McGuffin LJ. 2010. The binding site distance test score: a robust method for the assessment of predicted protein binding sites. *Bioinformatics* 26(22):2920-2921. (doi: 10.1093/bioinformatics/btq543)

[101] Chikata T, Carlson JM, Tamura Y, Borghan MA, Naruto T, Hashimoto M, Murakoshi H, Le AQ, Mallal S, John M, Gatanaga H, Oka S, Brumme ZL, Takiguchi M. 2014. Host-specific

adaptation of HIV-1 subtype B in the Japanese population. *Journal of Virology* 88(9):4763-4775. (doi: 10.1128/JVI.00147-14)

[102] Katoh K and Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* 30(4):772-780. (doi: 10.1093/molbev/mst010)

[103] Afgan E, Baker D, Van den Beek M, Blankenberg D, Bouvier D, Čech M, Chilton J, Clements D, Coraor N, Eberhard C, Grüning B, Guerler A, Hillman-Jackson J, Von Kuster G, Rasche E, Soranzo N, Turaga N, Taylor J, Nekrutenko A, Goecks J. 2016. The Galaxy platform for accessible, reproducible and collaborative biomedical analysis: 2016 update. *Nucleic Acids Research* 44(W1):W3-W10. (doi: 10.1093/nar/gkw343)

[104] Schneider TD and Stephens RM. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research* 18(20):6097-6100. (doi: 10.1093/nar/18.20.6097)

[105] R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (URL: https://www.R-project.org/)

[106] Wagih O. 2017. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* 33(22):3645-3647. (doi: 10.1093/bioinformatics/btx469)

[107] Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4. (https://ggplot2.tidyverse.org)

[108] Lui TF and Shafer RW. 2006. Web Resources for HIV type 1 Genotypic-Resistance Test Interpretation. *Clinical Infectious Diseases* 42(11):1608-1618. (doi: 10.1086/503914)

[109] Wolf FA, Angerer P, Theis FJ. 2018. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology* 19(15). (doi: 10.1186/s13059-017-1382-0)

[110] Waskom ML. 2021. seaborn: statistical data visualization. *The Journal of Open Source Software* 6(60):3021. (doi: 10.21105/joss.03021)

[111] McGuffin LJ, Edmunds NS, Genc AG, Alharbi SMA, Salehe BR, Adiyaman R. 2023. Prediction of protein structures, functions and interactions using the IntFOLD7, MultiFOLD and ModFOLDdock servers. *Nucleic Acids Research* 51(W1):W274-W280. (doi: 10.1093/nar/gkad297)

[112] Tamura K, Stecher G, Kumar S. 2021. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution* 38(7):3022-3027. (doi: 10.1093/molbev/msab120)

[113] Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, Morris JH, Ferrin TE. 2020. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Science* 30(1):70-82. (doi: 10.1002/pro.3943)

[114] Robin C, Haas J, Gumienny R, Smolisnki A, Tauriello G, Schwede T. 2021. Continuous Automated Model EvaluatiOn (CAMEO)—Perspectives on the future of fully automated evaluation of structure prediction methods. *Proteins: Structure, Function, and Bioinformatics* 89(12):1977-1986. (doi: 10.1002/prot.26213)

[115] Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* 32(1):268-274. (doi: 10.1093/molbev/msu300)

[116] Ishikawa SA, Zhukova A, Iwasaki W, Gascuel O. 2019. A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios. *Molecular Biology and Evolution* 36(9):2069-2085. (doi: 10.1093/molbev/msz131)

[117] García-González C, García-Bujalance S, Ruiz-Carrascoso G, Arribas JR, González-García J, Bernardino JI, Pascual-Pareja JF, Martínez-Prats L, Delgado R, Mingorance J. 2012. Detection and quantification of the K103N mutation in HIV reverse transcriptase by pyrosequencing. *Diagnostic Microbiology and Infectious Disease* 72(1): 90–96. (doi: 10.1016/j.diagmicrobio.2011.09.020)

[118] Von Wyl V, Yerly S, Böni J, Bürgisser P, Klimkait T, Battegay M, Bernasconi E, Cvassini M, Furrer H, Hirschel B, Vernazza PL, Rickenbach M, Lederberger B, Günthard HF. 2008. Factors associated with the emergence of K65R in patients with HIV-1 infection treated with combination antiretroviral therapy containing tenofovir. *Clinical Infectious Diseases* 46(8):1299-1309. (doi: 10.1086/528863)

[119] Parikh UM, Bacheler L, Koontz D, Mellors JW. 2006. The K65R mutation in human immunodeficiency virus type 1 reverse transcriptase exhibits bidirectional phenotypic antagonism with thymidine analog mutations. *Journal of Virology* 80(10):4971-4977. (doi: 10.1128/JVI.80.10.4971-4977.2006)

[120] Dwivedi R, Wang Y, Kline C, Fischer DK, Ambrose Z. 2022. APOBEC3 selects V179I in HIV-1 reverse transcriptase to provide selective advantage for non-nucleoside reverse transcriptase inhibitor-resistance mutants. *Frontiers in Virology* 2:919825. (doi: 10.3389/fviro.2022.919825)

[121] Smith SJ, Pauly GT, Akram A, Melody K, Rai G, Maloney DJ, Ambrose Z, Thomas CJ, Schneider JT, Hughes SH. 2016. Rilpivirine analogs potently inhibit drug-resistant HIV-1 mutants. *Retrovirology* 13:11. (doi: 10.1186/s12977-016-0244-2)

[122] Sanford M. 2012. Rilpivirine. *Drugs* 72(4):525-541. (doi: 10.2165/11208590-000000000-00000)

[123] Ilina T, LaBarge K, Sarafianos SG, Ishima R, Parniak MA. 2012. Inhibitors of HIV-1 Reverse Transcriptase—Associated Ribonuclease H Activity. *Biology* 1(3):521-541. (doi: 10.3390/biology1030521)

[124] Boyer PL, Smith SJ, Zhao XZ, Das K, Gruber K, Arnold E, Burke Jr. TR, Hughes SH. 2018. Developing and Evaluating Inhibitors against the RNase H Active Site of HIV-1 Reverse Transcriptase. *Journal of Virology* 92(13):e02203-17. (doi: 10.1128/JVI.02203-17)

[125] Lansdon EB, Liu Q, Leavitt SA, Balakrishnan M, Perry JK, Lancaster-Moyer C, Kutty N, Lui X, Squires NH, Watkins WJ, Kirschberg TA. 2011. Structural and Binding Analysis of Pyrimidinol Carboxylic Acid and *N*-Hydroxy Quinazolinedione HIV-1 RNase H Inhibitors. *Antimicrobial Agents and Chemotherapy* 55(6):2905-2915. (doi: 10.1128/AAC.01594-10)

[126] Álvarez M, Barrioluengo V, Afonso-Lehmann RN, Menéndez-Arias L. 2013. Altered error specificity of RNase H-deficient HIV-1 reverse transcriptases during DNA-dependent DNA synthesis. *Nucleic Acids Research* 41(8):4601-4612. (doi: 10.1093/nar/gkt109)

[127] Kirby KA, Myshakina NA, Christen MT, Chen Y-L, Schmidt HA, Huber AD, Xi Z, Kim S, Rao RK, Kramer ST, Yang Q, Singh K, Parniak MA, Wang Z, Ishima R, Sarafianos SG. 2017. A 2-Hydroxyisoquinoline-1,3-Dione Active-Site RNase H Inhibitor Binds in Multiple Modes to HIV-1 Reverse Transcriptase. *Antimicrobial Agents and Chemotherapy* 61(10):e01351-17. (doi: 10.1128/AAC.01351-17)

[128] Kirschberg TA, Balakrishnan M, Squires NH, Barnes T, Brendza KM, Chen X, Eisenberg EJ, Jin W, Kutty N, Leavitt S, Liclican A, Liu Q, Liu X, Mak J, Perry JK, Wang M, Watkins WJ, Lansdon EB. 2009. RNase H active site inhibitors of human immunodeficiency virus type 1 reverse transcriptase: design, biochemical activity, and structural information. *Journal of Medicinal Chemistry* 52(19):5781-5784. (doi: 10.1021/jm900597q)

APPENDICES

APPENDIX A

UMAP visualizations of all 22,838 sequences from HIV-1 infected individuals in Brazil from 2008 to 2017 based on treatment naivety, sex, age at the time of diagnosis, type of ART scheme used, years of treatment, year of diagnosis and virus subtype

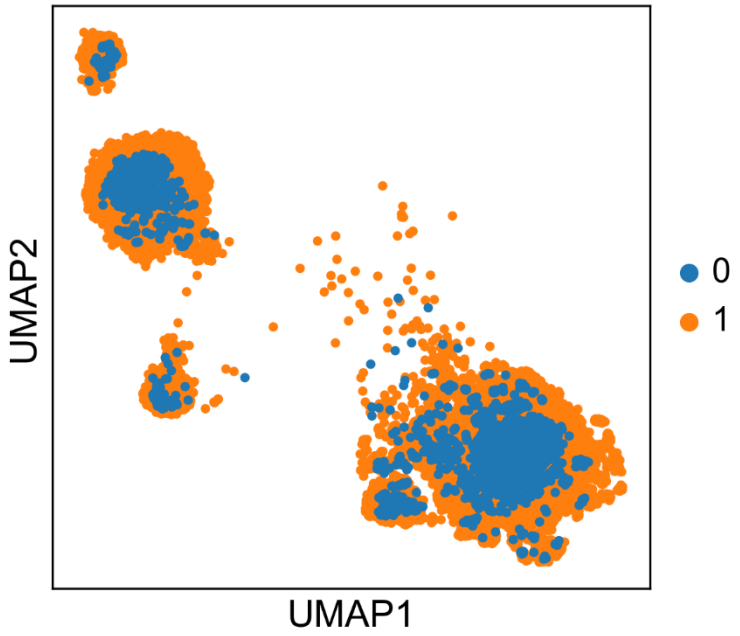


Figure A.1 - UMAP visualization of all 22,838 sequences from HIV-1 infected individuals in Brazil from 2008 to 2017, coloured based on the whether the patient in question was treatment-Naïve (represented by 0 in the blue dots) or treatment-NonNaive (represented by 1 in the orange dots)

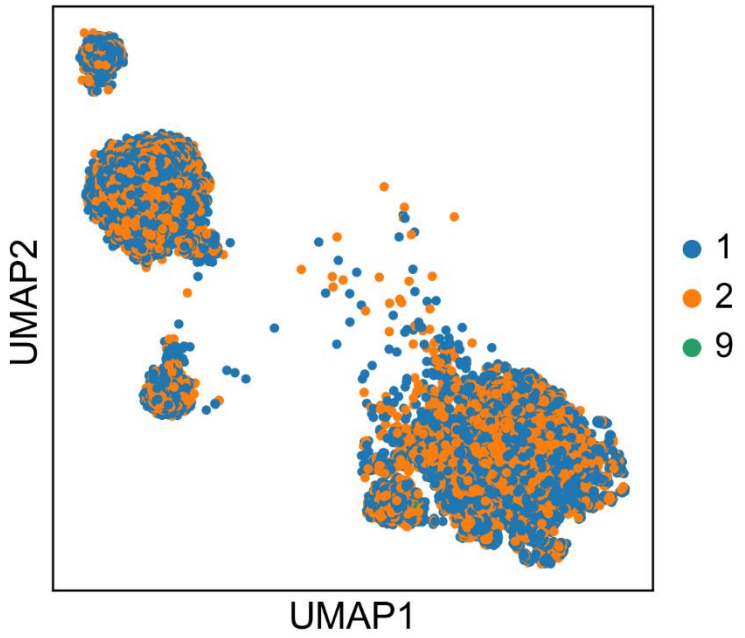


Figure A.2 - UMAP visualization of all 22,838 sequences from HIV-1 infected individuals in Brazil from 2008 to 2017, coloured based on the patient's sex (male is represented by 1 in the blue dots, female is represented by 2 in the orange dots and in the case of this data being unavailable is represented by 9 in the green dots)

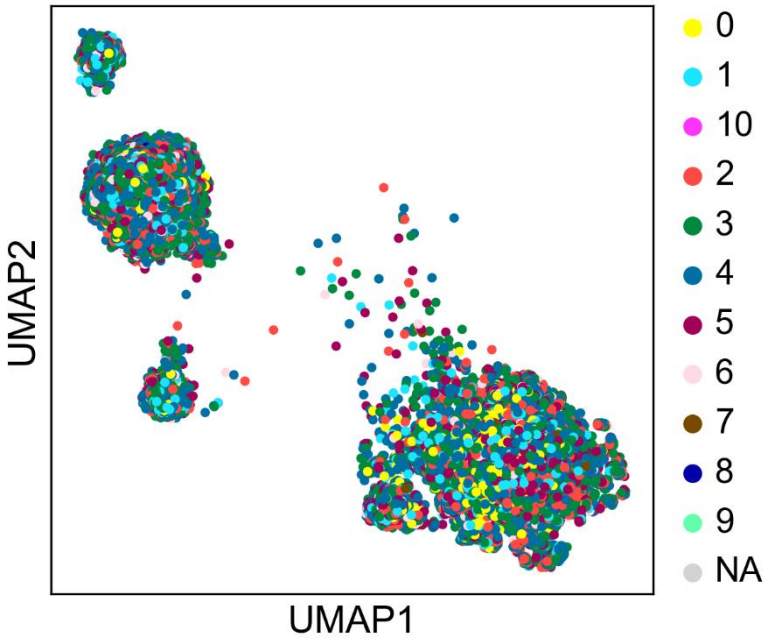


Figure A.3 - UMAP visualization of all 22,838 sequences from HIV-1 infected individuals in Brazil from 2008 to 2017, coloured based on the age of the patient at the time of diagnosis, where each number represents its respective decade, meaning 1 represents patients between the ages of 10 and 19 and so on, and NA represents unavailable data

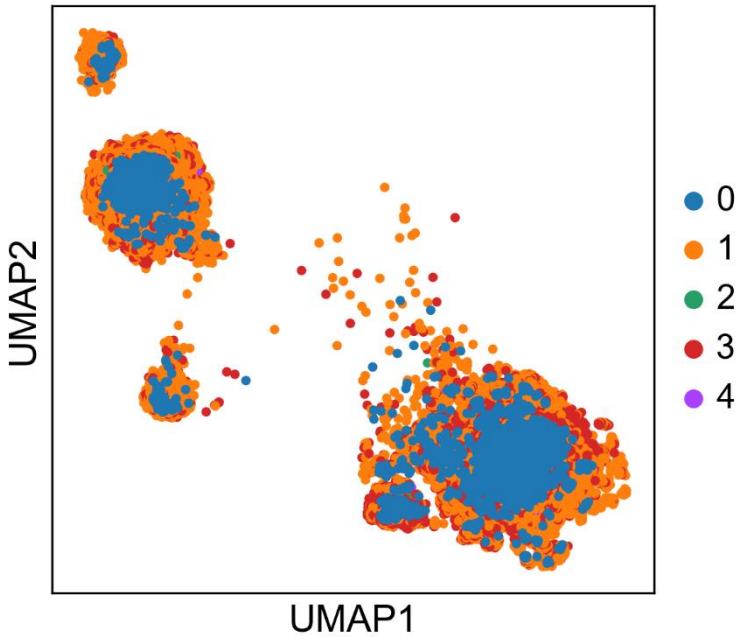


Figure A.4 - UMAP visualization of all 22,838 sequences from HIV-1 infected individuals in Brazil from 2008 to 2017, coloured based on the type of ART scheme used, where 0 in the blue dots represent the treatment-Naïve patients, 1 in the orange dots represent the NNRTI+NRTI treatment schemes, 2 in the green dots represent the only NNRTI treatment schemes, 3 in the red dots represent the only NRTI treatment schemes and 4 in the purple dots represent all other types of treatment schemes

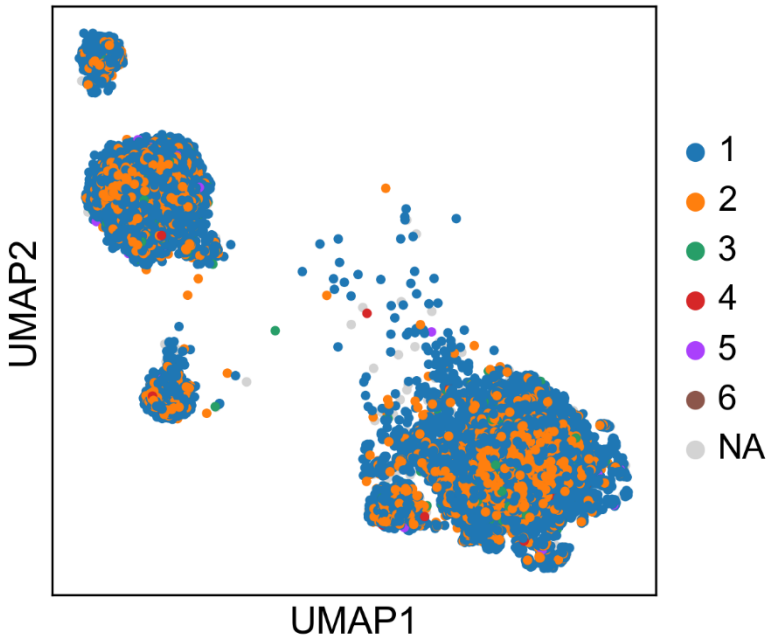


Figure A.5 - UMAP visualization of all 22,838 sequences from HIV-1 infected individuals in Brazil from 2008 to 2017, coloured based on the years of treatment each patient endured, in which 1 in the blue dots represent treatment years between 0 and 5, 2 in the orange dots represent treatment years between 5 and 10, 3 in the green dots represent years of treatment between 10 and 15, 4 in the red dots represent treatment years between 15 and 20, 5 in the purple dots represent treatment years between 20 and 25, 6 in the brown dots represent years of treatment greater than 25 and NA indicates unavailable data

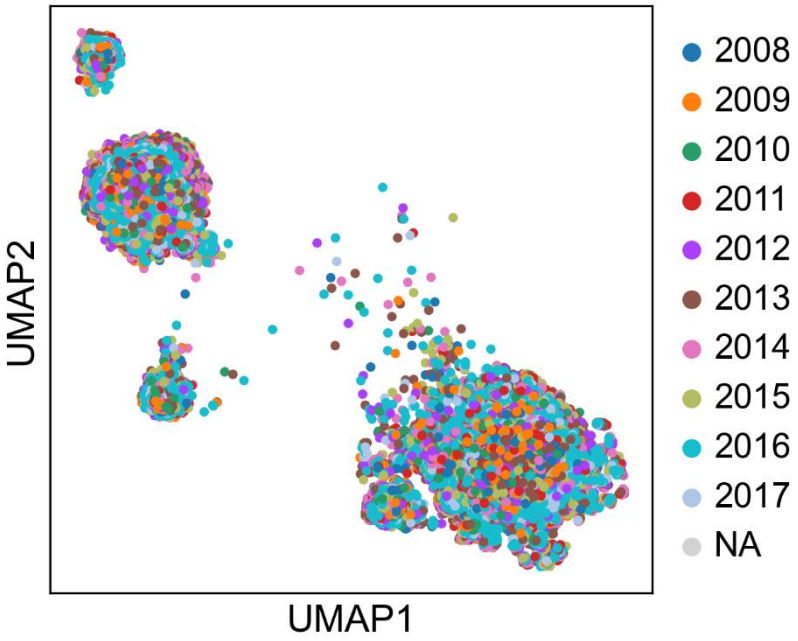


Figure A.6 - UMAP visualization of all 22,838 sequences from HIV-1 infected individuals in Brazil from 2008 to 2017, coloured based on the year of diagnosis – NA represents unavailable data

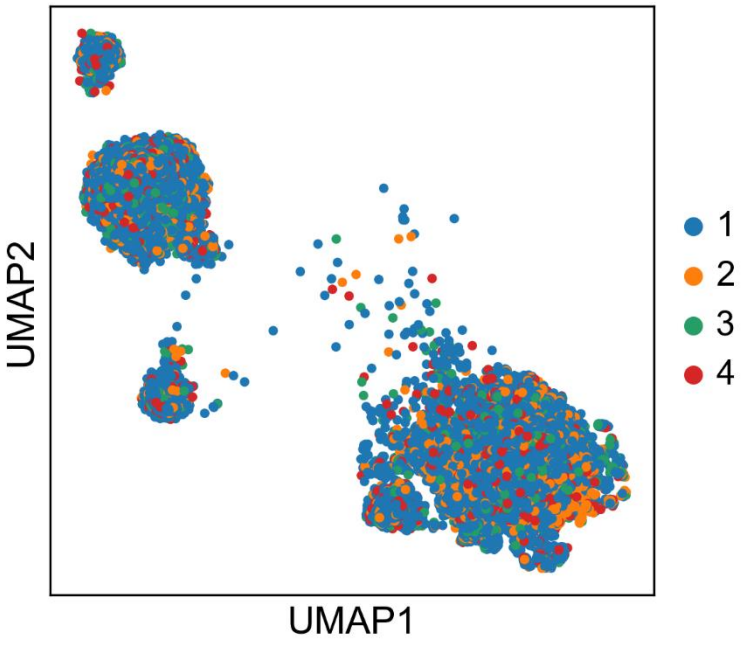


Figure A.7 - UMAP visualization of all 22,838 sequences from HIV-1 infected individuals in Brazil from 2008 to 2017, coloured based on the virus' subtype – 1 in the blue dots represent the B subtype, 2 in the orange dots represent the C subtype, 3 in the green dots represent the F1 subtype and 4 in the red dots represent other subtypes

APPENDIX B

Table of the CAMEO-LAB results showing the presence of each of the possible ligands to each of the ligand binding residues in the four different conformations (K103N, K103N+V108I, V179I and K103N+L100I) plus the wild-type (WT) conformation, as a way of comparison

Table B.1 - Presence of each of the possible 23 ligands to each of the putative ligand binding residues for each of the four conformations (K103N, K103N+V108I, V179I and K103N+L100I) and the wild-type (WT) for pair-wise comparison ("x" indicates that a certain ligand binds to the respective binding residue, whereas the cells highlighted in grey indicate there is no binding)

	1RY	5MA	7N1	BA5	DA	DC	DDG	DG	M9A	MG	MN	NVP	ON1	QID	R8E	RU	T27	T90	Y55	MUTATION
95		x	x						x			x			x	x	x	x		WT
		x							x			x			x	x	x			K103N
		x	x			x			x			x			x	x	x	x		K103N+V108I
		x	x			x			x			x			x	x	x	x		V179I
		x	x						x			x			x	x	x	x		K103N+L100I
100		x	x						x			x			x		x	x		WT
		x							x			x			x		x			K103N
		x	x						x			x			x		x	x		K103N+V108I
		x	x						x			x			x		x	x		V179I
		x	x						x			x			x		x	x		K103N+L100I
101		x	x						x			x			x		x	x		WT
		x							x			x			x		x			K103N
		x	x						x			x			x		x	x		K103N+V108I
		x	x						x			x			x		x	x		V179I
		x	x						x			x			x		x	x		K103N+L100I
102																				WT
																				K103N
		x	x						x			x			x		x	x		K103N+V108I
																				V179I
																				K103N+L100I
103		x	x						x			x			x		x	x		WT
		x							x			x			x		x			K103N
		x	x						x			x			x		x	x		K103N+V108I
		x	x						x			x			x		x	x		V179I
		x	x						x			x			x		x	x		K103N+L100I
106		x	x						x			x			x		x	x		WT
		x							x			x			x		x			K103N
		x	x						x			x			x		x	x		K103N+V108I

		x	x						x			x			x		x	x		V179I
		x	x						x			x			x		x	x		K103N+L100I
108		x	x				x		x			x			x		x	x		WT
		x					x		x			x			x		x			K103N
		x	x			x	x	x	x			x			x		x	x		K103N+V108I
		x	x			x	x	x	x			x			x		x	x		V179I
		x	x				x		x			x			x		x	x		K103N+L100I
179		x	x						x			x			x		x	x		WT
		x							x			x			x		x			K103N
		x	x						x			x			x		x	x		K103N+V108I
		x	x						x			x			x		x	x		V179I
		x	x						x			x			x		x	x		K103N+L100I
181		x	x				x		x			x			x		x	x		WT
		x							x			x			x		x			K103N
		x	x						x			x			x		x	x		K103N+V108I
		x	x						x			x			x		x	x		V179I
		x	x						x			x			x		x	x		K103N+L100I
186																				WT
																				K103N
	x	x	x			x	x	x	x	x		x			x		x	x		K103N+V108I
	x	x	x			x	x	x	x	x		x			x		x	x		V179I
																				K103N+L100I
187		x	x				x		x			x			x		x	x		WT
		x					x	x	x			x			x		x			K103N
		x	x				x	x	x	x		x			x		x	x		K103N+V108I
		x	x				x		x	x		x			x		x	x		V179I
		x	x				x		x			x			x		x	x		K103N+L100I
188		x	x						x			x			x		x	x		WT
		x							x			x			x		x			K103N
		x	x						x			x			x		x	x		K103N+V108I
		x	x						x			x			x		x	x		V179I

		x	x						x			x			x		x	x		K103N+L100I
189		x	x						x			x			x		x	x		WT
		x							x			x			x		x			K103N
		x	x						x			x			x		x	x		K103N+V108I
		x	x						x			x			x		x	x		V179I
		x	x						x			x			x		x	x		K103N+L100I
190		x	x						x			x			x		x	x		WT
		x							x			x			x		x			K103N
		x	x						x			x			x		x	x		K103N+V108I
		x	x						x			x			x		x	x		V179I
		x	x						x			x			x		x	x		K103N+L100I
222		x	x						x			x			x		x	x		WT
		x							x			x			x		x			K103N
		x	x			x	x	x	x			x			x		x	x		K103N+V108I
		x	x			x	x	x	x			x			x		x	x		V179I
		x	x						x			x			x		x	x		K103N+L100I
227		x	x						x			x			x		x	x		WT
		x							x			x			x		x			K103N
		x	x						x			x			x		x	x		K103N+V108I
		x	x						x			x			x		x	x		V179I
		x	x						x			x			x		x	x		K103N+L100I
228		x	x		x	x	x	x	x			x			x		x	x		WT
		x			x	x	x	x	x			x			x		x			K103N
		x	x		x	x	x	x	x			x			x		x	x		K103N+V108I
		x	x		x	x	x	x	x			x			x		x	x		V179I
		x	x			x	x	x	x			x			x		x	x		K103N+L100I
229		x	x		x	x	x	x	x			x			x		x	x		WT
		x			x	x	x	x	x			x			x		x			K103N
		x	x		x	x	x	x	x			x			x		x	x		K103N+V108I
		x	x		x	x	x	x	x			x			x		x	x		V179I
		x	x		x	x	x	x	x			x			x		x	x		K103N+L100I

234		x	x						x			x			x		x	x		WT
		x							x			x			x		x			K103N
		x	x						x			x			x		x	x		K103N+V108I
		x	x						x			x			x		x	x		V179I
		x	x						x			x			x		x	x		K103N+L100I
235		x	x						x			x			x		x	x		WT
		x							x			x			x		x			K103N
		x	x						x			x			x		x	x		K103N+V108I
		x	x						x			x			x		x	x		V179I
		x	x						x			x			x		x	x		K103N+L100I
236																				WT
		x							x			x			x		x			K103N
		x	x						x			x			x		x	x		K103N+V108I
		x	x						x			x			x		x	x		V179I
		x	x						x			x			x		x	x		K103N+L100I
318		x	x						x			x			x		x	x		WT
		x							x			x			x		x			K103N
		x	x						x			x			x		x	x		K103N+V108I
		x	x						x			x			x		x	x		V179I
		x	x						x			x			x		x	x		K103N+L100I
443																				WT
																				K103N
																				K103N+V108I
																				V179I
				x	x	x				x	x		x	x					x	K103N+L100I

APPENDIX C

Ancestral state reconstruction visualization on phylogenetic trees with subtype B and subtype C sequences

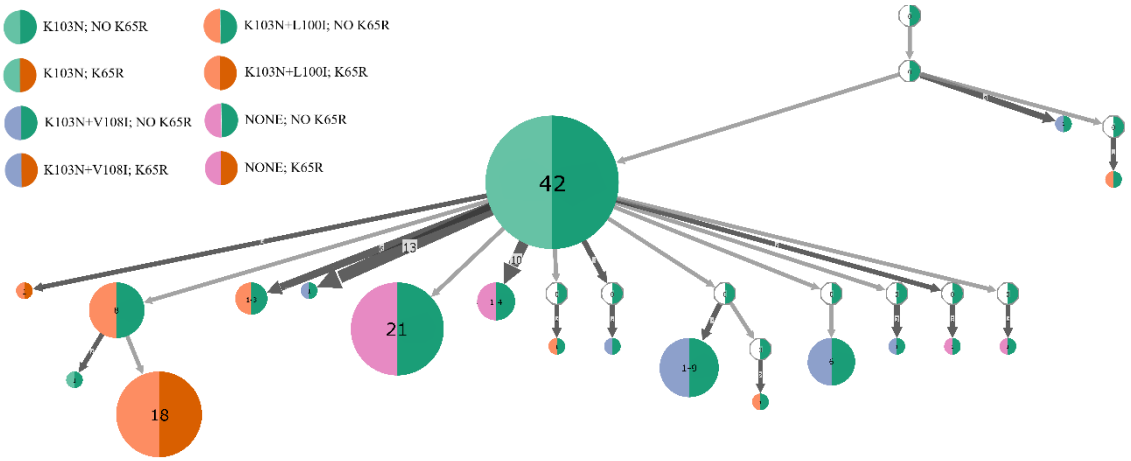


Figure C.8 - Ancestral state reconstruction (ACR) given the cluster's representative mutation (K103N, K103N+V108I or K103N+L100I) and presence/absence of the K65R mutation visualized by PastML on a phylogenetic tree inferred from 200 subtype B sequences. Circles denote clusters of samples (the left half represents which mutation present in the formed clusters (K103N (light green), K103N+V108I (blue), K103N+L100I (dark orange) or NONE of the above (pink) and the right half denotes the absence/presence of the K65R mutation, where dark green represents absence and dark orange, presence). The sample sizes of clusters are indicated in the labels (for example, the circle '42' encompasses 42 sequences – thus, the larger the circle, the more sequences it represents). Clusters with a '0' and white colour indicate internal tree nodes for which two different cluster mutations had similar marginal probabilities. Arrows between two circles denote transmissions from the top to the bottom cluster. The size and number on top of the arrows indicate that the arrows represent multiple transmission events leading to clusters of similar sizes [48]. The octagonal shaped nodes represent uncertain nodes, where the PastML software wasn't able to accurately predict the ancestral state due to a discrepancy between the two main approaches used in maximum likelihood-based ACR – maximum a posteriori (MAP) and joint ancestral scenario (Joint)

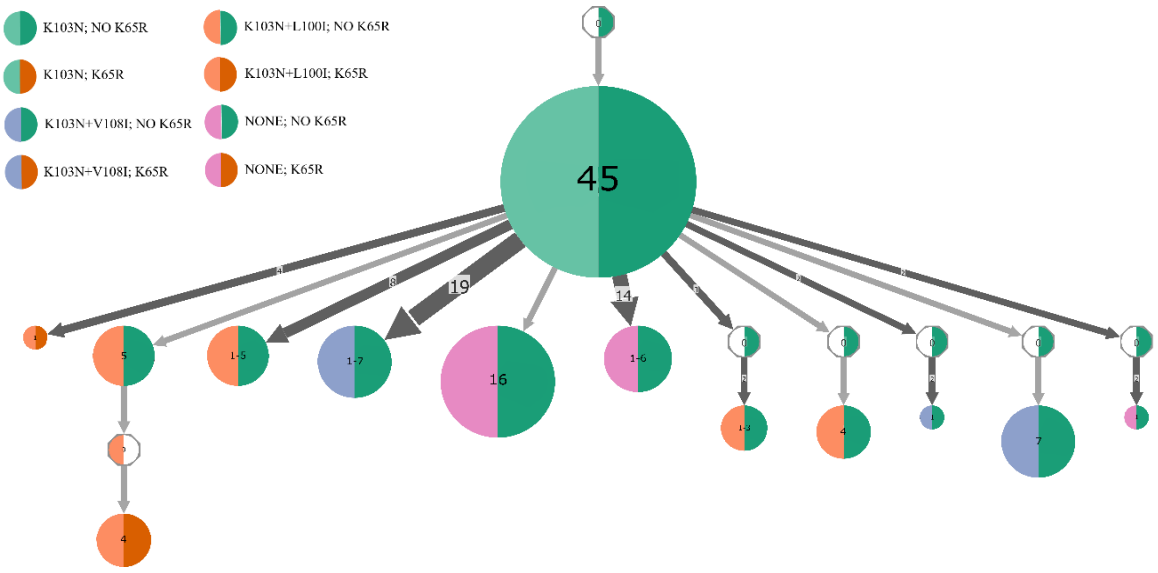


Figure C.9 - Ancestral state reconstruction (ACR) given the cluster's representative mutation (K103N, K103N+V108I or K103N+L100I) and presence/absence of the K65R mutation visualized by PastML on a phylogenetic tree inferred from 200 subtype B sequences. Circles denote clusters of samples (the left half represents which mutation present in the formed clusters (K103N (light green), K103N+V108I (blue), K103N+L100I (dark orange) or NONE of the above (pink) and the right half denotes the absence/presence of the K65R mutation, where dark green represents absence and dark orange, presence). The sample sizes of clusters are indicated in the labels (for example, the circle '45' encompasses 45 sequences – thus, the larger the circle, the more sequences it represents). Clusters with a '0' and white colour indicate internal tree nodes for which two different cluster mutations had similar marginal probabilities. Arrows between two circles denote transmissions from the top to the bottom cluster. The size and number on top of the arrows indicate that the arrows represent multiple transmission events leading to clusters of similar sizes [48]. The octagonal shaped nodes represent uncertain nodes, where the PastML software wasn't able to accurately predict the ancestral state due to a discrepancy between the two main approaches used in maximum likelihood-based ACR – maximum a posteriori (MAP) and joint ancestral scenario (Joint)

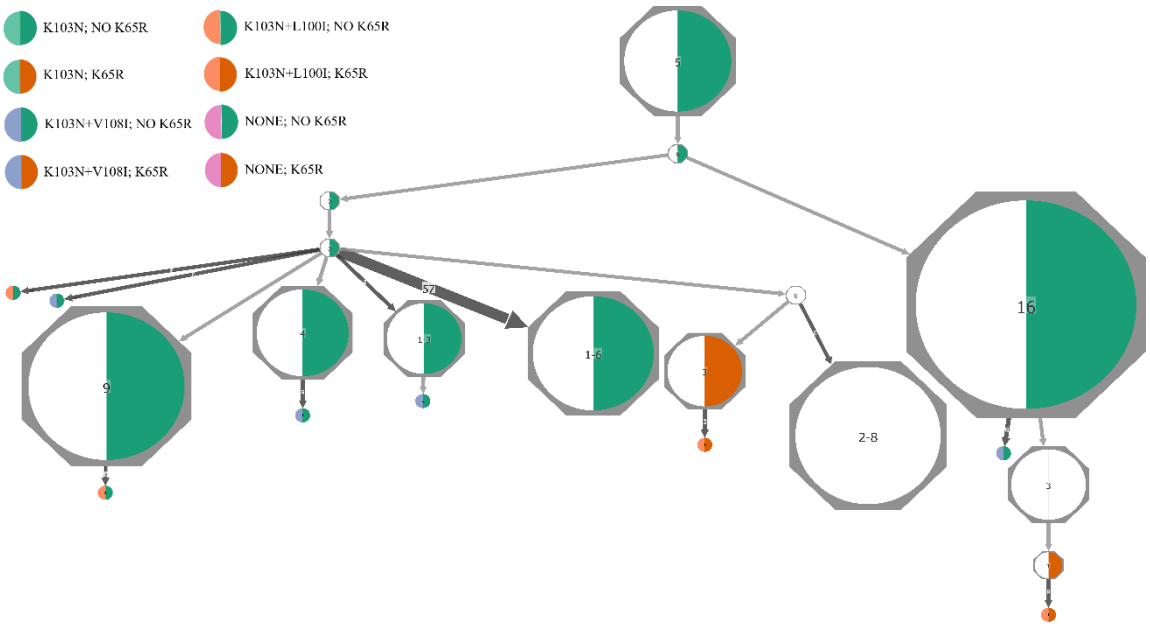


Figure C.10 - Ancestral state reconstruction (ACR) given the cluster's representative mutation (K103N, K103N+V108I or K103N+L100I) and presence/absence of the K65R mutation visualized by PastML on a phylogenetic tree inferred from 200 subtype C sequences. Circles denote clusters of samples (the left half represents which mutation present in the formed clusters (K103N (light green), K103N+V108I (blue), K103N+L100I (dark orange) or NONE of the above (pink) and the right half denotes the absence/presence of the K65R mutation, where dark green represents absence and dark orange, presence). The sample sizes of clusters are indicated in the labels (for example, the circle '45' encompasses 45 sequences – thus, the larger the circle, the more sequences it represents). Clusters with a '0' and white colour indicate internal tree nodes for which two different cluster mutations had similar marginal probabilities. Arrows between two circles denote transmissions from the top to the bottom cluster. The size and number on top of the arrows indicate that the arrows represent multiple transmission events leading to clusters of similar sizes [48]. The octagonal shaped nodes represent uncertain nodes, where the PastML software wasn't able to accurately predict the ancestral state due to a discrepancy between the two main approaches used in maximum likelihood-based ACR – maximum a posteriori (MAP) and joint ancestral scenario (Joint)

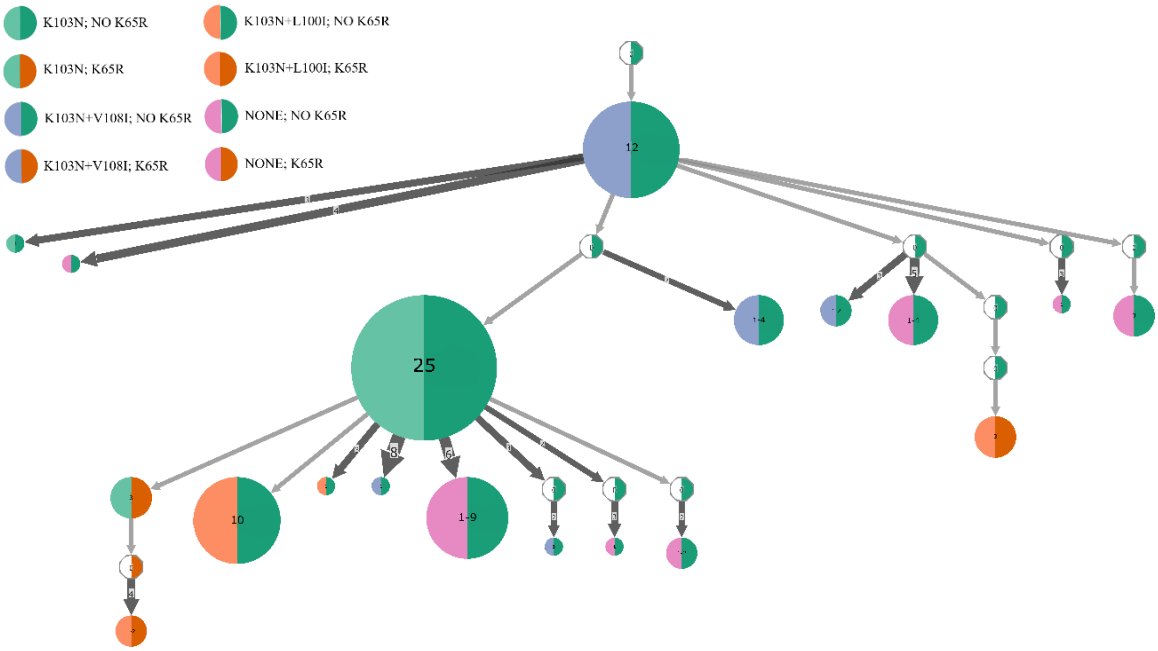


Figure C.11 - Ancestral state reconstruction (ACR) given the cluster's representative mutation (K103N, K103N+V108I or K103N+L100I) and presence/absence of the K65R mutation visualized by PastML on a phylogenetic tree inferred from 200 subtype C sequences. Circles denote clusters of samples (the left half represents which mutation present in the formed clusters (K103N (light green), K103N+V108I (blue), K103N+L100I (dark orange) or NONE of the above (pink) and the right half denotes the absence/presence of the K65R mutation, where dark green represents absence and dark orange, presence). The sample sizes of clusters are indicated in the labels (for example, the circle '25' encompasses 25 sequences – thus, the larger the circle, the more sequences it represents). Clusters with a '0' and white colour indicate internal tree nodes for which two different cluster mutations had similar marginal probabilities. Arrows between two circles denote transmissions from the top to the bottom cluster. The size and number on top of the arrows indicate that the arrows represent multiple transmission events leading to clusters of similar sizes [48]. The octagonal shaped nodes represent uncertain nodes, where the PastML software wasn't able to accurately predict the ancestral state due to a discrepancy between the two main approaches used in maximum likelihood-based ACR – maximum a posteriori (MAP) and joint ancestral scenario (Joint)

UNIVERSIDADE NOVA
DE LISBOA