Molecular Modeling Studies on Non-Nucleoside HIV-1 Reverse Transcriptase Inhibitors (NNRTIs)

By

Bilal Nizami

214573074

A thesis submitted in partial fulfillment for
the degree of Doctor of Philosophy
in the

Discipline of Pharmaceutical Sciences, School of Health Sciences

UKZN

Supervisor

Dr Bahareh Honarparvar

2016

Molecular Modeling Studies on Non-Nucleoside HIV-1 Reverse Transcriptase Inhibitors (NNRTIs)

BILAL NIZAMI

2016

A thesis submitted to the Discipline of Pharmaceutical Sciences, School of Health Science, University of KwaZulu-Natal, Westville, for the degree of Doctor of Philosophy.

This is a thesis in which the chapters are written as a set of discrete research papers, with an Overall Introduction and Final Discussion. Typically, these chapters will have been published in internationally recognized, peer-reviewed journals.

As the candidate's supervisors, I have approved this thesis for submission.

Supervisor:

Signed: -------------------------   Name: Dr B Honarparvar   Date: 02.12.02016

حصول علم کرو گہوارے سے لحد تک

“Seek knowledge from cradle to grave”

“Width of life is more important than length of life”

Ibn-Sina (Avicenna)

ABSTRACT

With the adaptation of UNAIDS recommended “Fast-Track approach” for tackling the AIDS epidemic, the world is committed to ending it by 2030. Nevertheless, enormous challenges lie ahead in ending the epidemic completely, with approx. 2.1 million new HIV infections case worldwide in 2015. Development of resistance against anti-retroviral drugs owing to mutations in the viral enzymes reduce the odds against AIDS.

Computational modeling techniques have emerged as an indispensable tool in modern drug discovery process and aid in understanding the complex biological phenomena. Comprehensive *in silico* investigation were performed in this work such as quantitative structure-activity relationship (QSAR), matched molecular pair analysis (MMPA), molecular docking, molecular dynamic (MD) simulations, dynamic pharmacophore (Dynophore), principle component analysis (PCA), MM-PBSA based binding free energy calculations.

As the first phase of this project, QSAR modeling and scaffold analysis of 289 pyrimidine derivatives were performed with non-nucleoside HIV RT inhibitory activities (NNRTI). The Associative Neural Network (ASNN) along with some other common machine learning methods were applied to develop a QSAR model for the anti-HIV RT activities. Scaffold-based analysis and molecular docking of the compounds used in the QSAR model identified a potential chemical scaffold. The results showed that scaffold-based analysis of the QSAR model could be helpful in identifying potent scaffolds for further exploration than analyzing the overall model. Matched molecular pair analysis (MMPA) was applied to the QSAR model to characterize molecular transformations causing a significant change in the anti-HIV activity. Interactions of few selected NNRTIs representing the identified scaffolds with HIV-1 RT were further studied in detail with MD simulations in our next stage of the project.

The K103N and E138K mutations in HIV RT are largely linked with treatment failure of the EFV (efavirenz) and RPV (rilpivirine), respectively when combined with tenofovir and emtricitabine. The K103N mutation had emerged as a clinical resistance mutation upon treatments with EFV, and it confers an almost uniform level of cross-resistance to most NNRTIs, except for the second generation of NNRTIs such as RPV and ETR (etravirine). RPV is a second-generation Di-aryl pyrimidine (DAPY) derivative, known to effectively inhibit the wild-type (WT) as well as various mutant RT such as K103N. The RT alongside protease has

been the main target of anti-HIV drugs used in multidrug combination therapy. In our second study, we performed a cumulative 240 ns of molecular dynamic (MD) simulations of WT HIV-1 RT and its corresponding K103N mutated form, complexed with RPV. Conformational analysis of the NNRTIs inside the binding pocket (NNIBP) revealed the ability of rilpivirine to adopt different conformations, which is possibly the reason for its reasonable activity against mutant HIV-1 RT. Binding free energy (MM-PB/GB SA) calculations of RPV with mutant HIV-1 RT were in agreement with experimental data. We also investigated the dynamics interaction patterns during the MD simulations using Dynophores, a novel approach for MD-based ligand-target interaction mapping. The results from this interaction profile analysis suggest an alternative interaction between the linker N atom of rilpivirine and Lys101, potentially providing the stability for ligand binding.

To further pursue the goal, in our next study, we performed MD simulations of WT and E138K mutant RT complexed with RPV, EFV, and ETR. The E138K is a non-polymorphic mutation in the p51 subunit of RT that is selected preferentially in patients receiving RPV and reduces its susceptibility up to 5-fold. In our previous study, we have explored the molecular level understanding of the binding of RPV to K103N mutated HIV-1 RT. Which has suggested that RPV's torsional flexibility (''wiggling'') and repositioning and reorientation within the pocket (''jiggling'') makes it withstand the drug resistance. In our third study, we present MD simulation of wild-type (WT) and E138K HIV-1 RT in complex with EFV, ETR, and RPV. Additionally, MD simulations of K103N RT complexed with RPV is also presented. The main motivation behind this study is to understand why second generation NNRTIs are able to bypass certain drug resistance mutation such as K103N, while at the same time being susceptible to E138K. Understanding of dynamics of NNRTI binding and effect of the mutation on drug binding is helpful in designing new inhibitors with improved resistance tolerance.

Acknowledgements

I thank Almighty Allah, the most beneficent, the most merciful, for giving me the patience and perseverance throughout my life to move in the right direction.

This thesis would not have been possible without the inspiration, wisdom and help provided by so many people and it is a pleasure to show my appreciation here. Firstly, I would like to thank my family and especially my parents Prof. Mohammed Riyazuddin and Anjum Riyaz and my siblings (Amir, Ammara, Yasir and Aisha) whose love, encouragement, and constant support are much appreciated. Further thanks are due to the friends who have kept me sane and made my time in Durban so enjoyable. In no particular order, I offer my heartfelt gratitude to Faiz Iqbal Khan, Jyotibon Dutta, Bheki Zwane, Faruk Tiltay, Mohammed Bouhadja and Hamza Akcit.

I offer my sincerest gratitude to my supervisor, Dr Bahareh Honarparvar, who has supported me throughout my Ph.D. whilst allowing me to work at my own pace. I am indebted to Prof Igor Tetko (Helmholtz Zentrum Muenchen) for his valuable guidance for QSAR study and Prof Neil Koorbanally (UKZN) for useful scientific discussion. I am also indebted to Prof Gerhard Wolber and his group members for hosting us at his lab at the Free University of Berlin.

I am indebted to Collage of Health Sciences, UKZN for the doctoral bursary and CHPC, an initiative supported by the Department of Science and Technology of South Africa, for the all the computational resources. I am also thankful to NRF for the KIC travel grants to attend the conference in Vienna, Austria in 2015 and Rhodes Universit, SA in 2016.

DECLARATION 1 – PLAGIARISM

I, **Bilal Nizami** declare that

1. The research reported in this thesis, except where otherwise indicated, is my original research.

2. This thesis has not been submitted for any degree or examination at any other university.

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

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

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

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

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

Signed

DECLARATION 2 - LIST OF PUBLICATIONS

RESEARCH OUTPUT

1. PUBLICATIONS IN PEER REVIEW JOURNALS

1.1. **Bilal Nizami**, Igor V. Tetko, Neil A. Koorbanally, Bahareh Honarparvar, *Chemometrics and Intelligent Laboratory Systems*, 2015, **148**, 134-144. 10.1016/j.chemolab.2015.09.011

1.2. **Nizami B**, Sydow D, Wolber G, Honarparvar B (2016) Molecular insight on the binding of NNRTI to K103N mutated HIV-1 RT: molecular dynamics simulations and dynamic pharmacophore analysis. Molecular bioSystems. doi:10.1039/C6MB00428H.

1.3. **Nizami B**, and Honarparvar B, Molecular dynamics simulations of various NNRTIs bound with E138K mutated HIV-1 RT, under submission.

2. CONFERENCE

2.1. **Bilal Nizami**, “Molecular insight on the binding of NNRTI to E138K mutated HIV-1 RT” oral presentation at Frank Warren Conference of the South African Chemical Institute at Rhodes university from 4th – 8th December 2016, Oral presentation.

2.2. **Bilal Nizami**, “QSAR and Molecular Docking of Non- Nucleoside HIV RT Inhibitors”, Vienna Summer School for Drug Design, University of Vienna, Austria, 2015, Poster presentation.

2.3. **Bilal Nizami**, Dynamics of rilpivirine binding with wild-type and K103N mutated HIV-1 RT, college of health sciences symposium, UKZN, 2016, Poster presentation.

2.4. **Bilal Nizami**, MD simulations of 1st and 2nd generation NNRTIs with mutant HIV-1 RT, CHPC National Meeting, 5th-9th Dec 2016, East London, South Africa, Poster presentation.

Dedicated to my parents and sister

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS

ABBREVIATIONS

DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
PDB	Protein Data Bank
MD	Molecular Dynamics
QSAR	Quantitative Structure Activity Relationship
MMPA	Matched Molecular Pair Analysis
HIV	Human Immunodeficiency Virus
SIV cpz	Simian immunodeficiency virus of chimpanzees
AIDS	Acquired Immunodeficiency Syndrome
HAART	Highly active antiretroviral therapy
NMR	Nuclear Magnetic Resonance
PME	Particle Mesh Ewald
GPU	Graphical Processing Unit
MMPBSA	Molecular Mechanics Poisson-Boltzmann Surface Area
PR	Protease
RT	Reverse Transcriptase
PI	Protease Inhibitor
NRTI	Nucleoside/nucleotide analogue RT Inhibitor
NNRTI	Non-Nucleoside/nucleotide analogue RT Inhibitor
NNIBP	NNRTI Binding Pocket

PHYSICAL CONSTANTS

Boltzmann Constant	kB	$= 3.2976268 \times 10^{21}\ kcal\ K^{-1}$
Plank Constant	h	$= 1.582611 \times 10^{-37}\ kcal\ s$
Gas Constant	R	$= 1.9858775 \times 10^{-3}\ kcal\ K^{-1}\ mol^{-1}$
Speed of Light	c	$= 299792458\ m\ s^{-1}$
Permittivity of Free Space	ε_0	$= 1/\left(4\pi c^2\right) \times 10^7\ F\ m^{-1}$

SYMBOLS

T	Temperature	K
U	Internal Energy	kcal mol^{-1}
H	Enthalpy	kcal mol^{-1}
S	Entropy	kcal mol^{-1} K^{-1}
G	Gibbs Free Energy	kcal mol $^{-1}$
A	Helmholtz Free Energy	kcal mol^{-1}
p	Pressure	bar
N	Number	---
L	Lagrangian	kcal mol^{-1}
H	Hamiltonian	kcal mol^{-1}
K	Kinetic Energy	kcal mol^{-1}
V	Potential Energy	kcal mol-^{1}
*K*a	Association Constant	M^{-1}
*K*d	Dissociation Constant	M
*K*d	Dissociation Constant	M
KM	Michaelis Constant	M

CHAPTER 1

1.1. Background and rationale

According to a 2012 WHO estimate, 35.3 million people were living with HIV/AIDS worldwide [1] , with a significant number of these infections being resistant to anti-retroviral therapies. With the adaptation of UNAIDS recommended "Fast-Track approach" for tackling the AIDS epidemic, the world is committed to ending it by 2030. Nevertheless, enormous challenges lie ahead in ending the epidemic completely, with approximately 2.1 million new HIV infections cases worldwide as per the WHO report in 2015 [2].

HIV utilizes Reverse Transcriptase (RT), an enzyme that makes copies of cDNA from RNA, a process called reverse transcription. RT, an important enzyme in HIV-1, catalyzes the transcription of the viral single-stranded (ss) RNA into double-stranded (ds) DNA. The HIV-1 RT consists of two subunits, the larger p66, and the smaller p51 [3], with the polymerase and ribonuclease H (RNase H) catalytic sites being located on the former [4]. The polymerase domain of HIV resembles the right hand with fingers, thumb, palm, and connection sub-domain [4]. The thumb and finger sub-domains of RT undergo conformational changes to perform the process of reverse transcription. The enzyme Reverse transcriptase (RT) alongside protease has been the main targets of anti-HIV drugs used in multidrug combination therapy. Anti HIV-1 RT agents are chemical compounds targeting the RT enzyme, thus effectively blocking the progression of the virus. There are two different class of drugs targeting the HIV-1 RT; NNRTI (non-nucleoside reverse transcriptase inhibitors) and NRTI (nucleoside reverse transcriptase inhibitors) both target a different aspect of the RT functioning. The NNRTIs bind in the binding pocket is approximately 10 Å away from the polymerase in RT and disrupts the conformational flexibility of the enzyme [3]. Pyrimidine derivatives were synthesized for decades and have been actively pursued as NNRTIs [5]. Two main series of pyrimidine derivatives are DABO (Dihydro-alkoxy-benzyl-oxo pyrimidine) and DAPY (Di aryl pyrimidine) [6].

The higher rate of mutation in HIV strains and the subsequent development of resistance to the NNRTIs is a major issue in managing HIV infection. This highlights the need for the rapid and rational development of NNRTIs. Development of resistance against anti-retroviral drugs owing to the mutations in the viral enzymes reduce the odds against AIDS. At present there are four approved drugs in the NNRTI class –nevirapine (NEV), efavirenz (EFV), etravirine (ETR), and rilpivirine (RPV) (See FIGURE 7)— while delavirdine (DLV) was approved in 1997, but now is not recommend as part of initial therapy.

A crucial role of RT in the life cycle of HIV-1 makes it the prime target of anti-retroviral therapy, such as non-nucleoside reverse transcriptase inhibitors (NNRTIs) [6]. This makes RT an attractive target for anti-retroviral drugs like Non-nucleoside Reverse Transcriptase Inhibitors (NNRTIs) [6]. This study deals with the molecular modeling studies of NNRTIs and HIV-1 RT using scaffold based QSAR, matched molecular pair analysis (MMPA), molecular docking, MD simulation studies and dynamic pharmacophore studies. An overview of the thesis and research presented in it is described briefly in section1.3.

1.2. Aims and objectives

The prime objective of this thesis is to understand the relationship between the chemical structure of NNRITs and its anti-HIV activity as well as its dynamic interaction with WT and various mutant RT structures. A detailed molecular level understanding is required for the design of novel and better HIV-1 RT inhibitors. There are three key goals:

1. To develop a robust QSAR model for selected NNRTIs that could predict anti-HIV RT activity. Moreover, to identify a potential chemical scaffold for further optimization, the aim being to understand the underlying structural changes that could contribute to improving the anti-HIV activity of the NNRTI. To achieve this goal QSAR modeling was combined with molecular docking studies and MMPA on the selected NNRTIs to provide a deeper insight into the computer-aided design of novel molecules against HIV RT.

2. Our scaffold-based QSAR study [7] identified two potential ligand scaffolds against HIV-1 RT. This motivated an investigation of the dynamics of HIV-1 RT sub-domains in WT and K103N mutant, complexed with rilpivirine using explicit MD simulation. To map the dynamic interaction pattern between ligand and RT, dynamic pharmacophores (dynophore) was performed on the MD snapshots.

3. To perform the MD simulation and analysis of wild-type (WT) and E138K HIV-1 RT in complex with efavirenz (EFV), etravirine (ETR), and rilpivirine (RPV), in order to gain

understanding the dynamics of NNRTI binding and effect of mutation on drug's binding affinity, which is helpful in designing new inhibitors with better resistance tolerance.

1.3. Overview of the study

Following is the general overview of the chapters in this thesis:

Chapter 1: Summarizes the background rationale of research, aims, and objectives of this work, and structure of the thesis.

Chapter 2: Describes the HIV and AIDS, different anti HIV-1 enzymes and its inhibitors and emergence of drug resistance in HIV.

Chapter 3: Deals with the theoretical framework of the various computational methods used in this study.

Chapter 4: This is the published work, presented in the final format as requested by the Journal of Chemometrics and Intelligent Laboratory Systems (ISSN: 0169-7439). The paper is entitled "QSAR models and scaffold-based analysis of non-nucleoside HIV RT inhibitors", and present some interesting insight into the different scaffolds of pyrimidine derivatives as NNRTI. In this chapter, we have described the scaffold based QSAR of selected NNRTIs and few potential chemical scaffold is identified for further optimization. Matched Molecular Pair analysis (MMPA) was also applied to characterize molecular transformations causing a significant change in the anti-HIV activity.

Chapter 5: This is the published work, presented in the final format as requested by the RSC Molecular Biosystems. The paper is entitled "Molecular insight on the binding of NNRTI to K103N mutated HIV-1 RT: Molecular dynamics simulations and dynamic pharmacophore analysis". In this chapter, a cumulative 240 ns of molecular dynamic (MD) simulations of WT HIV-1 RT and its corresponding K103N mutated form, complexed with rilpivirine, is presented. Conformational analysis of the NNRTI inside the binding pocket (NNIBP) as well as binding free energy (MM-PB/GB SA) calculations of rilpivirine with mutant HIV-1 RT is also presented.

Chapter 6: This work is under submission. In this study, a cumulative 1 µs of MD simulation of wild-type (WT) and E138K HIV-1 RT complexed with efavirenz (EFV), etravirine (ETR), and rilpivirine (RPV) is presented. Additionally, MD simulation of K103N RT complexed with RPV is also presented.

CHAPTER 2

2. About AIDS and HIV

2.1. Introduction

Acquired Immune Deficiency Syndrome – a spectrum of clinical conditions – caused by Human Immunodeficiency Virus (HIV) is characterized by the vulnerability of victim to opportunistic pathogens and increased risk of Kaposi's sarcoma and other rare forms of cancer [8]. As its name suggests, HIV largely infects human's CD4+ T cells, macrophages, and dendritic cells [9, 10]. After infecting the cells, HIV takes control of host cellular machinery to synthesize viral proteins and replicates rapidly. One of the key symptoms of HIV infection is the low count of circulating CD4+ T cells. When enough number of T cells have been infected by HIV, the host immune system can no longer perform its normal functions, and the victim becomes susceptible to other infections. Maximum cases of AIDS worldwide are caused by the more infectious subtype HIV-1. For the same reason, most studies of HIV (including this thesis) focus on the HIV-1 subtype.

This chapter briefly describes the historical background, lifecycle, prevention and treatment of AIDS. A brief discussion on different types of anti-retroviral drugs and their use is also presented. In this chapter, the focus will be on HIV-1 Reverse Transcriptase (RT) and drug discovery of anti-RT agents.

2.2. Discovery of AIDS/HIV

The first clinical occurrence of AIDS emerged between late 1980 and early 1981 when a group of gay men inexplicably presented with a rare condition known as Pneumocystic carinii pneumonia (PCP) [11] , followed by the several reports of rare skin cancer Kaposi's sarcoma (KP) in another group of men [12]. Both the conditions are associated with severely compromised immune systems and T cells were suspected to be the target of infection. The condition was soon named acquired immunodeficiency syndrome (AIDS). Two years later a novel virus was isolated from patients with AIDS [13, 14] and independently named lymphadenopathy-associated virus (LAV) and human T lymphotropic viruses type III (HTLV-III). Subsequently, in 1986, it was discovered that both the virus is same and belong to the *Retroviridae* family of viruses, thus renamed Human Immunodeficiency Virus (HIV) [15].

2.3. Origin and classification of HIV

HIV can be divided into two subtypes: HIV-1 and HIV-2, former being the major cause of AIDS. Current understanding places the appearance of HIV infection in humans via zoonosis that originated in primate-to-human species-jumping events. In the case of HIV-1 and HIV-2 strains, these transfer events occurred in Central and West Africa, most likely at several times. Simian immunodeficiency virus of chimpanzees (SIV cpz) –retroviruses able to infect at least 45 species of primates – is believed to be the immediate precursor to HIV-1 [16]. Pathogenic human retroviruses include lentiviruses HIV-1 and HIV-2 and onco-viruses includes HTLV-1 and HTLV-2. According to phylogenetic classification, HIV-1 can be sub-divided into four main groups designated as M, N, O and P group [17].

2.4. Epidemiology of HIV-1

The HIV infection was originally confined primarily to North America, Western Europe, and parts of sub-Saharan Africa, however, it has spread throughout the world, with increasing heterogeneity. **FIGURE 1** shows a world map with HIV prevalence rates by country, southern Africa share around 10-20 % of the global burden of HIV.

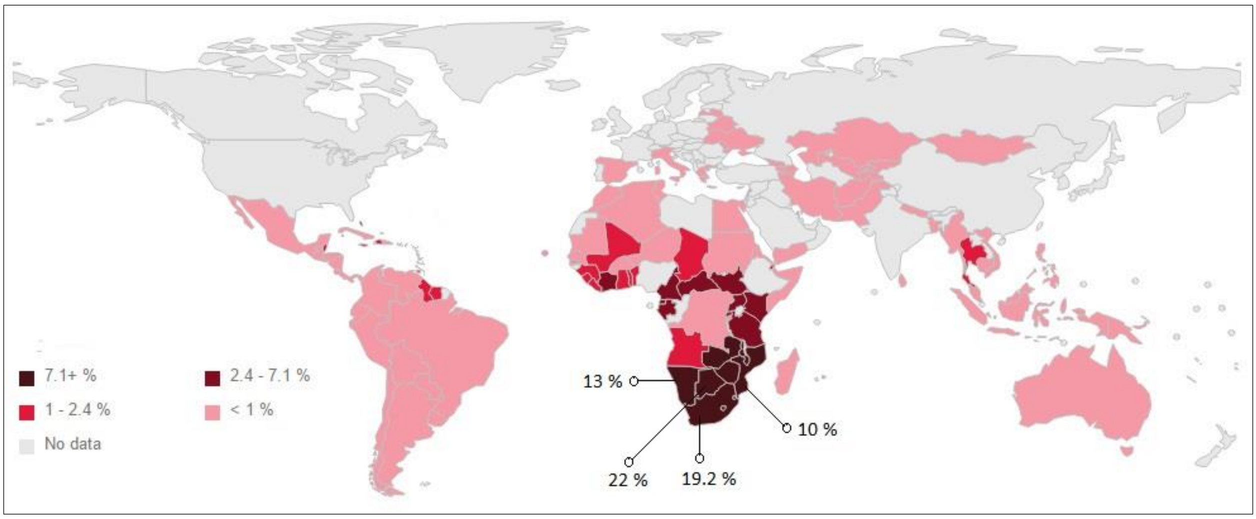


Figure 1: Geographical distribution of HIV.

Geographical distribution of HIV (prevalence rates by country), stats for highly affected sub-Saharan countries are shown additionally. Data from UNAIDS AIDS info epidemiological Status 2015 are available at http://aidsinfo.unaids.org/, accessed July 26, 2016.

As per UNAIDS fact sheet on AIDS, globally 36 million people were living with HIV in 2015, where 2.1 million being newly infected by HIV [18]. Sub-Saharan Africa is the most affected region with 66% of the global population of HIV-infected people. South Africa has the highest prevalence of HIV, with approximately 5.6 million people living with HIV [1, 18]. In Asia and the Pacific, there were 5.1 million people living with HIV in 2015, approximately 2.5 million of these cases are in India, where however the prevalence is only about 0.3%. Whereas, according to a recent estimate, there are 2.4 million people living with HIV in western and central Europe and North America [18].

The current AIDS pandemic can be described in two forms of HIV transmission. The first method involves sexual transmission, including both homo and heterosexual practice. The second asexual method includes mother-to-child transmission, transmission via infected blood and sharing injection needles among drug users. In the majority of countries in sub-Saharan Africa, the epidemic is attributed to heterosexual transmission. Similarly, the main mode of transmission in the Caribbean is via heterosexual means, whereas, in Asia and Eastern Europe, the most common transmission route is via heterosexual transmission and injecting drug users. The prevalence of sexually transmitted diseases, the practice of scarification, unsafe blood transfusions, and the poor state of hygiene and nutrition in some areas may all be assisting factors in the spread of HIV-1[19], whereas, some religious practices [20] and male circumcision [21] are shown to prevent the HIV prevalence.

2.5. The Pathophysiology of HIV Infection

The interactions between the HIV and the human immune system are unusually complex and follow a chronic course of the disease, as evinced by the variable rates of disease progression observed in HIV-infected persons. The CD4 and a chemokine co-receptor, typically either CCR5 or CXCR4, are indispensable for viral entry into target cells, whereas other receptors are believed to expedite infection or transmission of HIV. After the virus enters the body, there is a rapid migration to regional lymph nodes, followed by dissemination via the bloodstream to various lymphoid, leading to an abundance of virus in the peripheral blood. During primary infection, the level of viral load may reach several million virus particles per milliliter of blood and a significant drop in the count of circulating CD4+ T cells. Although it's debatable that which cells are infected first, nonetheless, dendritic cells (DCs), especially the Langerhans cells, as well as macrophages and resting CD4+ T cells are all potentially among the first cells to be targeted by HIV [22]. Initial infection with HIV-1 is often associated with `acute retroviral syndrome', which exhibits flu-like symptoms such as fevers, sore throats, swollen lymph nodes and rashes [23, 24]. This early stage of the infection leads to the loss of mucosal CD4+ T helper lymphocytes. These are the main targets of HIV, although the virus can infect several other cell types such as macrophages. It is the loss of CD4+ T cells which brings about the clampdown of the immune system, resulting in AIDS. The main function of the T helper cell is to regulate immune responses by the secretion of specialized factors that activate other white blood cells to fight infections. They control CD8+ lymphocytes which are responsible for directly killing certain tumor cells, cells infected by viruses and some parasites. Although these symptoms usually subside within 1 to 2 weeks, but main symptoms, characteristic of AIDS might not appear for years after a person is infected. This early stage of the infection leads to the loss of the bulk of mucosal CD4+ T cell [25]. The body unveils a strong immune defense to the initial high levels of virus, with infected CD4+ cells rapidly being removed. The initial results of this response are a lowering in the number of viral particles in circulation and the temporary recovery of CD4+ cell levels. The destruction of infected cells is balanced by the body's production of new CD4+ cells and a steady state, in which most CD4+ cells are uninfected. FIGURE 2 shows the initial infection and propagation of human HIV infection.

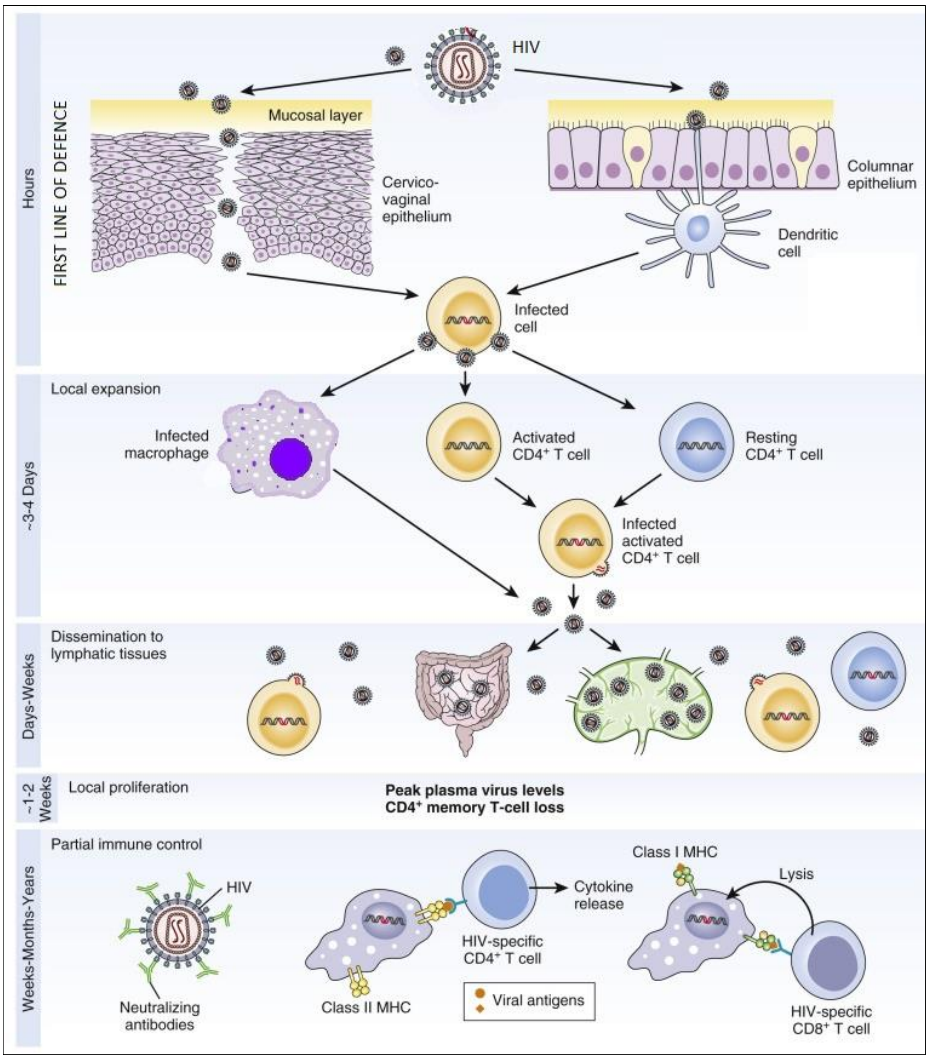


Figure 2: propagation of human HIV infection.

Initial infection and propagation of human HIV infection. Figure adapted from “The Immunology of Human Immunodeficiency Virus Infection”, accessed July 27, 2016 [26].

2.6. The Structure and Organisation of HIV Genome

HIV is a lentivirus, a member of a subfamily of Retroviridae with some difference in structure from other retroviruses, and complex gene expression and replication. HIVs are initially assembled and released from infected cells as spherical immature particles (virion) containing precursors of Gag and Gag-Pol proteins that ultimately make up the mature virus. Mature HIV is enveloped by a lipid bilayer and roughly spherical in shape with a diameter of about 120 nm [27] (Figure 3). Lipid bilayer of viral envelope is dotted with spikes of the glycoproteins gp120 and gp41, which are responsible for binding to the host cell. The envelope surrounds the nucleocapsid core which contains viral genetic material, three essential enzymes; integrase (IN), reverse transcriptase (RT) and protease (PR); accessory proteins and some cellular factors. The genetic material of HIV consists of two copies of non-covalently linked, single-stranded RNA and is transcribed it into DNA by the enzyme reverse transcriptase (RT) [28]. Viral RNA is enclosed within a conical nucleocapsid with approximately 2000 molecules of p24 protein [29].

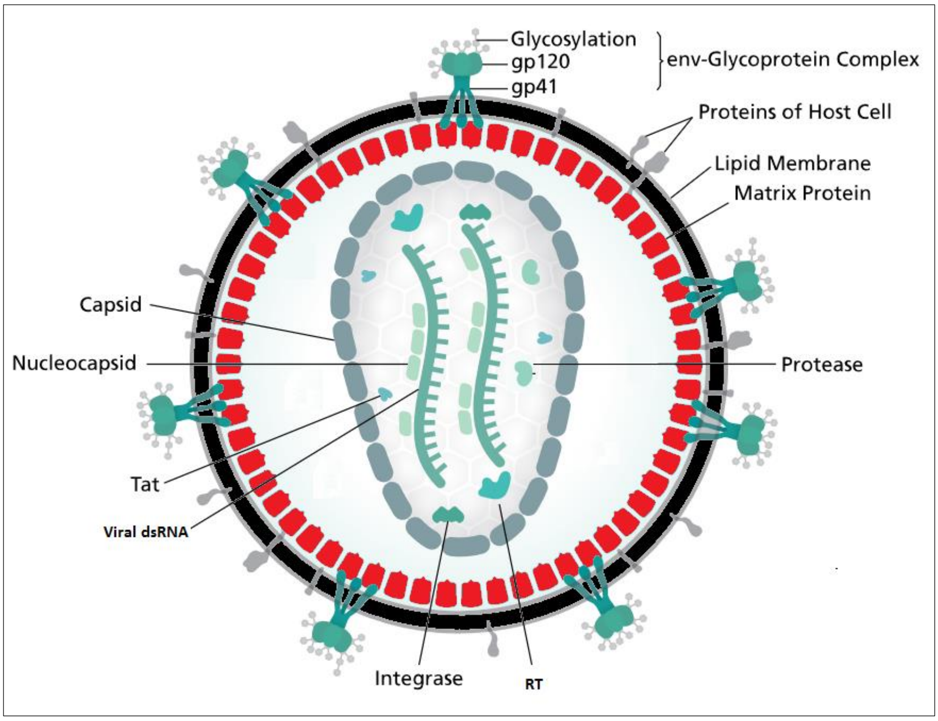


Figure 3: Structure of HIV with its capsid and important enzymes

HIV genome encodes multiple viral proteins, belonging to three classes viz. structural proteins, essential regulatory proteins, and accessory regulatory elements. The name and function of these encoded proteins are given in TABLE 1. The HIV genome contains three main genes arranged from 5' to 3' end in order starting from, gag (group-specific antigen), pol (polymerase) and ending in *env* (envelope) (see FIGURE 4). These genes encode for the vital structural proteins and enzymes necessary for replication of HIV [30]. The gag encodes a polyprotein that is cleaved into three proteins *i.e.* matrix (MA or p17), capsid (CA or p24), and nucleocapsid (NC or p7). Together, these three proteins provide the basic physical infrastructure of the retrovirus.

The pol gene also encodes a polyprotein, which is cleaved into the three functional enzymatic proteins, viz. protease (PR), reverse transcriptase (RT) and integrase (IN). Together, these enzymes provide the basic cellular machinery by which retroviruses replicate. Protease catalyzes the proteolytic cleavage of the polypeptide chains (like gag and pol polyprotein) into functional proteins. RT, an important enzyme in HIV-1 which is undertaken in this thesis, catalyzes the transcription of the viral single-stranded (ss) RNA into double-stranded (ds) DNA. HIV RT consists of two subunits, the larger p66 and the smaller p51 [3], with the polymerase and ribonuclease H (RNase H) catalytic sites being located on the former. A detailed account of structure and function of HIV-1 RT is given in section 2.9. Integrase catalyzes the incorporation of the reverse transcribed viral DNA into the host chromosomes, in order to utilize the host cellular machinery to transcribe the other viral proteins.

Table 1: protein products of the HIV genes

Entire protein products of the HIV genes with their structural information and function [29, 31].

Class	***Encoding gene***	***Chain***	***Protein***	***Function***
Structural	*gag*	*GAG POLYPROTEIN*	*MA (p17)*	*Stabilizes the viral particle*
			CA (p24)	*Core antigen capsid protein*
			p6	*Mediates interactions between Gag and Vpr*
			NC (p7)	*Nucleocapsid protein*
	pol	*POL POLYPROTEIN*	*PR (11) (protease)*	*Catalyzes the cleavage of the polypeptide chains into the functional proteins*
			RT (p51/p66)	*Reverse transcribe the viral RNA into DNA*
			IN (p34)	*Integration of viral DNA into the host genome*
	env	*GP160*	*SU (gp120)*	*Facilitates the binding to CD4, macrophages and T lymphocytes*
			TM (gp41)	*Transmembrane glycoprotein, along with SU helps in the viral entry to the host cell.*
	tat	*TAT*	*TAT*	*Regulation of reverse transcription*

Essential regulatory elements	*rev*	*REV*	*REV*	*Helps in synthesis of major viral proteins*
Accessory regulatory elements	*nef*	*NEF*	*NEF*	*Main role in nuclear import of viral DNA to the host nucleus*
	vpr	*VPR*	*VPR*	
	vif	*VIF*	*VIF*	
	vpu	*VPU*	*VPU*	

The *env* gene encodes for the gp160 protein, which is cleaved into gp120 and gp41 by a host protein named furin belonging to the family of pro-protein convertase. Gp120 and gp41 help the virus to bind and enter to the host CD4, macrophages, and T lymphocytes cells [32]. The genomes of HIV and other lentiviruses (Simian Immunodeficiency Virus (SIV) and Feline Immunodeficiency Virus (FIV) comprises regulatory genes in addition to the gag, pol, and env which are believed to be responsible for their increased pathogenicity [30]. Regulatory elements in the HIV-1 includes essential elements*:* *tat*, *rev*, and accessory element: *nef*, *vpr*, *vif*, and *vpu*. The *tat* and *rev* are important elements responsible for changing host gene expression and are essential for viral replication *in vitro*, whereas accessory regulatory elements are not essential for replication in certain tissues [31]. To be incorporated into a host cell's genome, HIV ds RNA strands must be translated into DNA through reverse transcription [28]. Beside transcribing DNA, some regions of 3' and 5' end of RNA are also encoded at the two ends of the genome known as LTR (long terminal Repeats, see **Figure 4**) [30]. As evident from the **Figure 4,** some regions of the genome is encoded more than once for different products which lead to the considerable increase in genetic information due to this phenomenon known as frames shifting [30].

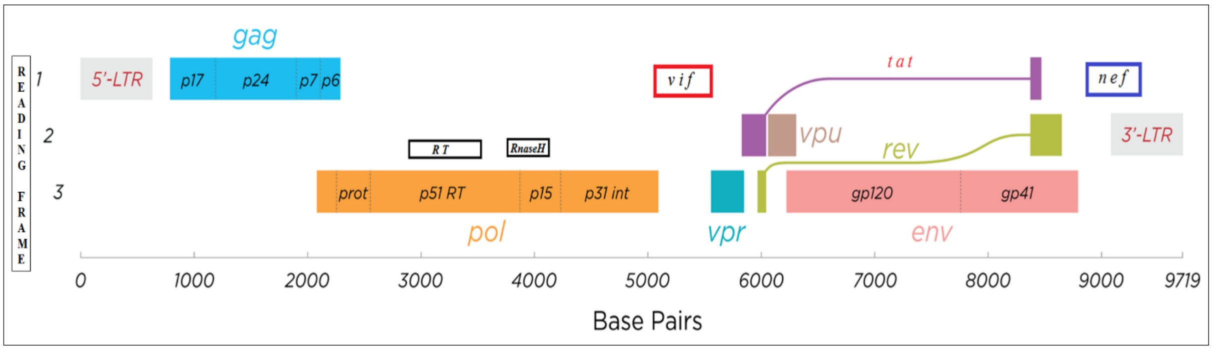


Figure 4: The Organization of HIV genome

2.7. Life cycle of HIV-1

Viruses are unable to reproduce (replicate) by themselves, instead, they need a host cell for it. Detailed understanding of the viral life cycle is a crucial factor in facilitating the development of anti-HIV treatment strategies. An overview of the HIV replication cycle and the roles played by some of the proteins described in the previous section within this cycle is given here. Important stages in the life cycle of HIV is shown in **FIGURE 5**. The events in the life cycle of HIV-1 can be divided distinct stages, viz. entry to the host cell which involves binding and fusion to the CD4 receptor, reverse transcription, integration, viral replication, assembly, release, and maturation.

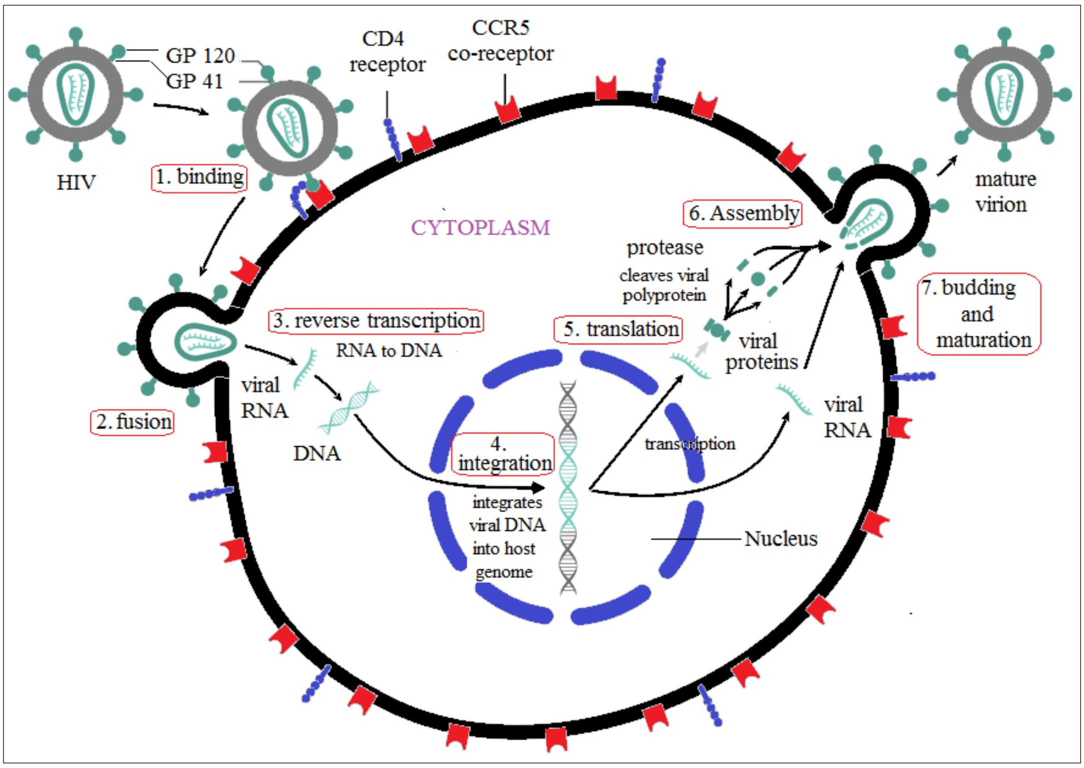


Figure 5: Stages in the life cycle of HIV-1.

Important stages are numbered and encircled in the red box. It begins with binding of the gp120 to the CD4 receptor followed by fusion of the viral membrane and the host cell membrane. The viral genome is released and RT transcribes the RNA to DNA. The viral DNA is transported into the nucleus via to be integrated into the host chromosome and the provirus is formed. The provirus serves as a template for the production of new viral proteins and RNA genomes. The newly transcribed viral proteins and RNA are transported out of the nucleus where they are assembled into new virions. The virions are then released, killing the host cells int the process. The final stage is the maturation of the viral particle, marked by the cleavage of viral polypeptides and virion is ready to infect new cells.

2.7.1. Entry to the cell

HIV-1 is an enveloped virus and its envelope proteins play an important part in its entry into the host cells. The process starts with the binding of gp120 to CD4+ receptor on the target cell's surface. Subsequently, gp41 facilitates the fusion of viral envelope with the cell membrane of target cells [33]. Viral entry to the cell begins by the interaction of the gp160 spike (gp120+ gp41) and both CD4 and a CCR5/CXCR4 chemokine receptor on the cell surface. The first step in fusion involves the strong attachment of the gp120 to CD4. Once gp120 is bound to the CD4 protein, the envelope complex undergoes a structural change and gp41 folds into a hairpin-like structure. This causes the virus and cell membranes to fuse and the creation of the fusion pore in the cell membrane is initiated. The viral genome and enzymes are then released into the host cell's cytoplasm through these pores [33].

2.7.2. Reverse transcription

Once the viral core is within the host cell, the next key stage in the life cycle is the conversion of the single-stranded viral RNA genome into the double stranded DNA which can be incorporated into the host cell chromosomes. Before this stage, the core of the virion is re-organized to create a complex known as the reverse transcription complex (RTC). Viral RNA is transcribed into ds DNA by the viral enzyme reverse transcriptase (RT) –a multifunction viral enzyme with a distinct polymerase and RNaseH active sites. RT first uses a single strand of RNA as a template and create a single strand of DNA, then this DNA strand is transcribed into double stranded DNA molecule. At the polymerase active site, incoming nucleotides complementary to the RNA or DNA template are added into the growing complementary DNA chain. The RNaseH site degrades the viral RNA genome, freeing the DNA copy to act as a template for the transcription of the second stranded of DNA. The reverse transcription is an extremely errant process with a mutation rate of 3×10^{-5} per base pair per replicative cycle and the resulting mutations are attributed as the leading cause of drug resistance [34].

2.7.3. Integration

Now the viral DNA is ready for the integration with the host chromosome. For the integration, viral genome needs to be transported to the host cell's nucleus, regulatory protein *vpr* being the most important viral factor in this process [35]. The integration of the viral DNA into the host cell's genome is carried out by HIV integrase. Integrated viral DNA is also known as "provirus", which might remain dormant for the duration of HIV latent stage [36].

2.7.4. Viral replication, release, and maturation

Once the genetic material of HIV is integrated into the host genome, it is ready to replicate and make more copies of viral particles. The integrated DNA provirus acts as a template which the host cell translates into RNA, encoding more than 30 RNAs including the HIV genome. Some of the newly transcribed viral RNA, undergo splicing to produce mature messenger RNAs (mRNAs). These mRNAs are translated into the regulatory proteins Tat and Rev in the cytoplasm. As the Rev protein start to accumulates, it binds to copies of un-spliced viral RNAs inside the nucleus and makes them leave the nucleus [37]. Some of these full-length RNAs function as the genome of new viruses, while others function as mRNAs that are translated to produce the structural proteins Gag and Env. It is believed that the Gag chain is responsible for the formation of multimers with Gag-Pol which then bind to copies of the virus RNA genome to package them into new virus particles [38]. The new virus particle is then transported to the cell membrane for release by various host factors and cellular machinery [39]. The essential viral proteins remain inactive within the Gag and Gag-Pol chains of the new virion, rendering it non-infectious. It is only after the final step of maturation, involving the cleavage of polypeptide chain into active proteins; HIV adopts its fully mature form. Viral protease plays its vital role at this stage by cleaving the polypeptide chains into active enzymes [40]. Failure of protease to performs its task properly makes virion unable to infect new cells, and HIV can't replicate [41]. This fact has been exploited in anti-retroviral therapy by including protease inhibitors in the drug regimen.

2.8. HIV-1 Antiretroviral Drug Discovery

The primary aim of antiretroviral therapy (ART) is to decrease morbidity, improve the quality of life, prolong life, restore and preserve immunity and prevent transmission. Before 1996, only a few anti-retroviral options were available against HIV-1 infection. Management of HIV-1 infection was involved of mainly treatment against the opportunistic infections. Understanding of the life cycle of HIV-1 has led to the advance in the anti-retroviral therapy by exploiting multiple drugs targets at the specific steps in the viral replication cycle. The first drug developed was anti-HIV RT agent named Zidovudine, which was approved in 1987 [42]. The major breakthrough in the treatment of HIV-1 infection came in the mid-1990s with the development of RT and PR inhibitors. Later on, it was realized that combining several drugs, targeted at enzymes involved in various stages of viral life cycle, was the most effective way to treat HIV-1 infection [43]. This approach of combining three or more drugs from at least two different classes is referred to as highly active antiretroviral therapy (HAART). Currently, there are 25 anti-retroviral agents approved by the US (see TABLE 2), classified into 6 groups viz. entry inhibitors, nucleoside reverse transcriptase inhibitors (NRTI), non-nucleotide reverse transcriptase inhibitors (NNRTIs), Integrase inhibitors and protease inhibitors (PI) [44]. Usually, these drugs are used in combination as the part of HAART. Usual combinations include 2 NRTIs along with 1 NNRTI, PI or integrase inhibitors (also known as integrase nuclear strand transfer inhibitors or INSTIs).

Table 2: FDA-approved anti-HIV drugs.

list of FDA approved anti-retroviral agents classified into drug classes according to their viral targets[45].

Drug class	Drug name	FDA approval year
NRTI	Abacavir	1998
	didanosine	1991
	Emtricitabine	2003
	Lamivudine	1995
	Stavudine	1994
	tenofovir	2001
	Zidovudine	1987
NNRTI	Efavirenz	1998
	Etravirine	2008
	Nevirapine	1996
	Rilpivirine	2011
Protease Inhibitors	Atazanavir	2003
	Darunavir	2006
	Fosamprenavir (prodrug of Amprenavir)	2003
	Indinavir	1996
	Nelfinavir	1997
	Ritonavir	1996
	Saquinavir	1995
	Tipranavir	2005
Fusion inhibitors/Entry Inhibitors	enfuvirtide	2003
	maraviroc	2007
Integrase Inhibitors	dolutegravir	2013
	elvitegravir	2014
	raltegravir	2007
Pharmacokinetic Enhancers (Increase the effectiveness of an HIV medicine)	cobicistat	2014

Entry or fusion inhibitors works by interfering with the HIV entry or fusion by blocking different viral proteins. Currently, available drugs in this class are maraviroc and enfuvirtide. Maraviroc targets the CCR5, a co-receptor located on human helper T-cells. Enfuvirtide –a peptide molecule— works by binding to gp41 surface protein of HIV and prevent infection of host cells [46]. Enfuvirtide must be administered as IV injection to avoid degradation by gut enzymes.

NRTIs inhibit the elongation of transcribed product by replacing the normal nucleoside with a modified nucleoside, which lacks a 3' OH group, thus terminating the chain elongation during the reverse transcription. NNRTIs bind to an allosteric site ~10 Å away from polymerase site of the HIV RT and alters its conformational landscape, thus preventing the enzyme from correctly performing the reverse transcription. Protease inhibitor (PI) prevent the HIV infection by blocking the protease necessary for the cleavage of gag and gag/pol precursor proteins [47]. Virion formed in the presence of PI lacks mature gag and pol polyprotein and subsequent protein products. Such virus particles are defective and mostly non-infectious. Integrase inhibitors prevent the integration of viral DNA with the host chromosome by blocking the integrase enzyme.

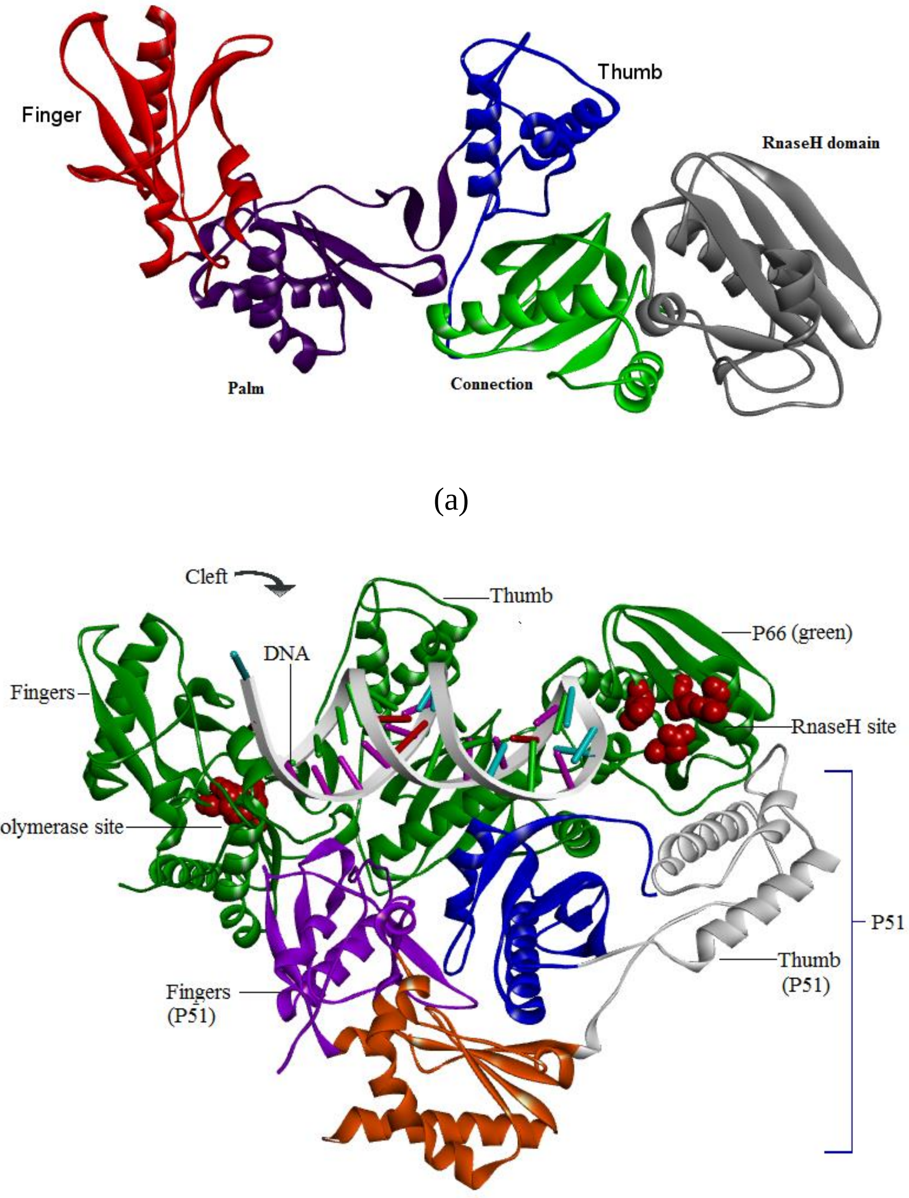
2.9. HIV-1 RT structure, function and inhibition

With the discovery of reverse transcriptase (RT) in 1970, the central dogma of life was changed to accommodate the prospect that genetic information can proceed in reverse direction *i.e.* from RNA to DNA [48]. The biological process of passage of genetic information from RNA to DNA is known as reverse transcription[1], mostly observed in retroviruses. Enzyme HIV-1 RT catalyzes the conversion of genomic ssRNA into dsDNA after the viral entry into the host cell. A brief detail of structure and function of HIV-1 RT is presented in this section.

[1] Reverse transcription is the reverse flow of genetic information. Transcription is the first step in the gene expression where DNA is copied into its complementary nucleotide chain of mRNA, with the help of RNA polymerase enzyme. Transcribed mRNA is then move to cytoplasm, where it takes part in the translation: the process of bio-synthesis of proteins. Information contained within the genes (DNA) is expressed via the proteins, which control various cellular processes.

2.9.1. Structure

HIV-1 RT is an asymmetric heterodimer consist of two subunits. HIV-1 RT starts as a homodimer of two 66 kDa subunits with 560 amino acid residues, but later during the protein processing, one of the subunits is cleaved by the proteolytic enzyme, leaving it without 120 residues at the C-terminal. The two subunit are, the larger P66 and the smaller P51, later being different than former in its conformation, as the result of proteolytic cleavage [49]. The 3D crystal structure pf HIV-1 RT is analogous to the human right hand, with fingers, thumb, palm and connection sub-domains (see FIGURE 6). Although subunits p51 and p66 have the same amino acid sequences, there are some significant conformational differences. The p51 has no cleft and the residues that are involved in the catalytic functions of the enzyme are buried deep. The finger of P51 is situated towards the palm sub-domain and the palm is positioned further away from the fingers and palm sub-domains as compared to p66. The p51 subunit mainly carries out a structural role by serving as a support to p66 and stabilize RT. It is interesting to note that the RNase H sub-domain is cleaved from P51 subunit. It has been speculated that the proteolytic cleavage of a part of P51 subunit gives an evolutionary advantage to HIV by enabling it to encode two protein subunits from the same gene. Thus HIV is able to produce the proteins with more than one structure and function [50]. HIV-1 RT is a multifunction enzyme with two distinct active sites viz. DNA polymerase and RNase H, both located on P66 subunit. The discarded portion of P51 unit consisted of RNase H domain. Numerous crystal structure of free HIV-1 RT as well complex with drug molecules and RNA/DNA template has been solved and studied.



(b)

Figure 6: Structure of HIV-1 RT

(a) P66 subunit (b) both P66 and P51 subunits along with a DNA template. The structure is homologous to the human right hand. P51 subunit also has these sub-domains, but lack the RnaseH part. The catalytic amino acids of the Polymerase and RnaseH sites are shown in red CPK. Figure is drawn using PDB code 1RT1.

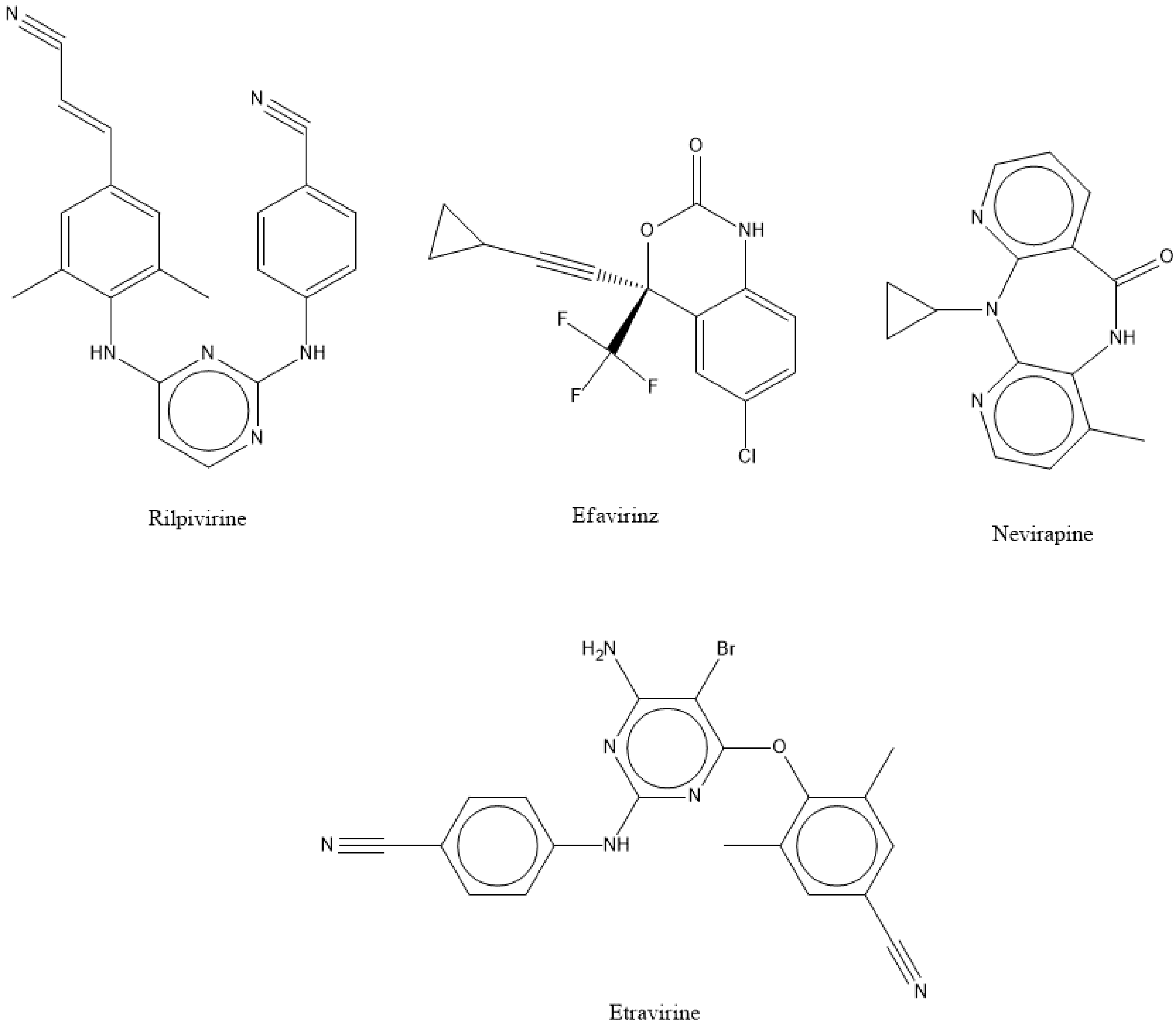
2.9.2. Function and molecular mechanism of polymerization

RT has three enzymatic functions; a RNA-dependent DNA polymerase activity to create a copy of ssDNA from the RNA template, a DNA-dependent DNA polymerase activity to make second DNA template from ssDNA, and a RNase H site to cleave the RNA template from the RNA-DNA hybrid. The polymerase and RNase H site of HIV-1 are located on P61 subunit and spatially separated by approximately 60 Å from each other. Polymerase site is situated in the palm subdomain between thumb and fingers and RNase H at the end of the P66 subunit. The polymerase domain comprises of four sub-domains: fingers (residue 1–85 and 118–155), palm (86–117 and 156–236), thumb (237–318), and connection (319–426). Three catalytic aspartic acids (D110, D185, and D186) forms the active site of the polymerase in the palm region (see FIGURE 6) [51]. P51 folds into the same four sub-domains as the polymerase domain of p66 (fingers, palm, thumb, and connection); yet, the positions of the sub-domains in relation to each other are different in p66 and p51 (see FIGURE 6). The RNase H active site is composed of three aspartic acids and one glutamic acid (D443, E478, D498, and D549) [51].

Extensive biochemical and crystallographic studies have led to an understanding of the mechanism of the DNA polymerization carried out by HIV-1 RT. The process begins with the binding of RT to the viral RNA template, which led to the opening of P66 thumb-finger cleft. Next step after the binding of RT with nucleotide sequence is the nucleotide incorporation, which starts with the binding of dNTP (nucleoside triphosphates containing deoxyribose) at the nucleotide binding site (N site) [52]. Subsequently, thumb of P66 closed down on incoming dNTP, in order to accurately line up the 3′-OH of the primer, the phosphate of the dNTP, and the polymerase active site [52, 53]. It has been shown that upon binding of the dNTP, certain residues in a loop between 60 and 75 of the P66 bends inwardly towards the active site. In particular, residues K65 and R72 interact with the incoming dNTP, forming salt bridges with the phosphates [52]. Phosphodiester bond is established between the dNTP and the primer with the associated release of pyrophosphate. Afterward, the finger of P66 sub-domain opens up to let the pyrophosphate leave the active site. The nucleic acid substrate translocates to free the nucleotide-binding site in RT for the next incoming dNTP.

2.9.3. Inhibition of HIV-1 RT

As HIV-1 is a retrovirus that requires RT enzyme to replicates, it is one of the popular targets of anti-retroviral drug development [54]. There are two different class of drugs targeting the HIV-1 RT. NNRTI and NRTI; both target a different aspect of RT functioning. A brief description of drugs inhibiting the HIV-1 RT is given in the following section, with an emphasis on NNRTI, as it is the main focus of this thesis. FIGURE 7 shows chemical structure of US FDA-approved NNRTI agents.



Rilpivirine

Efavirinz

Nevirapine

Etravirine

Figure 7: Chemical structure of FDA-approved NNRTI

2.9.3.1. Non-nucleoside RT inhibitors (NNRTIs)

NNRTIs are, in general, small (< 600 Da), hydrophobic compounds having a diverse range of chemical scaffolds. They inhibit the virus by binding non-competitively at the allosteric site [55] known as NNRTI binding pocket (NNIBP), located approximately 10 Å from the polymerase active site. NNIBP is neither present in apo nor in substrate-bound RT but comes into existence after drug binds to RT due to the rotation of side chains of Y181 and Y188. The hydrophobic binding pocket is in the hinge between the thumb and palm sub-domains and consists of L100, K101, K102, K103, V106, T107, V108, V179, Y181, Y188, V189, G190, F227, W229, L234 of p66 and E138 of p51 [56]. Several modes of entry to the pocket has been proposed, most common being near the p66/p51 interface surrounded by K101, K103, and V179. As a result of NNRTI binding, the normal conformational landscape of RT is disrupted, thus blocking its enzymatic activity [54]. There are four US FDA approved drugs in the NNRTI class (see FIGURE 7, TABLE 2). Various theories have been put forward explaining the method of RT inhibition by NNRTIs. A brief account of main theories is given below:

According to one theory, NNRTI binding disrupts the conformation of palm domain of RT, effectually altering the geometry of polymerase site. It is suggested that the process of polymerization by RT is extremely reliant on the alignment of the catalytic residue 185 and 186, thus any distortion of the active site inhibits its catalytic function [56].

Another model is known as ‘arthritic thumb model’ suggests that NNRTI binding interrupts the movement of thumb thus adversely affecting the RT function [57]. NNRTIs are also shown to block the RT by altering the dimerization of its two subunits and subsequently affect the structural stability [58].

2.9.3.2. Nucleoside RT inhibitors (NRTIs)

NRTIs are competitive inhibitors that compete with the natural dNTP substrate for incorporation into the growing DNA chain. NRTI inhibit the elongation of transcribed product by replacing the normal nucleoside. They share a similar structure to dNTP’s in that they both have a nitrogenous base and are attached to a ribose sugar, but NRTIs lacks a 3' OH group, thus terminating the chain elongation during the reverse transcription. There are currently seven NRTIs approved by US FDA to be used as the part of the anti-retroviral regimen (see TABLE 2 and FIGURE 8). Once the drug enters cells they need to be phosphorylated by cellular kinases to become active, hence NRTIs are administered as pro-drugs.

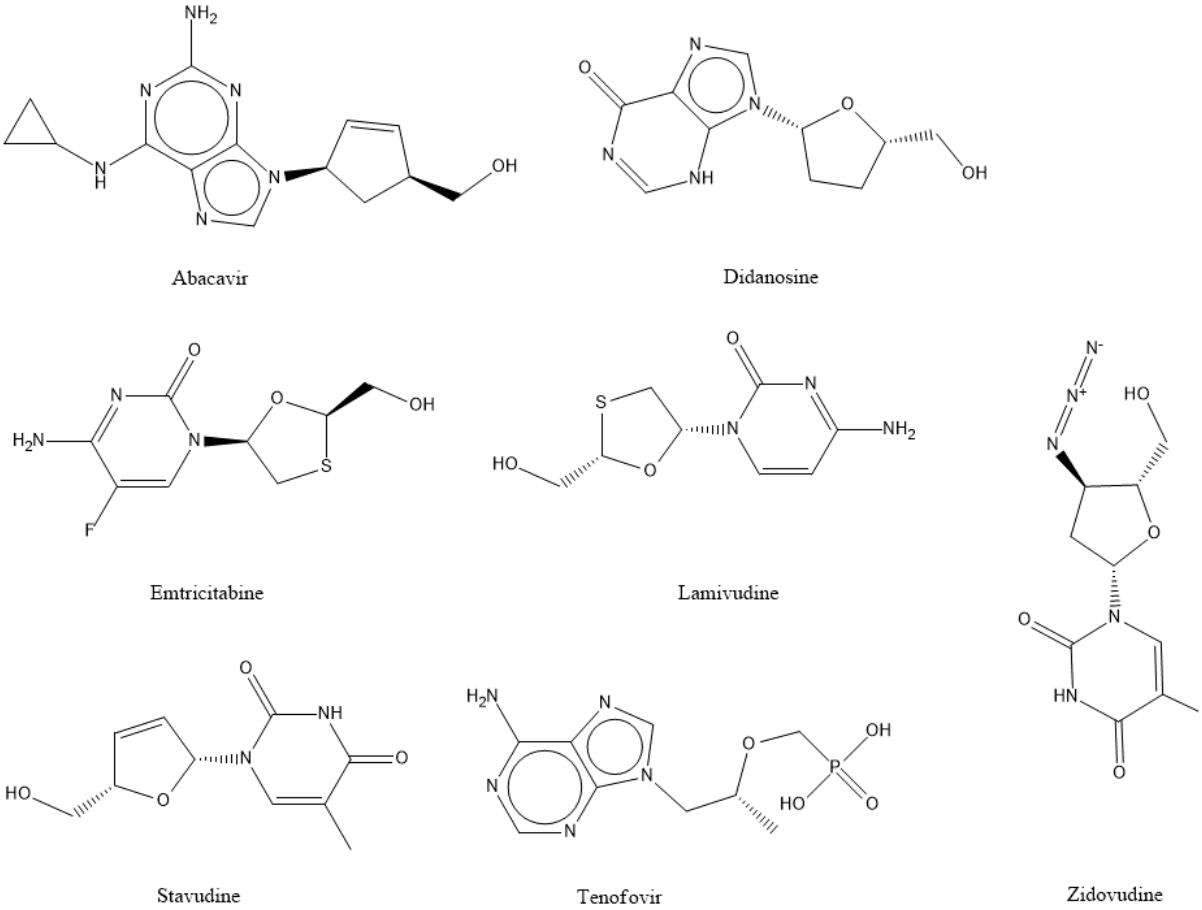


Figure 8: Chemical FDA structure of approved NRTI

2.10. Drug resistance in HIV

Regardless of advances in anti-HIV therapy, HIV infection remains an immense challenge due to the rapid onset of mutation instigating drug resistance. Though most HIV infections seem to be started by a single virus particle, enough mutations occur within a few years of infection to generate a group of related viruses with the differing genome. An HIV-1 infected person harbors a group of HIV-1 variants primarily originated from a single virus that had spread the infection [59]. There are various factors that contribute to the development of resistance with the possibility that resistant HIV can spread from person to person. The major reasons for the rapid emergence of drug resistance are:

1. The HIV RT is highly error-prone and lacks a proofreading ability, thus it has high mutation frequency.

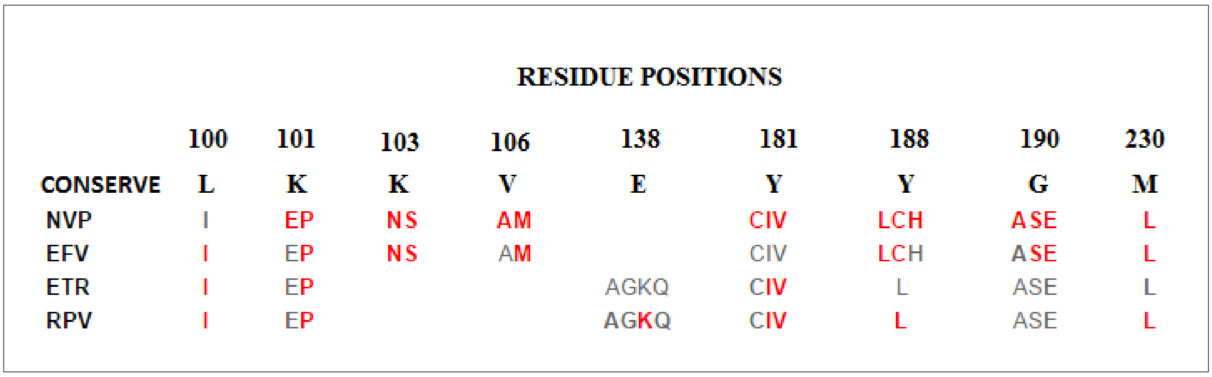
2. High replication rate of HIV (10^9-10^{12} new virus produced daily in untreated patients)

Due to this fact, many HIV mutants are produced, which differ in their susceptibility to ward anti-retroviral drugs, some develop into drug-resistant HIV strains. The positions of mutations that confer resistance to either NRTI or NNRTI drugs are mainly situated in the polymerase domain of RT [60].

2.10.1. Drug resistance in Reverse Transcriptase (RT)

Mutations in RT leading to drug resistance against NNRTI and NRTIs are shown in FIGURE 9. In the case of NRTIs, mutation either causes the enzyme to evolve greater specificity for the natural substrates [61] or it leads to rise in the efficiency of an excision reaction [62, 63]. The heightened specificity for the substrate is caused by mutations close to the dNTP binding site, whereas mutation in the distal region usually results in an increase in the efficacy of the template removal reaction.

Single residue changes are usually enough to confer high-level resistance to the NNRTIs. In the case of NNRTIs, almost all the drug resistance mutations are seen in and around the NNRTI binding pocket [64-66], particularly mutations at positions K103, Y181, Y188 and G190 cause significant effects on resistance (see FIGURE 9). Except for the E138 A/G/K/Q/R, all of the mutation is observed in the P66 subunit of RT. Unlike the polymerase site or dNTP-binding site of RT, the residues forming the NNIBP are not highly conserved, therefore, HIV-1 has a comparatively lower genetic barrier for developing NNRTI-resistance mutations than NRTI-resistance mutations. These mutations are assumed to be sterically altering the NNRTI interactions with the NNIBP residues[58]. However, as reported in crystal structures, K100N doesn’t alter the drug-binding pocket interactions [67], rather it is believed to affect the drug binding by raising the energetic barrier of the NNIBP formation [68].



RESIDUE POSITIONS

	100	101	103	106	138	181	188	190	230
CONSERVE	L	K	K	V	E	Y	Y	G	M
NVP	I	EP	NS	AM		CIV	LCH	ASE	L
EFV	I	EP	NS	AM		CIV	LCH	ASE	L
ETR	I	EP			AGKQ	CIV	L	ASE	L
RPV	I	EP			AGKQ	CIV	L	ASE	L

Figure 9: Mutations in HIV RT [69].

Mutations associated with drug resistance are shown in red.

However, mutations in NNIBP are not the only reason for the development of drug resistance toward the NNRTIs. It has been shown that besides the primary mutations in the binding pocket, drug resistance is also caused by the change of amino acids at other positions known as accessory mutations [70], which don’t directly interact with the drug. Mutations in the connection region of HIV-1 RT are also shown to cause resistance, such as N384I has been linked with NVP resistance [71] and the D549N, Q475A, and Y501A mutants have been observed to produce resistance to certain NNRTIs [72].

HIV-1 could develop resistance against the nevirapine, in the very beginning of treatment. First generation NNRTIs drastically lose their potency against a single common NNRTI-resistance mutation such as K103N or Y181C. Second generation DAPY analogues could adopt multiple conformations inside the binding pocket, thus maintaining efficacy against mutant HIV-1. However, as new drug-resistance forms emerge, understanding the molecular mechanism of NNRTI inhibition and resistance caused by different mutations is helpful for designing better anti-retroviral agents. The goal of such efforts being the discovery of effective NNRTI, which should overcome the impacts of common drug-resistance mutations.

CHAPTER 3

Computational methods in drug design and protein modelling

3.1.Molecular Modeling

The field of molecular modelling of Biosystems by computer has been gradually getting attention from scientists from diverse backgrounds. Specifically, modelling huge biological polymers like proteins, nucleic acids, and lipids require a highly multidisciplinary approach. Molecular modeling involves the study of molecular structure and function through theoretical and computational methods. The computational modeling involves a range of methods, for example, ab initio, semi-empirical quantum mechanics, empirical (molecular) mechanics, molecular dynamics, Monte Carlo, free energy methods, quantitative structure/activity relationships (QSAR), molecular docking, homology modeling, and many other conventional methods [73, 74]. The contemporary research problem being addressed by molecular modeling are as fascinating and as complex as the biological systems themselves. An assortment of issues is being addressed such as the dynamic structure of a biomolecule, energetics of hydrogen-bond formation in proteins and nucleic acids, protein folding, the complex functioning of a supramolecular aggregate and energetics of ligand-protein binding. Modeling of biomolecules offers a systematic way to understand structural/dynamical/thermodynamic phenomena, test and develop hypotheses, interpret and extend experimental data, and help better comprehend and extend basic laws that rule molecular wonders [73]. The concept of molecular modeling started with the idea that geometry, structure, energy, and various molecular properties can be computed from physical models. Such models consider atoms as solid sphere connected by springs (bonds) with each other. The molecule rotates, vibrates, and translates to take energetically preferred conformations as a combined result of the inter and intramolecular forces acting upon it. In the next section, the applied molecular modeling techniques in this thesis will be briefly discussed.

3.2.Quantitative Structure-Activity Relationship (QSAR)

An emerging field in cheminformatics, that deals with the prediction of physicochemical and biological properties of molecules with computational models, is referred to as QSAR (Quantitative Structure-Activity Relationship). Beside its application in prediction of various properties, such as solubility, lipophilicity, toxicity, mutagenicity [75-77], QSAR is now routinely employed as a tool in drug design workflow. QSAR is the mathematical modeling of chemical structures of compounds and their relationship with biological activity and is actively used in drug design [78, 79].

$$Activity\,/\,Toxicity \;= f\left(physiochemical\,/\,structural\ properties\right) + error \qquad 1$$

The key notion of QSAR is that molecules with similar chemical structures have similar properties and a change in molecular structures results into a change in its biological activities and physiochemical properties. There are three main components of the QSAR modeling:

1. The property to be modeled

2. Chemical information

3. The algorithm to model the relationship between the biological end point and the structural properties.

Prediction of the property of a chemical molecule depends upon the acquisition of knowledge of property values for a set of similar molecules referred to as the training set, which usually contains the results of experimental measurements. An essential concept is that any experimental value is associated with certain uncertainty. This is especially true for biological data, where often experimental values have a probabilistic meaning. Regrettably, the information about the uncertainty of the experimental data is not always available, and often users ignore the fact that this assessment is fundamental.

In the paradigm of drug design, knowledge of the relationship between structural properties of chemical compounds and their biological activities is crucial in optimizing lead molecules. The construction of QSAR models typically consists of two main steps: (i) calculation and representation of structural features (molecular descriptors) of the selected compounds; (ii) multivariate analysis for correlating molecular descriptors with the measured activities (biological, physico-chemical and ADMET properties). Numerous methods, ranging from simple linear regression to complex machine learning algorithms are applied to explore the structure-activity relationships. The linear regression models, in general, allow a relatively straightforward interpretation in terms of linear regression coefficients, provided that descriptors used in the equation are not correlated. However, the models obtained by machine learning methods, such as Neural Network and Support Vector Machine, are more difficult to interpret, due to the non-linear nature of the algorithms. The machine learning algorithms could be divided into two categories: those for regressions and those for classification. Regression methods get a continuous value, whereas classifiers find the category, *e.g.* the active or non-active status, the toxicity class etc. FIGURE 10 lists some of the common classification and regression methods used in QSAR. In next sections, some of the important technical aspects of QSAR methodology will be briefly described.

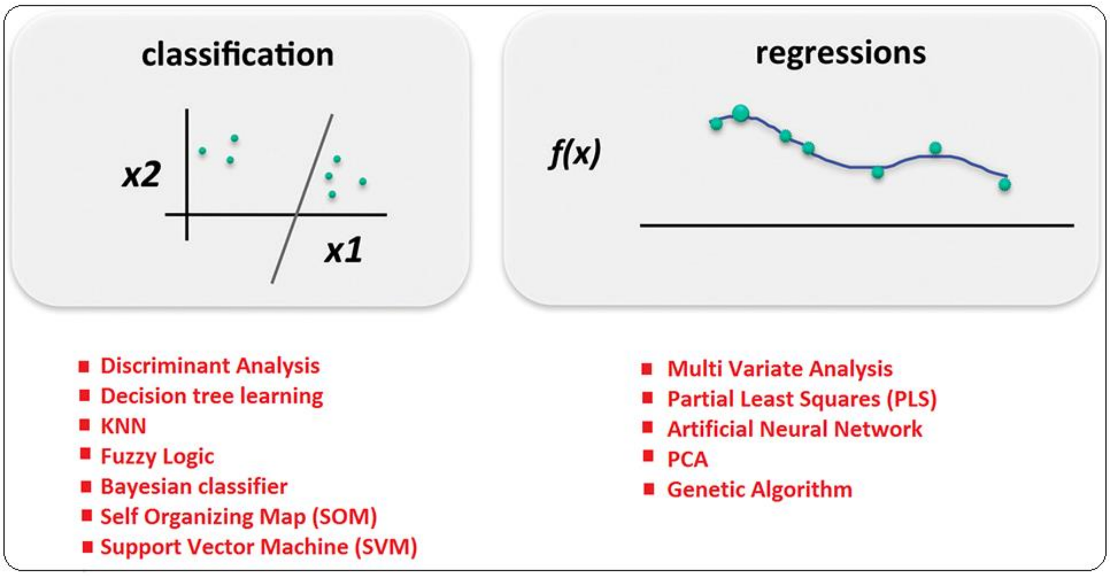


Figure 10: Machine learning techniques are commonly used to build QSAR models.

3.2.1. Molecular descriptors

To employ machine learning methods – abstract mathematical methods– in QSAR modeling, chemical structural information must be represented in numerical form. A set (vector) of the numerical description of a chemical compound’s features are referred to as molecular descriptors. Due to the complex nature of the molecular structure, it can be represented numerically in a fundamentally unlimited number of ways. Thus, the selection of a set of optimal descriptors is crucial for successful modeling. The molecular descriptors are separated into two main categories: experimental measurements, such as log*P*, molar refractivity, dipole moment, polarizability, etc. and theoretical molecular descriptors, which are derived from a figurative representation of the molecule. The theoretical molecular descriptors can be further classified into the three broad categories based on the degree of structural information they encode as:

- 0 D descriptors – for example, constitutional descriptors, count descriptors
- Linear or one-dimensional descriptors - such as molecular weight, the number of particular types of atoms or functional groups, the number of fragments etc.
- 2 D descriptors – based on graph theory.
- 3 D descriptors – based on the three-dimensional structure of a molecule.

An example of some of the descriptors used in this study is given in **TABLE 5**.

3.2.1.1.Dragon descriptors

Dragon descriptors encompass a vast variety of 1D, 2D and 3D descriptors grouped into 20 logical blocks. Spanning a vast variety of descriptor types, the Dragon descriptors are very prevalent and are often the first choice for QSAR modeling of various properties [80].

3.2.1.2.QNPR descriptors

These descriptors are used for Quantitative Name Property Relationship, thus giving the acronym QNPR [81]. The descriptors are calculated straight from the compounds name or SMILES representation. For each chemical molecule, either canonical SMILES or IUPAC name are split into the fragments of a specified length, which is determined by the configuration.

3.2.1.3.Chemaxon Descriptors

Chemaxon descriptors are a group of molecular descriptors implemented in OCHEM platform[82] and can be calculated for any set of molecules. The implemented descriptors are divided into 7 groups: Elemental analysis, charge, geometry, partitioning, protonation, isomers, and others.

3.2.1.4.Inductive Descriptors

Inductive descriptors are based on the Linear Free Energy Relationships (LFER) equations for inductive and steric substituent constants. Calculation of the descriptors is based on the inductive and steric effects, inductive electronegativity and molecular capacitance of the molecule. Such parameters can be easily computed from electro negativities and covalent radii of the constituent atoms and interatomic distances, which reflect the different aspects of intra- and intermolecular interactions[83].

3.2.1.5.Fragmentor

The ISIDA Fragmentor descriptors are the part of ISIDA Fragmentor2015 program developed at Université de Strasbourg, France. This program is a part of the ISIDA project, which stands for "In SIlico Design and data Analysis" and aims to develop tools for the calculation of descriptors, the navigation in chemical space, QSAR and virtual screening. The descriptors include molecular fragment count based on a series of graph algorithm. They are a type of fragments descriptors, which uses 2D Lewis graph representation of the compounds but do not consider stereoisomerism [84-88].

3.2.1.6.ALogPS

It includes two descriptors; (1) ALogPS_log*P*: octanol/water partition coefficient. (2) ALogPS_logS: solubility in water.

3.2.1.7.GSFrag

The GSFrag program calculates the occurrence numbers of certain special fragments on k=2 to 10, vertices in a molecular graph *G*. The molecular fragments consisting of one or more disconnected components, where each component is a path (of length 9 or less), a cycle (on 10 or fewer vertices), or a path (cycle) with a number of attached chains of unit length [89, 90].

3.2.1.8.MerSy (MERA Symmetry)

MerSy descriptors are calculated using a 3D representation of molecules in the framework of MERA algorithm and include the quantitative estimations of molecular symmetry with respect to symmetry axes from C2 to C6 and to the inversion-rotational axis from S1 to S6 in the space of principal rotational invariants about each orthogonal component.

3.2.1.9. Adriana

Adriana is a group of descriptors calculated by ADRIANA.code – an Algorithms for the Encoding of Molecular Structures. It contains a unique group of empirical methods for calculating molecular descriptors on a comprehensive geometric and physicochemical basis. A multilevel sophisticated hierarchy is used to represent 3D molecular structure. Adriana consists of a different set of descriptors such as topological, shape based, and 3D property-weighted autocorrelation descriptors.

3.2.1.10. Spectrophores

Spectrophores are 1D descriptors calculated from the property fields neighboring the molecules. The computation of the Spectrophores descriptors is independent of the geometry of the molecule which enables the rapid assessment of different molecules.

3.2.2. Machine learning methods in QSAR

The machine learning methods are typically used for QSAR predictions are based on supervised learning.

3.3. Matched Molecular Pair Analysis (MMPA)

MMPA is a method in cheminformatics in which significant structural changes, within a database of drug-like molecules, is established based on experimentally measured data. The term was first coined by Kenny and Sadowski [91] in their book titled "Chemoinformatics in Drug Discovery". The basis of MMPA-based analysis is the analysis of the chemical datasets dealing with pairs of compounds. Such pairs of compounds are known as Matched Molecular Pair (MMP), defined as a pair of molecules that differ in only a minor single point change (see TABLE 7).

3.4. Atomistic Modeling of Proteins

Molecular modeling includes all the theoretical and computational methods, used to study or model the molecules at the atomistic level. The molecular modeling is used in varied areas of computational chemistry, drug design, computational biology and materials science to study diverse molecular systems. Experimental techniques, for instance, x-ray crystallography and NMR spectroscopy allow us to determine the 3D protein structures. With the aid of NMR spectroscopy, even motions of proteins can be investigated. Nevertheless, in many cases, it is not feasible to explore protein structure and dynamics using experimental techniques. *In silico* techniques like homology modeling and molecular dynamics could provide insight in such cases. In the following sections, a brief description is given for a variety of such modeling techniques, focusing on molecular dynamics (MD). Experimental data are required to model the starting atomic configuration of a bio-molecular system for all the simulation approaches. A brief overview of experimental techniques is given in appendix I.IV. Thus we begin with a short note on the most frequently used techniques for determination of protein structure.

3.5. Computational modeling approaches used for enhancing structural understanding

The molecular modelling methods deals with the atomistic level description of the molecular systems. Different level of details is employed, such as atoms as the smallest individual unit (molecular mechanics approach), explicit treatment of electronic waves functions each atom (quantum mechanical approach), or a hybrid approach of QM/MM modeling. More details about these approaches are given in appendix I.I.

3.5.1. Molecular dynamic (MD) simulations

Bio-molecular dynamics occur over a wide range of time and space, and the choice of method to study them is influenced by the questions asked. MD simulation is a very powerful technique in modern molecular modeling, which enables to explore structure and dynamics in depth detail—basically, the motion of individual atoms can be traced. Macroscopic properties determined in an experiment are not direct observations, but averages over billions of molecules. This collection of molecules representing a set of measurable properties is termed as statistical mechanics *ensemble*. MD and other classic simulation methods are dependent on *force fields* – an empirical calculation of atomic interaction and subsequent evaluation of the potential energy of the system as a function of point like atomic coordinates. A force field represents both the set of equations used to calculate the potential energy as well as force constants between atoms. Force fields also consist of a set of parameters used to fit the equations. In the formalism of MD, the atoms are treated as a small solid ball with a given mass and charge. The charges are used to compute an electrostatic force field from which the force on each atom in the system can be estimated. The force on each atom can then be used to compute the motion of the atoms according to the Newton's equation of motion. After time *t* new positions of atoms in the systems is updated. This process is then iterated to evolve the system configuration. The MD simulation involves a certain level of approximation and for most cases, these approximations work well, but they fail in modeling the quantum effects such as bond formation or breaking [92, 93]. Nevertheless, MD has been used to study wide varieties of biological [94-98] and non-biological [99] systems[100].

3.5.2. Theory of the MD Simulations

The molecular dynamics method was initially introduced by Alder and Wainwright in late fifties [101, 102] to study the interactions of hard spheres systems. The next key development was when the simulation using a realistic potential for liquid argon was carried out by Aneesur Rahman in 1964 [103]. The foremost MD simulation of a realistic system was on the liquid water in 1974 [104]. The first MD simulations of protein appeared in 1977 with the simulation of the bovine pancreatic trypsin inhibitor (BPTI) [105]. In 2013 Martin Karplus, Michael Levitt and Arieh Warshel were awarded the Nobel prize in chemistry for "the development of multiscale models for complex chemical systems". Now, one regularly finds MD simulations of solvated proteins, protein-DNA complexes as well as lipid systems addressing a range of subjects including the thermodynamics of ligand binding and the folding of small proteins. In a molecular dynamics simulation, the time-dependent behavior of the molecular system is obtained by integrating the Newton's equations of motion using a suitable numerical integrator and the potential energy function. The result of the MD simulation is a time series of conformations or atomic positions; called MD trajectory. Most molecular dynamics simulations are performed under conditions of constant N, V, E (Microcanonical ensemble). FIGURE 11 gives an overview of the general methodology of a MD simulation.

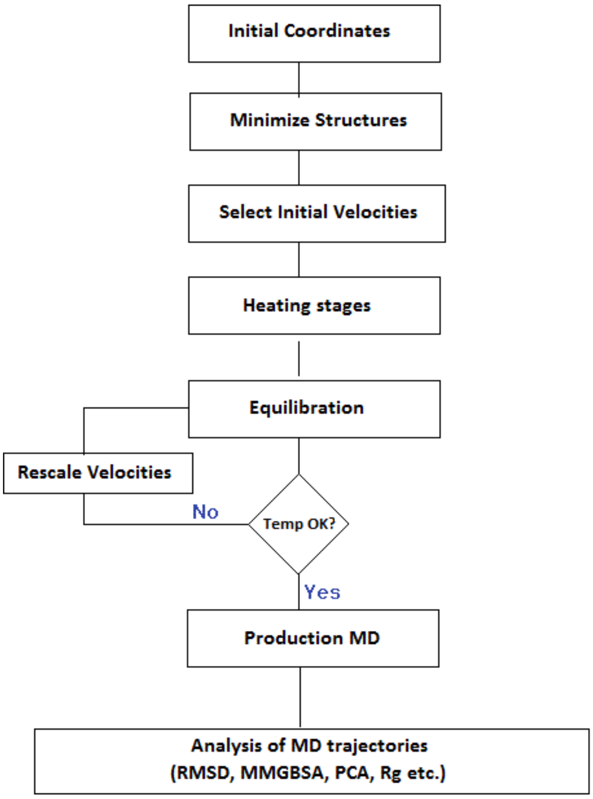


Figure 11: The protocol of setting up and running a molecular dynamics Simulation.

In this section, we describe in some detail the steps taken to setup and run a molecular dynamics simulation. In the following section, the theoretical basis of MD simulation will be discussed.

Fundamentally the methodology of MD is very simple. Firstly, all the atoms in a given system are assigned coordinates, velocities, and charges. The positions and charges of atoms are then used to calculate a potential. This calculated potential is used to compute the force experienced by each of the atoms in the simulation. By integrating Newton's laws of motion over a short time step a new set of positions and velocities is determined for each of the atoms. The updated values can now be re-used into the first step of the calculation and the process is repeated, creating a trajectory that describes the positions, velocities and accelerations of the particles as they change with time. MD produce information of the atomic systems at the microscopic level, including atomic positions and velocities. The conversion of this microscopic information to macroscopic observables such as pressure, energy, heat capacities, etc., requires statistical mechanics. It is a branch of theoretical physics that studies the average behavior of macroscopic systems, using probability theory, where the state of the system is uncertain. Statistical mechanics is essential for the study of biological systems by molecular dynamics simulation. Extensive details on the subject could be found in textbooks dealings with statistical mechanics [106].

From a given MD trajectory, the average values of macroscopic properties can be determined. MD is deterministic in nature *i.e.* once the positions and velocities of each atom are known, the state of the system can be predicted at any point in time. Molecular dynamics simulations can be time-consuming and computationally expensive for a large system. However, with the advances in parallel computing and supercomputing clusters, biological systems are being investigated at a larger time scale. Simulations of solvated proteins are calculated up to the nanosecond time scale, however, simulations into the millisecond regime have been reported.

3.5.3. Classical Mechanics

MD simulation consists of the numerical, step-by-step, solution of the classical Newtonian equations for N-particle system. A system of N-particle can be completely described by 3N generalized coordinates q_i (where i = 1,2,3. . . 3N), 3N generalized velocities q_i and a potential energy function $V(q_i)$.

Force acting on an atom can be given by Newton's equation of motion.

$$F_i = m_i a_i \qquad 2$$

where F_i is the force exerted on particle *i*, m_i is the mass of particle *i* and a_i is the acceleration of particle *i.* The force is also given by the gradient of the potential energy as:

$$F_i = -\nabla V_i \qquad 3$$

By combining the two equations, we have:

$$-\frac{dV}{dr_i} = m_i \frac{d^2 r_i}{dt^2} \qquad 4$$

where V is the potential energy of the system.

The acceleration of an atom is given as the derivative of the potential energy with respect to the position, r.

$$a = -\frac{1}{m}\frac{dV}{dr} \qquad 5$$

We can see that, to calculate a MD trajectory, one only needs the initial positions of the atoms, an initial distribution of velocities and the acceleration, which is determined using the gradient of the potential energy function. The initial atomic positions can be obtained from experimental structures, such as the x-ray crystallography or NMR spectroscopy. The initial distribution of velocities is determined randomly from a Maxwell-Boltzmann or Gaussian distribution at a given temperature and corrected so that the overall momentum is 0. It gives the probability that an atom *i* has a velocity V_x in the *x* direction at a temperature T.

$$p(V_{ix}) = \left(\frac{m_i}{2\pi k_B T}\right)^{1/2} \exp\left[-\frac{1}{2}\frac{m_i v_{ix}^2}{k_B T}\right] \qquad 6$$

The potential energy is a function of the atomic positions (3N) of all the atoms in the system. Due to the complicated nature of this function, there is no analytical solution to the equations of motion; they must be solved numerically. Numerous numerical algorithms have been developed for integrating the equations of motion such as Verlet algorithm, Leap-frog algorithm, Velocity Verlet and Beeman's algorithm.

3.5.4. Potential Energy Function and Force Fields

To estimate the force on each atom, it is first essential to compute the potential energy function. Although a precise calculation of the potential energy of a N atom system would have to reflect the contribution of each individual atom, pair, triplet and so on, most MD packages define the potential energy in terms of five components. The energy, E, is a function of the atomic positions, R

$$V(R) = E_{bonded} + E_{nonbonded} \qquad 7$$

The E_{bonded} has three terms (see FIGURE 12)

$$E_{bonded} = E_{stretch} + E_{bend} + E_{rotate} \qquad 8$$

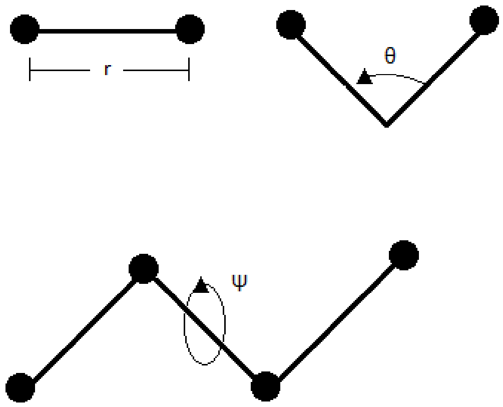


Figure 12: The different components of bonded interactions

different components of bonded interactions in the interatomic interaction potential of an MD force field. where r governs bond stretching, θ the bond angle and ψ the dihedral angle between two atoms separated by three bonds.

The bonded components of bond stretching, bending and torsional bonded interactions can be represented in terms of the deviation of the bond length r, angle θ and dihedral angle ψ from a reference, or equilibrium value (see Eq. 9).

$$\begin{aligned} E_{bond} &= \sum_{bonds} k_r \left(r - r_{eq}\right)^2 \\ E_{angle} &= \sum_{angles} k_\phi \left(\phi - \phi_{eq}\right)^2 \\ E_{dihedral} &= \sum_{dihed} \frac{E_n}{2}\left(1 + \cos(n\phi - \gamma)\right) \end{aligned} \qquad 9$$

The contribution of non-bonded interactions has two components, the Van der Waals interaction energy and the electrostatic interaction energy.

$$E_{non-bonded} = E_{vdw} + E_{electrostatic} \qquad 10$$

The van der Waal's forces are approximated as a Lennard-Jones 6-12 potential (Eq. 11), and the electrostatic interaction energy is given by the Coulomb potential (Eq. 12).

$$E_{vdw} = \sum_{non-bond} \left(\frac{A_{ik}}{r_{ik}^{12}} - \frac{C_{ik}}{r_{ik}^{6}} \right) \qquad 11$$

$$E_{elec} = \sum_{nonbond} \frac{q_i q_k}{Dr_{ik}} \qquad 12$$

The constants k_r, r_{eq}, k_θ, θ_{eq}, etc. are constants which are estimated from standard parameterization schemes such as CHARMM [107] GROMOS [108] OPLS-AA [109] and AMBER [110]. The requisite parameters are determined by a combination of atoms of varying types and fitting to either experimental or ab initio quantum mechanical calculations. This approach assumes that the parameters derived from these small subsets of atoms can be applied with sufficient accuracy for a larger molecular system of the same set of atoms. Force fields may differ in their functional form and in the systems and physical conditions (such as temperature and pressure) for which they are parameterized. Two key differences exist in the treatment of bonded terms of the force fields. The first is the variable use of "improper" dihedrals, which can be used to retain chirality or planarity at an atom center. The second dissimilarity is that the CHARMM force field adds an Urey-Bradly angle term, which treats the two terminal atoms in an angle with a quadratic term that is subjected to the inter-atomic distance. Similar to the bonded terms, force fields also vary in the treatment of non-bonded interactions. A detailed comparison of different force fields can be found in the book titled "Molecular Modeling of Proteins" [93].

AMBER (Assisted Model Building with Energy Refinement) is a family of force fields for molecular dynamics of biomolecules. All the MD simulations performed in this work use the ff99SB and ff12SB (AMBER force fields), which are parameterized to be suitable for studying proteins. For the ligands in the MD systems, General AMBER force field (GAFF) was used, which provides parameters for small organic molecules to facilitate simulations of drugs and small molecule ligands in conjunction with biomolecules [111].

There are some intrinsic limitations with the use of any force field. The parameters in the frequently used force fields, for example, AMBER, have been validated for equilibrium structures over a small timescale but imprecisions might arise when systems drift away from the equilibrium or for pressure or temperature conditions very different from those used in the standard parameterisation. Additionally, a limited number of atomic combinations are used to create the parameter set, which is not sufficient for many molecular systems of interest.

A variety of MD softwares are available designed especially for the simulation of biomolecular systems such as AMBER [112], GROMACS [113], CHARM [107], NAMD [114], and

LAMMPS [115]. Of late the Desmond [116] package has been created for accelerated parallel MD simulations. These packages allow the use of existing force fields including variants of AMBER and CHARM force fields.

3.6. Binding Affinities

Chemical entities which bind to proteins are called ligands. Though in some cases ligands form irreversible covalent bonds with proteins, most bind via non-covalent bonds in a reversible manner. In the latter case, the bound and unbound states of protein and ligand achieve an equilibrium, which is expressed as the following chemical reaction, where k_1 and k_2 are rate constants.

$$A + B \underset{k_2}{\overset{k_1}{\rightleftarrows}} AB \qquad 13$$

In equilibrium, the concentration of free protein A, free ligand B and the complex AB is constant. In the thermodynamic term, reactions are driven by the minimization of a potential, the appropriate thermodynamic potentials are determined by the conditions in which the reaction occurs. In typical experimental conditions (also known as the NPT ensemble because the number of molecules, pressure, and temperature are kept constant) the potential is known as the Gibbs free energy denoted by G.

$$\begin{aligned} G &\equiv H - TS \\ A &\equiv U - TS \end{aligned} \qquad 14$$

where H is the enthalpy, S the entropy, T the temperature and U is the internal energy of the system. There are several formally exact methods (having no empirically fitted parameters), for calculation of free energy differences from molecular simulation. Two such methods are (a) free energy perturbation (FEP) which is based on exponential averages of the change in the potential energy, and (b) thermodynamic integration (TI), which is based on integrating the change in energy as one state of the system is steadily changed into another. Free energy is a function of the state of the system, and in a close thermodynamic system, the net change is zero. **FIGURE 13** represents a thermodynamic cycle which could be used to calculate the relative binding affinity (Δ*Δ*Gbind) of two different ligands to the same target by following Eq. 15

$$\Delta\Delta G_{bind} = \Delta G_B - \Delta G_A \qquad 15$$

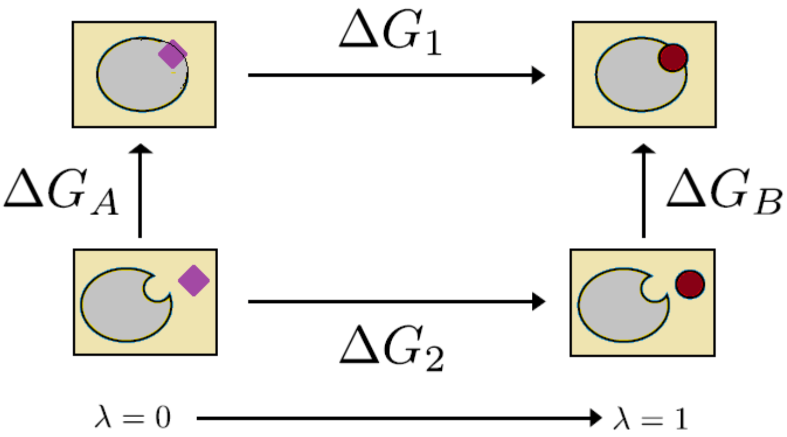


Figure 13: A thermodynamic cycle

A thermodynamic cycle demonstrating the indirect computation of the relative binding free energy difference.

The use of *ab initio* methods for calculating free energies is computationally expensive and time-consuming. Moreover, achieving convergence is also an issue with such methods. A variety of approximate methods has been developed to overcome the shortcomings of the exact methods. These methods implement less precise physical models and empirically fitted parameters. An example of such faster methods includes molecular docking, linear Interaction energy, molecular mechanics Poisson-Boltzmann surface area (MMPBSA) and its simpler form using the analytic generalized Born (MMGBSA) method [74].

3.7. Molecular Docking

Molecular docking is an important technique in structural molecular biology and computer-assisted drug design. The main objective of docking is to predict the principal binding mode(s) of a ligand with a protein's 3D structure. The binding modes of a ligand with its target protein can be distinctively defined by its state variables; position (*x*-, *y*-, and *z*-translations), orientation (Euler angles, axis-angle, or a quaternion), and, its conformation (the torsion angles for rotatable bond) if the ligand is flexible. All docking approaches involve a scoring function to rank the several possible binding modes and a search algorithm to efficiently explore the ligand's state variables. The scoring functions can be empirical, force-field based, or knowledge-based. A wide variation of 'scoring functions' both physical and empirical, are available for this purpose, such as AutoDock [117], X-Score [118], DrugScore [119], ChemScore [120], GOLD [121], FlexX [122], LigScore [123] and LUDI [124]. Search methods can also be classified into local and global. Search space is relatively small, in the case of inflexible ligand and/or rigid binding site in the receptor. A quite few machine learning algorithms are employed to search the ligands state variables space; for instance, Pattern Search[125], Monte Carlo Simulated Annealing (SA)[126], Genetic Algorithm (GA)[127], and Lamarckian GA[128]. Docking methods often ignore solvation contributions in the process of ligand binding.

3.8. Molecular Mechanics Poisson-Boltzmann Surface Area (MMPBSA)

The MMPBSA [129, 130] –a continuum/implicit solvation electrostatic method is the most rigorous of the approximate methods for binding free energy calculations on a range of biological systems. A reasonably detailed physical model and fast computation (compared to *ab initio* methods) have led to its extensive usage. The MMPBSA method involves the calculation of absolute binding free energies by computing average free energies of the enzyme-inhibitor complex, and the free enzyme, inhibitor separately. These free energy values are then used to calculate the change in free energies by the following equation, which is average over an ensemble of frames from MD trajectories.

$$\Delta G = G_{complex} - G_{enzyme} - G_{ligand} \qquad 16$$

The solvent and counter ions are removed from the MD trajectories and they are swapped by a continuum solvent representation in case the MD was performed with explicit solvent model and, a thermodynamic cycle is employed. The eventual objective of any free energy calculations method is the absolute free energy of binding in a solvent. However, in MD of solvated protein-ligand complex, most the energy contributions would come from solvent-solvent interactions, resulting in variations in total energy of an order of magnitude larger than the binding energy. Hence, direct calculations would require a very large number of snapshots to converge. The solution to this employed by MMPBSA is to use the thermodynamic cycle as shown in FIGURE 14. Where the binding free energy for system in vacuum is also calculated along with solvation energies of the complex, enzyme and ligand and the free energies for solvated system is given by:

$$\begin{aligned}\Delta G_b^{aq} &= \Delta G_b^{vac} + \Delta G_{com}^{sol} - \left(\Delta G_{enz}^{sol} + \Delta G_{lig}^{sol}\right) \\ &= \Delta G_b^{vac} + \Delta G^{sol}\end{aligned} \qquad 17$$

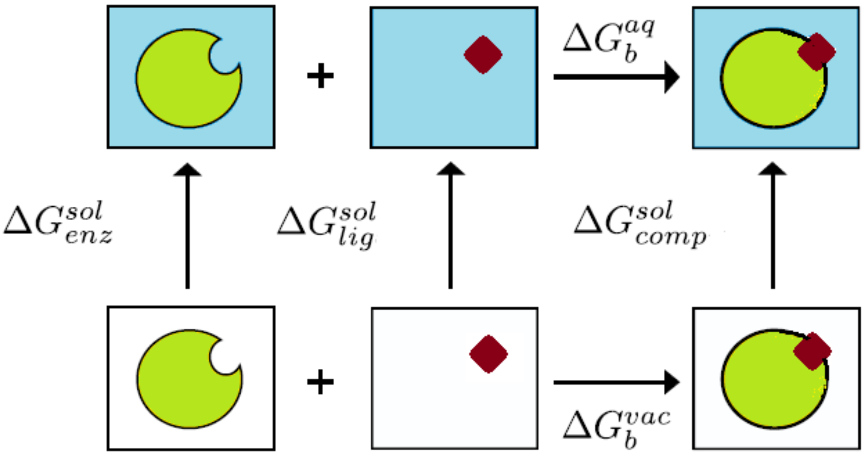


Figure 14: Thermodynamic cycle used in MMPBSA

The thermodynamic cycle used in to indirectly calculate the binding free energy in MMPBSA. The binding free energy changes in vacuo (ΔG_b^{vac}) and the solvation free energies for the complex (ΔG_{comp}^{sol}), enzyme (ΔG_{enz}^{sol}) and ligand (ΔG_{lig}^{sol}) are calculated and then ΔG_b^{aq} computed is computed by equation described above.

The ΔG_b^{vac} and ΔG^{sol} components of the binding free energies are calculated independently, using different methodologies. The free energies in vacuum ΔG_b^{vac}, is calculated using the molecular mechanics, which can be decomposed into a sum of electrostatic, van der Waals and internal molecular mechanics interactions by Eq.18.

$$\Delta G_b^{vac} = \Delta G_{ele}^{MM} + \Delta G_{vdW}^{MM} + \Delta G_{\text{int}}^{MM} \qquad 18$$

The term ΔG^{sol} (solvation free energy) represents the free energy change associated with the movement of the solute from vacuum into a solvent environment. This can be decomposed into contributions from polar and non-polar interactions between the solute and solvent as:

$$\Delta G^{sol} = \Delta G_{polar}^{sol} = \Delta G_{nonpolar}^{sol} \qquad 19$$

The polar component of solvation energy is computed by numerically solving either linearized Poisson-Boltzmann or Generalized Born (MMGBSA) equation with an implicit solvent modeled as a high dielectric constant. The non-polar component is calculated using an empirical term related to the solvent accessible surface area.

3.8.1. Polar solvation energy using Poisson-Boltzmann (PB) equation

To compute the polar solvation energy (ΔG_{polar}^{sol}), it is required to model the electrostatic potential surrounding the system in an appropriate solvent. The Poisson-Boltzmann equation uses a continuum implicit solvent model with high dielectric constant, aqueous ions as a "diffuse charge cloud" and the solute as a collection of fixed point charges implanted in a lower dielectric continuum. Although there are many statistical mechanics based derivation of the PB equation. It may be derived very easily from Poisson's equation [131, 132] also. Electric potential (φ) for a given charge distribution (ρ_f) can be estimated by solving the Poisson equation, where φ(r) at a point r generated by a charge distribution $\rho_f(r)$ in an environment of dielectric coefficient ε(r) (relative to the permittivity of free space, ε_0) is given by:

$$\nabla.\left[\varepsilon_o \varepsilon(r) \nabla \varphi(r)\right] = -4\pi\rho(r) \quad 20$$

For the simulation of biomolecules, the functional form of ε(r) is subjected to the molecular geometry with the biomolecule represented as continuum region of low polarizability embedded in a surrounding continuum solvent of higher polarizability. For a biomolecular system, usually the dielectric constant of the solute is chosen to be in the range of 1 to 4 and a value of 80 is used to represent water.

The charge distribution (ρ_f) comes from solute charge density (ρ_{f}(r)), and from the ions present in the solvent (c(r)). The ρ_{f}(r) can be defined as a set of delta functions centered on each solute atom's center and scaled by the atom's charge. The ion contribution is modeled as a continuum with charge distributed according to the Boltzmann distribution. For N ion species with charges q_n and bulk concentrations of c_n^{∞}, the ion charge distribution is given by:

$$c(r) = 4\pi \sum_{i=1}^{N} q_n c_n^{\infty} e^{-\beta q_n \psi(r)} \quad 21$$

where $\beta = 1/k_B T$

By combining the Eq. $\nabla.\left[\varepsilon_o\varepsilon(r)\nabla\varphi(r)\right]=-4\pi\rho(r)$ 20 and 21, we get the following equation:

$$\nabla.\left[\varepsilon_0\varepsilon(r)\nabla\phi(r)\right]+4\pi\sum_{i=1}^{N}q_n c_n^{\infty}e^{-\beta q_n\psi(r)}=-4\pi\rho_f(r) \qquad 22$$

In the case of an electrostatic neutral system, this can be simplified for two ions of equal bulk concentration, c ∞, with opposite charges of equal magnitude, q as follows:

$$\nabla.\left[\varepsilon_0\varepsilon(r)\nabla\phi(r)\right]+8\pi qc^{\infty}\sinh\left[\beta q\psi(r)\right]=-4\pi\rho_f(r) \qquad 23$$

This equation can be rewritten in a linearized form by expanding the hyperbolic sine function as a Taylor series as:

$$\nabla.\varepsilon_0\varepsilon(r)\nabla\phi(r)+8\pi q^2c^{\infty}\beta\psi(r)=-4\pi\rho\ (r) \qquad 24$$

Eq. 24 is the linearized Poisson-Boltzmann equation, which can be solved by wide verities of numerical methods. In MD packages, designed for simulation of biomolecules such as AMBER[112], the finite difference approach (FDM) is the most widely implemented method [133, 134]. A FDM comprises of succeeding steps of: (1) mapping atomic charges to the FD grid points; (2) assigning non-periodic/periodic boundary conditions; and (3) applying a dielectric model to define the high-dielectric (*i.e.*, water) and low-dielectric (*i.e.*, solute interior) regions and mapping it to the FD grid edges. These steps permit the partial differential equation to be transformed into a linear or nonlinear system. More details could be found in section 5.1.1 of AMBER14 manual [135].

3.9. Overview of molecular modelling studies on HIV-1 RT and NNRTIs

From the ligand-based drug design approach, the technique of QSAR has been used to study the relationship between the structure of NNRTIs and its anti-HIV activity [136]. There are many QSAR studies on the different chemical scaffold of anti-HIV agents [137, 138]. Such efforts have led to the realization of the importance of hydrophobicity, steric, butterfly-like structure [139], aryl, hetero-aryl moieties and their structure for the anti-HIV activity [140, 141]. Nevertheless, most earlier QSAR models for NNRTIs were developed on smaller data sets and a single class of compounds [142-145]. There was thus a need for a larger and more diverse dataset for QSAR studies. One of the important objectives of this thesis was to develop a robust QSAR model for selected NNRTIs that could predict anti-HIV RT activity. Another objective was to identify a potential chemical scaffold for further optimization, the aim being to understand the underlying structural changes that could contribute to improving the anti-HIV activity of the NNRTI.

Regardless of its importance as a drug target, RT has been less widely computationally modelled than other HIV targets, especially PR. This is due to the large size of RT (it contains approximately 1000+ residues). Several simulation studies conducted have been performed in implicit solvents on shorter timescale [68, 146], with restrained atoms [147] or only consider a part of the enzyme [148] in order to decrease the computational expense. While it has been stated that about one-third of the residues of RT are immovable [147] it is not clear that omitting certain part of RT during the simulation does not affect the dynamics of the system. Previous studies have dealt with the molecular basis of NNRTI resistance due to K103N mutation in RT [68, 146] using 500 ps explicit MD simulation. Rilpivirine is a Di-aryl pyrimidine (DAPY) derivative with potent anti-HIV-1 RT activity against both WT and mutant HIV-1 RT (FIGURE 21). Its ability to reasonably adapt to the K103N mutation in RT is assumed to be due to the structural flexibility and the hydrogen bond formed by the linker N atoms [149]. This has motivated us to investigate the dynamics of HIV-1 RT sub-domains in WT and K103N mutant, complexed with rilpivirine. The details of the MD study are discussed in 3.5.1. Regardless of improvement in anti-HIV therapy, HIV remains a challenge due to the rapid onset of mutations instigating drug resistance. Despite being the prime target of anti-HIV therapy, RT is responsible for emerging resistance to other drugs in the class: first, directly to RT inhibitors and/or second, indirectly as a key basis for instigating genetic variations [60]. Residues K101, K103, and E138 (p51) are situated at the rim of the NNRTI binding pocket (NNIBP) entrance for most NNRTIs. The mutations in NNIBP can lead to loss of aromatic ring stacking interactions (Y181C or Y188L), steric hindrance (L100I or G190A/S), and alteration of hydrophobic interactions (V106A or V179D). The effects of drug resistance mutations are rather severe on the inflexible first-generation NNRTIs, for instance, high level of resistance by Y181C to nevirapine. The K103N and E138K mutations are largely linked with treatment failure of the efavirenz and rilpivirine, respectively, when combined with tenofovir and emtricitabine [150, 151]. CHAPTER 6 present MD simulation of wild-type (WT) and E138K HIV-1 RT in complex with efavirenz (EFV), etravirine (ETR), and rilpivirine (RPV).

CHAPTER 4

QSAR models and scaffold-based analysis of non-nucleoside HIV RT inhibitors

Bilal Nizami[a], Igor V. Tetko[b], Neil A. Koorbanally[c], Bahareh Honarparvar[a]*

[a]School of Pharmacy and Pharmacology, University of KwaZulu-Natal, Durban 4000, South Africa.

[b]Helmholtz-Zentrum München - German Research Centre for Environmental Health (GmbH), Institute of Structural Biology, Ingolstaedter Landstrasse 1, D-85764, Neuherberg, Germany.

[c]School of Chemistry, University of KwaZulu-Natal, Private Bag X54001, Durban, 4000, South Africa.

*Corresponding author: Honarparvar@ukzn.ac.za (Dr. Bahareh Honarparvar), Telephone: + 27 31 2608482, School of Pharmacy and Pharmacology, University of KwaZulu-Natal, Durban 4001, South Africa.

4.1. Abstract

QSAR modeling and analysis of 289 pyrimidine derivatives with non-nucleoside HIV RT inhibitory activity (NNRTI) was carried out in this work. The Associative Neural Network (ASNN) method was applied to develop a Quantitative Structure-Activity relationship (QSAR) for anti-HIV RT activity. The calculated models were validated using the bagging approach. A consensus model with $R^2 = 0.87$ and RMSE = 0.5 was obtained from 10 individual models. Scaffold analysis and molecular docking of the compounds used in the QSAR model identified a potential chemical scaffold. The results showed that scaffold-based analysis of the QSAR model could be helpful in identifying potent scaffolds for further exploration than analyzing the overall model. Matched Molecular Pair analysis (MMPA) was applied in the QSAR model to characterize molecular transformations causing a significant change in the anti-HIV activity. The linear QSAR model was calculated to explore the structural features important for NNRTI activity. The results revealed that the activity of NNRT inhibitors is strongly dependent on their aromaticity and structural flexibility. The scaffold-based analysis of QSAR models with molecular docking and MMPA was found to be helpful in characterizing potential scaffolds for anti-HIV RT derivatives. The outcome of this study provides a deeper insight on the computer aided scaffold-based design of novel molecules with HIV RT activities. Moreover, we clearly showed that the model's failure to correctly predict new chemical series could be due to the limitation of its applicability domain (AD). Redevelopment of models using new measurements can dramatically increase their ADs and performance.

Keywords:

Non-nucleoside Reverse Transcriptase (NNRT) Pyrimidine derivatives; HIV Reverse transcriptase (HIV-RT); Quantitative Structure-Activity relationship (QSAR); Matched Molecular Pair analysis (MMPA); Molecular Docking.

4.2. Introduction

According to a recent WHO estimate, 35.3 million people were living with HIV/AIDS worldwide in 2012 [135], with a significant number of these infections being resistant to antiretroviral therapies. HIV utilizes Reverse Transcriptase (RT), an enzyme that makes copies of cDNA from RNA, a process called reverse transcription. This makes RT an attractive target for anti-retroviral drugs like Non-nucleoside Reverse Transcriptase Inhibitors (NNRTIs) [6].

The higher rate of mutation in HIV strains, and the subsequent development of resistance to the NNRTIs is a major issue in managing HIV infection. This highlights the need for rapid and rational development of NNRTIs. Pyrimidine derivatives were synthesized for decades and have been actively pursued as NNRTIs [5]. Two main series of pyrimidine derivatives are DABO (Dihydro-alkoxy-benzyl-oxopyrimidine) and DAPY (Di aryl pyrimidine) [6]. Owing to NNRTIs importance in targeting HIV RT, QSAR studies have been used to understand the relationship between its structure and anti-HIV RT activity.

Quantitative Structure-Activity relationship (QSAR) is the mathematical modeling of chemical structures of compounds and their relationship with biological activity, and is actively used in drug design [78, 79]. Knowledge of the relationship between structural properties of chemical compounds and their biological activity is crucial in optimizing lead molecules. The construction of QSAR models typically consists of two main steps: (i) calculation and representation of structural features (molecular descriptors) of the selected compounds; (ii) multivariate analysis for correlating molecular descriptors with observed activities (biological, physico-chemical and ADMET properties). Numerous methods, ranging from simple linear regression to complex machine learning algorithms are applied to explore structure-activity relationships. The linear regression models in general allow a relatively straightforward interpretation in terms of linear regression coefficients, provided that descriptors used in the equation are not correlated. However, the models obtained by machine learning methods, such as Neural Network and Support Vector Machine, are more difficult to interpret, due to the non-linear nature of the algorithms.

A recently developed Matched Molecular Pair Analysis (MMPA) approach has the capability to address the ‘black box’ nature of QSAR models [152]. Matched Molecular pair (MMP) is defined as a pair of molecules that differ by a minor structural change at a single point [153]. An MMP associated with a significant change in activity is known as ‘activity cliff’ and is of particular interest. Due to its nature and the way it uses structural information, MMPA can be utilized as a complementary method to QSAR modeling.

Molecular docking techniques have also been widely used to discover new small molecules targeting bacterial/viral proteins [154, 155]. Docking is often used to predict the suitable pose and affinity of drug molecules in the binding pocket of protein targets to rationalize the active and inactive lead compounds.

Earlier QSAR models for NNRTIs were developed on smaller data sets and a single class of compounds [142-145]. There was thus a need for a larger and more diverse dataset for QSAR studies. An important objective was to develop a robust QSAR model for selected NNRTIs that could predict anti-HIV RT activity. Another objective was to identify a potential chemical scaffold for further optimization, the aim being to understand the underlying structural changes that could contribute to improving the anti-HIV activity of the NNRTI. QSAR modeling was combined with molecular docking studies and MMPA on the selected NNRTIs to provide a deeper insight into the computer-aided design of novel molecules against HIV RT.

4.3. Methods

4.3.1. Dataset

Publicly available NNRTIs with a pyrimidine ring in their structure shown to possess anti-HIVRT activity were obtained from the ChemDB database [47] and published articles [156-171]. To maintain homogeneity in the dataset, only molecules with reported IC_{50} values (against HIV RT) were considered. Online Chemical Modeling Environment (OCHEM) software was used to develop the QSAR model [82].

It is argued that activity data originating from different sources should not be mixed as these measurements are highly dependent on experimental conditions, therefore the publications used in this work reported the Anti HIV-RT assay using the same protocol as described in Tramontano [35] and Balzarini [172, 173], with the experimental conditions in all the articles being comparable. It has also been shown that collecting the IC_{50} data from different sources adds only a moderate amount of noise [174]. Combining data from different sources, after careful consideration of experimental conditions, therefore appears to be a valid approach.

Duplicate molecules were removed from the dataset, with 289 molecules being used to develop the QSAR model. An additional 47 structures were obtained [175-177] and used as external validation set.

4.3.2. QSAR models

QSAR models for anti-HIV RT activity were developed using ASNN (Associative Neural Networks). ASNN is an extension of Artificial Neural Network (ANN), in which multiple neuronal responses are combined in a single neuron (network ensemble). In ASNN, an ensemble of feed forward neural network is combined with kNN. It corrects the bias in the ensemble by utilizing the correlation between ensemble responses as a measure of distance [178, 179]. The best models with ASNN were obtained using Supersab [180] training method with iterations = 5000, neurons = 9 and a set of 64 ensembles. We also have used kNN (k nearest neighbour), Neural Network and the SVM algorithm to compare with the employed approach. Molecular descriptors such as Dragon, CDK, AlogP and Estate, Adriana, Chemaxon, ISIDA Fragments, GS Fragments, Inductive descriptors, Spectrophores, Mera, Mersy and QNPR were utilized in this work, with details about the descriptors found elsewhere [181]. Before calculating the descriptors, 3D structures of molecules were optimized by Corina [182].

Descriptors that had less than two unique values, large absolute values, less than 0.01 co-variance, and infinite values were excluded. Unsupervised forward selection [183] and simple pairwise correlation (descriptors having correlation coefficient r < 0.95 with any other descriptors are removed) methods were used to filter descriptors. Bagging (Bootstrap aggregating) with 64 models ("bags") and 5-fold cross validation technique were used for model assessment. Better results were obtained with bagging, hence the models with this technique were retained. Bagging is a meta-learning method that involves creation of an ensemble of models (64 in this case) based on random training sets [184]. These training sets are randomly drawn from the original dataset by sampling with replacement. The "out-of-the-bag" samples, which are not selected in the training sets (approximately 33% of the original dataset size), are used for estimating the prediction power of the model. The bagging is a useful approach for avoiding overfitting since estimation of the model's performance is done using "out-of-the-bag" samples, which do not participate in the model development. Due to sampling with replacement, "out-of-the-bag" samples span all the original training set of different bags. As compared to the methods where a single training and test set is involved, bagging estimates the prediction for the whole original set. The final ensemble model was the simple average of the individual bagging models.

IC_{50} values of 17 molecules in the final dataset were reported in ranges, for example *N*-(2-chloro-4-sulfamoylphenyl)-3-[[4-(4-cyano-2,6-dimethylphenoxy)pyrimidin-2-yl] amino] propanamide is reported to have a IC_{50} value of >10. Such records can be handled effectively by the OCHEM. There are two alternatives for such cases, one is to use the boundary values and another is to handle them as ranges, the latter more informative option being used in the present study.

4.3.3. Scaffold analysis

To identify the scaffold contributing to anti-HIV activity, the data were divided into active and inactive molecules. The dataset was discretized using an average activity value over the entire dataset, *i.e.* -6.7 log (mol L^{-1}) as the threshold between active and inactive molecules. Molecules having a value between -6.7 to -4 log (mol $L^{-}1$) were classified as inactive, and molecules with values between -6.7 to -9.3 log (mol L^{-1}) were categorized as active.

These active and inactive sets of molecules were analyzed using the SetCompare tool [185], which identifies whether the probability of a particular scaffold overrepresented in one of the two sets is by chance or not. It uses a hyper-geometric distribution for the analysis, which applies to sampling without replacement from a finite population whose components can be grouped into two mutually exclusive categories. Scaffold hunter descriptors [22] were chosen for comparison of the active and inactive sets.

4.3.4. Molecular Docking

Molecular docking simulation of all the 289 molecules in the binding site of HIV RT was performed to assess how well different scaffolds performed in the QSAR model. The crystal structure of HIV-1 reverse transcriptase co-crystallized with the MKC422 ligand (PBD: 1RT1, resolution: 2.5 Å, R value: 0.197) [186] was retrieved from the Protein Data Bank (PDB) [187] and used for docking. To validate the docking protocol, the ligand (MKC422) was removed from the experimentally reported HIV1 RT- MKC422 complex (1RT1) and re-docked into the binding site of HIV -1 reverse transcriptase. All the selected molecules were geometry optimized by Guassian09 software [188] using the PM6 semi-empirical method [189]. Gasteiger charges were added and nonpolar hydrogen atoms were merged to carbon atoms in all the ligand structures. Selected molecules were docked into the NNRTI binding pocket in the HIV-1 reverse transcriptase (by defining the grid box with spacing of 1Å and size of 24 ×24 × 24 pointing in x, y and z directions around the MKC422 ligand present in the PDB crystal). Water molecules were removed and polar hydrogen was added to the crystal structure of the receptor protein. Molecular docking was performed by Raccoon AutoDock [190] using AutoDockTools (ADT)[191] and AutoDockVina [117] with default docking parameters. The Lamarckian Genetic algorithm [128] was used as the search algorithm with default parameter values. For different ligands, the docked confirmation with the most negative binding energy values was undertaken for further analysis.

4.3.5. Matched Molecular Pair Analysis (MMPA)

The basis of MMPA is the identification of pairs of molecules bearing a specific structural relationship to each other, with each molecular pair being associated with a chemical transformation, where a minor structural modification is seen between them. MMPs were identified, as described by Hussain *et.al* [192], where transformations with p value <0.01 were identified as significant. To address where a transformation passes the *p* value filtered by mere chance, the Holm-Bonferroni method [193] was applied. To avoid very dissimilar structures in Matched Pairs, only molecular pairs with a minimal Tanimoto similarity index of 25 were retained.

4.4. Results and discussion

4.4.1. QSAR models

QSAR models were calculated by ASNN, kNN, Neural Network and SVM with different molecular descriptors (TABLE 3). Models with the ASNN method had the best quality and hence were used for further analysis. Studies have shown that the consensus model calculated out of individual models performed best [185, 194, 195]. An average consensus model was built using 10 QSAR models with the exception of that with spectrophore descriptors, due to its low performance. The consensus model with $R^2 = 0.86 \pm 0.02$, $Q^2 = 0.87 \pm 0.02$ and RMSE $= 0.5 \pm 0.03$ was obtained. A plot between the measured and predicted activity values of the consensus model is shown in FIGURE 15, suggesting a robust and reliable model. This model was chosen for further analysis after removal of the outliers (discussed later) with $R^2 = 0.87 \pm 0.02$, $Q^2 = 0.87 \pm 0.02$ and RMSE $= 0.48 \pm 0.03$. Additionally, a QSAR model based on all descriptors was calculated with $R^2 = 0.86 \pm 0.02$, $Q^2 = 0.87 \pm 0.02$ and RMSE $= 0.49 \pm 0.04$. It had a similar performance to the consensus model. We decided to use the consensus model, which also estimated its applicability domain.

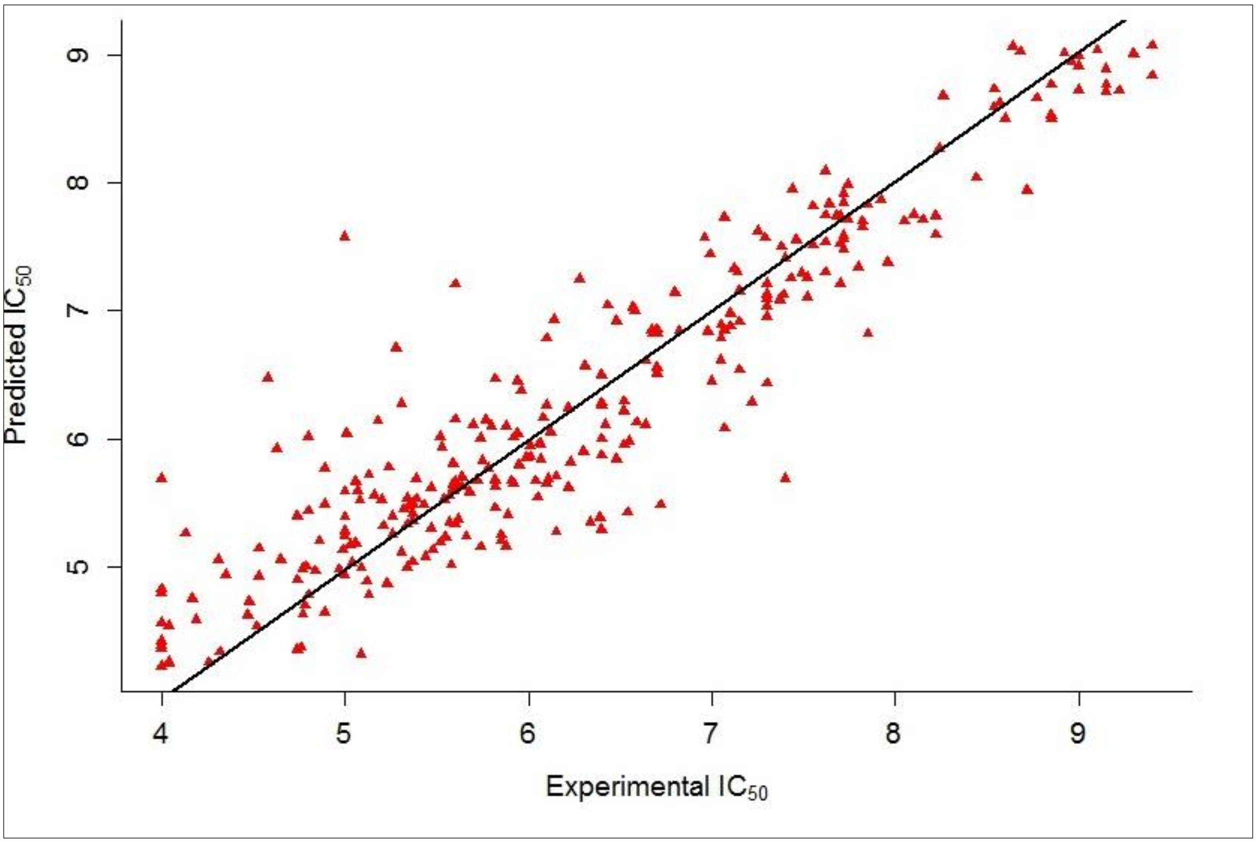


Figure 15: Consensus QSAR Model.

Plot of measured versus predicted IC_{50} values in logarithmic scale of the consensus model ($R^2 = 0.87$). The Consensus model line is plotted in black color. Measured IC_{50} are plotted on the X axis, and predicted IC_{50} values on the y axis. Most of the data points lie close to the model line

The applicability domain (AD) of a QSAR model is the chemical structure subspace in which it makes predictions with a given reliability [196]. AD is required as the QSAR model may have different accuracies of predictions for compounds based on their similarity to the training set molecules, *i.e.* distance to model (DM). In the current study, the CONSENSUS-STD (standard deviation of predictions of the ensemble of models in the consensus model) was used as a measure of DM. This DM provided the best separation of molecules with low and high accuracy of predictions in several benchmarking studies [197, 198]. A threshold value of 95% of compounds from the training set was used to determine the qualitative ADs of models. It is assumed that 5% of compounds outside the AD have little effect on the prediction confidence [194]. A Williams's plot of the consensus model with CONSENSUS-STD as a distance to model is shown in **FIGURE 16**, and the AD defined above was used to warn users about unreliable predictions.

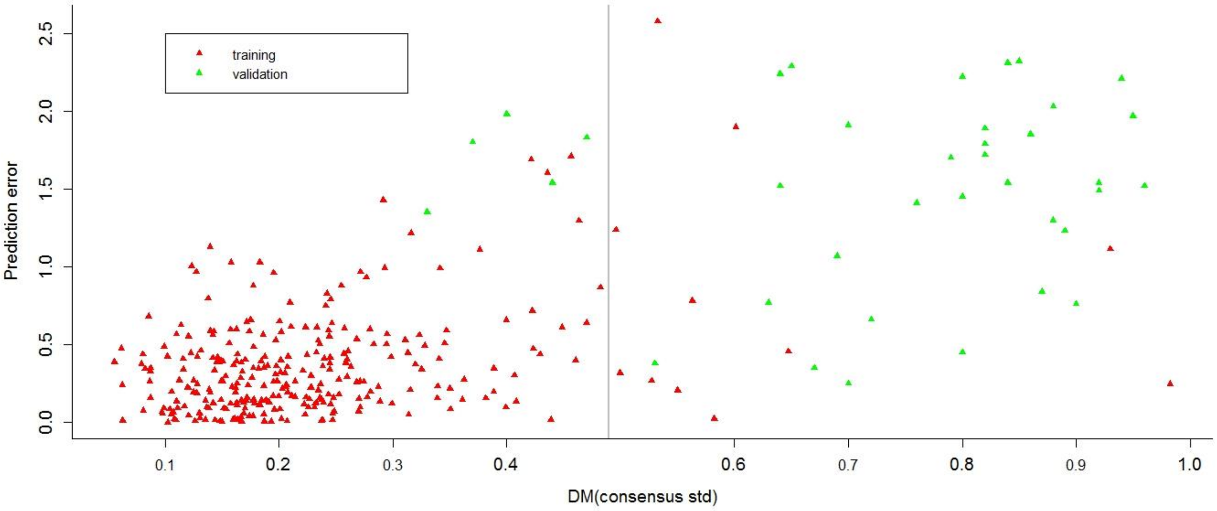


Figure 16: Applicability domain.

AD Plot with CONSENSUS-STD as a measure of distance to model (DM). Threshold value of 95% of compounds from the training set is selected (vertical line).

4.4.2. Model validation

The external validation set was used to evaluate the consensus QSAR model. The model calculated a poor prediction for the validation set ($R^2 = 0.16$, RMSE = 2 ± 0.1), as most of the samples were outside the AD of the model (see **FIGURE 16**). Another validation set was obtained from random splitting of a combined set of original training and validation sets. The size of the validation set was kept at 20% of the combined dataset. A new consensus models was developed using the same workflow as the model based on the initial training set. This model had comparable performance for the training ($R^2 = 0.85 \pm 0.02$, $Q^2 = 0.84 \pm 0.02$, RMSE = 0.51 ± 0.03) and the validation set ($R^2 = 0.83 \pm 0.04$, $Q^2 = 0.82 \pm 0.04$, RMSE = 0.61 ± 0.06).

The extrapolation is a difficult problem for the QSAR approaches. The prediction using QSAR model is valid only if the molecules being predicted are within its applicability. Because the original consensus model did not cover the chemical compounds from the new series (see FIGURE 16), it failed to predict their activity. The extension of the chemical space by including new molecules extended its AD as is shown by Williams plot with CONSENSUS-STD (FIGURE 17), where most of the validation set compounds are within the model's AD. It explains the aforementioned observation. The extension of model's AD made the correct prediction possible for these series of compounds. This is an important result, which exemplifies how QSAR models can fail for new data due to limitations of their ADs. It also clearly demonstrates that the same models re-developed with new measurements can become an important tool to predict activities of compounds within these series.

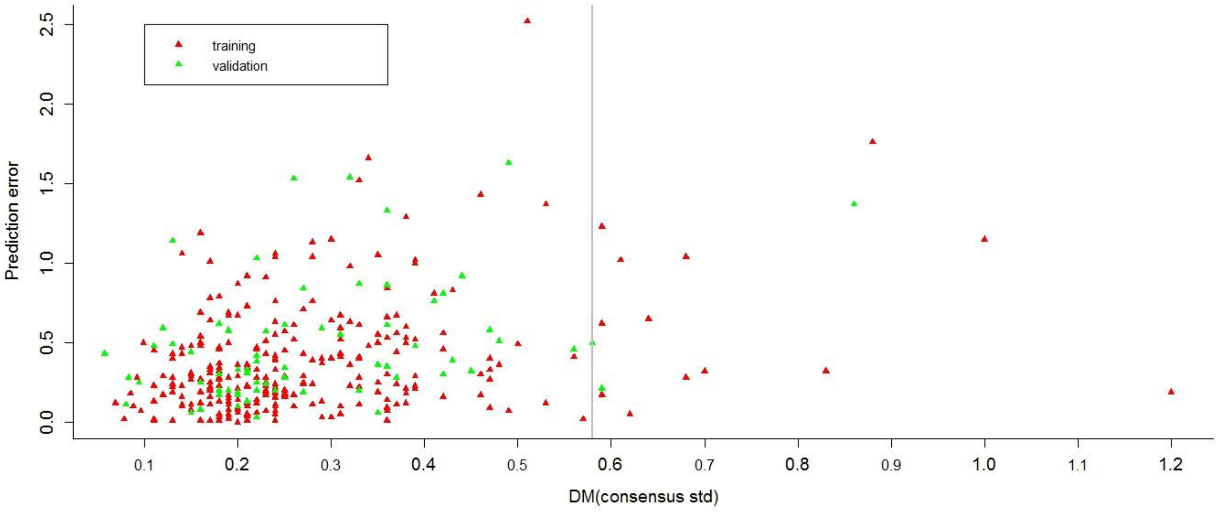


Figure 17: Applicability domain.

Plot with CONSENSUS-STD as a measure of distance to model (DM) for new QSAR. Threshold value of 95% of compounds from the training set is selected (vertical line).

Table 3: Comparison of QSAR models built using different algorithms and sets of descriptors.

Pearson correlation coefficient (R2) is reported for each model

Descriptors	**ASNN**	**kNN**	**SVM**	**ANN**
Dragon6	**0.86**	**0.86**	0.13	**0.86**
Fragmentor	0.85	0.77	**0.86**	0.84
CDK	**0.83**	**0.83**	0.06	**0.83**
ALogPS, OEstate	**0.83**	0.81	**0.83**	0.82
GSFrag	**0.84**	0.8	0.08	0.83
Mera, Mersy	0.80	**0.83**	0.18	0.79
ChemaxonDescriptors	**0.81**	0.77	0.09	0.79
InductiveDescriptors	**0.80**	0.78	0.64	0.77
Adriana	0.83	**0.84**	0.18	0.81
Spectrophores	0.71	**0.73**	0.02	0.66
QNPR	**0.84**	0.82	0.87	0.83
Average Consensus Model	0.86			

4.4.3. QSAR model analysis

The consensus QSAR model was analyzed for individual chemical scaffolds in the data set. A combinatorial approach was taken to interpret the QSAR model in terms of chemical scaffold, molecular docking and MMP analysis. The objective was to identify interesting scaffolds with better QSAR performance and a large range of activity values, thus allowing their application in the design of new molecules. General chemical scaffolds of the molecules from the training set are shown in Figure 18 along with Q^2 (coefficient of determination) values, number of molecules (N), mean of IC_{50} values (µM) and standard deviations (SD) calculated for each scaffold. Eq. 25 was used to calculate the coefficient of determination (Q^2) value. The higher the Q^2 value, the better the model at explaining the variation of data.

$$Q^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - y_i'\right)^2}{\sum_{i=1}^{n}\left(y_i - Y\right)^2} \qquad 25$$

Where y_i= i^{th} measured IC_{50},y_i'= i^{th} predicted activity, n= population size and Y= mean of measured IC_{50}

The chemical scaffolds A and C shown in **FIGURE 18** have the highest Q2 values 0.76 and 0.80, respectively. It should be mentioned that scaffold A is a subset of scaffold C, with the analysis concentrating on scaffold C as the larger one. Scaffold C covers a large range of activity values as depicted by mean IC_{50} values and SD. The Q2 value of the consensus model was 0.87 ± 0.02, which includes scaffold A to G in its training set. However, individually, some of these scaffolds had a lower Q2 or even negative Q2 values. For scaffolds with low or negative Q2 values (e.g., B, D, E, F and G), the model differentiated between different scaffolds (i.e., to some extent classified them into active and inactive classes) rather than explaining changes in the activities of molecules due to various substituents. The model may not be able to predict changes in activities of new compounds based on such scaffolds. The high accuracy of the QSAR model for molecules with scaffold C suggests that it had properly learnt the chemical features responsible for the anti-HIV activity and thus may accurately predict new molecules. The analysis in the next sections provides a deeper insight into this scaffold.

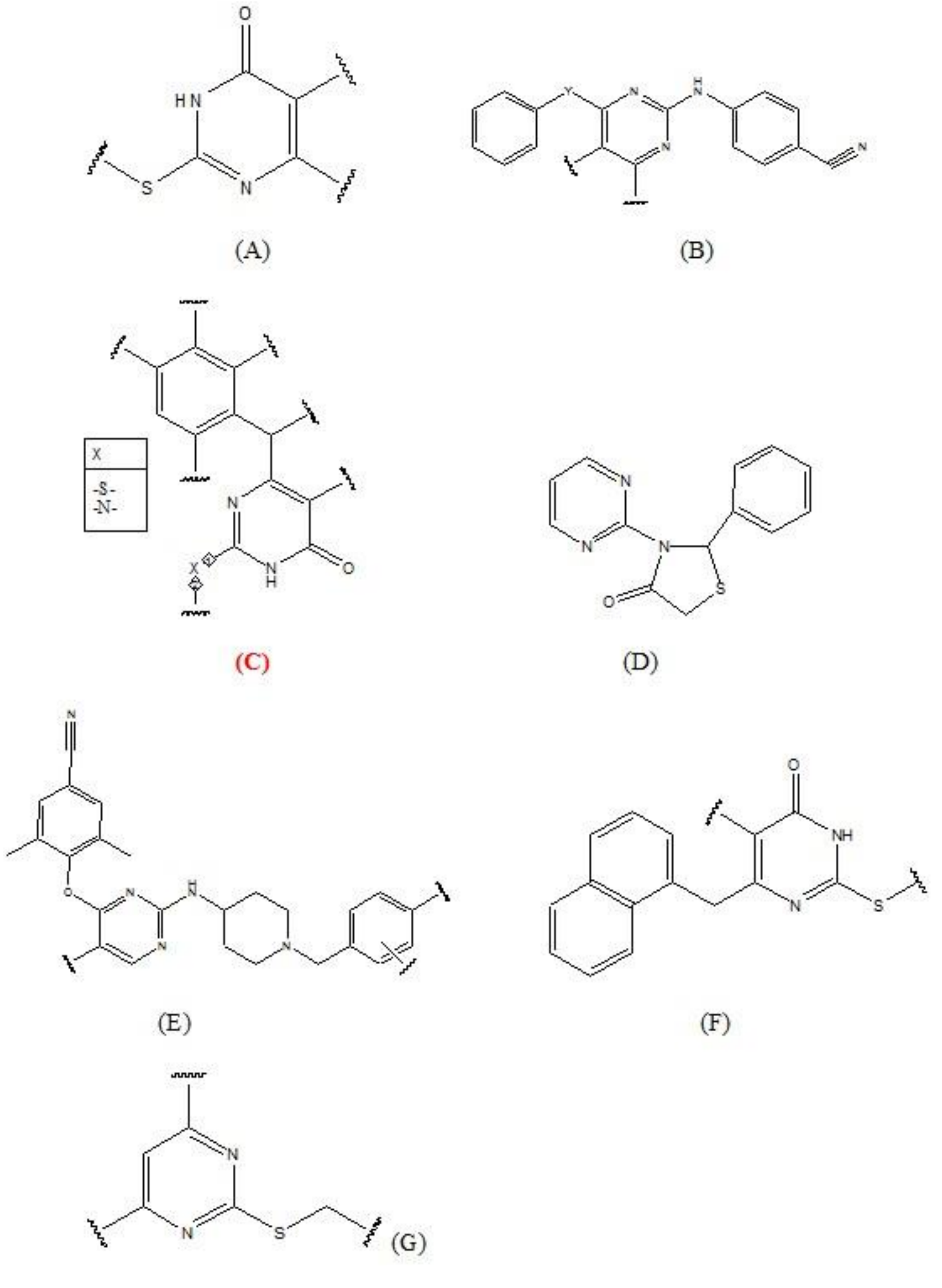


Figure 18: Scaffold used in the QSAR model.

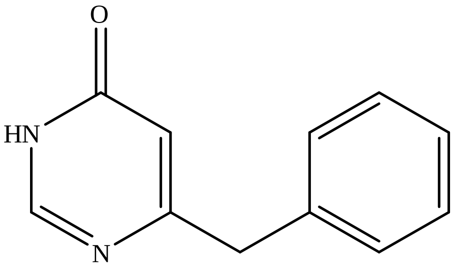
*Scaffold used in the QSAR model development and their model Coefficient of determination (Q^2). N= number of molecules representing the particular scaffold, Mean = average IC_{50} values in µM, SD = standard deviation of IC_{50} values. (**A**) Q^2 = 0.76, N= 15, Mean = 1.5 µM, SD = 1.6 (**B**) Q^2 = 0.4, N= 30, Mean IC_{50} = 0.003 µM, SD= 0.005(**C**) Q^2 = 0.80, N= 142, Mean = 8.46 µM, SD= 17.06 (**D**) Q^2 = 0.41, N= 23, Mean = 6.3 µM, SD = 15.52 (**E**) Q^2 = 0.27, N= 53, Mean = 0.1 µM, SD= 0.16 (**F**) Q^2 = 0.33, N= 9, Mean = 24.3 µM, SD= 16.8 (**G**) Q^2 = -0.60, N= 14, Mean 2.2 µM, SD= 4.2*

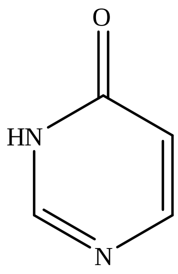
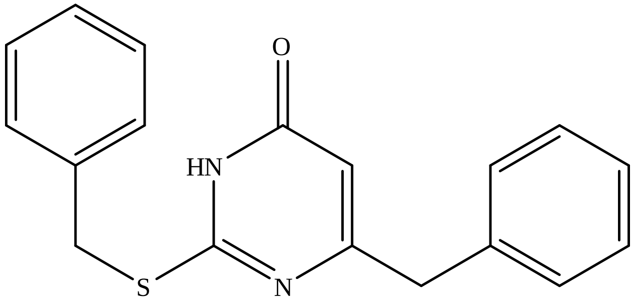
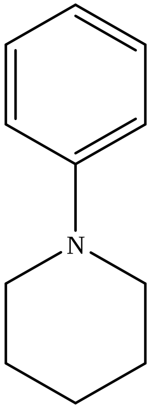
4.4.4. Scaffold analysis

The outcome of the Set Compare task using scaffold hunter descriptors is summarized in TABLE 4. For simplicity and to differentiate these scaffolds from those in FIGURE 18, we labeled them S1 to S4. Scaffold C, with the highest Q^2 values in the QSAR model, has the same general structure to that of S1 and S2. It should also be noted that S2 is a subpart of S1, with both these scaffolds being overrepresented in the set of inactive compounds (55.3%) compared to the active set (14.1%). However, entire representative molecules of the particular scaffold in the inactive set contain an attached S atom at position 2 of the oxopyrimidine ring. Furthermore, all of the S atoms in the inactive set have aromatic substitution. If the S atom is present at this position in the active set, it either has an aliphatic chain or aliphatic cyclic substitution, which is further corroborated by the overrepresentation of S3 in the Inactive set. Scaffold C has the highest Q^2 value in the consensus QSAR model. The model correctly captured the structural features of the scaffold C required for activity, suggesting that this scaffold can potentially be used to design novel anti-HIV molecules with improved activity.

Table 4: Chemical scaffolds overrepresented in the datasets.

Appearance counts as well as percentage of representation are listed along with the p-value of the respective scaffold.

Scaffold No.	**Scaffold**	**Active set (99 molecules)**	**Inactive set (199 molecules)**	**P-value**
S1		14 (14.1%)	105 (55.3%)	-2.7E-12

S2		17 (17.2%)	113 (59.5%)	-1.75E-12
S3		0 (0.0%)	28 (14.7%)	-3.78E-6
S4		33 (33.3%)	1 (0.5%)	1.8E-16

In addition, S4 is a part of scaffold E that is represented in 33.3% of active molecules compared to 0.5 % of the inactive set. Nevertheless, the contribution of S4 in the active set of molecules is not well captured by the QSAR model as depicted by Q^2 = 0.27 of scaffold E. This means that the model predicted all molecules with this scaffold as active ones and was unable to identify structural changes within this scaffold that contribute to a change in the activity of the molecules. The predictions for such scaffolds would apparently not be promising to identify new potent molecules.

4.4.5. Interpretable QSAR model

An interpretable linear QSAR model was built to understand the structural features contributing to the anti-HIV activity. Topological, Geometrical and Constitutional descriptors were calculated using Dragon [80], and the model was built using multiple linear regression (MLR), which can be described by Equation 2. Although its corresponding R^2 (0.69 ± 0.03) was less than the consensus model (by about 17%), the structural features affecting the activity could be explained easily with this model.

$$Normalised\ Y = 6.29 + 1.27 * nN - 1.13 * RBN + 1.12 * RBF + 0.915 * AROM - 0.645 * nDB - 0.54 * HNar + 0.486 * CMBL + 0.46 * nCsp3 + 0.353 * DISPs - 0.293 * LOC + 0.258 * PW4 - 0.225 * DISPe - 0.209 * DISPi - 0.194 * BLI - 0.144 * SPH + 0.138 * MAXDN + 0.136 * AMW - 0.135 * HOMA \qquad (2)$$

R^2= 0.69 ± 0.03 Q^2 = 0.68 ± 0.03 RMSE = 0.76 ± 0.03

Details about the descriptors in Equation 2 are provided in **TABLE 5**.

NNRTI binds inside the hydrophobic pocket of RT, known as the Non-nucleoside inhibitor binding pocket (NNIBP). The NNIBP is flexible and does not exist in the absence of a bound ligand. Conformational flexibility of the inhibitor plays a crucial role in its adaptation in the binding pocket. It has been shown that a flexible ligand can adapt very well to the changes in the binding pocket, with most of the NNRTIs either taking the “horseshoe” or “butterfly” shape [199].

The main contributing descriptor to the NNRTIs activity is the number of nitrogen atoms, which can form hydrogen bonds with oxygen in the active residues of the binding pocket [199]. The number of rotatable bonds also seem to be an essential factor in activity. An increase in the number of double bonds (nDB) has a negative impact on the activity. Higher number of double bonds are correlated with lower fraction of rotatable bonds, which can decrease the molecule`s flexibility.

Descriptors such as AROM (aromaticity index), BLI (Kier benzene-likeliness index) and HOMA (Harmonic Oscillator Model of Aromaticity index) represent the aromaticity of the molecule. In Equation 2, AROM has the highest positive impact amongst all aromaticity indices, whereas others have a slight negative impact. Overall aromaticity can be considered favorable for activity. The interaction between two aromatic rings is known as π-π stacking. The aromatic rings in the NNRTI are known to make a π–π stacking interaction with Y181 and Y188 of the target enzyme [199].

Descriptors such as HNar (Narumi harmonic topological index) and LOC (lopping centric index) are indices for branching in the molecule. In Equation 2, the HNar and LOC have a negative sign, which implies that more branching in the NNRTI adversely affects the activity by reducing the hydrophobic interaction between the binding pocket residues and ligand through steric hindrance.

In chemical graph theory, the molecules are represented as graphs, where vertices correspond to the atoms and the edges to the chemical bonds. This representation is used in calculating various descriptors like Randic shape index that range from linear molecules to completely connected graph. The Randic shape index has a positive effect on the activity, which suggests that a molecule with more linear shape will have lower anti-HIV activity. The number of sp^3 hybridized C atoms (descriptor nCsp3) favors the anti-RT activity, which is consistent with the earlier discussion on the favorable role of two sp^3 hybridized carbon atoms in the anti-retroviral activity of the triazolo[4,5-g]quinoline scaffold [200].

Spherocity appeared in the QSAR model with the index SPH, and is an unfavorable feature for biological activity. The SPH values range from 0 for flat molecules (such as benzene) to 1 for total spherical molecules. A possible explanation might be that NNRTI takes a non-flat shape inside the NNIBP and relatively more spherical structures might move away from the energetically favorable shape.

MAXDN (maximal electrotopological negative variation, an index of nucleophilicity) has a positive correlation with the activity, and indicates that polar interactions between the protein and the NNRTIs could stabilize the ligand inside the active binding pocket. Other important molecular properties for anti-HIV activity are molecular shape and charge distribution (DISPs, DISPe and DISPi), and molecular weight.

Table 5: List of molecular descriptors from the linear QSAR model

Sr. No	Symbol	Descriptor	Explanation
1.	nN	Constitutional	Number of Nitrogen atoms
2.	RBN	Constitutional	Number of rotatable bonds
3.	RBF	Constitutional	Rotatable bond fraction
4.	AROM	Geometrical	Aromaticity index
5.	nDB	Constitutional	Number of double bonds
6.	HNar	Topological	Narumi harmonic topological index that relates to molecular branching
7.	CMBL	Geometrical	Conjugated maximum bond length
8.	nCsp3	Constitutional	Number of sp^3 hybridized Carbon atoms
9.	DISPs	Geometrical	Displacement value / weighted by I-state
10.	LOC	Topological	Lopping centric index provides the level of branching in the molecule. The higher the value, the more branched the molecule.
11.	PW4	Topological	Randic shape index. Range from linear molecules to completely connected graph
12.	DISPe	Geometrical	Displacement value / weighted by atomic Sanderson electronegativities
13.	DISPi	Geometrical	Displacement value / weighted by ionization potential
14.	BLI	Topological	Kier benzene-likeliness index. A measure of molecular aromaticity
15.	SPH	Geometrical	Spherosity index varies from zero for flat molecules, such as benzene, to unity for totally spherical molecules
16.	MAXDN	Topological	Maximal electrotopological negative variation related to the nucleophilicity of the molecule
17.	AMW	Constitutional	Average molecular weight
18.	HOMA	Geometrical	Harmonic Oscillator Model of Aromaticity index

4.4.6. Molecular docking

All the molecules in the training set were docked inside the NNRTI binding pocket of HIV-RT. Due to the reasonable structural similarity between the MKC422 and selected NNRTIs, PDB 1RT1 (MKC422 co-crystallized with HIV RT) was chosen for the current study. The superimposed structure of docked MKC422 and PDB crystal MKC422 has the RMSD of 0.064Å (Supplementary FIGURE S1). The low RMSD value validate the docking procedure adopted in this study. MKC422 binds in the same binding site with very similar orientation as that of the crystal MKC422. The interaction between the NNRTI molecule and the residue of binding pocket is demonstrated in FIGURE 19.

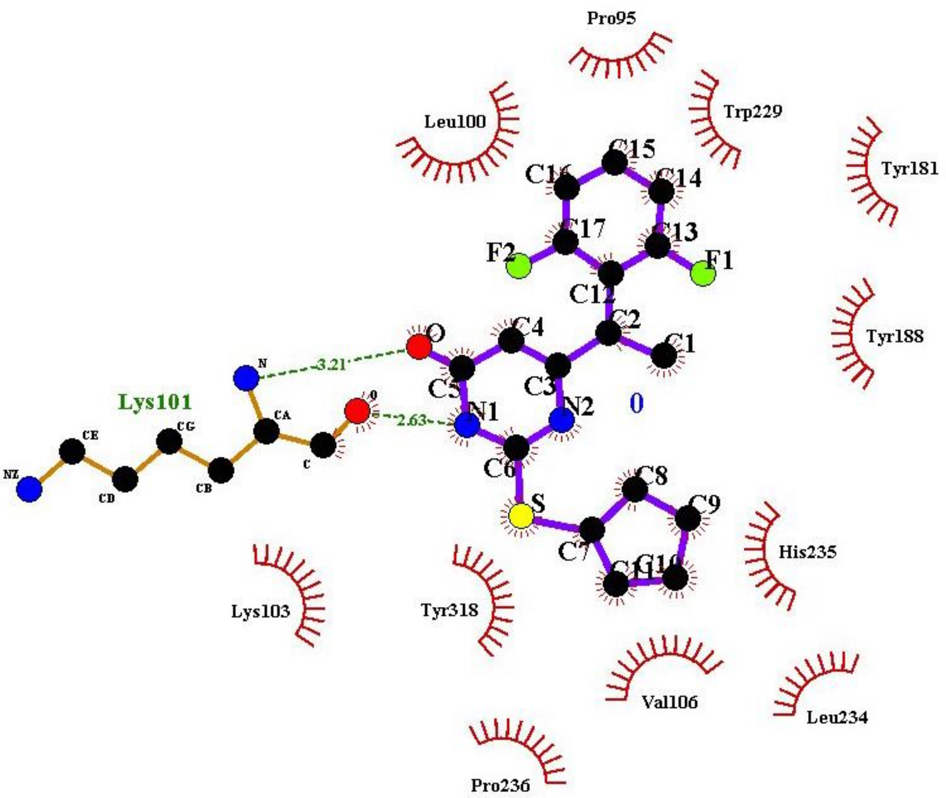


Figure 19: Interaction between NNRTI molecule (blue) and residue of binding pocket.

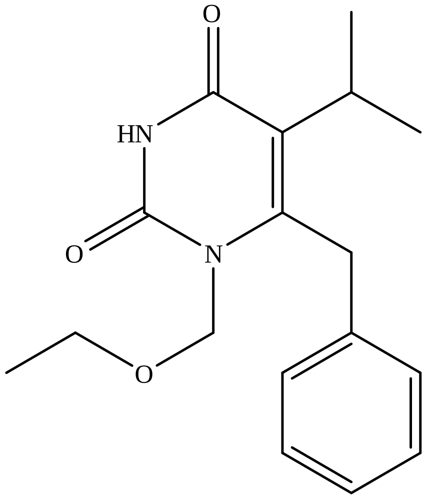
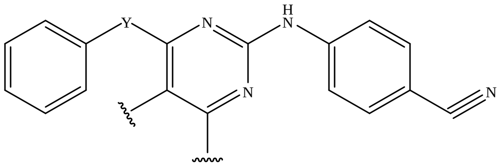
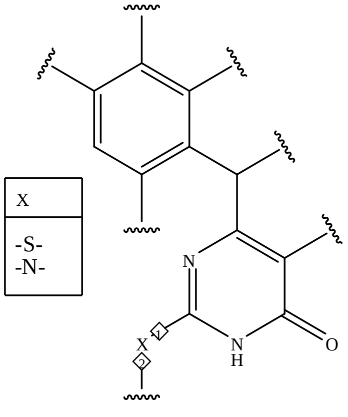
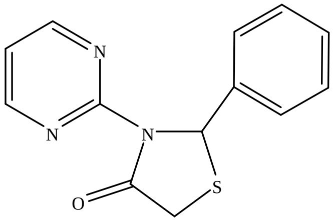
Hydrogen bonds are shown with green dotted lines. Red semicircles represents the hydrophobic interaction.

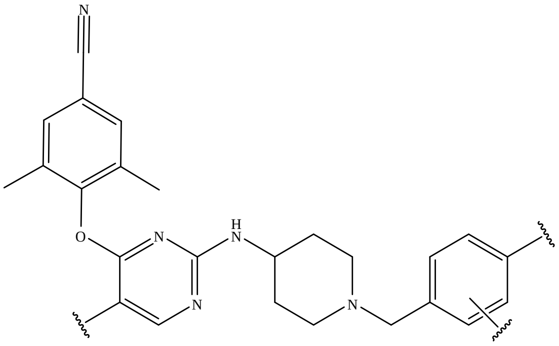
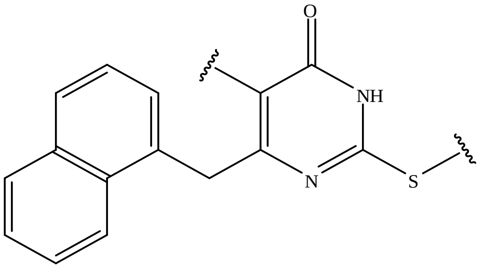
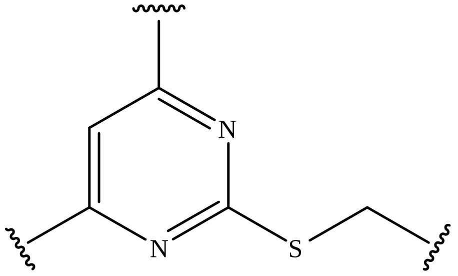
Average docked binding energies and Q^2 values of the entire chemical scaffolds from the training set and outliers are given in TABLE 6. Of note is that all the scaffolds with lower Q^2 values (B, D, E, F and G) showed higher (less negative) average binding energies (higher than the MKC422), whereas scaffold C in the QSAR model showed lower (more negative) average binding energy. The reasonably good docked binding energy of scaffold C (-10.28 Kcal/mol) could be attributed to the possibility of hydrophobic interaction, as well as to the hydrogen bond formation between ligand and the active residues of RT (FIGURE 19). The significance of hydrophobic interactions, π-π stacking (between aromatic rings) and hydrogen bonding (between N of NNRTI and O of active residue) in anti-RT activity was also confirmed in the linear QSAR model (Equation 2).

Table 6: Docked binding energy.

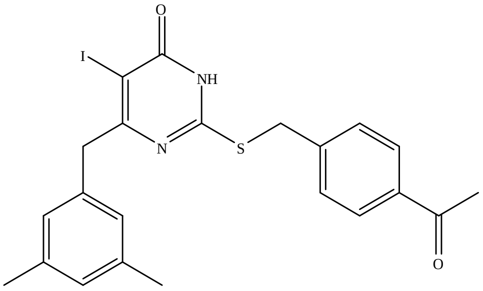
Average docked binding energies (Kcal/mol) of the selected NNRTIs inside the binding pocket of the HIV RT enzyme. (A) Average docked binding energies of scaffolds and model Q^2 values. (B) QSAR model's outlier molecules.

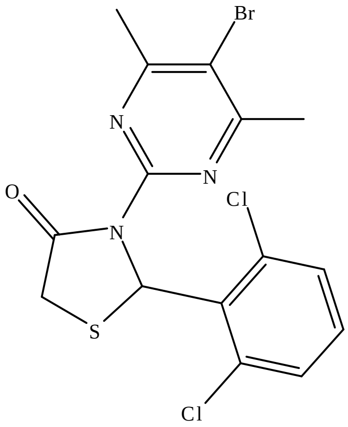
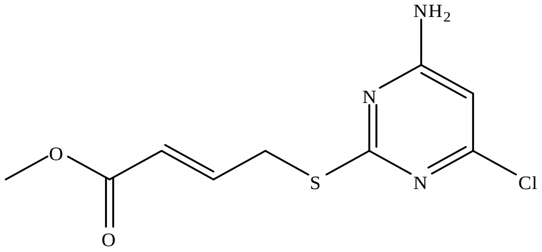
(A)

Scaffolds	**Average docked binding energy (Kcal mol^{-1})**	**Q^2**	**N**	**Chemical structure**
MKC 422	-10.2			
B	-9.9	0.41	30	
C	**-10.28**	**0.80**	**142**	
D	-8.4	0.41	23	

E	-8.67	0.27	53	
F	-10.12	0.16	09	
G	-8.65	-0.6	14	

(B)

Outlier	**Docked Binding Energy (kcal mol^{-1})**	**IC$_{50}$ (μM)**	**Structure**	**Scaffold**
O1	-7.9	0.41		C

O2	-7.8	9.82		D
O3	-7.0	5.25		G

4.4.7. Analysis of outliers based on QSAR and docking

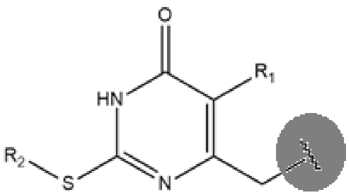
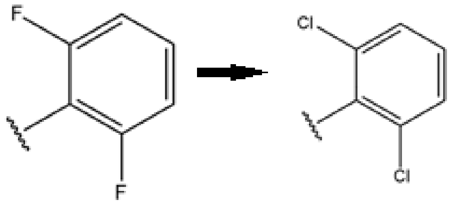
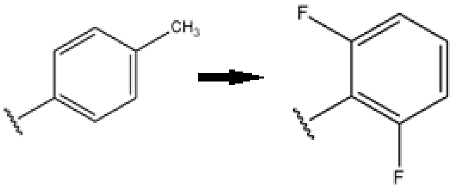
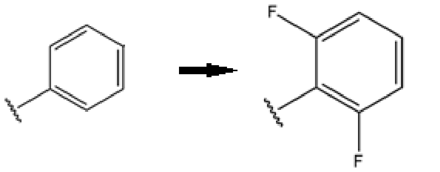
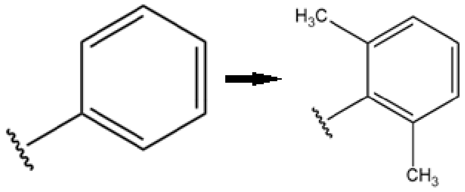
The detection and interpretation of outliers in the QSAR model is crucial to a satisfactory fit and predicting ability of the QSAR model [165]. The reason for a compound to be an outlier could be referred to as different Mechanisms Of Action (MOA), different modes of interaction with target molecules [201], conformational flexibility of the receptor binding site [202] and unusual binding mode [203]. Due to the noise in the data and experimental measurement errors, a molecule may also be an outlier. In general, one should not exclude any molecule merely because it has a high error in the QSAR model. Regarding the analysis of molecules with large prediction errors in the consensus model, in a few cases high docked binding energies were observed. Indeed, three out of 11 molecules with deviations between predicted and calculated values more than 1 log unit were also included in the outliers of the docking procedure. These molecules had at least two units higher docking scores than the reference MKC422 (-10.2 Kcal mol^{-1}), which might be attributed to the different MOAs of the outlying molecules. Excluding these three outliers improved the consensus model R^2 from 0.86 to 0.87.

4.4.8. MMP analysis

The consensus QSAR model was analyzed for matched molecular pairs and 'activity cliff'. Identified molecular transformations, which have the mean activity change of 0.7 log units or more, in addition to their core structure, are reported in TABLE 7. It is evident that substitution with fluorine or methyl in the *meta* position had a positive effect on anti-HIV activity. TR4 is basically the introduction of two sp^3 hybridized carbon atoms at the *meta* position of aromatic rings. The Linear QSAR model has also shown the positive effect of sp^3 hybridized carbon on the anti-HIV activity.

Table 7: The effect of transformation on anti-HIV activity (–log (mol L^{-1})).

Point of transformation is encircled in grey (a) Mean activity change within all the pairs of transformation (c) Number of matched pairs in the particular transformation, (d) Ratio of compound increasing and decreasing the activity value.

Core molecule	**Transformation**		**Δ IC_{50}[a] (μM)**	**# matched Pair [c]**	**Inc/Dec [d]**
	TR1		-0.9 ± 0.54	12	0/12
	TR2		1.5 ± 0.24	12	12/0
	TR3		1.6 ± 0.29	12	12/0
	TR4		0.74 ± 0.51	15	15/1

4.5. Conclusions

Non-nucleoside reverse transcriptase inhibitors (NNRTIs) were collected from the literature. Several QSAR models were built using ASNN algorithms and various molecular descriptors. For further application of different representations of chemical structures, a consensus model was obtained with a R^2 value of 0.87 and analyzed for the scaffold's performance. Some scaffolds had a lower coefficient of determination value. Molecular docking was helpful in identifying potential chemical scaffolds. Identifying outliers based on prediction error and molecular docking improved the QSAR model by excluding molecules with different modes of action. The linear QSAR model highlights the structural features affecting anti-HIV activity. The QSAR model was analyzed using MMP to understand structural features, which were correctly learned by the model. MMPA was shown as a powerful method for addressing the 'black box' nature of QSAR, and enable medicinal chemists to choose molecules for further optimization. Significant transformations in the backbone structure were identified using this method.

The current work serves as a computer-aided strategy for further optimization of a lead molecule. It is also speculated that the scaffolds with high Q^2 in the QSAR model have significant structural features correctly learnt by the model. Thus, predicting structures of potential compounds based on these scaffolds would be accurate. However, the model statistics for predicting new molecules should not be the only approach considered. The scaffold-based analysis is a better approach to identify chemical scaffolds for further optimization.

To the best of our knowledge, this is one of the first QSAR studies on diverse anti-HIV pyrimidine derivatives, where the combination of QSAR, molecular docking and MMP were applied to understand the structure activity relationship. We have shown how the complementary nature of these approaches help to better understand and interpret biological data and propose a design of new inhibitors. Finally, the used data, the consensus model and its sub-models are published on the OCHEM web site http://ochem.eu/article/93085 and are freely accessible for interested users. Their public availability will contribute to the widespread use of the computational chemistry tools on the Web[204].

Last, but not least we exemplified the problem with extrapolation of QSAR models for new chemical series. Despite the failure of original consensus model to correctly predict new data, the re-calculated model provided good performance for new compounds from the same series. The Williams plots clearly showed that the reason of the initial model's failure, and the success of the re-developed model was due to the change in the AD of the new model. The AD of re-constructed model covered the compounds from new chemical series and thus could correctly predict their inhibitory activities (IC_{50}).

Compliance with Ethical Standards

The authors declare that no human or animal subjects were involved in the research.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

This work was supported by a doctoral bursary grant from the College of Health Sciences (CHS) at UKZN in South Africa. We are grateful for the helpful support of CHPC (www.chpc.ac.za) for providing computational resources. We are also thankful to Mrs Carren Martin for her editorial support.

4.6. Supporting information

QSAR models and scaffold-based analysis of non-nucleoside HIV RT inhibitors

Bilal Nizami[a], Igor V. Tetko[b], Neil A. Koorbanally[c], Bahareh Honarparvar[a]*

[a]School of Pharmacy and Pharmacology, University of KwaZulu-Natal, Durban 4000, South Africa.

[b]Helmholtz-Zentrum München - German Research Centre for Environmental Health (GmbH), Institute of Structural Biology, Ingolstaedter Landstrasse 1, D-85764, Neuherberg, Germany.

[c]School of Chemistry, University of KwaZulu-Natal, Private Bag X54001, Durban, 4000, South Africa.

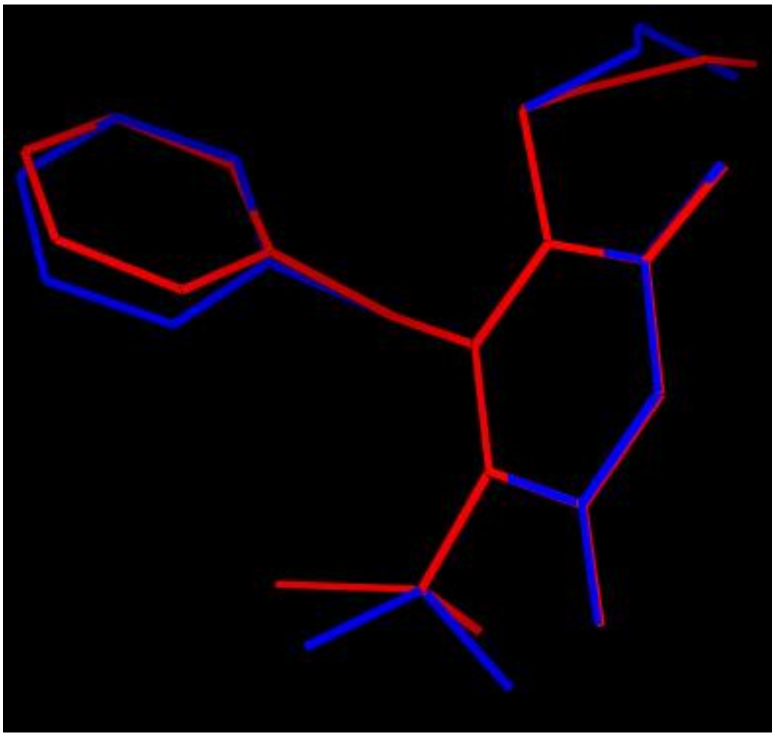


Figure S1: Superimposition of docked MKC422 (Coloured Blue) and MKC422 from experimental HIV1 RT- MKC422 complex (PDB: 1RT1) (Coloured Red).

CHAPTER 5

Molecular insight on the binding of NNRTI to K103N mutated HIV-1 RT: Molecular dynamics simulations and dynamic pharmacophore analysis

Bilal Nizami[1], Dominique Sydow[2], Gerhard Wolber[2], and Bahareh Honarparvar*,[1]

[1]School of Pharmacy and Pharmacology, University of KwaZulu-Natal, Durban 4000, South Africa.
[2]Institute of Pharmacy, Freie Universität Berlin, Königin-Luise-Str. 2+4, 12175 Berlin, Germany.

*Corresponding author: Honarparvar@ukzn.ac.za (Dr. Bahareh Honarparvar), Telephone: + 27 31 2608482, School of Pharmacy and Pharmacology, University of KwaZulu-Natal, Durban 4001, South Africa.

5.1. Abstract

Regardless of advances in anti-HIV therapy, HIV infection remains an immense challenge due to the rapid onset of mutation instigating drug resistance. Rilpivirine is a second-generation Di-aryl pyrimidine (DAPY) derivative, known to effectively inhibit wild-type (WT) as well as various mutant HIV-1 reverse transcriptase (RT). In this study, a cumulative 240 ns of molecular dynamic (MD) simulations of WT HIV-1 RT and its corresponding K103N mutated form, complexed with rilpivirine, were performed in solution. Conformational analysis of the NNRTI inside the binding pocket (NNIBP) revealed the ability of rilpivirine to adopt different conformations, which is possibly the reason for its reasonable activity against mutant HIV-1 RT. Binding free energy (MM-PB/GB SA) calculations of rilpivirine with mutant HIV-1 RT agree with experimental data. The dynamics of interaction patterns were investigated based on the MD simulations using dynophores, a novel approach for MD-based ligand-target interaction mapping. The results from this interaction profile analysis suggest an alternate interaction between the linker N atom of rilpivirine and Lys 101, potentially providing the stability for ligand binding. PCA analysis and per residue fluctuation has highlighted the significant role of flexible thumb and finger sub-domains of RT in its biological activity. This study investigated the underlying reason for rilpivirine's improved inhibitory profile against mutant RT, which could be helpful to understand the molecular basis of HIV-1 RT drug resistance and design novel NNRTIs with improved drug resistance tolerance.

Keywords: HIV-1 RT, Drug resistance, NNRTI, MD, Molecular dynamics, PCA, Dynophore, MMPBSA, Binding free energy.

5.2. Introduction

Sub-Saharan Africa has 66% of the global population of HIV-infected people, with South Africa having the highest incidence, with approximately 5.6 million people living with HIV [1, 18]. The country's extensive rollout of antiretroviral therapy (ART) has resulted in the disease no longer being a death sentence, with a decrease in mortality rates over the last decade [205, 206]. However, drug resistance to ART has become a serious challenge, as approximately 2 million people became newly infected with HIV [18] .

Reverse transcriptase (RT), an important enzyme in HIV-1, catalyzes the transcription of the viral single-stranded (ss) RNA into double-stranded (ds) DNA. HIV RT consists of two subunits, the larger p66 and the smaller p51 [3], with the polymerase and ribonuclease H (Rnase H) catalytic sites being located on the former. The polymerase domain of HIV resembles the right hand with fingers, thumb, palm, and connection sub-domain (FIGURE 20) [4]. A crucial role of RT in the life cycle of HIV-1 makes it the prime target of anti-retroviral therapy, such as non-nucleoside reverse transcriptase inhibitors (NNRTIs) [6]. The thumb and finger sub-domains of RT undergo conformational changes to perform the process of reverse transcription. The NNRTIs bind in the binding pocket is approximately 10 Å away from the polymerase in RT and disrupts the conformational flexibility of the enzyme [3].

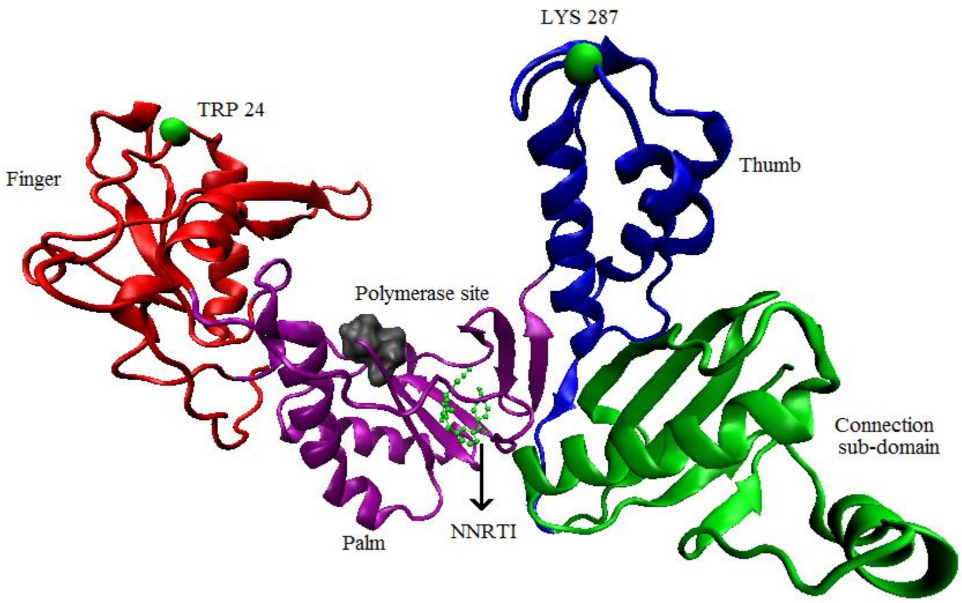


Figure 20: Sub-domains of HIV RT

Different sub-domains of HIV RT, i.e. thumb (blue), finger (red), palm (violet), and connection sub-domain (green) along with bound NNRTI rilpivirine. Index for the distance between finger and thumb, i.e. Trp 24 in the fingers and Lys 287 in the thumb are depicted with green spheres.

Regardless of improvement in anti-HIV therapy, HIV remains a challenge due to the rapid onset of mutations instigating drug resistance. Continuous efforts are underway to understand the effect of mutations on drug binding at the molecular level. Crystallographic studies have revealed that mutations causing NNRTI-resistance are mainly located in the vicinity of non-nucleoside inhibitor binding pocket (NNIBP) [64-66]. Mutations close to the NNRTI binding pocket appear to convene resistance to NNRTIs by either reducing the binding affinity of ligand or altering the dynamics of entry/exit of a drug. Mutation of Lys to Asn at position 103 (K103N) in the binding pocket is the most common NNRTI mutation in the KwaZulu-Natal province of South Africa, [207] and it is associated with anti-HIV treatment failure [208]. The importance of this mutation in the onset of NNRTI resistance motivated a comparison of the conformational dynamics of K103N mutated HIV-1 RT with WT HIV-1 RT bound to the NNRTI.

Our earlier scaffold-based QSAR study [7] identified two potential ligand scaffolds against HIV-1 RT (FIGURE 21). In that study, it was shown that some NNRTIs scaffolds had a higher coefficient of determination (Q^2) values, which accounts for higher structural variations appearing in the QSAR model.

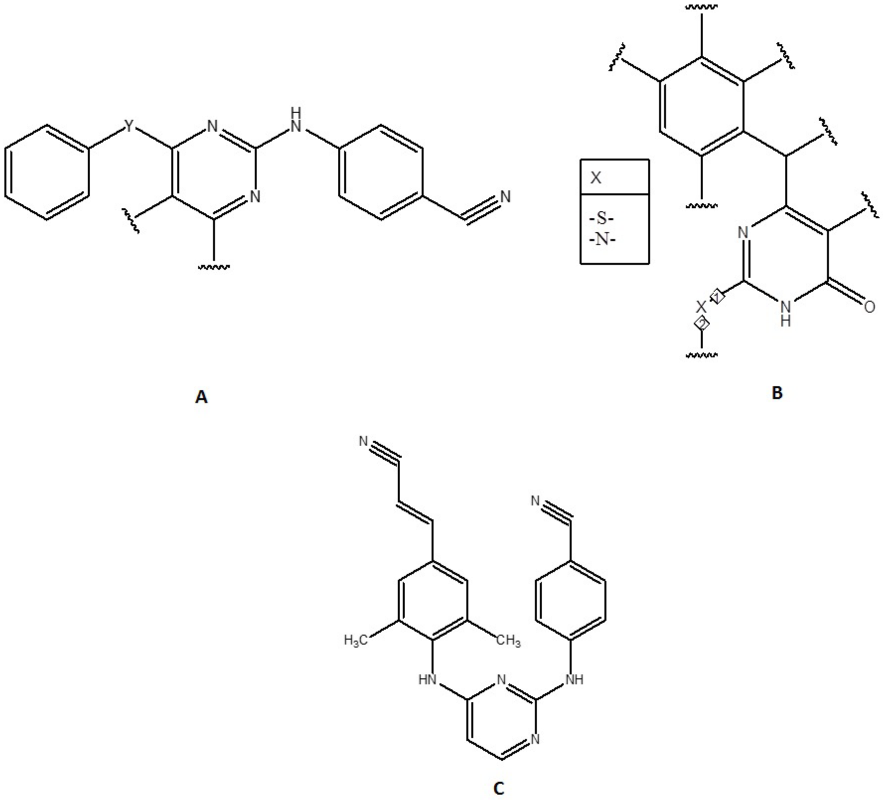


Figure 21: Strcuture of NNRTI.

*The structures of two identified potential chemical scaffolds of NNRTI (**A**) DAPY (Di-aryl pyrimidine) (**B**) DABO (Dihydro-alkoxy-benzyl-oxopyrimidine) and (**C**) rilpivirine a DAPY derivative [7]*

This study also indicated the importance of aromaticity, the number of nitrogen atoms and the structural flexibility of pyrimidine derivatives NNRTIs [7].

Previous studies have dealt with the molecular basis of NNRTI resistance due to K103N mutation in RT [68, 146] using 500 ps explicit MD simulation. Rilpivirine is a Di-aryl pyrimidine (DAPY) derivative with potent anti-HIV-1 RT activity against both WT and mutant HIV-1 RT (Figure 21). Its ability to reasonably adapt to the K103N mutation in RT is assumed to be due to the structural flexibility and the hydrogen bond formed by the linker N atoms [149]. This motivated an investigation of the dynamics of HIV-1 RT sub-domains in WT and K103N mutant, complexed with rilpivirine.

Due to a limited understanding of the conformational dynamics of K103N mutant HIV-1 RT, a cumulative 240 ns explicit MD simulation for WT and K103N mutant RT-rilpivirine complexes were performed in this study. The trajectories were analyzed based on backbone root mean square deviation (RMSD), root mean square fluctuation (RMSF), the radius of gyration (*Rg*), hydrogen bonding, analysis of distances between thumb and finger sub-domains, essential dynamics based on PCA, and dynamic pharmacophores (dynophore). Ligand binding often involves dynamic conformational transitions that may not be evident from a single, static structure [209]. Molecular geometry and chemical characteristics of the binding pocket in protein molecules play a crucial role in ligand binding [210]. The nature of the binding surface [211], as well as the complementary shape and polarity [212] are also assumed to be contributing factors in ligand-protein binding. An analysis of NNIBP volume over 30 ns MD simulation is also discussed.

5.3. Methods

5.3.1. HIV RT and NNRTI model preparation

An initial 3D model of free and bound WT HIV-1 RT was taken from the crystal structure of HIV-1 reverse transcriptase in complex with rilpivirine (PDB: 4G1Q) [213]. For the K103N mutant HIV-1 RT enzyme, HIV-1 K103N reverse transcriptase in complex with etravirine (PDB: 3MED) was undertaken [66]. Both etravirine and rilpivirine correspond to the same class of NNRTIs chemical scaffolds and adopted the similar pose inside the NNRTI binding pocket (superimposed crystal structure is shown in S1). The crystallographic waters were removed, and the correct protonation state was predicted and assigned using Propka [214, 215]. The structure of rilpivirine was sketched using ACD/ChemSketch [216] and the geometry was optimized with HF/6-31G*. The restrained electrostatic potential (RESP) charges [217, 218] were calculated using Gaussian09 [188] and fitted using the antechamber tool of Amber [111, 219]. The ligand was docked into the binding pocket in the K103N HIV-1 RT by making a grid box (spacing of 1Å and size of 24 ×24 × 24) around catalytic residue D110, D185 and D186 as well other residues of the pocket. Molecular docking was performed by Raccoon AutoDock [190] using AutoDockTools (ADT) [191] and AutoDockVina [117] with default docking parameters. The Lamarckian Genetic algorithm [128] was used as the search algorithm with default parameter values. The docking protocol was followed as described in our previous work[7].

5.3.2. Molecular dynamic simulation

Rilpivirine was parametrized using general amber force field (GAFF) [111, 219] with Antechamber. MD simulations were performed in Amber 12 [112] using the ff99SB force field [220] with the explicit TIP3P water model [221] box, keeping a minimum distance of 8 Å between the solute and each face. Missing hydrogen atoms were predicted and added using propka3.1 [214, 215] and Cl^- ions were utilized to neutralize the system using the Leap program of Amber12. Propka assign protonation state of an amino acid in protein based on empirical pKa prediction of titratable residues while considering its microenvironment. The long-range electrostatic force was treated using the particle mesh Ewald (PME) method [222], with a direct space and vdW (van der Waals) cutoff of 12 Å. Prior to the MD runs, the systems were partially minimized with a restrained force of 500 kcal/mol on the solute molecule using 750 cycles of the steepest descent, followed by 2500 cycles of the conjugate gradient method. The system was further minimized for 1500 cycles of conjugate gradient method, after which the systems were heated gradually from 0 to 300 *K* with a harmonic restraint of 10 kcal/mol to hold the solute fixed. Langevin dynamics was used to control the temperature using a collision frequency of 1.0 ps^{-1} and constant volume MD simulation. Before the production phase, MD systems were equilibrated for 2 ns at 300 *K* with a constant pressure of 1 bar. The SHAKE algorithm [223] was used to constrain bonds involving hydrogen atoms. A total of eight MD systems were prepared in this fashion, two for free RT (WT and K103N RT) and six for RIL-RT complex. The production phase of NPT MD was run for 30 ns with a time step of 2 fs using GPU version of Amber 12 [224], thus a total of 240 ns cumulative MD simulation was performed and analyzed in this work. The MD trajectories were analyzed for RMSD, RMSF, *Rg*, distance between thumb and fingers, and number of hydrogen bonds, for which the Ptraj and cpptraj [225] module of Amber 12 was used. The trajectories were also analyzed for pocket volume using MDpocket [226].

5.3.3. Principal component analysis (PCA)

To identify the correlated motion of the HIV-1 RT sub-domain, PCA analysis of the MD trajectories was performed. PCA is a multivariate statistical approach to reduce the dimensions of data, the intention being to remove the rotational and translational movement, and to excerpt the important motion from a MD simulation. This can be used to compare the conformational changes over the two MD trajectories. In PCA of the MD, eigenvalues and eigenvectors are obtained by diagonalizing the covariance matrix of atomic fluctuation. The eigenvector with the largest possible variance in the dataset is called first principal component (PC1). Each succeeding eigenvectors or PC_n (where n =2, 3…, n) in turn has the highest possible variance under the condition that is perpendicular to the previous PC. The eigenvalues represent the extent of motion, while the eigenvector defines the direction of motion [227]. The motion defined by the principal components was visualized by projecting the trajectory onto it. PCA was performed on backbone atoms of all the 30 ns MD trajectory. The ions and solvent molecules were stripped and the trajectory was rms fitted to the first frame. Ptraj from Amber 12 suite was used to perform the PCA and the porcupine plot of protein motion was created by NMWiz GUI for ProDyPrody [228] in VMD [229].

5.3.4. Binding free energy calculation

To assess the impact of the mutation on the ligand-protein affinity, binding free energy calculations for WT RT-RIL and K103N RT-RIL complex were performed using the MM-PB/GB SA method implemented in Amber 12. The free energy of ligand binding is the difference in free energy between two states, *i.e.* bound and unbound states of two solvated protein molecules (**equation 1**).

$$[L]_{aq} + [P]_{aq} \xrightleftharpoons[\Delta G(bind)]{} [LP]_{aq} \qquad (1)$$

[L]=ligand concentration, [P]=protein concentration and [LP]= complex concentration

Oweing to practical reasons, **equation 1** is not ideal for free energy calculation. An effective way is to divide the calculation as below:

$$\Delta G(bind,aq) = \Delta G(bind,vacuum) + \Delta G_{(aq,\,complex)} - \Delta G(aq,lig) - \Delta G(aq,receptor) \qquad (2)$$

In the MM-GBSA approach, the electrostatic component of the solvation free energy is calculated by solving the Generalized Born (GB) equation and adding an empirical term for hydrophobic contributions as:

$$\Delta G_{aq} = \Delta G_{gb} + \Delta G_{hydrophobic} \quad (3)$$

Linearised Poisson Boltzman equation is used for calculating solvation free energy in MM-PBSA method. *ΔG*$_{\text{vacuum}}$ is obtained using the following equation by calculating the average interaction energy between receptor and ligand, and by taking the entropy change upon binding into account:

$$\Delta G_{vacuum} = \Delta H - T\Delta S \quad (4)$$

Where T is the absolute temperature and ΔS is the change in entropy.

The average interaction energy of ligand with WT and mutated RT was calculated from 6000 snapshots extracted from 30 ns trajectories.

5.3.5. Dynophores

The mode of interaction between a ligand and its target can be represented by a structure-based 3D pharmacophore that describes the ensemble of electronic and steric features responsible for the interaction [230]. However, such pharmacophores only provide a static view of the ligand-target-interactions derived from a single conformational state of the complex, *e.g.* a crystal structure or a docking pose. In a novel dynamic pharmacophore approach, termed dynophores, which is an extension of the classic 3D pharmacophores, with statistical and sequential information about the conformational flexibility of a molecular system derived from MD simulations. For dynophore generation, pharmacophores are automatically created from each snapshot of an MD simulation. Pharmacophore features reoccurring over time, which share (i) the same feature type (*e.g.* hydrogen bond donor or hydrophobic regions) and (ii) the same atoms on ligand-site, are grouped into so-called dynophore "superfeatures". They contain information about the feature types and interaction partners between ligand and target, which are monitored in terms of their occurrence frequency (statistical behavior) and occurrence pattern (sequential behavior). The program used to generate dynophores and analyze the interactions pattern is DynophoreApp [231], which uses the API of the ilib/LigandScout framework [232, 233], a software for 3D pharmacophore modeling.

Dynophores provide a broad applicability scope as a MD analysis tool used in molecular modeling studies. While the analysis of MD trajectories typically focuses on selected inter-atomic distances and angles, statistics on the occurrence of specific chemical interactions (*i.e.* chemical and steric complementarity of certain chemical moieties on ligand and target side) during a MD trajectory yield a more comprehensive view on the dynamics of ligand binding by evaluating the importance and spatial evolution of chemical interactions. Recently dynophores analysis has been used to highlight subtle but relevant binding differences of highly similar muscarinic M_2 acetylcholine receptor modulators [234].

5.4. Results and discussion

In this section, MD trajectories of free and bound WT HIV-1 RT were analyzed to understand how the binding of rilpivirine alters the dynamics of HIV-1 RT, as well as the reason behind the ability of rilpivirine to withstand the drug resistance. The results from the RMSD and RMSF analysis, the ligand's mode of binding inside the NNRTI binding pocket, radius of gyration, analysis of distance between thumb and fingers, analysis of NNIBP volume, PCA and dynophore analysis are presented here.

5.4.1. RMSD and RMSF analysis

The structural stability of all the trajectories is depicted by the RMSD plot, the root means square deviation (in Å) of backbone atoms, in reference to the first frame of 30 ns production MD, is shown in **FIGURE 22**. The average value of the protein backbone RMSD over the 30000 frames for free K013N RT, K103N RT-RIL, free WT and WT-RPV were 3.2, 3.7, 2.9 and 3.5 Å, respectively. RMSD plots of additional MD runs are provided with supplementary information (**FIGURE S2**). All the RMSD plots have plateaued during the 30-ns simulation time scale. Reasonable convergence was achieved, as the RMSD curve become stable after 12 ns for all the structures, though K103N RIL shows more fluctuation in RMSD values between 20 – 21 ns in one of the simulations (FIGURE 22). It is also evident that a single mutation from LYS to ASN at position 103 does not bring any significant changes in the overall conformation of RT over 30 ns MD trajectories. This observation is consistent with the earlier crystal structure study, where it was found that the overall conformation of mutant RT does not differ significantly from the wild-type [235].

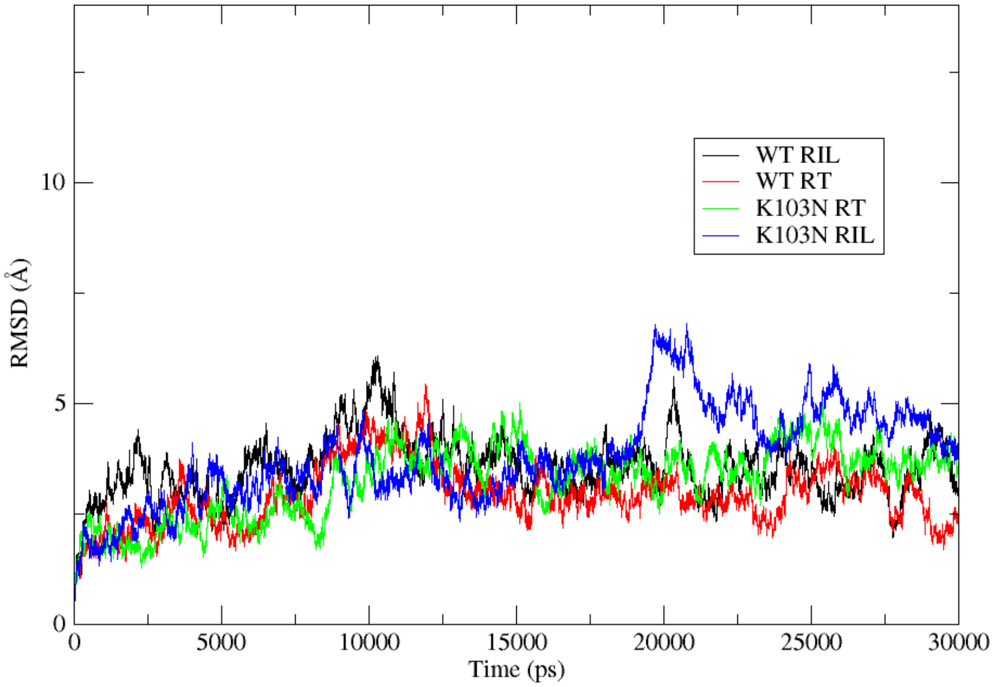


Figure 22: RMSD plot of backbone atoms of HIV-1 RT.

RMSD plot of backbone atoms of HIV-1 RT over 30 ns MD trajectories. WT RT in red, WT RT-RIL complex in black, K103N RT in green and K103N RT-RIL complex is shown in blue.

5.4.2. Ligand's mode of binding inside the NNRTI binding pocket

To assess the mutation effect on the induced ligand conformation, the RMSD of ligand inside the NNRTI binding pocket in RT were calculated. FIGURE 23 shows the ligand RMSD bound to WT RT and K103N RT. Ligand RMSD for additional simulation is shown in supplementary figures (FIGURE S4).

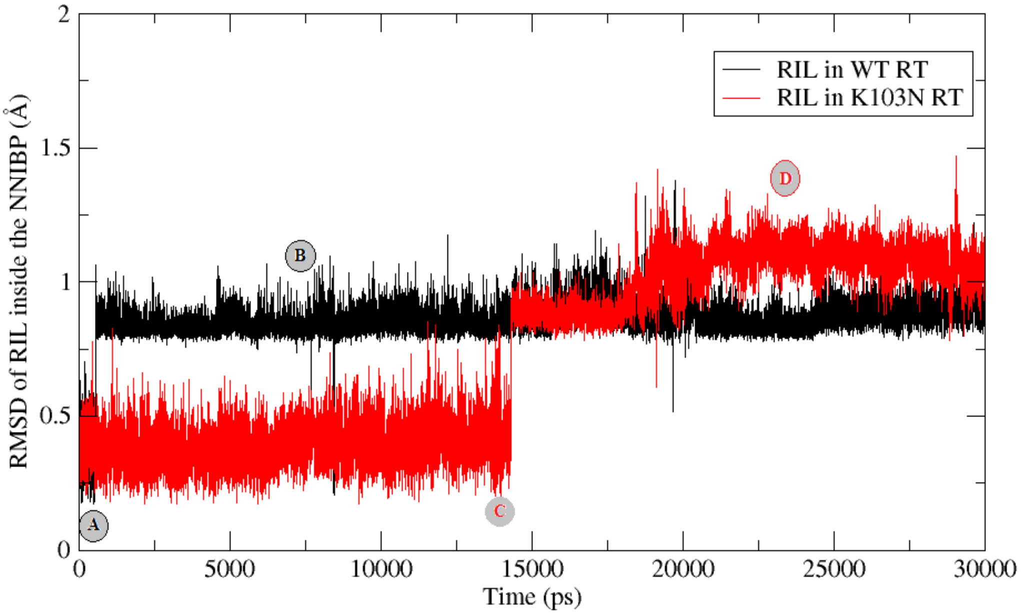


Figure 23: RMSD of rilpivirine inside the NNIBP.

*RMSD of rilpivirine inside the NNIBP of WT RT and K103N RT over 30ns MD trajectories with different part marked as A, B, C and D. The corresponding structure of rilpivirine is presented in **Figure 5***

It is interesting to note that rilpivirine, takes two different conformations inside the binding pocket during the simulation time (FIGURE 23, S3). FIGURE 24 shows the average ligand structure calculated over different parts of the trajectory for WT and the mutated systems marked as A, B, C and D in FIGURE 23. NNRTIs can take different modes inside the binding pocket, such as 'horseshoe' or 'butterfly' shapes [236], with rilpivirine being known to bind RT in multiple modes. It was observed that rilpivirine adopts the butterfly conformer inside the binding pocket of K103N RT, compared to the horseshoe shape in WT RT. However, when we repeat the MD simulation 2 more times with different starting structures, rilpivirine takes only horseshoe shape. Nonetheless, all the three RMSD plots of rilpivirine supports the notion of torsional flexibility (''wiggling'') and repositioning and reorientation within the pocket (''jiggling''). Such adaptations seem to be critical for the ability of the DAPY analoges to preserve their potency against drug-resistant mutant HIV-1 RTs[199].

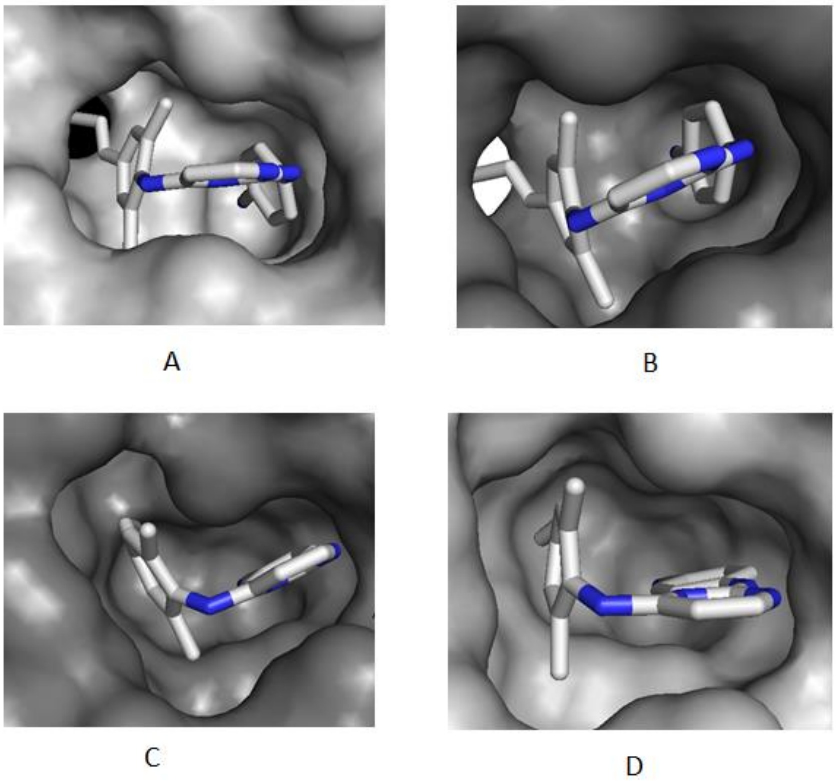


Figure 24: 3D structure of rilpivirine.

Structure of rilpivirine inside the NNRTI binding pocket in WT RT (A) frame 1 to 56 (B) frame 57 to 2000. Inside K103N mutated RT (C) frame 1 to 1422 (D) frame 1432 to 1855. Hydrogen atoms are not shown for clarity.

FIGURE S5 shows the rotation of the benzonitrile ring of rilpivirine inside the binding pocket of WT RT and K103N RT over the MD trajectory. The dihedral angle was measured using C8, N2, C1 and C2 atoms (FIGURE S5A), with plot of the dihedral angle vs simulation time corresponding with the changes in the ligand RMSD, as is shown in FIGURE 23**.** In the case of WT RT, the plane of the benzonitrile ring mainly forms an angle greater than 50˚ to give the horseshoe shape. Inside the binding pocket of K103N RT, it attains an angle of around ±180˚ over more than half of the simulation time, which gives the butterfly shape to the ligand. This reconfirms our earlier observation regarding the multiple conformations that rilpivirine adopts inside the WT and mutated HIV-1 RT.

Fluctuations in the Cα atom of each residue is shown in FIGURE 25**,** where it is evident that a single mutation K103N in the NNRTI binding pocket does not alter the overall structural flexibility of HIV-1 RT. This observation is in line with a previous study on RT, where no effect of the N348I/T369I double mutation on RMSF of RT was found [237].

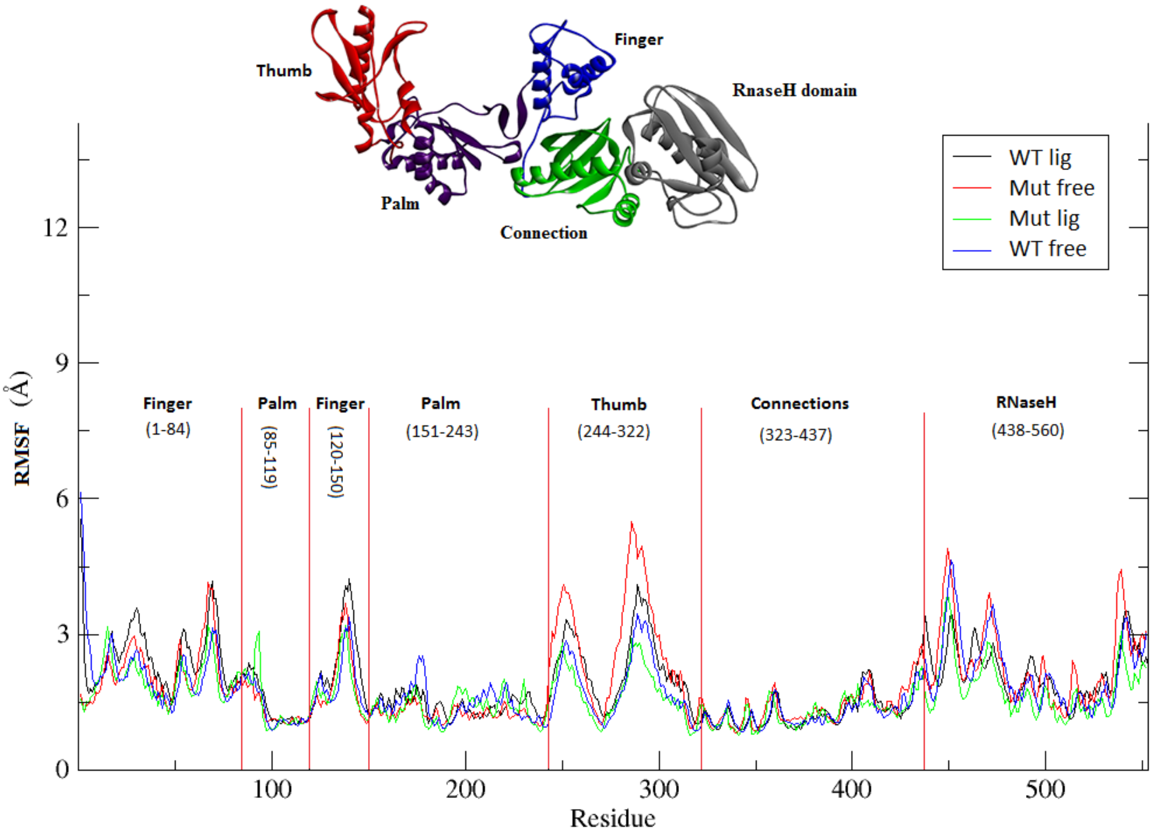


Figure 25: Root mean square fluctuation.

RMSF of Cα atoms of residues in HIV-1 RT. Residues are grouped into RT sub-domain they are the part of i.e. finger, palm, thumb, connection, and RNaseH.

5.4.3. Radius of gyration

The radius of gyration (Rg) refers to the distribution of atoms from the protein's center of mass (COM), indicating the level of protein compactness [238]. **FIGURE 26** shows the plot of the Rg evolution during the MD simulation for WT and K103N mutated HIV-1 RT.

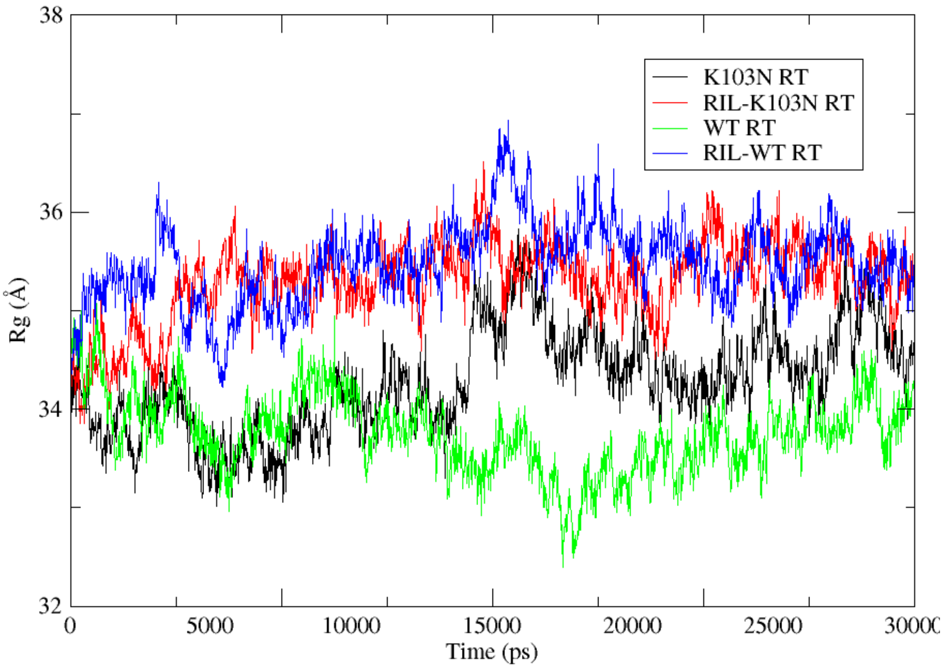


Figure 26: Radius of gyration (Rg)

Plot of Rg of Cα atoms of HIV RT over the simulation time.

For the brief initial time of 250 ps, all four trajectories showed similar Rg values, with the free WT and K103N RT had lower Rg values compared to the two RT-RIL complex. The WT and mutated RT-RIL complexes show the same Rg over the 30 ns, suggesting that there is no change in their structural compactness. A higher Rg, as in the case of RT-RIL, suggest the opening of the binding pocket in the protein to accommodate the entry of the ligand. This is supported by the histogram of Rg in **FIGURE S6**, where the free enzymes have a higher population of compact conformations compare to the RT-RIL complex. It is notable that there is an increase in Rg of the free mutated RT compared to the free WT RT around 9000 ps, where the mutated free protein takes less compact conformations. This encouraged us to explore the dynamics of thumb and fingers by calculating the distance between thumb and finger sub-domains of HIV RT (**FIGURE 20**). This analysis is discussed in the later section.

5.4.4. Analysis of distance between thumb and fingers

A double stranded nucleotide is held at the polymerase site by the thumb and fingers of RT. Their dynamics plays a major role in the RT functioning, and the NNRTI binding disrupts it

[3]. The opening and closing of the thumb and finger sub-domains can be described by the relative distance between them. The correct placement of 3′ end of RNA/DNA primer in the polymerase cavity is crucial for reverse transcription process. To understand the impact of K103N mutation and ligand binding on thumb and fingers dynamics, the relative distance between the COM of 24 TRP and 287 LYS were measured. It has been shown that the distance between these two residues clearly reflects the opening/closing of the thumb and finger sub-domains [239]. FIGURE 27 shows the plot of the distance measured over the entire 30 ns of simulation.

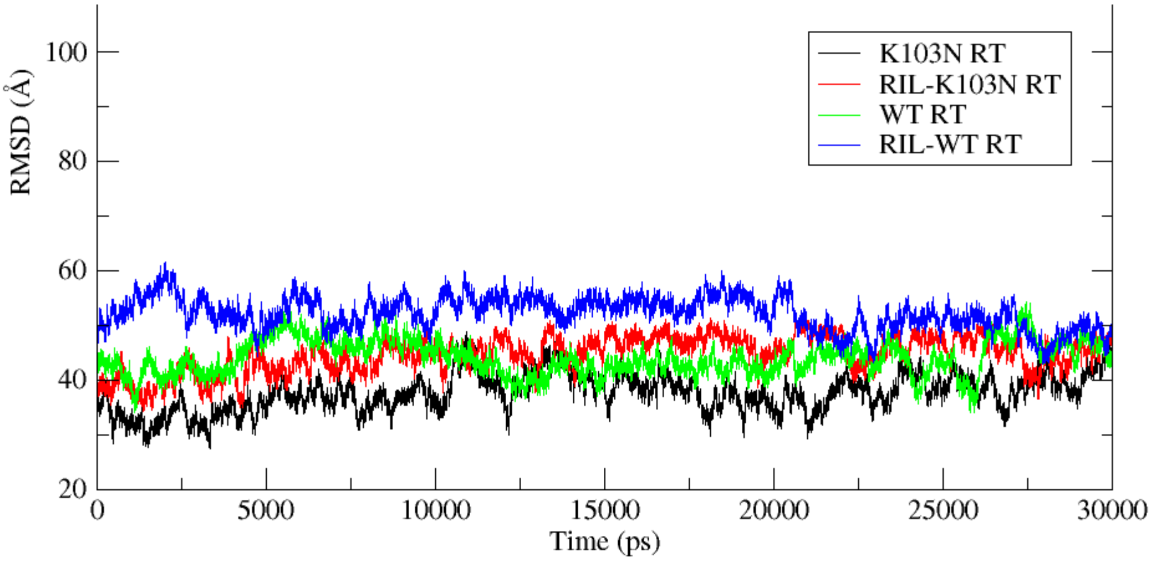


Figure 27: The plot of distance (in Å)

The plot of distance between COM of 24 TRP of fingers and 287 LYS of thumb sub-domain of HIV-1 RT.

The average distance between the COM of 24 TRP and 287 LYS in the WT RT-RIL and mutated RT-RIL is 52.2 Å and 44.6 Å respectively and does not show significant fluctuation. However, the average distances in case of the free WT and mutated RT changes at approximately 43.9 Å and 37.5 Å respectively. It should be noted that the increase in distance (approximately 2 Å) between thumb and finger sub-domains after 9.5 ns is also supported by the increase in Rg (around 1 Å) of free mutated protein around the same frame.

5.4.5. Analysis of NNRTI binding pocket (NNIBP) volume

The NNIBP (**Figure S7**) volume in Å^3 for 30 ns trajectories is plotted in **Figure S8,** with the average volume of the NNIBP for WT RT-RIL and mutated RT-RIL complex being 550.4 Å^3 and 527.8 Å^3 respectively. The free WT (191.1 Å^3) and mutated RT (432.9 Å^3) NNIBP volume differs considerably. It is interesting to note that the free K103N mutated enzyme has very large NNIBP compared to the free WT. While a wide open NNIBP can fit a larger variety of inhibitors without large steric clashes, a completely open binding site is not necessarily suitable to efficiently fit a given ligand [226]. This is possibly the reason that mutated enzyme binding pocket is energetically unfavorable for ligand binding, as the required active residues to interact with the ligand is not available.

5.4.6. Principal component analysis

To investigate the major motions in the HIV-1 RT enzyme, principal component analysis (PCA) on 30 ns trajectories was carried out. **Figure S9** in supplementary shows the plot of the structural variance explained by the first 20 eigenvalues (principal component). The 2D projection of MD trajectory on to the first two PCs obtained from diagonalization of the covariance matrix of the atomic fluctuations is shown in **Figure 28**.

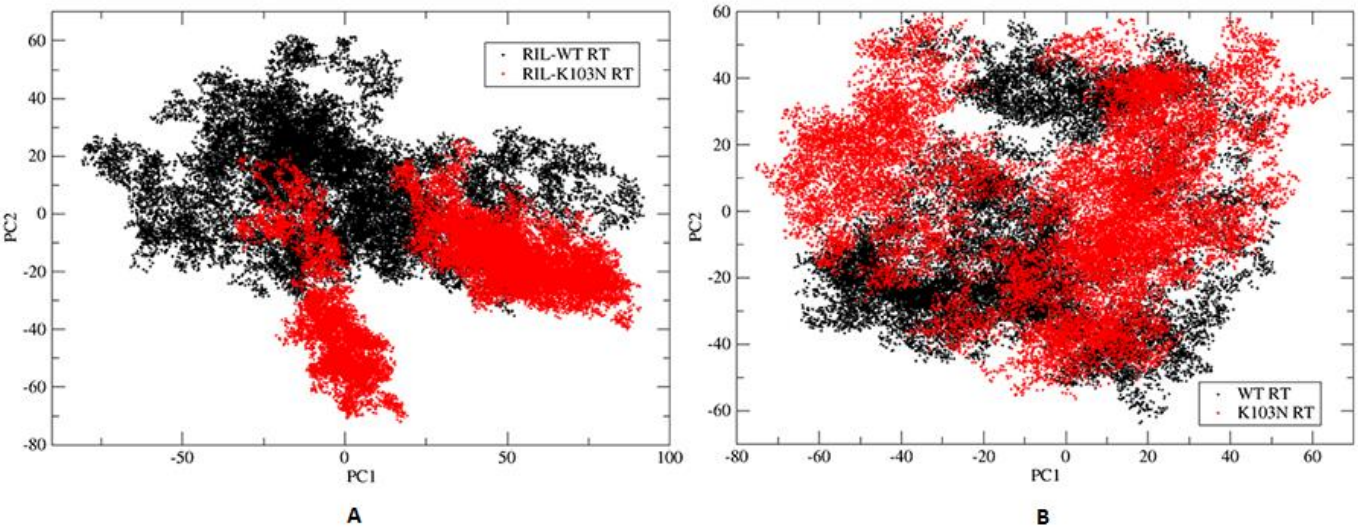


Figure 28: PCA plot

*Projection of MD trajectories on the first two PCs (**A**) WT and K103N mutated complex and (**B**) free WT and K103N mutated.*

K103N RT shows some conformational subspace overlap with the WT RT trajectories, implying that K103N RT samples only a part of the conformational subspace covered by WT

RT. It is noteworthy that mutated RT, when bound with rilpivirine, spans two distinct conformational sub-space in two out of three simulations. However, the free WT and mutated proteins have similar variances along the first and second PCs as well as higher overlap, indicating that they take the same conformations. Based on this observation, it could be inferred that conformational ensembles explored during the simulation time by RIL- K103N RT is different than its WT counterpart.

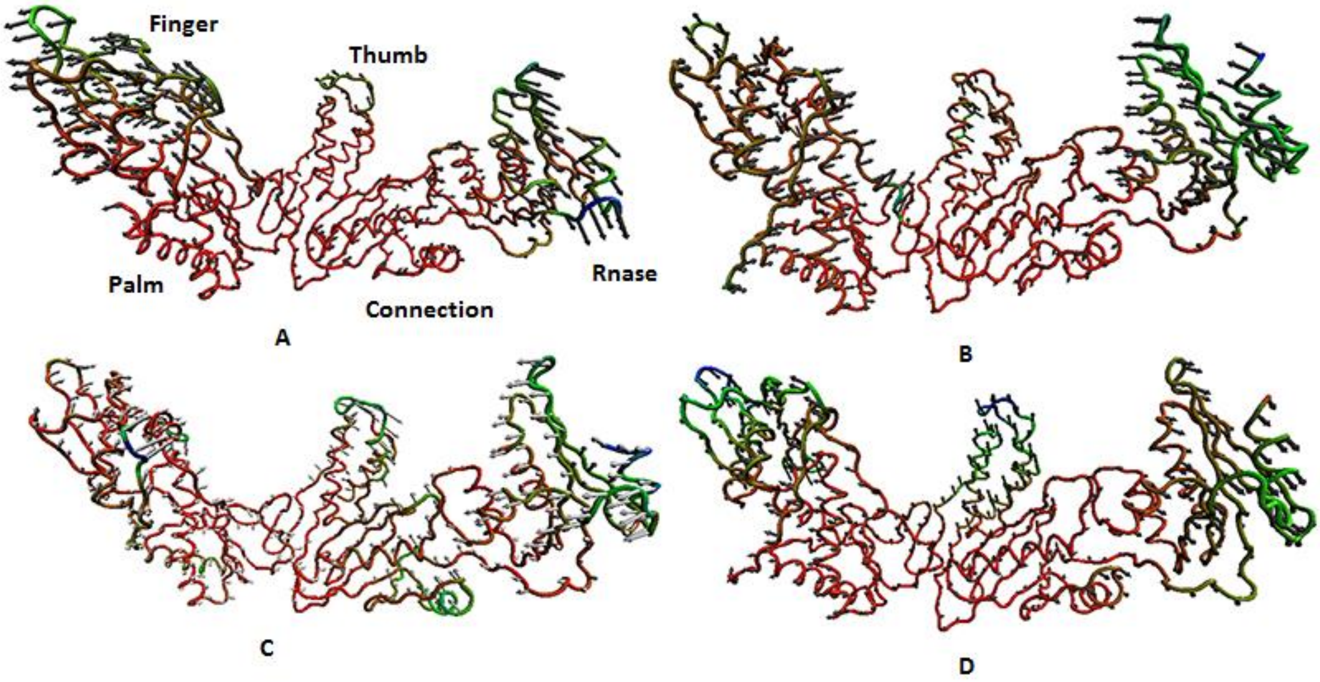


Figure 29: Porcupine plot of significant motion in HIV-1 RT

*Porcupine plot showing the significant motion across the first PC in the apo (**A**) K103N mutated RT and (**B**) WT RT, (**C**) K103N RT-RIL complex (**D**) WT RT-RIL complex. The color of the protein strands signifies the extent of motion and arrows shows the direction of correlated motion. Blue color reflects the highest movement followed by green, whereas red depict least moving parts of the protein.*

Based on the RMSF plot (FIGURE 25), it is evident that finger (residues 1-84 and 120-150) and thumb (residues 244-322) are the highly flexible regions, whereas the connection and palm sub-domains show fewer fluctuations. Interestingly, the porcupine plot (FIGURE 29) reconfirmed our observation that the prominent motion in RT is seen in both the thumb and finger sub-domains, whereas the connection sub-domain is rigid. The motion of sub-domains differs between the WT RT and K103N mutated RT, which suggests that the dynamics of the RT sub-domains and the impact of the mutation on the conformational variation of the RT enzyme could be of great significance.

5.4.7. Binding free energy of rilpivirine with HIV-1 RT

The average free energies of the solvation and other MM energy components for 30 ns MD trajectories of HIV RT complex with rilpivirine are given in TABLE 8.

Table 8: Binding free energy (ΔG_{bind} in kcal/mol).

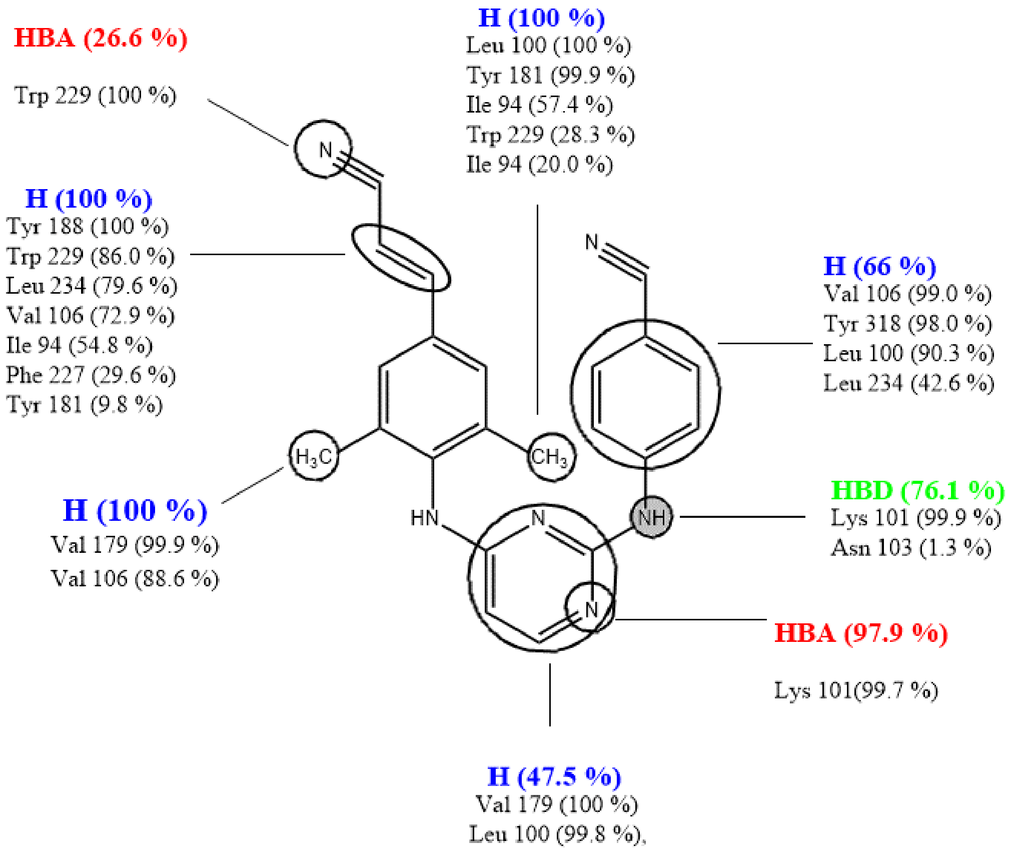
Free energy and different energy components, such as van der Waals (ΔE_{VDW}), electrostatic energy (ΔE_{Ele}) and solvation energy (ΔE_{Sol} GB and PB) for mutant and WT complex.

Complex	ΔE_{VDW}	ΔE_{Ele}	ΔE_{Sol} (GB)	ΔE_{Sol} (PB)	ΔG_{Bind}(GB)	ΔG_{Bind}(PB)
K103N RT	-63.43	- 12.08	25.82	67.00	-49.70	-8.51
WT RT	-67.66	-13.70	28.60	69.26	-52.76	-12.10

The contribution of MM van der Waals and electrostatic energy to the ΔG_{bind} is higher (more negative) in the WT-RPV complex, whereas the calculated solvation energy is slightly higher for binding of rilpivirine with WT-RPV. The ΔG_{bind} (GB) for the RIL-WT RT complex is -52.76 kcal/mol, and for K103N RT-RIL complex -49.70 kcal/mol. There is only a minor of -3.06 kcal/mol in the ΔG_{bind} of rilpivirine with WT and mutant HIV-1 RT, suggesting its similar biological activity against both strains of HIV-1 RT. Similar, trend is also seen with ΔGBind(PB). This observation is in agreement with the experimental biological activity measurement, where rilpivirine shows close activity against WT (0.4 nM) and mutant HIV-1 RT (0.3 nM) [240].

5.4.8. Dynophore analysis

A pharmacophore is composed of the abstract chemical and electronic features of a ligand that are responsible for its interaction with the binding pocket residues. To further understand the dynamic interaction of rilpivirine with the NNRTI binding pocket residues, novel dynamic pharmacophores, so-termed dynophores, were calculated for 2,000 equidistant snapshots of the 30 ns MD trajectories. FIGURE 30 shows the calculated dynophores for WT RT and K103N RT bound with rilpivirine.


HBA (26.6 %)
Trp 229 (100 %)
H (100 %)
Leu 100 (100 %)
Tyr 181 (99.9 %)
Ile 94 (57.4 %)
Trp 229 (28.3 %)
Ile 94 (20.0 %)
H (100 %)
Tyr 188 (100 %)
Trp 229 (86.0 %)
Leu 234 (79.6 %)
Val 106 (72.9 %)
Ile 94 (54.8 %)
Phe 227 (29.6 %)
Tyr 181 (9.8 %)
H (66 %)
Val 106 (99.0 %)
Tyr 318 (98.0 %)
Leu 100 (90.3 %)
Leu 234 (42.6 %)
H (100 %)
Val 179 (99.9 %)
Val 106 (88.6 %)
HBD (76.1 %)
Lys 101 (99.9 %)
Asn 103 (1.3 %)
HBA (97.9 %)
Lys 101(99.7 %)
H (47.5 %)
Val 179 (100 %)
Leu 100 (99.8 %),
N
N
H_3C
CH_3
HN
N
NH
N

A

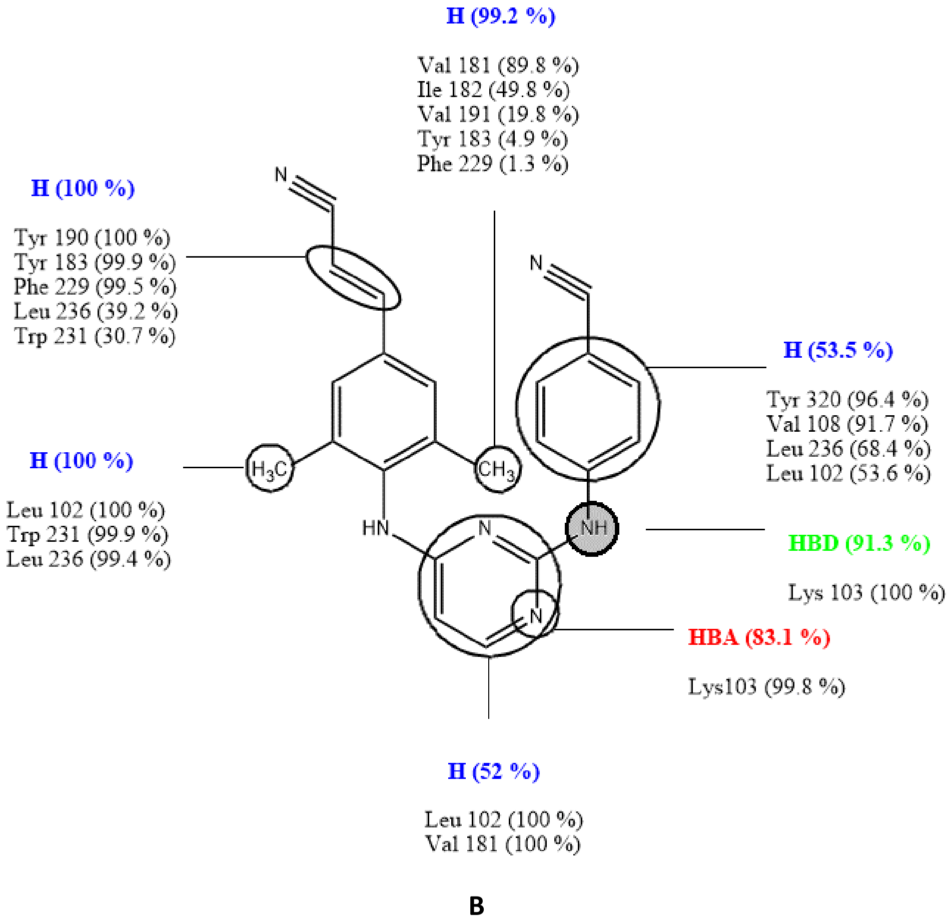


Figure 30: Comparison of the dynophores

*Dynaphores of (**A**) K103N mutated and (**B**) WT RT. Dynophores were generated from 2,000 equidistant snapshots MD trajectory. Occurrence percentages for the pharmacophore feature types, such as hydrogen bond acceptor and donor (HBA and HBD), as well as hydrophobic area (H), are shown with their associated interaction partners on the target site. The linker N atoms are encircled in gray.*

Rilpivirine belongs to the recent class of NNRTI, which has better tolerance for K103N mutation compared to the previous NNRTIs by interacting with the Asn103 in the mutant RT [241]. Dynophore-based analysis of the MD trajectories showed that the linker N atom of rilpivirine is constantly forming a hydrogen bond to Lys 103 in the wild type, whereas it permanently forms hydrogen bonds with Lys 101 (99.9%), but rarely with Asn 103 (1.3%) in the mutant enzyme. Our results, therefore suggest an interaction via hydrogen bonding between the linker N atom of rilpivirine and Lys 101, rather than the previously reported interaction with Asn 103 for the K103N mutant. This observation indicates an alternative interaction after K103N mutation, potentially stabilizing the ligand in the binding site. Alternatively, for further analysis, hydrogen bonds that are formed between the ligand and surrounding residues in the range of 10 Å for more than 5% time of trajectory were calculated with ptraj. It was observed that in WT RT, the H bond was formed between the N2 of rilpivirine and the oxygen atom of Lys 103, with 95% occupancy time. However, N2 of rilpivirine formed H bond with the Oxygen of Lys 101 over 92% of the K103N mutated RT's trajectory. These results are in good agreement with the dynophore findings, with differences in occurrence frequencies possibly being attributed to the different hydrogen bonding distance and angle cutoffs in the ptraj [225] and LigandScout/DynophoreApp [233].

5.5. Implication of MD results in drug resistance of K013N HIV-1 RT

Knowing that rilpivirine inhibits mutant HIV-1 RT [240], investigating the effect of mutations on drug resistance at the molecular level is of considerable significance. This motivated us to perform explicit MD simulation of rilpivirine bound to RT to investigate the molecular basis of its ability to withstand the binding site mutation of K103N in HIV-1 RT. The following insights could be gained from these results:

Conformational adaptability of rilpivirine, which belongs to the novel class of NNRTIs, assists their effectiveness against the mutated HIV RT. RMSD and dihedral analysis revealed that rilpivirine takes distinct conformations inside the binding pocket in the WT RT and K103N mutated RT. It takes a horseshoe shape inside the binding pocket of WT RT, whereas a butterfly conformation appeared in the K103N mutated RT. This could explain the reasonable biological activity profile of rilpivirine against various mutants, including WT HIV-1 RT [240].

Binding free energies calculated using the MMGBSA method do not differ significantly between WT RT-RIL (-52.81 kcal/mol) and K103N RT-RIL complexes (-51.54 kcal/mol), suggesting rilpivirine's ability to block both the

MD simulations of rilpivirine bound WT and K103N mutant RT were analyzed in terms of ligand-protein interaction profiles using dynophores, suggesting an alternative interaction for the mutant RT compared to previous reports. The linker N atom of rilpivirine is constantly forming a hydrogen bond to Lys 103 in the WT, while permanently hydrogen bonding with Lys 101 (99.9%), but only rarely with Asn 103 (1.3%) in the mutant enzyme. The outcome of this analysis suggests hydrogen bonding interaction between the linker N atom of rilpivirine and Lys 101 rather than the previously reported interaction with Asn 103 for the K103N mutant.

5.6. Conclusion

Dynamics-based studies of enzyme systems are crucial for the molecular understanding of conformational changes and mechanism of drug resistance. To achieve this insight, a cumulative 240 ns molecular dynamics simulations of apo and rilpivirine bound WT HIV-1 RT and K103N mutated HIV-1 RT in explicit solvent were performed. The findings showed that rilpivirine adopts different conformation due to "jiggling" and "wiggling" inside the NNRTI binding pocket of HIV-1 RT, indicating structural flexibility, thereby bypassing the drug resistance. Per residue fluctuations and PCA revealed that the major motion in RT was observed in both thumb and finger sub-domains, whereas the connection region was more rigid. It, therefore appears that investigating the dynamics of the thumb and finger sub-domains, and altering the rigidity of the connection domain, could also be an approach for HIV RT inhibition. Similar values of ΔG_{bind} of rilpivirine with WT and mutant RT correlated with the good inhibitory profile against WT, as well as with the mutant HIV-1 RT. Dynophore-based analysis of interaction patterns during conducted MD simulations of HIV-1 wild type and K103N mutant suggest that the better resistance tolerance of rilpivirine comes from binding to LYS 101 in K103N mutated RT. K103N mutation increased the NNIBP volume, which might be the reason for the poor binding of other ligands with RT. This study as well as previous crystallographic studies[199] suggests that any future development of novel NNRTI should be undertaken keeping the ligand's ability to undergo conformational changes within the pocket, while improving the efficacy. Extension of this study dealing with investigations on similar line on other NNRTIs and mutant are currently in progress.

Acknowledgements

This work was supported by a doctoral bursary grant from the Collage of Health Sciences (CHS) at UKZN in South Africa. We are grateful to the CHPC (www.chpc.ac.za) for providing computational resources and tools. BN and BH are also grateful to the National Research Foundation (NRF), South Africa, for the KIC travel grant (grant no 97110 and 97107) to visit Dr. Wolber's lab. We are also thankful to Mrs. Carren Martin for her editorial support.

5.7. Supporting information

Molecular insight on the binding of NNRTI to K103N mutated HIV-1 RT: Molecular dynamics simulations and dynamic pharmacophore analysis

Bilal Nizami[1], Dominique Sydow[2], Gerhard Wolber[2], and Bahareh Honarparvar*,[1]

[1]School of Pharmacy and Pharmacology, University of KwaZulu-Natal, Durban 4000, South Africa.
[2]Institute of Pharmacy, Freie Universität Berlin, Königin-Luise-Str. 2+4, 12175 Berlin, Germany.

*Corresponding author: Honarparvar@ukzn.ac.za (Dr. Bahareh Honarparvar), Telephone: + 27 31 2608482, School of Pharmacy and Pharmacology, University of KwaZulu-Natal, Durban 4001, South Africa.

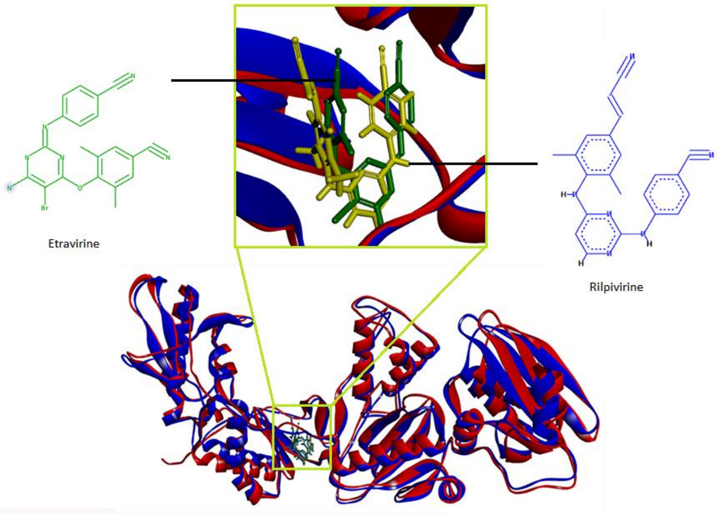


Figure S2: Superimposed structures of 3MED (blue) and 4G1Q (red). Superimposed structure of etravirine and rilpivirine inside the NNRTI binding pocket in their respective PDB is also shown in box.

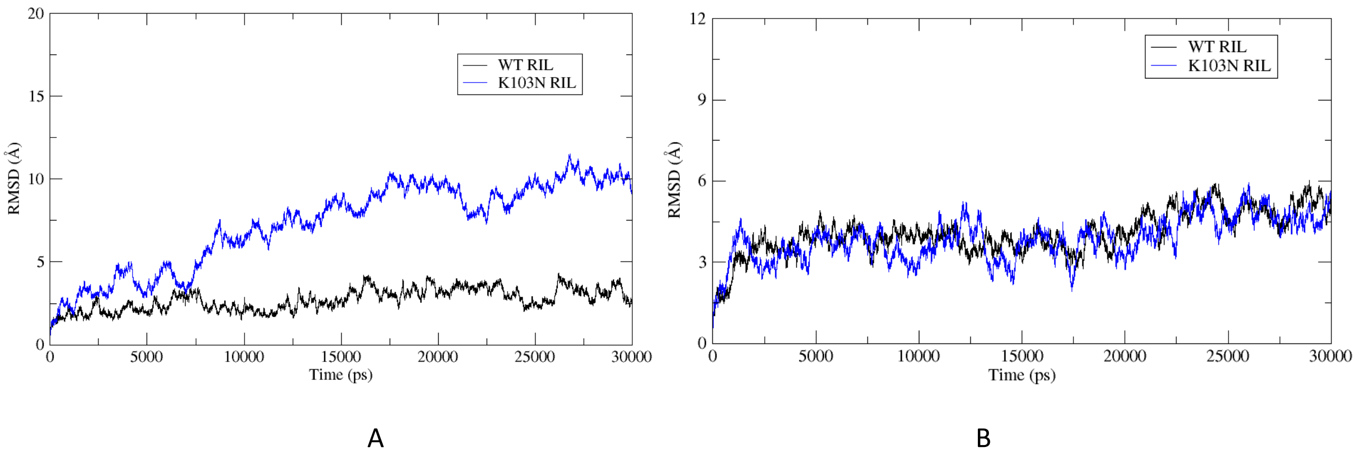


Figure S3: RMSD plot of backbone atoms of HIV-1 RT over two 30 ns MD trajectories

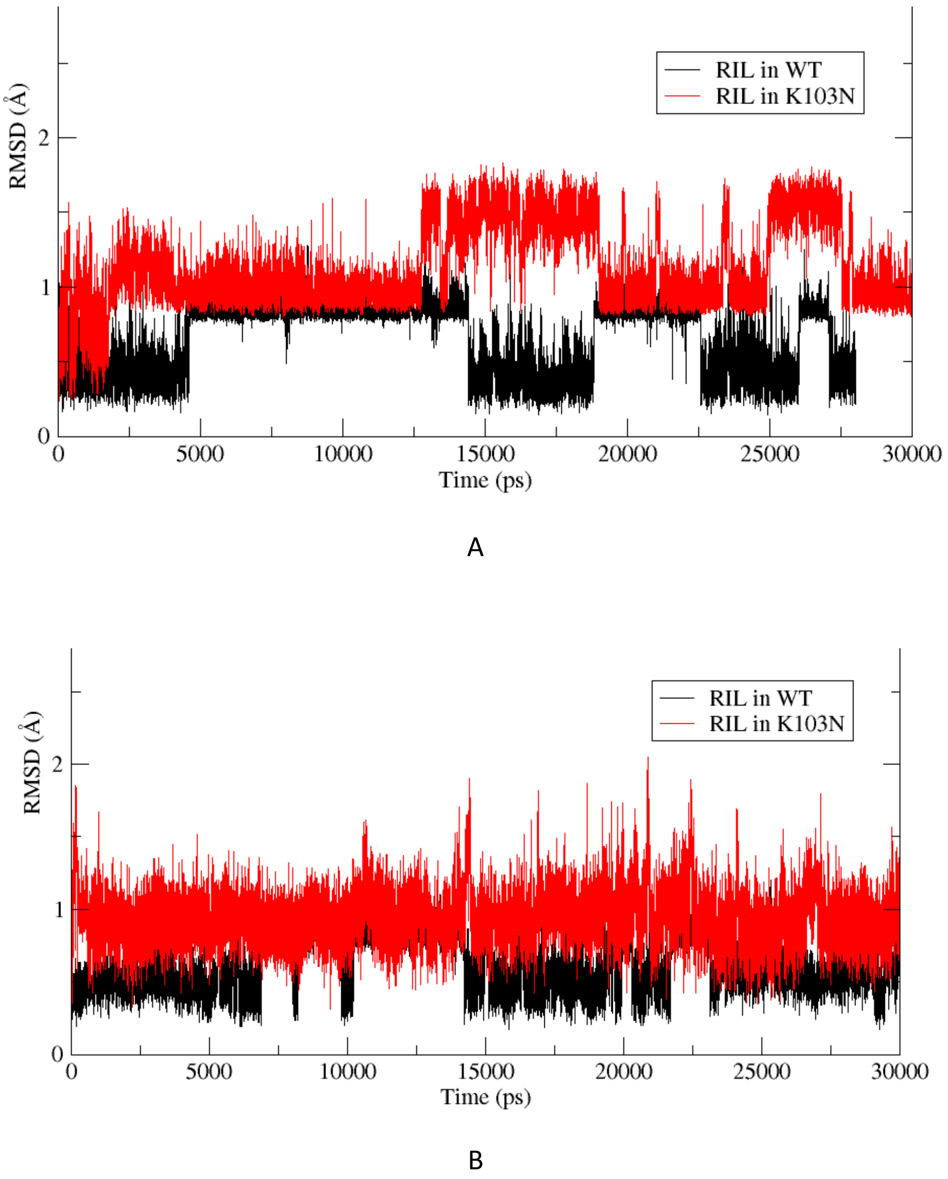


Figure S4: RMSD of rilpivirine inside the NNIBP of WT RT and K103N RT over two MD run of 30ns

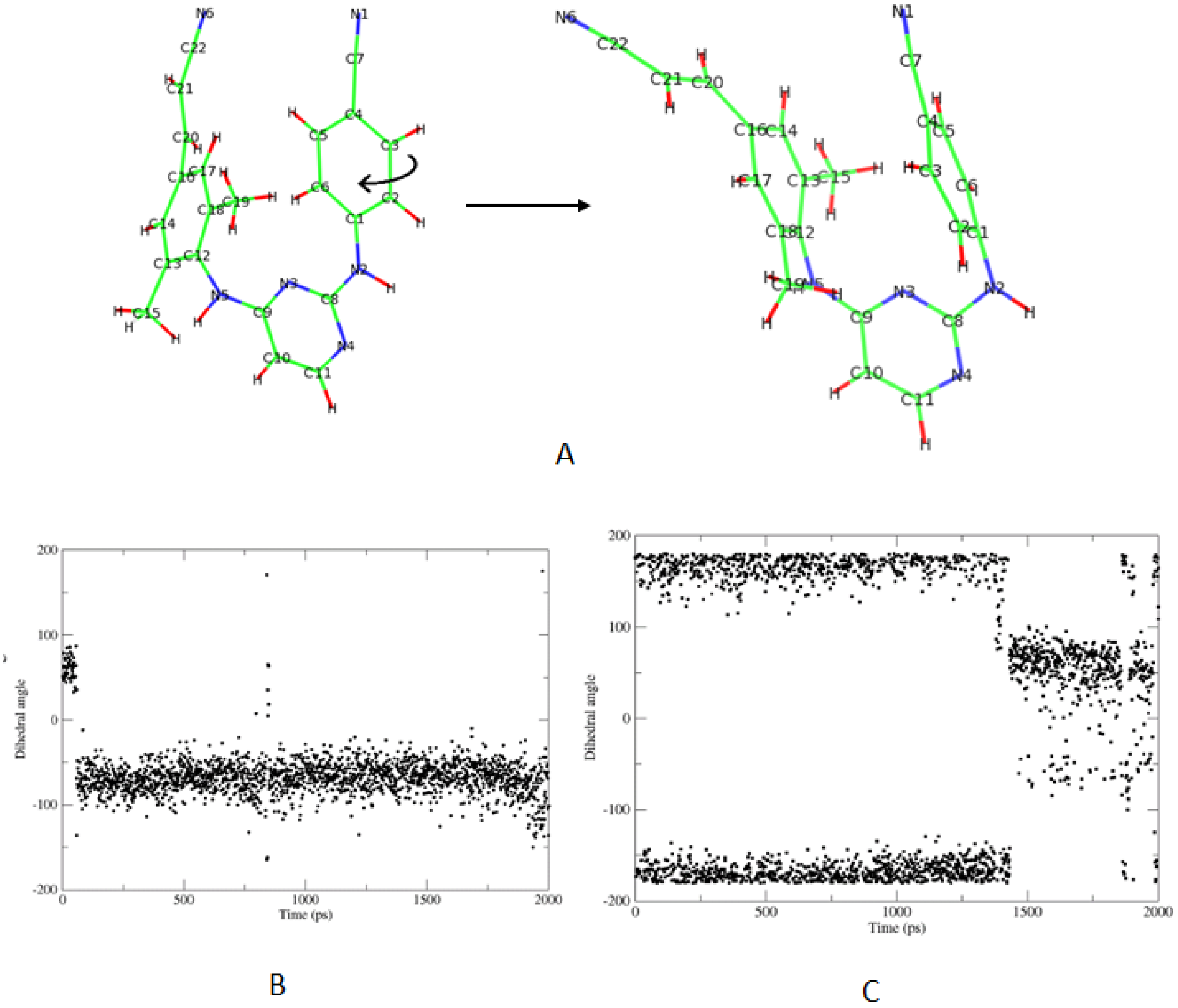


Figure S5: (A) rotation of benzonitrile ring in rilpivirine during the course of simulation (B) plot of rotation angle of rilpivirine in WT-RT (C) in K103N RT

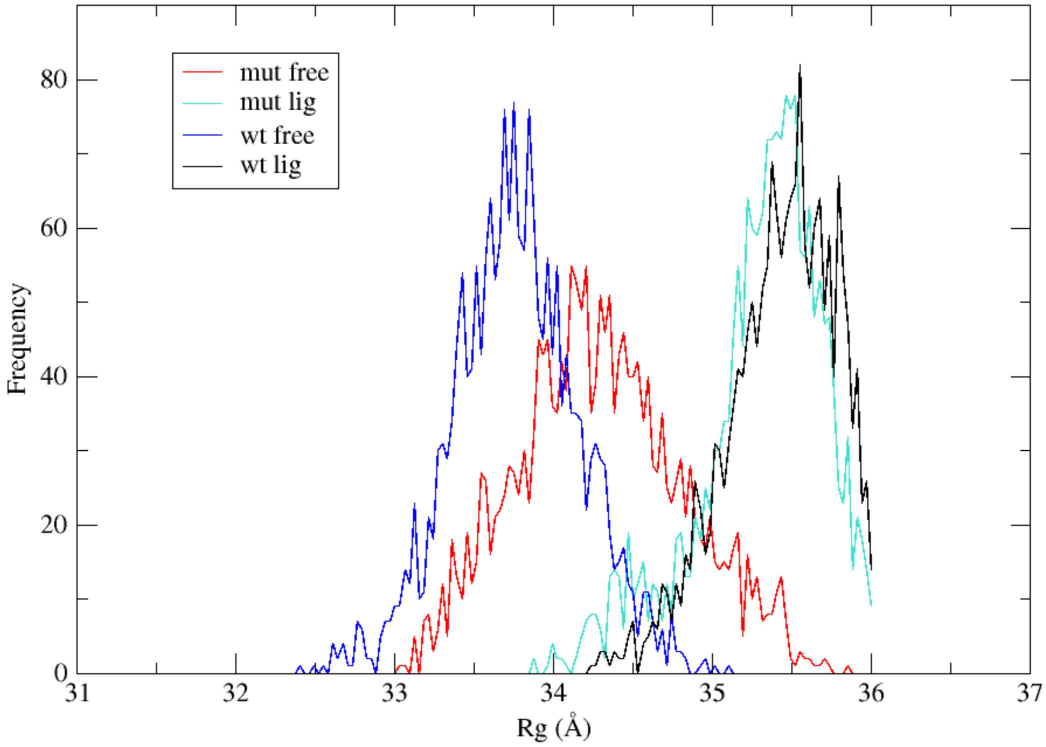


Figure S6: Frequency histogram of Rg (Å) calculated for HIV RT over 30 ns MD simulation.

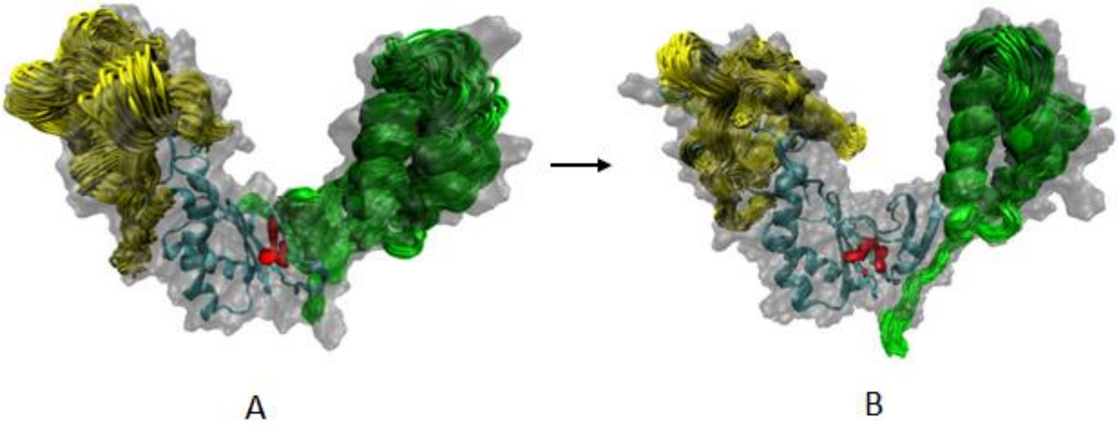


Figure S7: A view of NNRTI binding pocket (red surface) inside the ensemble of (A) WT HIV-1 RT and (B) K103N RT, along with transparent grey external molecular surface, thumb is shown in green and fingers in yellow. The figure is generated from 20 ns MD trajectory. Binding cavity surface is shown at isovalue of 8 in VMD. The unit of isovalue can be expressed as number of alpha sphere centers in an 8 Å3 cube around each grid point per snapshot. The more a cavity is conserved (or dense) the higher this value.

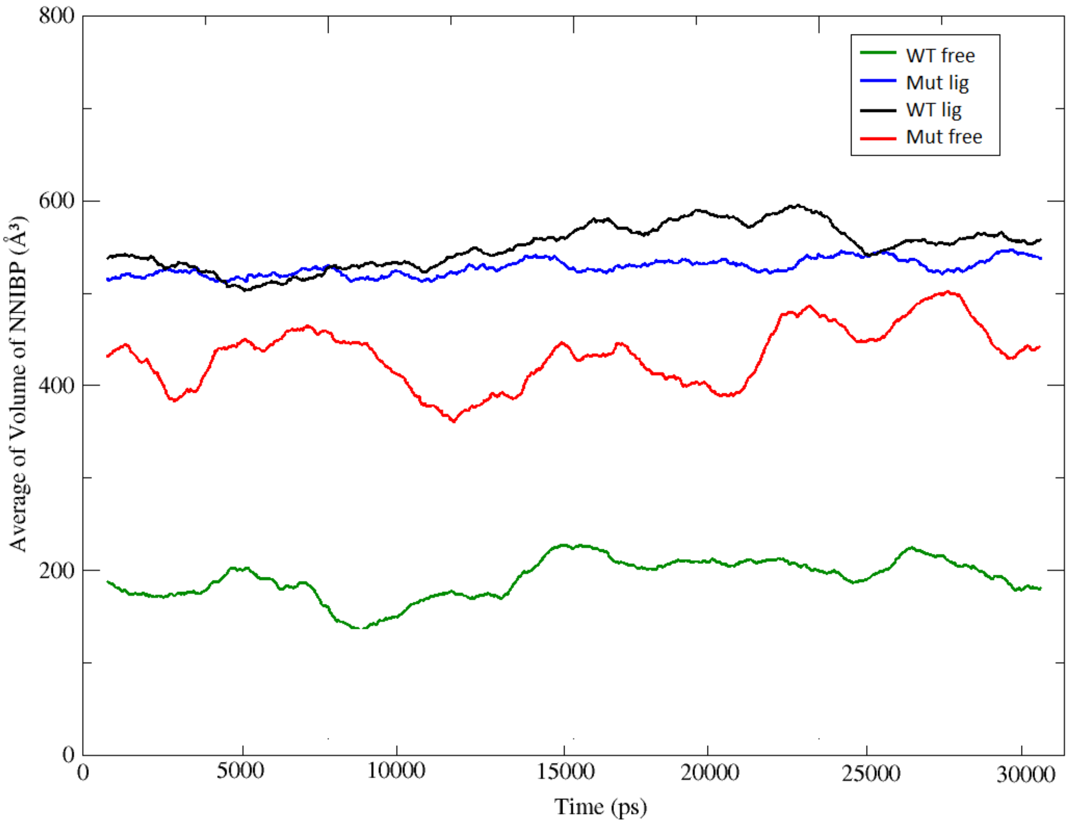


Figure S8: Plot of cumulative average of NNIBP pocket volume (Å) calculated from 30000 snapshots for RT, RT-RIL, K103N RT and K103N RT-RIL. Sliding window of 200 was used to calculate the running average and is plotted against time (ps).

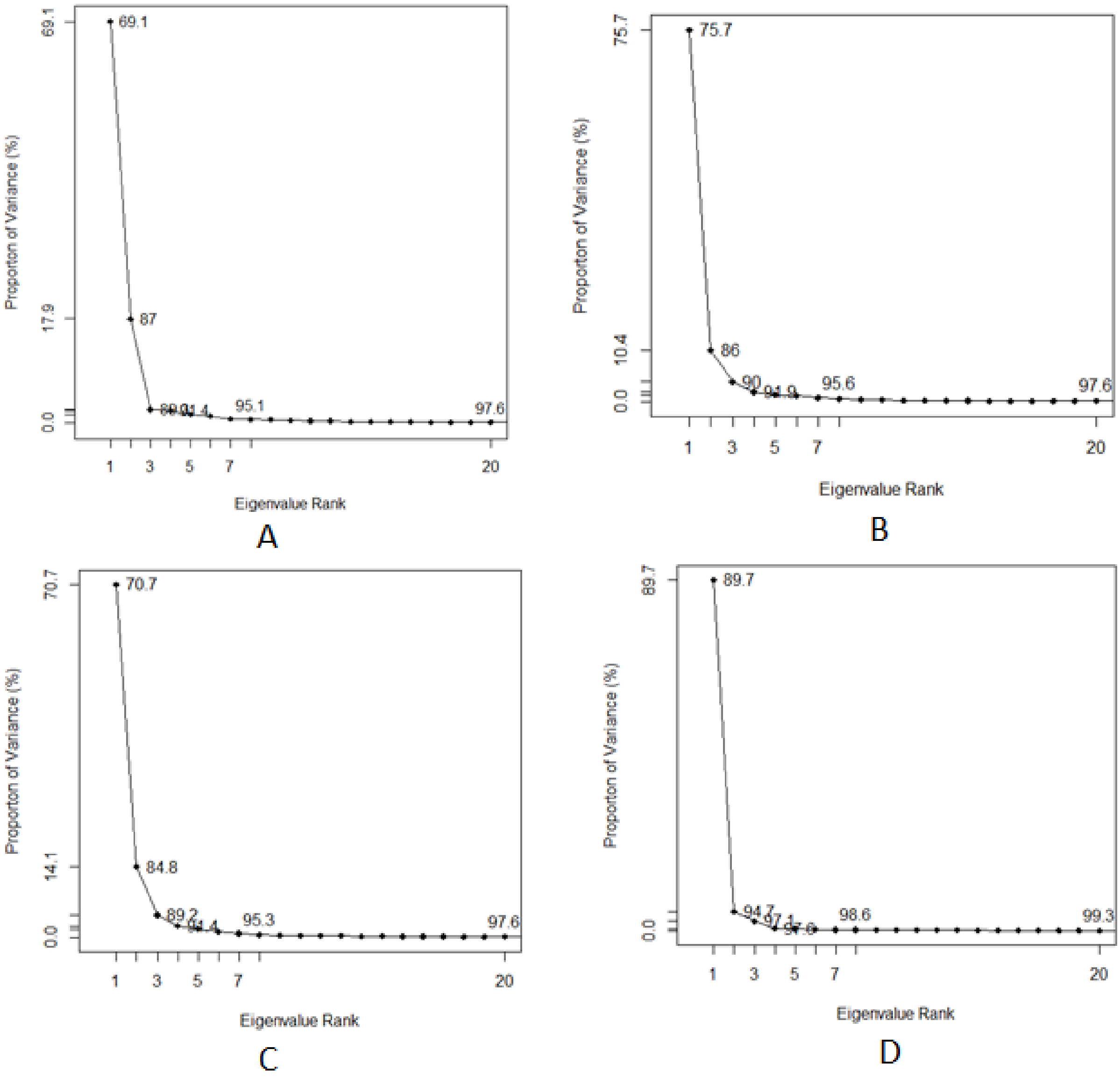
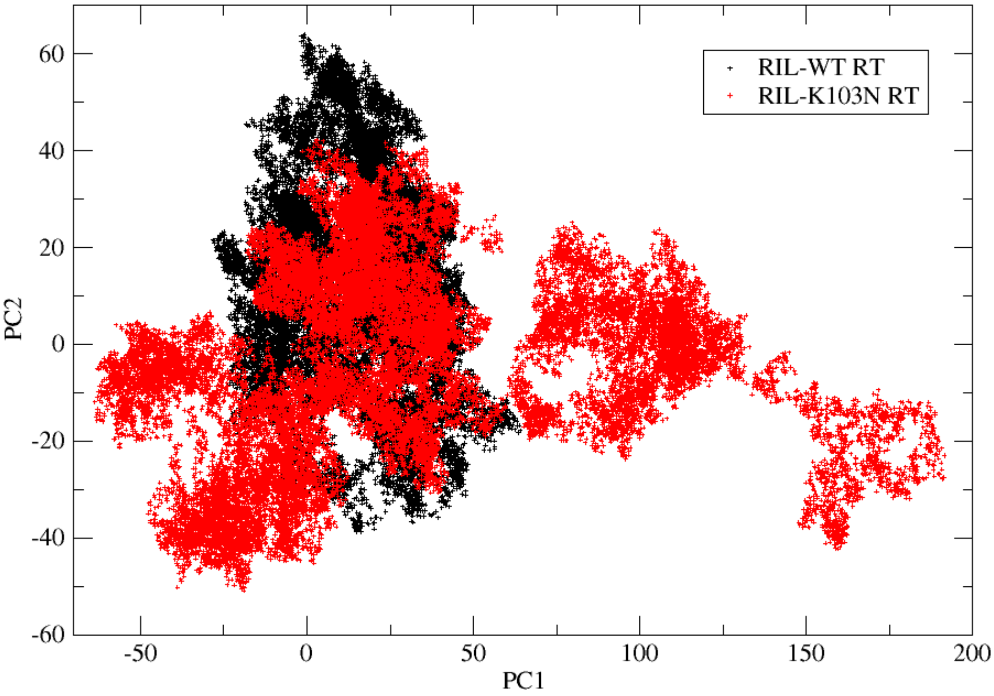
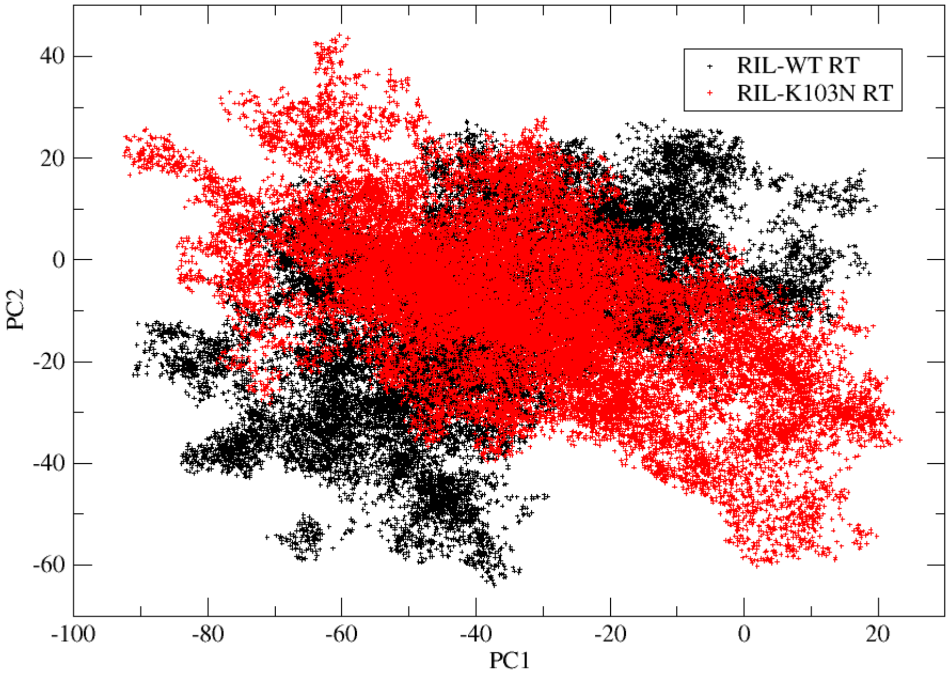


Figure S9: Proportion of variation explained by each PC. Values along the points in plot are cumulative variance. (A) K103N free HIV-1 RT, (B) WT free (C) WT-RPV (D) K103N-RIL



(A)



(B)

Figure S10: PCA plot of two MD simulation

CHAPTER 6

Molecular dynamics simulations of various NNRTIs bound with E138K mutated HIV-1 RT

Bilal Nizami[1] and Bahareh Honarparvar*[,1]

[1]School of Pharmacy and Pharmacology, University of KwaZulu-Natal, Durban 4000, South Africa.

*Corresponding author: Honarparvar@ukzn.ac.za (Dr. Bahareh Honarparvar), Telephone: + 27 31 2608482, School of Pharmacy and Pharmacology, University of KwaZulu-Natal, Durban 4001, South Africa.

6.1. Abstract

Irrespective of improvement in HIV treatment, challenges remain due to the rapid development of drug resistance due to mutations in HIV genomw. In this study, a cumulative 900 ns of molecular dynamic (MD) simulations of rilpivirine (RPV), efavirenz (EFV), and etravirine (ETR), bound with wild-type (WT) and GLU138LYS (E138K) mutated HIV-1 RT, were performed in solution. Binding free energy (MM-PB/GB SA) calculations were performed for all the ligand bound HIV-1 RT complexes. Calculated binding energies suggested the loss of potency of selected ligands against E138K-RT. Movement of the side chain of mutant residue (LYS 138) away from the binding pocket lead to loss of bonding with ligand, causing the drop in biological activity. This study investigated the underlying reason for selected NNRTIs failure to inhibit E138K mutant RT, which could be helpful to understand the molecular basis of HIV-1 RT drug resistance and design novel NNRTIs with improved drug resistance tolerance. Additionally, RPV binding with K103N HIV-1 RT is also investigated in this study.

Keywords

HIV-1, RT, Reverse Transcriptase, MD, Molecular Dynamics, Free energy, MM-PBSA, AIDS

6.2. Introduction

With the adaptation of UNAIDS recommended "Fast-Track approach" for tackling the AIDS epidemic, the world is committed to ending it by 2030. Nevertheless, enormous challenges lie ahead in ending the epidemic completely, with approx. 2.1 million new HIV infections cases worldwide in 2015 [2]. Development of resistance against anti-retroviral drugs owing to mutations in the viral enzymes reduce the odds against AIDS. The enzyme Reverse transcriptase (RT) alongside protease has been the main targets of anti-HIV drugs used in multi-drug combination therapy. RT, an important enzyme in HIV-1, catalyzes the transcription of the viral single-stranded (ss) RNA into double-stranded (ds) DNA. The HIV-1 RT consists of two subunits, the larger p66 and the smaller p51 [3], with polymerase and ribonuclease H (RNase H) catalytic binding sites [4]. There are two different class of drugs targeting the HIV-1 RT; NNRTI (non-nucleoside reverse transcriptase inhibitors) and NRTI (nucleoside reverse transcriptase inhibitors) both target different aspects of the RT functioning. There are currently four approved drugs in the NNRTI class –nevirapine (NEV), efavirenz (EFV), etravirine (ETR), and rilpivirine (RPV) (See **FIGURE 31**) — while delavirdine (DLV) was approved in 1997, but now is not recommend as part of initial therapy.

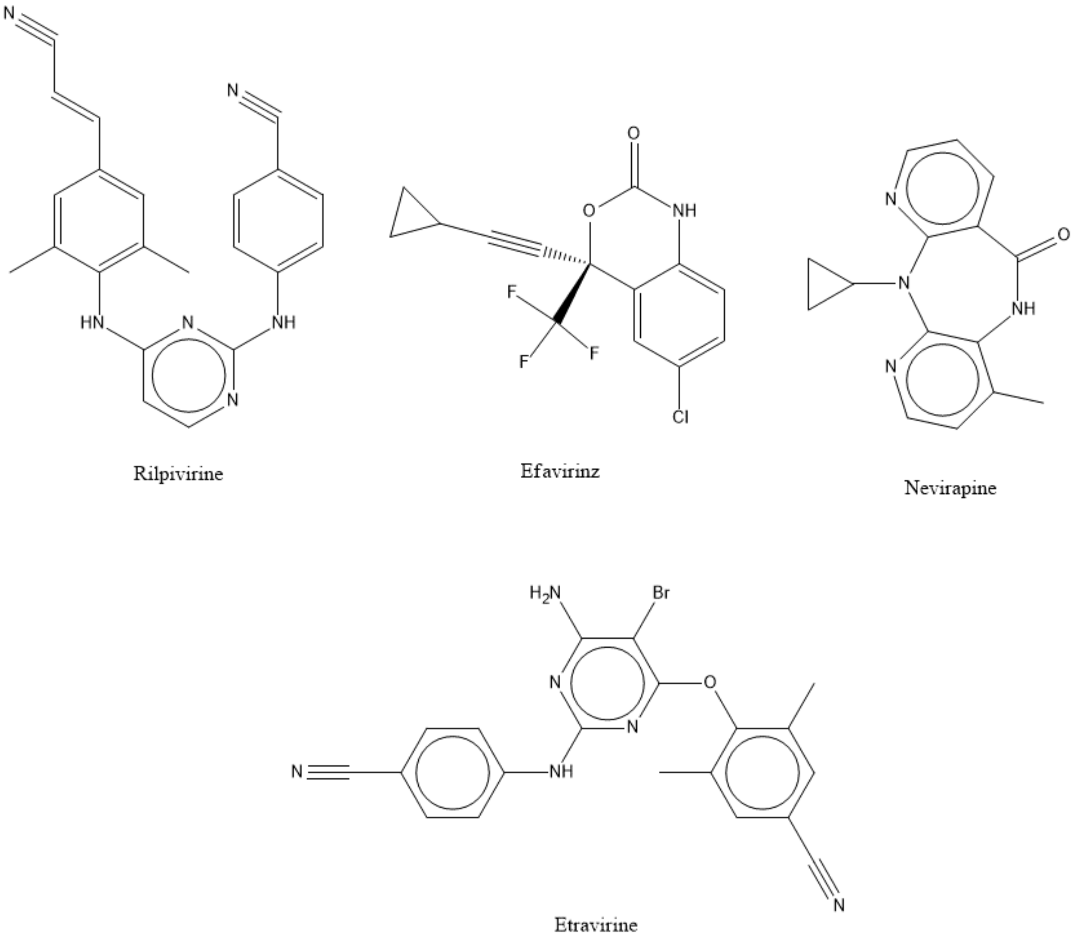


Figure 31: Structures of US FDA approved NNRTIs.

Regardless of improvement in anti-HIV therapy, HIV remains a challenge due to the rapid onset of mutations instigating drug resistance. Despite being the prime target of anti-HIV therapy, RT is responsible for emerging resistance to other drugs in the class: first, directly to RT inhibitors and second, indirectly as a key basis for instigating genetic variations [60]. Crystallographic studies have revealed that mutations causing NNRTI-resistance are mainly located in the vicinity of non-nucleoside inhibitor binding pocket (NNIBP) [64-66]. Some of the observed NNRTI-resistance mutations are L100I, K101E, V106A, K103N, V179D, Y181C, Y188L, G190A, and E138K (in p51 sub-domain); K103N and Y181C being the most common in patients receiving NNRTI treatments [60]. Residues K101, K103, and E138 (in p51) are situated at the rim of the NNRTI binding pocket (NNIBP) entrance for most NNRTIs. The mutations in NNIBP leads to the loss of aromatic ring stacking interactions (Y181C or Y188L), steric hindrance (L100I or G190A/S), and alteration of hydrophobic interactions (V106A or V179D). Effects of drug resistance mutations are rather severe on the first-generation NNRTIs (nevirapine, delavirdine, and efavirenz), for instance, high level of resistance by Y181C to nevirapine. The K103N and E138K mutations are largely linked with treatment failure of the efavirenz and rilpivirine, respectively, when combined with tenofovir and emtricitabine [150, 151]. The K103N mutation had emerged as a clinical resistance mutation upon treatments with efavirenz, and it confers an almost uniform level of cross-resistance to most NNRTIs [242-244], with the exception of the second-generation NNRTIs such as RPV and ETR [149]. The E138K is a non-polymorphic mutation in the p51 subunit of RT that is selected preferentially in patients receiving RPV and reduces its susceptibility up to 5-fold [245]. In our previous study, we have explored the molecular level understanding of the binding of RPV to K103N mutated HIV-1 RT [246]. The study has suggested that RPV's torsional flexibility (‘‘wiggling’’) and repositioning and reorientation within the pocket (‘‘jiggling’’) makes it withstand the drug resistance. Moreover, it was also proposed that the better resistance tolerance of rilpivirine is due to the formation of an alternate binding of linker N atom of rilpivirine to LYS 101 in K103N mutated RT [246]. The present study deals with E138K mutation in the p51 subunit of HIV-1 RT. Here we present MD simulation of wild-type (WT) and E138K HIV-1 RT complexed with efavirenz (EFV), etravirine (ETR), and rilpivirine (RPV). Additionally, MD simulation of K103N RT complexed with RPV is also presented. The main motivation behind this study is to understand why the second generation NNRTIs are able to bypass certain drug resistance mutation such as K103N, while at the same time being susceptible to E138K. Understanding of dynamics of NNRTI binding and effect of the mutation on drug binding is helpful in designing new inhibitors with better resistance tolerance.

Due to the large size of HIV-1 RT (over 1000 amino acids) and the consequent increase in the computational cost of MD simulations has led to exclusion of the p51 subunit from most MD studies. In addition, the structural complexity and flexibility of the RT made the MD simulations more formidable, though potentially more rewarding. As GLU 138 (p51) is part of the NNRTI binding pocket, HIV-1 RT with a part of the p51 subunit was considered for cumulative 1 µs MD simulation. To understand the binding affinities of NVP, EFV, and ETR against mutant RT at the molecular level, this study is the first account of MD simulation for E138K HIV-1 RT.

6.3. Materials and methods

6.3.1. System preparation and MD simulations

The initial starting structure of WT and mutants' HIV-1 RT (E138K and K103N) was taken from X-ray structure. PDB 4G1Q [66] for wild-type (WT), 2HNY [247] for E138K and 3MED

[213] for K103N was used to model initial starting structure. The protonation of the HIV-1 RT was assigned at physiological PH by H++ web server [248]. The whole enzyme-inhibitor complexes were solvated in a cubic box of water 8 Å from protein surface and neutralized with counter ions using tleap suite of Amber14 [249]. We used the ff12SB [250] force field for protein, GAFF [111, 219] for ligands and the TIP3P model [221] for water. The selected NNRTIs were first modelled and docked inside the NNRTI binding pocket of RT as described in our previous work [246, 251]. A total of 9 such systems were prepared (See TABLE 9). All the systems were equilibrated in the NPT ensemble at 300 *K* for 2 ns, and then the 100 ns of production simulation for each system was performed in the NVT ensemble at 300 *K* using Amber14 [249]. The same protocol was followed for MD simulations as described in our previous work [246]. A total of 900 ns of cumulative MD simulations were carried out and analyzed in this work. Approximately over 93000 atoms were present in the solvated system. The MD trajectories were analyzed in terms of RMSD, RMSF, the distance between thumb and fingers, and number of hydrogen bonds using ptraj and cpptraj modules [225] of Amber 14.

Table 9: A total of 9 MD systems were prepared as shown in the table below.

WT HIV-1 RT was modelled from PDB 4G1Q, E138K RT was modelled from PDB 2HNY.

S. No.	Ligand	Protein (PDB)
1	Rilpivirine	WT RT (4G1Q)
2		K103N mutant RT (3MED)
3		E138K mutant RT (2HNY)
4	Etravirine	WT RT (4G1Q)
5		E138K mutant RT (2HNY)
6	Efavirenz	WT RT (4G1Q)
7		E138K mutant RT (2HNY)
8	Free enzyme	WT RT (4G1Q)
9		E138K mutant RT (2HNY)

6.3.2. Binding free energy calculations

Binding free energy calculations for all nine MD systems were performed using the MM-PB/GB SA method implemented in Amber 14. The binding free energy of ligand is the difference in free energy between two states, *i.e.* bound and unbound states of two solvated protein molecules as:

$$[L]_{aq}+[P]_{aq}\underset{\Delta G(bind)}{\xleftrightarrow{\hspace{3em}}}[LP]_{aq} \qquad (1)$$

[L]=ligand concentration, [P]=protein concentration and [LP]= complex concentration

Due to practical reasons, **equation 1** is not ideal for free energy calculations. An effective approach is to divide the calculations as follows:

$$\Delta G(bind,aq)=\Delta G(bind,vacuum)+\Delta G_{(aq,complex)}-\Delta G(aq,lig)-\Delta G(aq,receptor) \qquad (2)$$

In the MM-GBSA method, the electrostatic component of the solvation free energy is calculated by solving the Generalized Born (GB) equation and adding an empirical term for hydrophobic contributions as:

$$\Delta G_{aq}=\Delta G_{gb}+\Delta G_{hydrophobic} \qquad (3)$$

Linearised Poisson Boltzmann equation is used for calculations of the solvation free energies with MM-PBSA method. ΔG_{vacuum} is obtained using the following equation by calculating the average interaction energy between a receptor and ligand, and taking the entropy change upon binding into account as:

$$\Delta G_{vacuum}=\Delta H-T\Delta S \qquad (4)$$

Where T is the absolute temperature and ΔS is the change in entropy.

The average interaction energy of the selected NNRTIs with WT and mutated RT was calculated from 20000 snapshots extracted from each 100 ns MD trajectories.

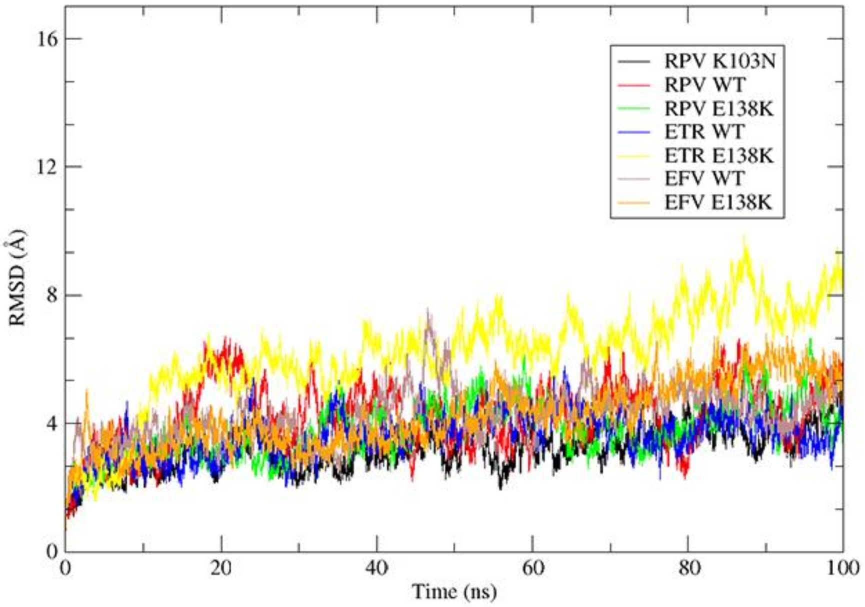
6.4. Results and Discussion

In this section, 100 ns MD trajectories of all the systems mentioned in TABLE 9 are analyzed to understand how the binding of NNRTIs alters the conformational dynamics of HIV-1 RT. The results from the RMSD and RMSF analysis, the ligand’s mode of binding inside the

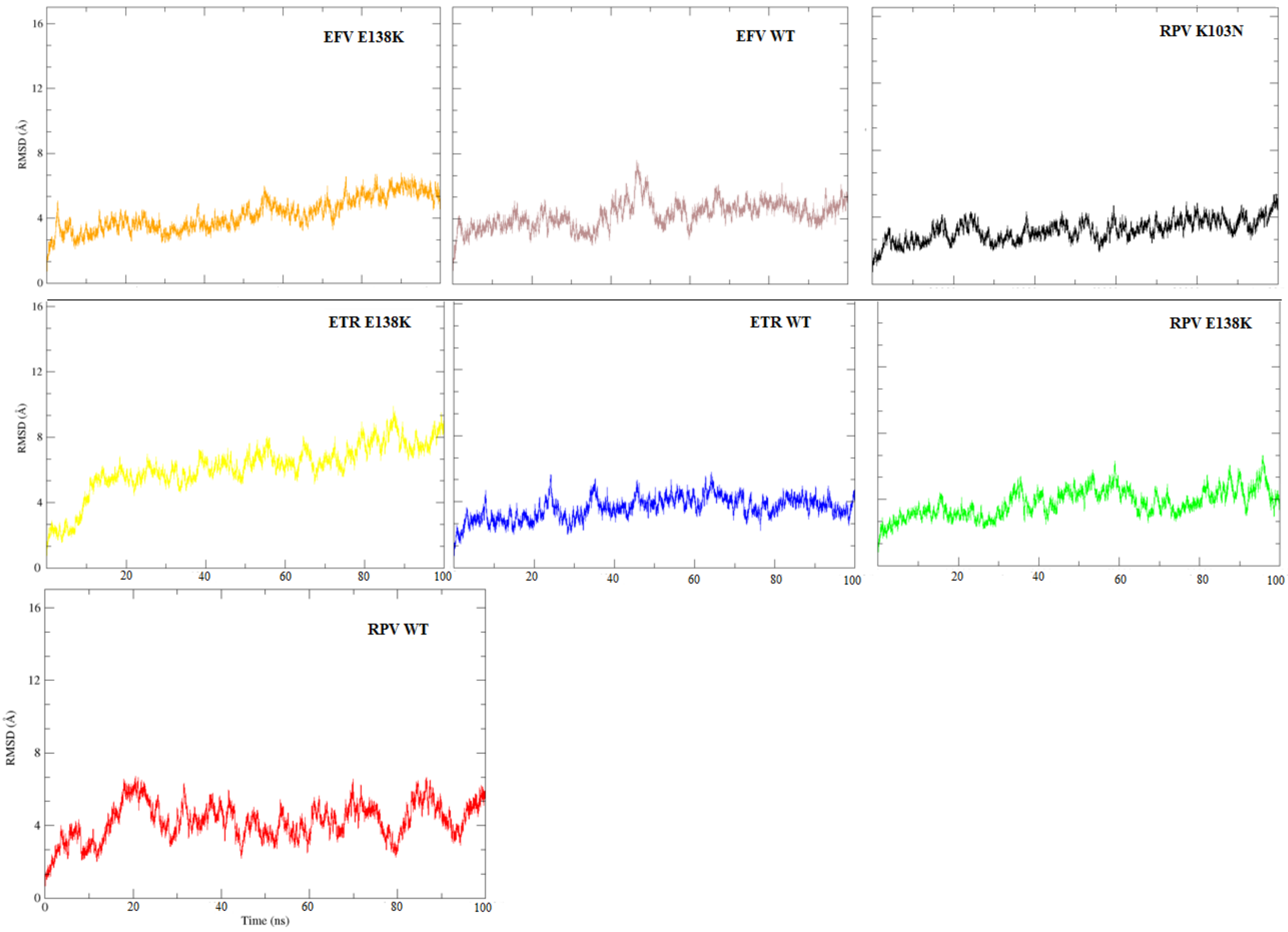
NNRTI binding pocket, radius of gyration**,** analysis of distance between thumb and fingers, PCA, and binding free energy analysis are presented here.

6.4.1. RMSD and RMSF

The root means square deviations (RMSD in Å) of the backbone atoms in reference to first frame for the MD trajectories are shown in **FIGURE 32**. A reasonable convergence is achieved as all the systems have been plateaued during the 100 ns production MD. It is also evident that a single mutation from GLU to LYS at position 138 in the p51 subdomain of RT does not bring any significant changes in the overall conformation of RT over 100 ns MD trajectories. This observation is in accordant with the earlier crystal structure study, where it was found that the E138K RT has phenomenal similar protein conformations to WT RT, except side-chain of mutant residue (LYS138) moves away from the NNRTI pocket [247]. Although, we see the difference in the case of ETR-E138K, where the structure of ETR-E138K deviates more from the reference structure as compared to its WT counterpart and other RT-NNRTI complexes (**FIGURE 32**). Deviation of backbone atoms of each residue (RMSF) from the reference structure over 100 ns trajectories is shown in **FIGURE 33**. It reconfirms that the single mutation (E138K) doesn't impact the fluctuation of any region of RT in any significant way except that of ETR-E138K, where movement of each residue seems to be amplified. RMSF is also plotted for free WT and E138K in **FIGURE 33B**.



(a)



(b)

Figure 32: RMSD of backbone atoms of HIV RT in Å.

(a) of all the systems (b) individual backbone RMSD of individual MD systems.

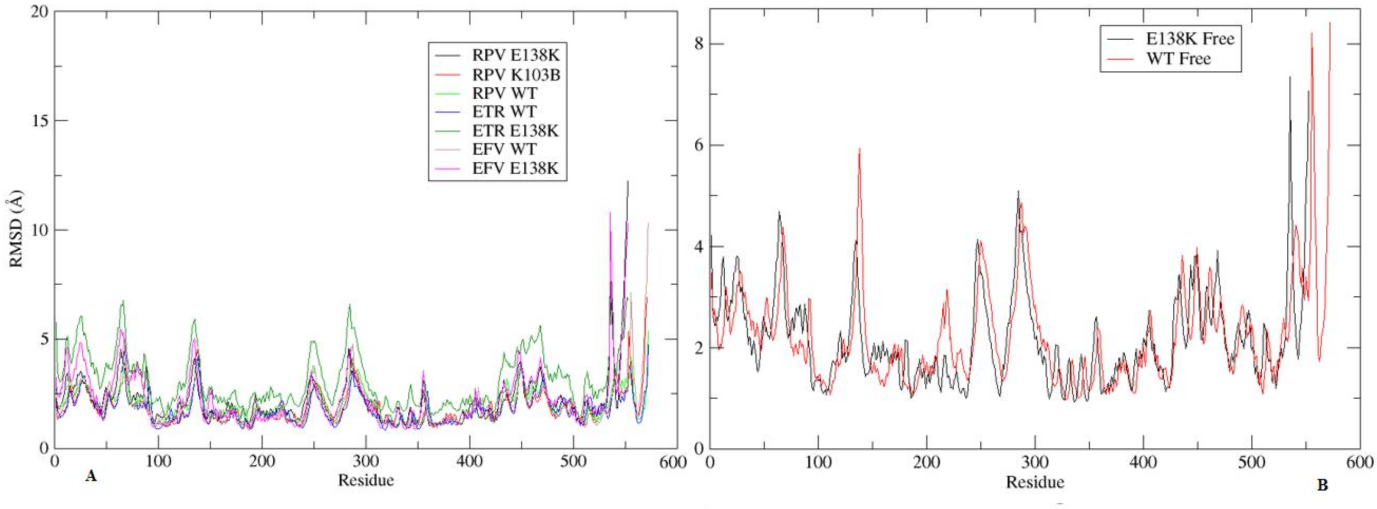


Figure 33: Root mean square deviation per residue (RMSF)

***A**: for all the RT-ligand complexes, **B**: Free HIV RT.*

6.4.2. Distance of GLU/LYS 138 from binding pocket

FIGURE 34 shows the superimposed crystal structure of WT and E138K mutant RT. It is shown that side chain of LYS 138 moves away from the binding pocket resulting in loss of potential contact point for the ligand to bind inside the NNIBP which might lead to decrease or even loss of activity. For further assessment, LYS/GLU 138 sidechain distance was measured from the bound ligand. The plot of the distance between the geometric center of side chain of GLU/LYS 138 and of ligand is shown in FIGURE 35. For all the E138K systems, the side chain moved away from the binding pocket, leading to a reduction in potency. In the case of WT-RT, the distance generally is less than 10 Å, whereas in E138K-RT it is often close to or greater than 15 Å. Although, the movement of the side chain of LYS is more prominent for ETR-RT complex.

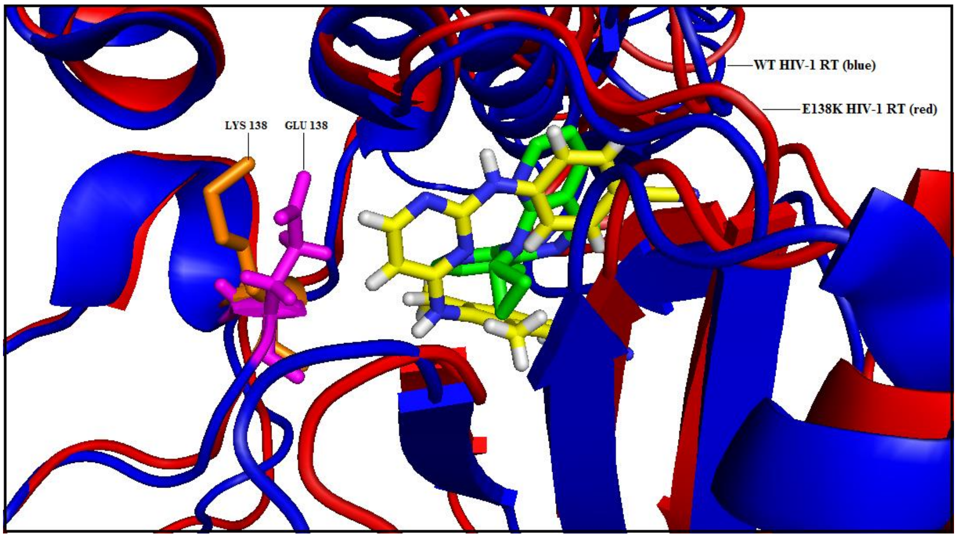


Figure 34: Structural superimposition of WT (blue) and E138K (red) HIV-1 RT along with crystal bound rilpivirine (yellow) and nevirapine (green).

GLU 138 of WT RT and LYS 138 of mutant RT is also shown as stick. Side chains of LYS move away from the binding pocket as compared to the GLU in its WT counterpart.

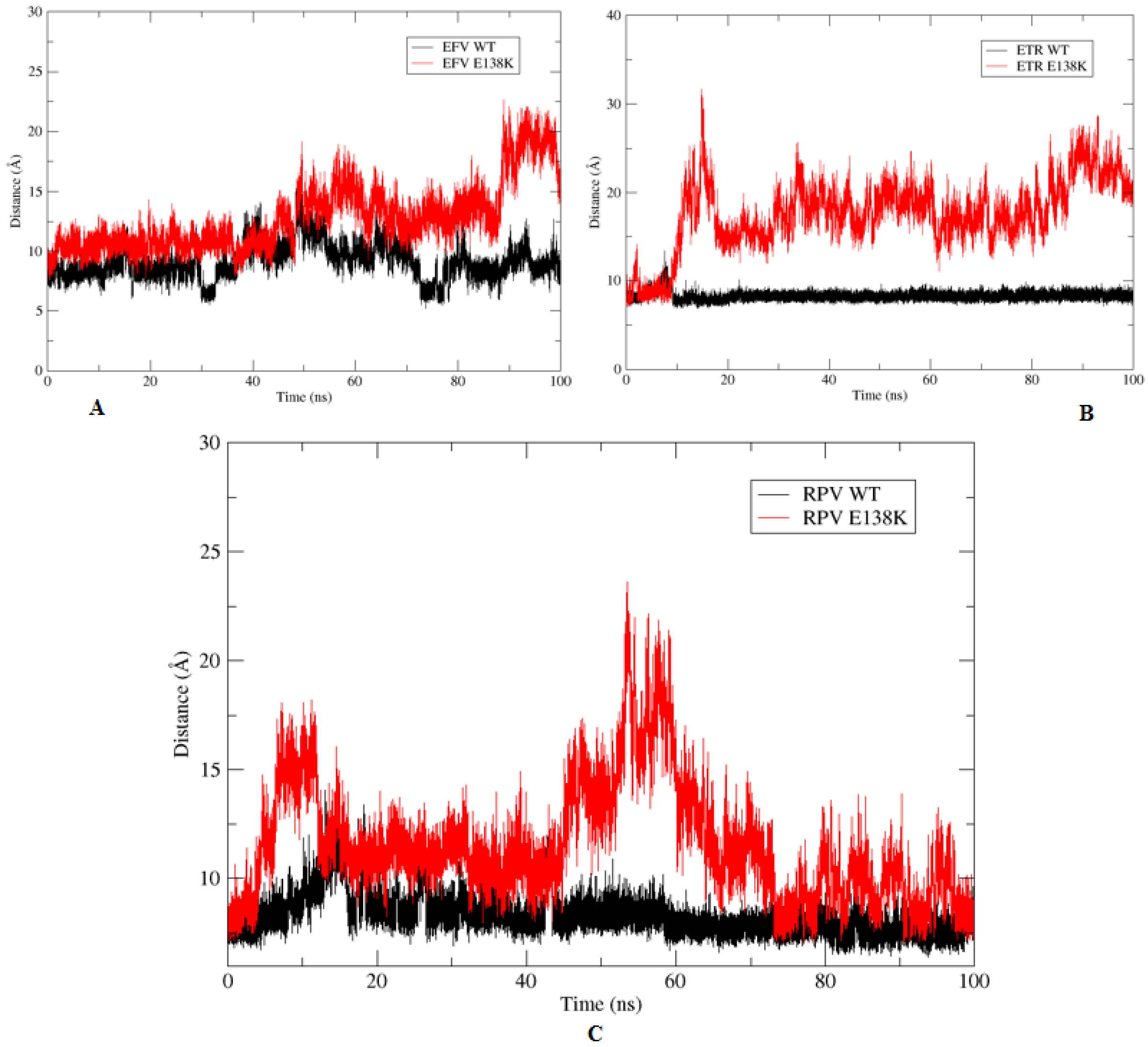


Figure 35: Distance of side chains of GLU 138 (WT-RT) and LYS 138 (E138K-RT) from the geometric center of bound NNRTI.

(A) EFV-RT, (B) ETR-RT, and (C) RPV-RT complex.

6.4.3. Hydrogen bond analysis

Hydrogen bonds formed between the ligand and residues of NNRTI binding pocket is listed in TABLE 10, where H bond observed over 5 % of the MD trajectories are considered. H Bond between ligand-binding site residues formed over only a few frames doesn’t contribute significantly to stabilizing the ligand inside its binding pocket of protein.

Table 10: List of hydrogen bonds formed between the selected ligands and NNIBP residues.

Only H bonds observed more than 5 % of simulation time is enlisted here. Acceptors are electronegative atoms whereas donors are the atoms connected to H atoms forming the bond with the acceptor. Atoms involved in the H bonds are given in parenthesis.

NNRTI-RT complex	Acceptor	Donor	% Time seen
RPV-WT	LYS 101(O)	RPV (N5)	80
	RPV (N3)	LYS 101(N)	10
RPV-E138K	LYS 101 (O)	RPV (N5)	72
	RPV (N3)	LYS 101 (N)	42
EFV-WT	LYS 101 (O)	EFV (N1)	57
	GLU 138 (OE1)	EFV(O2)	7
EFV-E138K	LYS 101 (O)	EFV (N1)	75
	EFV (O2)	LYS 103 (NZ)	5
ETR-WT	LYS 101 (O)	ETR (N2)	87
	GLU 138 (OE1)	ETR (N6)	5
ETR-E138K	ETR (N5)	TYR 188 (OH)	16
	LYS 101 (O)	ETR (N2)	14
	ETR (N5)	TYR 181 (OH)	5

RPV forms two H bonds, one with N5 and another with N3 of LYS 101 of WT and mutant RT as shown in FIGURE 36. EFV forms two H-bonds with WT RT, one with LYS 101 and second with GLU 138, whereas it is unable to make the bond with LYS 138 in the mutant. Same is with ETR, wherein mutant RT LYS 138 is unable to form any H-bond with it (see TABLE 10, FIGURE 37, and FIGURE 38). Disruption in hydrogen bonding network due to mutations seems to have the major contribution into the loss of NNRTIs potency against E138K HIV-1 RT. This phenomenon also causes the reduction in efficacy of nevirapine against other RT mutants such as K101E [247].

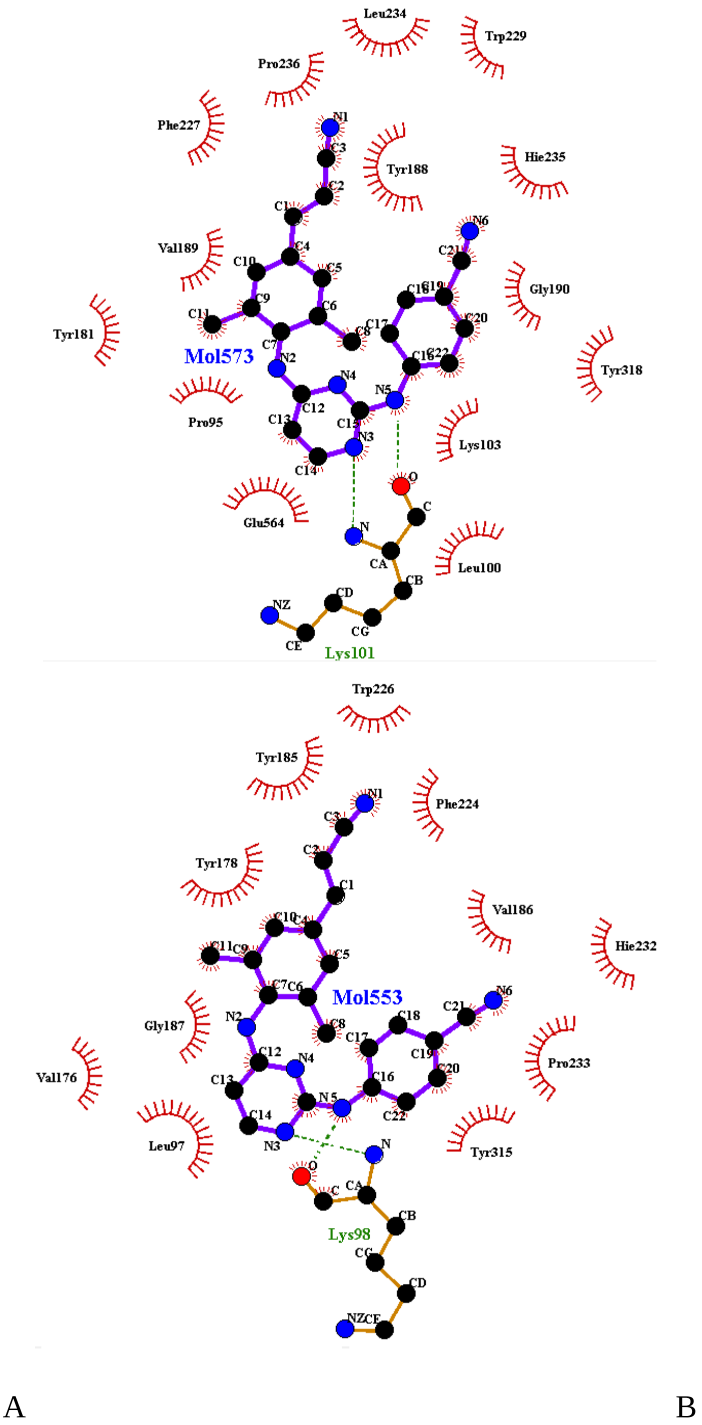


Figure 36: Plots of ligand interactions with the close contact residues of NNRTIs binding pocket

*(**A**) RPV-RT WT-RT (**B**) E138K-R*

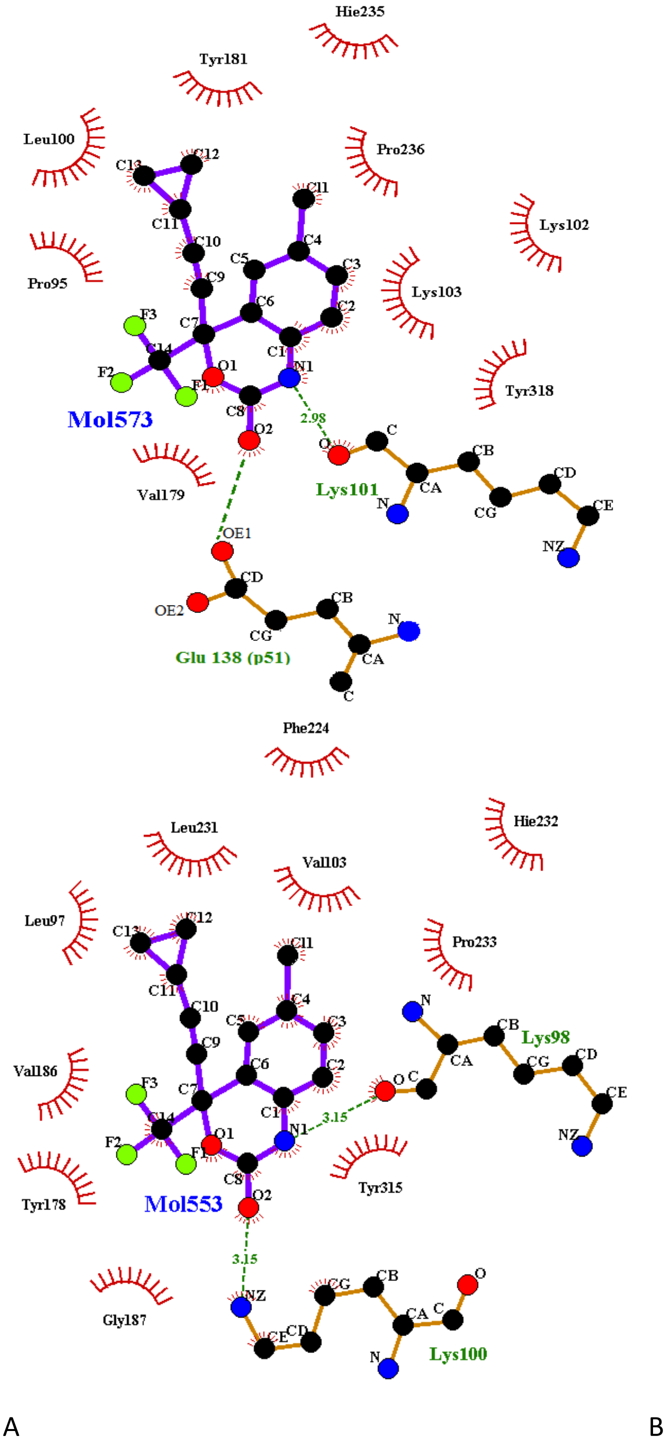


Figure 37: Plots of ligand interactions with the residues of NNRTIs binding pocket

*(**A**) EFV-RT WT RT (**B**): E138K RT*

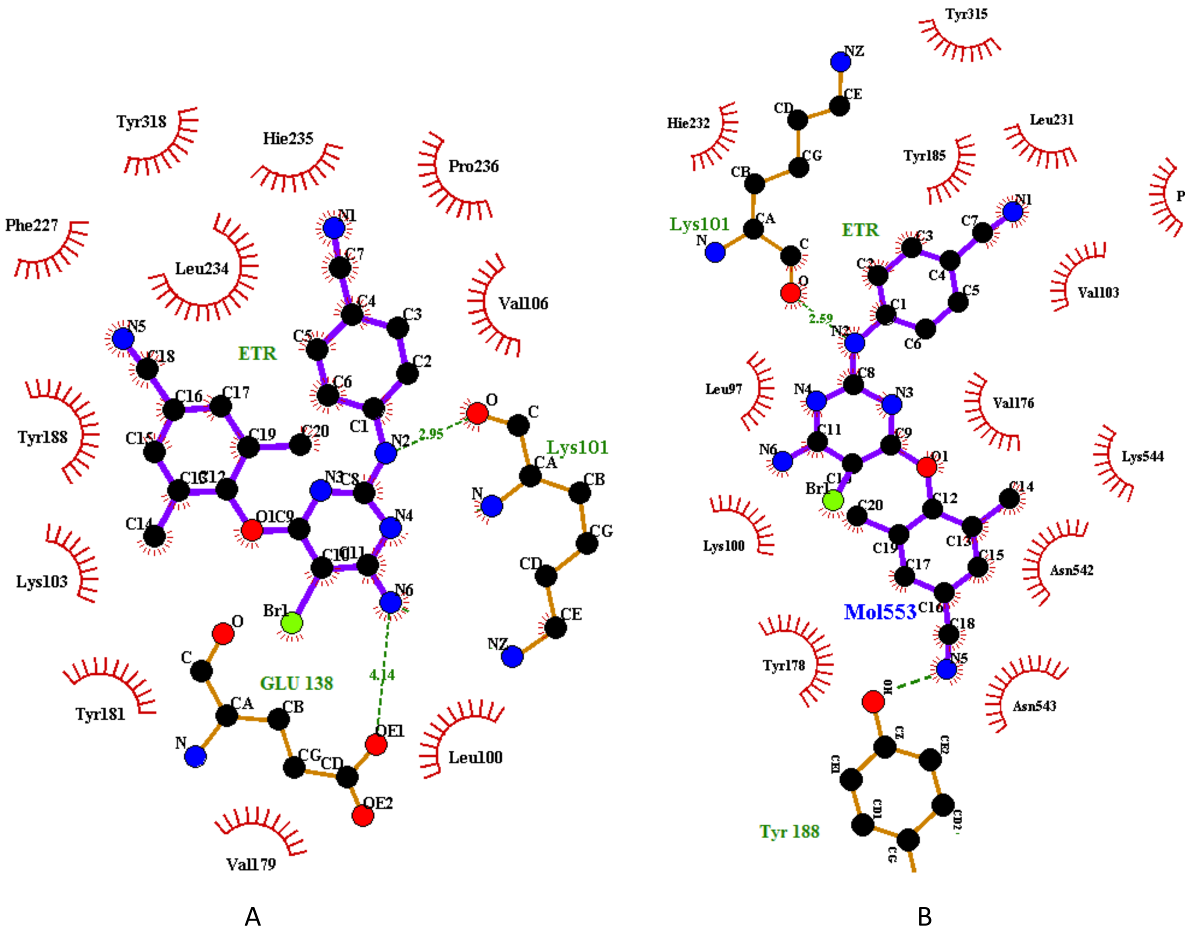


Figure 38: Plot of ligand interaction with the residues of NNRTIs binding pocket

***(A)**: ETR-RT (A) WT RT **(B)** E138K RT*

6.4.4. Distance analysis between thumb and fingers of HIV RT

A double stranded nucleotide is held at the polymerase site by the thumb and fingers of RT. Their dynamics plays a major role in the RT functioning, and the NNRTI binding disrupts it [3]. The opening and closing of the thumb and finger sub-domains can be described by the relative distance between them. The correct placement of 3′ end of RNA/DNA primer in the polymerase cavity is crucial for reverse transcription process. FIGURE 39 shows the time evolution of distance between thumb and finger of RT. The distance between the center of mass (COM) of 24 TRP and 287 LYS was used as an index to measure the distance between thumb and ligand.

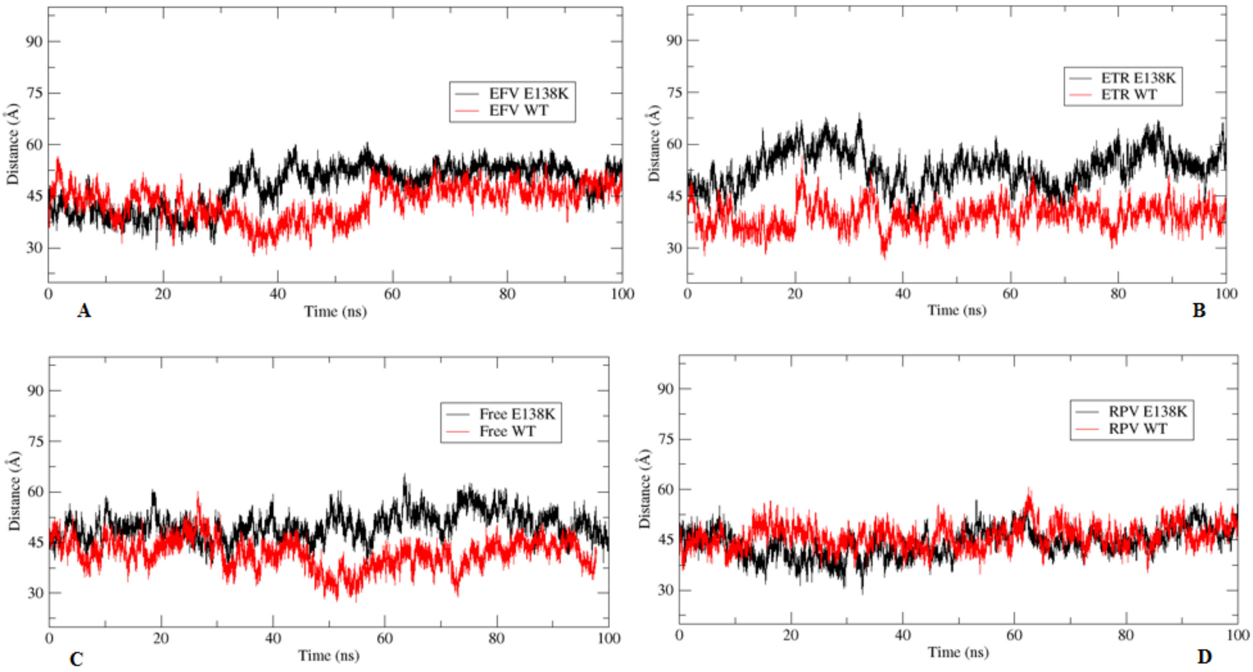


Figure 39: Plots of time evolution of distance between COM of 24 TRP in finger and 287 LYS located in thumb of HIV-1 RT of

(A) EFV-RT, (B) ETR-RT, (C) Free RT, and (D) RPV-RT complexes.

Frequency histograms of the distance between thumb and finger sub-domain of RT over the MD trajectories are demonstrated in FIGURE 40.

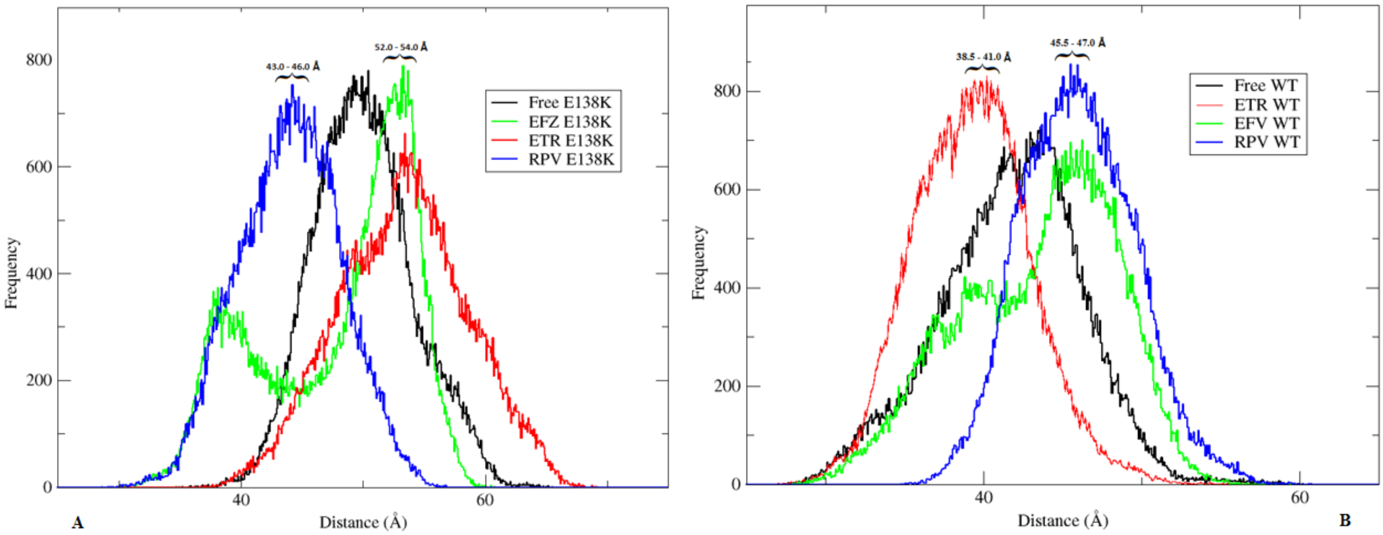


Figure 40: Histogram plots of the distance between thumb and finger sub-domain of HIV-1 RT for

*(**A**) E138K RT and (**B**) WT-RT*

6.4.5. MM-PBSA Binding free energies

The average binding free energies for six 100 ns MD trajectories of HIV RT complex with EFV, ETR and RPV are reported in TABLE 11. It is evident from the calculated binding free energies that mutation at position 138 in p51 sub-domain leads to loss of efficacy of EFV, RPV, and ETR against E138K-RT, thus causing drug resistance. Loss of interactions due to the movement of side chain of LYS 138 away from the NNRTI binding pocket could be the possible reason behind this. This observation could also be inferred from the lower value of electrostatic contribution (ΔE_{Ele}), in the case of E138K-RT complexes.

Table 11: Binding free energies (ΔG_{bind} in kcal/mol).

Free energy and different energy components, such as van der Waals (ΔE_{VDW}), electrostatic energy (ΔE_{Ele}) and solvation energy (ΔE_{Sol} PB) for mutant and WT complexes.

Complex	**ΔE_{VDW}**	**ΔE_{Ele}**	**ΔE_{Sol} (PB)**	**ΔG_{Bind}(PB)**
RPV WT	-60.24	-26.89	48.40	-9.65
RPV E138K	-57.34	-4.83	26.50	-8.07
EFV WT	-40.00	-16.62	29.71	-9.55
EFV E138K	-41.98	-14.10	26.31	-6.40
ETR WT	-62.42	-20.77	43.29	-9.60
ETR E138K	-32.87	-11.21	66.39	-5.59

6.5. Conclusions

Atomistic level simulation studies of enzyme systems are vital for the through a molecular understanding of conformational changes and mechanism of drug resistance. To achieve this insight, a cumulative 900 ns molecular dynamics simulations of apo, RPV, EFV, and ETR bound WT HIV-1 RT and E138K mutated HIV-1 RT in explicit solvent were performed. RMSD and RMSF plots showed that the all the MD systems were converged as well as the single mutation (E138K) don't change the degree of fluctuation in any region of RT in any significant way except that of ETR-E138K. The findings showed that E138K mutation in HIV-1 RT leads to loss of decrease in efficacy of even second generation NNRTIs such as RPV. Nevertheless, RPV adopts different conformations inside the NNIBP in K103N-RT, due to “jiggling” and “wiggling”, indicating structural flexibility, thereby bypassing the drug resistance. Binding free energy calculations showed that the mutation at position 138 in p51 sub-domain leads to loss of efficacy of EFV, RPV, and ETR against E138K-RT. Movement of the side chain of mutant residue away from the NNRTI binding pocket could be the possible reason behind the loss of protein-ligand interactions.

Chapter 7

CONCLUSION

In this thesis, scaffold based QSAR modeling have been used to model the structure-activity relationship and to identify the potential chemical scaffold with anti HIV-1 RT activity. Further, fully atomistic MD simulations have been used to investigate the binding of NNRTI to RT enzymes in HIV-1.

NNRTIs collected from the literature were used to calculate several QSAR models using ASNN algorithms and various molecular descriptors. For further application of different representations of chemical structures, a consensus QSAR model was calculated using selected individual model and analyzed for the scaffold's performance. Some scaffolds had a lower coefficient of determination value. Suggesting that the scaffolds with high Q^2 in the QSAR model have significant structural features correctly learned by the model. Thus, predicting structures of potential compounds based on these scaffolds would be accurate. The linear QSAR model was developed to highlight the structural features affecting anti-HIV activity. MMPA was shown as a powerful method for addressing the 'black box' nature of QSAR, and enable medicinal chemists to choose molecules for further optimization. Significant transformations in the backbone structure were identified using this method. We have shown that the model statistics for predicting new molecules should not be the only approach considered. The scaffold-based analysis is a better approach to identify chemical scaffolds for further optimization. The calculated QSAR model and its sub-models are published on the OCHEM web site http://ochem.eu/article/93085 and are freely accessible for interested users. Their public availability will contribute to the widespread use of the computational chemistry tools on the Web. In this work, we also demonstrated the problem with extrapolation of QSAR models for new chemical series. Despite the failure of original consensus model to appropriately predict new data, the re-calculated model provided good performance for new compounds from the same series due to the change in the applicability domain (AD) of the new model. The AD of re-built model covered the compounds from new chemical series and thus was able to correctly predict their inhibitory activities (IC_{50}).

Further, fully atomistic MD simulations have been used to investigate the binding of NNRTI to RT enzymes in HIV-1. In particular, changes to both structure and thermodynamic properties in the RT caused by mutations associated with drug resistance have been explained. The development of such mutations in response to therapy is well known and represents the main hindrance to treatment success. The primary cause of drug resistance is the lowering of the binding affinity between drug and target protein. While experimental techniques exist that measure this quantity, they cannot provide detailed molecular insight into the causes of resistance.

In CHAPTER 5 MD simulation of Rilpivirine with WT HIV-RT and K103N HIV-RT is presented. Analysis of MD trajectories suggested that rilpivirine takes distinct conformations inside the binding pocket in the WT RT and K103N mutated RT. It takes a horseshoe shape inside the binding pocket of WT RT, whereas a butterfly conformation appeared in the K103N mutated RT. This elucidates the reasonable biological activity profile of rilpivirine against various mutants, including WT HIV-1 RT [240]. MD simulation of rilpivirine bound WT and K103N mutant RT was analyzed in terms of ligand-protein interaction profiles using novel dynophores method, suggesting an alternative interaction for the mutant RT compared to previous reports. The outcome of this analysis suggests hydrogen bonding interaction between the linker N atom of rilpivirine and Lys 101 rather than the previously reported interaction with Asn 103 for the K103N mutant. Further, MD simulation and binding free energy calculations of ligand bound WT and E138K HIV-RT suggested that the mutation at position 138 in p51 sub-domain leads to loss of efficacy of EFV, RPV, and ETR against E138K-RT. Movement of the side chain of mutant residue away from the NNRTI binding pocket could be the possible reason behind the loss of interactions. The method adopted in this thesis is theoretically applicable to any system in which drugs bind to proteins and could be used to investigate the impact of mutations in a wide range of systems.

APPENDIX

Parameters used in MD with amber

Following are the script depicting the paraments used to generate MD trajectory

1. **Minimization of system**

```
Initial minimization of MMP3 (MMMM): solvent molecules and added ions
 &cntrl
  imin   = 1,
  maxcyc = 1000,
  ncyc   = 500,
  ntb    = 1,
  ntr    = 1,
  cut    = 8.0,
 /
Hold the Protein fixed
500.0
RES 1 570
END
END
```

2. **Heating in stages**

```
Heating Step of MMP3 (MMMM): stage-1
 &cntrl
  imin= 0,
  irest=0,
  NTX=1,
  ntb= 1,
  NTPR=500,
  NTWX=500,
  NTWR=500,
  ntr=1,
  Tempi=00.0,
  Temp0=50.0,
  NTT=3,
  gamma_ln=1.0,
  NTC=2,
  NTF=2,
  cut= 8.0,
  nstlim=5000,
  dt=0.002,
/
Keep Protein and inhibitor fixed with weak restraints
10.0
RES 1 571
END
END
```

3. Equilibration and production MD

```
Equilibration Step of MMP3 (MMMM): stage-1
 &cntrl
  imin= 0,
  irest=1,
  NTX=7,
  ntb=2,
  ntp=1,
  PRES0=1.0,
  TAUP=2.0,
  NTPR=500,
  NTWX=500,
  ntr=0,
  Tempi=300.0,
  Temp0=300.0,
  NTT=3,
  gamma_ln=1.0,
  NTC=2,
 NTF=2,
 cut=8.0,
 nstlim=1000000,
 dt=0.002
/
```

I. Quantum Mechanics (QM)

The word “quantum” in QM, it refers to a discrete unit assigned to certain physical quantities. In the quantum realm, particles are discrete packets of energy with wave-like properties. The mathematical formulations of quantum mechanics provided the basic framework of many fields such as computational chemistry, quantum chemistry, solid-state physics, atomic physics, particle physics, molecular physics, computational physics, nuclear chemistry, condensed matter physics, and nuclear physics.

In the quantum mechanics, the state of a system at a time could be defined by a wave function. This allows for the calculation of probabilities of finding an electron in a particular region around the nucleus at a particular time. The region of high probability around the nucleus of an atom often referred to as "clouds", could be drawn where the electron might be located with the most probability. QM methods are based on the solution of the time-dependent Schrödinger wave equation. Like the Newton's equation in classical mechanics, the Schrödinger equation describes the change in wave function with time. This approach is appealing since many molecular properties (3D structure, energies etc.) can be computed directly from the electronic and nuclear structures, which are fundamental physical entities. The Schrödinger wave equation describes the motions of the electrons and nuclei and can be given as:

$$\mathrm{H}\psi_n = \mathrm{E}_n\psi_n \qquad 26$$

where the Hamiltonian operator ($\hat{\mathrm{H}}$) is the summation of the kinetic (E_k) and potential (E_p) energy of the system and can be written as:

$$\mathrm{H}(\mathrm{P},\mathrm{X}) = \mathrm{E}_k(\mathrm{P}) + \mathrm{E}_p(\mathrm{X}) \qquad 27$$

Where P and X denote the joint momentum and position vectors for all the nuclei and electrons in the molecule. The potential energy $\mathrm{E}_p(\mathrm{X})$ originates from electrostatic interactions. Per this, the quantum states E_n (eigenvalues) form a discrete set, corresponding to the Eigen functions ψ_n, for the system of electrons and nuclei. The Schrödinger equation thus defines the spatial probability distributions equivalent to the energy states in a stationary quantum system[73]. A solution of the wave equation for a protein is computationally very expensive. Thus, various semi-empirical and ab initio approximation methods are used. A very common approach employed is the Born-Oppenheimer approximation, which presumes that the nuclei of atoms are stationary with respect to the electrons due to their comparatively large mass. This cuts the problem to the computation of the wave function of the electrons in the field of the fixed nuclei only. This wave function can then be used to compute the forces on the nuclei, whose positions are updated using classical mechanics. The procedure is then repeated with the assumptions that the electrons move instantaneously with the nuclei to continue the evolution of the system. More computationally intensive methods, such as Car-Parrinello, are also available which can calculate the coupled nuclear and electron motions. More details of QM could be accessed in the literature[252].

II. Molecular Mechanics (MM)

Molecular mechanics also known as *force-field* or *potential energy* method employ classical mechanics to model molecular systems. A fundamental principle in molecular mechanics is that collective forces can be used to define molecular geometries and energies. The 3D structure eventually computed is could then be considered stable and at the lowest total internal energy. In MM, a molecule is treated as a collection of masses centered at the atoms linked by bonds (springs). The molecule can stretch, bend, and rotate about those bonds due to the inter and intramolecular forces (**FIGURE 12**). Molecular mechanics can be used to study molecule systems ranging in size and complexity from small to large biological systems or material assemblies with many thousands to millions of atoms.

III. Hybrid approaches (QM/MM)

Although classical MD can investigate many of the phenomena of biological systems with great detail, its inability to handle bond breaking/formation and evaluation of the transition states during reactions is a serious limitation in its part. A full QM calculation is computationally quite intensive, usually impractical for even moderately large systems. This has led to the advances in hybrid QM/MM methodology, where most of the system is treated by MM but a critical part is treated by QM.

IV. Experimental methods for structure determination

There are two main experimental techniques for the structure determination, X-ray crystallography, and nuclear magnetic resonance [253]. In cases, where structural determination by experimental methods are not possible, a blend of the existing related structures and protein sequence can be used to construct a model of the unknown structure. The technique of the elucidation of the 3D structure of the protein of unknown structure is known as homology modeling.

I. X-RAY Crystallography

It is a tool used for ascertaining the 3D atomic and molecular structure of a molecule, in which the crystalline atoms cause a beam of incident X-rays to diffract. By computing the angles and intensities of the diffracted beams, the 3D structure of the density of electrons within the crystal can be determined. From this electron density, the mean positions of the atoms, their chemical bonds, as well as their disorder can be determined. In practice X-rays with wavelengths between 0.4 Å to 1.6 Å are used to image proteins. The construction of high-quality atomic images needs an ordered array of objects to diffract the incident x-rays. Hence, instead of using proteins in a solution, it is necessary to crystallize them first.

Most of the X-rays scattered by a crystal atom destructively interfere but some constructively interfere also and form a diffraction pattern which can then be recorded on a photographic film. Analysis of this pattern using Bragg's law permits the spacing of the diffraction peaks to be related to the spacing of the atoms within the sample.

II. Nuclear Magnetic Resonance Spectroscopy

Nuclear magnetic resonance (NMR) spectroscopy is a technique in which the intrinsic magnetic moment of the atomic nuclei (1H, 13C, 15N or 31P) are used to probe their chemical environment [254]. Large magnetic fields are applied to a sample in order to align the nuclear spins of the atoms. Subsequently, the atoms are exposed to varying radio frequency which excites them into higher energy state. This excited state is transient, and atom comes down to its ground state by losing the extra energy in the form of radio frequency radiation. The specific frequency of the emitted wave is influenced by both the atom type and its environment. This resonant frequency is compared to a reference signal, the shift in the frequency is called the chemical shift and is measured in parts per million (ppm). By varying the frequency of radiation to which the sample is exposed different properties can be probed. In terms of 3D protein structures, the most important types of the technique used are correlation spectroscopy (COSY) and nuclear Overhauser effect (NOE). This gives information on H atoms which are covalently connected through one or two other atoms (i.e. they are very close in the protein sequence) and atoms which are close in space irrelevant of where they occur in the sequence respectively. Merging information from these two methods with knowledge of the protein sequence allows distances between atoms to be computed.

BIBLIOGRAPHY

1. UNAIDS., *Joint United Nations Programme on HIV/AIDS (UNAIDS) (2011) World AIDS Day Report.* Geneva, Switzerland: UNAIDS, 2011.
2. UNAIDS, *GLOBAL AIDS UPDATE*. 2016.
3. Kohlstaedt, L.A., et al., *Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor.* Science, 1992. **256**(5065): p. 1783-90.
4. Steitz, T.A., *DNA polymerases: structural diversity and common mechanisms.* J Biol Chem, 1999. **274**(25): p. 17395-8.
5. Reynolds, C., et al., *In search of a treatment for HIV--current therapies and the role of non-nucleoside reverse transcriptase inhibitors (NNRTIs).* Chem Soc Rev, 2012. **41**(13): p. 4657-70.
6. Tronchet, J.M. and M. Seman, *Nonnucleoside inhibitors of HIV-1 reverse transcriptase: from the biology of reverse transcription to molecular design.* Curr Top Med Chem, 2003. **3**(13): p. 1496-511.
7. Bilal Nizami, I.T., Neil A. Koorbanally, Bahareh Honarparvar, *QSAR models and scaffold-based analysis of non-nucleoside HIV RT inhibitors.* Chemometrics and Intelligent Laboratory Systems, 2015. **148**: p. 134-144.
8. Weiss, R.A., *How does HIV cause AIDS?* Science, 1993. **260**(5112): p. 1273-9.
9. Cunningham, A.L., et al., *Manipulation of dendritic cell function by viruses.* Curr Opin Microbiol, 2010. **13**(4): p. 524-9.
10. Douek, D.C., M. Roederer, and R.A. Koup, *Emerging concepts in the immunopathogenesis of AIDS.* Annu Rev Med, 2009. **60**: p. 471-84.
11. Centers for Disease, C., *Pneumocystis pneumonia--Los Angeles.* MMWR Morb Mortal Wkly Rep, 1981. **30**(21): p. 250-2.
12. Centers for Disease, C., *Kaposi's sarcoma and Pneumocystis pneumonia among homosexual men--New York City and California.* MMWR Morb Mortal Wkly Rep, 1981. **30**(25): p. 305-8.
13. Gallo, R.C., et al., *Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS).* Science, 1983. **220**(4599): p. 865-7.
14. Barre-Sinoussi, F., et al., *Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS).* Science, 1983. **220**(4599): p. 868-71.
15. Coffin, J., et al., *What to call the AIDS virus?* Nature, 1986. **321**(6065): p. 10.
16. Santiago, M.L., et al., *Foci of endemic simian immunodeficiency virus infection in wild-living eastern chimpanzees (Pan troglodytes schweinfurthii).* J Virol, 2003. **77**(13): p. 7545-62.
17. Sharp, P.M., et al., *The origins of acquired immune deficiency syndrome viruses: where and when?* Philos Trans R Soc Lond B Biol Sci, 2001. **356**(1410): p. 867-76.
18. UNAIDS. *AIDS fact sheet*. 2015 [cited 2016 30 May]; Available from: http://www.unaids.org/en/resources/campaigns/HowAIDSchangedeverything/factsheet.
19. Bentwich, Z., et al., *Immune activation in the context of HIV infection.* Clin Exp Immunol, 1998. **111**(1): p. 1-2.
20. Gray, P.B., *HIV and Islam: is HIV prevalence lower among Muslims?* Soc Sci Med, 2004. **58**(9): p. 1751-6.
21. Templeton, D.J., *Male circumcision to reduce sexual transmission of HIV.* Curr Opin HIV AIDS, 2010. **5**(4): p. 344-9.
22. Haase, A.T., *Early events in sexual transmission of HIV and SIV and opportunities for interventions.* Annu Rev Med, 2011. **62**: p. 127-39.
23. Tindall, B. and D.A. Cooper, *Primary HIV infection: host responses and intervention strategies.* AIDS, 1991. **5**(1): p. 1-14.
24. Kahn, J.O. and B.D. Walker, *Acute human immunodeficiency virus type 1 infection.* N Engl J Med, 1998. **339**(1): p. 33-9.

25. Mehandru, S., et al., *Primary HIV-1 infection is associated with preferential depletion of CD4+ T lymphocytes from effector sites in the gastrointestinal tract.* J Exp Med, 2004. **200**(6): p. 761-70.
26. Bennett, J.E., R. Dolin, and M.J. Blaser, *Principles and practice of infectious diseases*. Vol. 1. 2014: Elsevier Health Sciences.
27. McGovern, S.L., et al., *A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening.* J Med Chem, 2002. **45**(8): p. 1712-22.
28. Fields, B.N., D.M. Knipe, and P.M. Howley, *Fields virology*. 2007, Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins.
29. Mahy, B.W.J. and M.H.V. Van Regenmortel, *Encyclopedia of virology*. 2008, Amsterdam; Boston: Academic Press.
30. Coffin, J.M., S.H. Hughes, and H. Varmus, *Retroviruses*. 1997, Plainview, N.Y.: Cold Spring Harbor Laboratory Press.
31. Frankel, A.D. and J.A. Young, *HIV-1: fifteen proteins and an RNA.* Annu Rev Biochem, 1998. **67**: p. 1-25.
32. King, S.R., *HIV: virology and mechanisms of disease.* Ann Emerg Med, 1994. **24**(3): p. 443-9.
33. Chan, D.C. and P.S. Kim, *HIV entry and its inhibition.* Cell, 1998. **93**(5): p. 681-4.
34. Hu, W.S. and S.H. Hughes, *HIV-1 reverse transcription.* Cold Spring Harb Perspect Med, 2012. **2**(10).
35. Zhao, R.Y. and M.I. Bukrinsky, *HIV-1 accessory proteins: VpR.* Methods Mol Biol, 2014. **1087**: p. 125-34.
36. Zheng, Y.H., N. Lovsin, and B.M. Peterlin, *Newly identified host factors modulate HIV replication.* Immunol Lett, 2005. **97**(2): p. 225-34.
37. Pollard, V.W. and M.H. Malim, *The HIV-1 Rev protein.* Annu Rev Microbiol, 1998. **52**: p. 491-532.
38. Butsch, M. and K. Boris-Lawrie, *Destiny of unspliced retroviral RNA: ribosome and/or virion?* J Virol, 2002. **76**(7): p. 3089-94.
39. Hill, M., G. Tachedjian, and J. Mak, *The packaging and maturation of the HIV-1 Pol proteins.* Curr HIV Res, 2005. **3**(1): p. 73-85.
40. Alfano, M. and G. Poli, *The HIV Life Cycle: Multiple Targets for Antiretroviral Agents.* Drug Design Reviews - Online, 2004. **1**(1): p. 83-92.
41. Kohl, N.E., et al., *Active human immunodeficiency virus protease is required for viral infectivity.* Proc Natl Acad Sci U S A, 1988. **85**(13): p. 4686-90.
42. Warnke, D., J. Barreto, and Z. Temesgen, *Antiretroviral Drugs.* The Journal of Clinical Pharmacology, 2007. **47**(12): p. 1570-1579.
43. Cohen, C.J., *Successful HIV treatment: lessons learned.* J Manag Care Pharm, 2006. **12**(7 Suppl B): p. S6-11.
44. De Clercq, E., *The history of antiretrovirals: key discoveries over the past 25 years.* Reviews in Medical Virology, 2009. **19**(5): p. 287-299.
45. *AIDS info*. 2017 [cited 2017 March 2017]; Available from: https://aidsinfo.nih.gov/.
46. Bai, Y., et al., *Covalent fusion inhibitors targeting HIV-1 gp41 deep pocket.* Amino Acids, 2013. **44**(2): p. 701-13.
47. Wensing, A.M., N.M. van Maarseveen, and M. Nijhuis, *Fifteen years of HIV Protease Inhibitors: raising the barrier to resistance.* Antiviral Res, 2010. **85**(1): p. 59-74.
48. *Central dogma reversed.* Nature, 1970. **226**(5252): p. 1198-9.
49. Chattopadhyay, D., et al., *Purification and characterization of heterodimeric human immunodeficiency virus type 1 (HIV-1) reverse transcriptase produced by in vitro processing of p66 with recombinant HIV-1 protease.* J Biol Chem, 1992. **267**(20): p. 14227-32.
50. Castro, H.C., et al., *HIV-1 reverse transcriptase: a therapeutical target in the spotlight.* Curr Med Chem, 2006. **13**(3): p. 313-24.

51. Sarafianos, S.G., et al., *Structure and function of HIV-1 reverse transcriptase: molecular mechanisms of polymerization and inhibition.* J Mol Biol, 2009. **385**(3): p. 693-713.
52. Huang, H., et al., *Structure of a covalently trapped catalytic complex of HIV-1 reverse transcriptase: implications for drug resistance.* Science, 1998. **282**(5394): p. 1669-75.
53. Reardon, J.E., *Human immunodeficiency virus reverse transcriptase: steady-state and pre-steady-state kinetics of nucleotide incorporation.* Biochemistry, 1992. **31**(18): p. 4473-9.
54. Ivetac, A. and J.A. McCammon, *Elucidating the inhibition mechanism of HIV-1 non-nucleoside reverse transcriptase inhibitors through multicopy molecular dynamics simulations.* J Mol Biol, 2009. **388**(3): p. 644-58.
55. De Clercq, E., *Non-nucleoside reverse transcriptase inhibitors (NNRTIs): past, present, and future.* Chem Biodivers, 2004. **1**(1): p. 44-64.
56. Hsiou, Y., et al., *Structure of unliganded HIV-1 reverse transcriptase at 2.7 A resolution: implications of conformational changes for polymerization and inhibition mechanisms.* Structure, 1996. **4**(7): p. 853-60.
57. Tantillo, C., et al., *Locations of anti-AIDS drug binding sites and resistance mutations in the three-dimensional structure of HIV-1 reverse transcriptase. Implications for mechanisms of drug inhibition and resistance.* J Mol Biol, 1994. **243**(3): p. 369-87.
58. Sluis-Cremer, N., N.A. Temiz, and I. Bahar, *Conformational changes in HIV-1 reverse transcriptase induced by nonnucleoside reverse transcriptase inhibitor binding.* Curr HIV Res, 2004. **2**(4): p. 323-32.
59. Keele, B.F., et al., *Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection.* Proc Natl Acad Sci U S A, 2008. **105**(21): p. 7552-7.
60. Das, K. and E. Arnold, *HIV-1 reverse transcriptase and antiviral drug resistance. Part 1.* Curr Opin Virol, 2013. **3**(2): p. 111-8.
61. Wainberg, M.A., et al., *Enhanced fidelity of 3TC-selected mutant HIV-1 reverse transcriptase.* Science, 1996. **271**(5253): p. 1282-5.
62. Boyer, P.L., et al., *Selective excision of AZTMP by drug-resistant human immunodeficiency virus reverse transcriptase.* J Virol, 2001. **75**(10): p. 4832-42.
63. Sarafianos, S.G., et al., *Trapping HIV-1 reverse transcriptase before and after translocation on DNA.* J Biol Chem, 2003. **278**(18): p. 16280-8.
64. Das, K., et al., *Crystal structures of 8-Cl and 9-Cl TIBO complexed with wild-type HIV-1 RT and 8-Cl TIBO complexed with the Tyr181Cys HIV-1 RT drug-resistant mutant.* J Mol Biol, 1996. **264**(5): p. 1085-100.
65. Ren, J., et al., *Structural basis for the resilience of efavirenz (DMP-266) to drug resistance mutations in HIV-1 reverse transcriptase.* Structure, 2000. **8**(10): p. 1089-94.
66. Kuroda, D.G., et al., *Snapshot of the equilibrium dynamics of a drug bound to HIV-1 reverse transcriptase.* Nat Chem, 2013. **5**(3): p. 174-81.
67. Hsiou, Y., et al., *The Lys103Asn mutation of HIV-1 RT: a novel mechanism of drug resistance.* J Mol Biol, 2001. **309**(2): p. 437-45.
68. Rodriguez-Barrios, F., J. Balzarini, and F. Gago, *The molecular basis of resilience to the effect of the Lys103Asn mutation in non-nucleoside HIV-1 reverse transcriptase inhibitors studied by targeted molecular dynamics simulations.* J Am Chem Soc, 2005. **127**(20): p. 7570-8.
69. *HIV drug resistance databse*. 2017 [cited 2017 March]; Available from: https://hivdb.stanford.edu/.
70. Cane, P.A., et al., *Identification of accessory mutations associated with high-level resistance in HIV-1 reverse transcriptase.* AIDS, 2007. **21**(4): p. 447-55.
71. Schuckmann, M.M., et al., *The N348I mutation at the connection subdomain of HIV-1 reverse transcriptase decreases binding to nevirapine.* J Biol Chem, 2010. **285**(49): p. 38700-9.
72. Nikolenko, G.N., K.A. Delviks-Frankenberry, and V.K. Pathak, *A novel molecular mechanism of dual resistance to nucleoside and nonnucleoside reverse transcriptase inhibitors.* J Virol, 2010. **84**(10): p. 5238-49.

73. Schlick, T., *Molecular Modeling and Simulation An Interdisciplinary Guide* 2nd Edition ed. Vol. 21. 2010: Springer.

74. Honarparvar, B., et al., *Integrated Approach to Structure-Based Enzymatic Drug Design: Molecular Modeling, Spectroscopy, and Experimental Bioactivity.* Chemical Reviews, 2014. **114**(1): p. 493-537.

75. Kubinyi, H., *QSAR and 3D QSAR in drug design Part 2: applications and problems.* Drug Discovery Today, 1997. **2**(12): p. 538-546.

76. Ekins, S., et al., *Progress in predicting human ADME parameters in silico.* J Pharmacol Toxicol Methods, 2000. **44**(1): p. 251-72.

77. Perkins, R., et al., *Quantitative structure-activity relationship methods: perspectives on drug discovery and toxicology.* Environ Toxicol Chem, 2003. **22**(8): p. 1666-79.

78. Patel, H.M., et al., *Quantitative structure–activity relationship (QSAR) studies as strategic approach in drug discovery.* Medicinal Chemistry Research, 2014.

79. Nantasenamat, C., et al., *A Practical Overview of Quantitative Structure-Activity Relationship.* Excli Journal, 2009. **8**: p. 74-88.

80. Consonni, V. and R. Todeschini, *Handbook of molecular descriptors*. Methods and principles in medicinal chemistry. 2000, Weinheim ; New York: Wiley-VCH. xxi, 667 p.

81. Thormann, M., et al., *Nomen Est Omen: Quantitative Prediction of Molecular Properties Directly from IUPAC Names.* The Open Applied Informatics Journal, 2007. **1**(1): p. 28-32.

82. Sushko, I., et al., *Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information.* J Comput Aided Mol Des, 2011. **25**(6): p. 533-54.

83. Cherkasov, A., *Inductive QSAR Descriptors. Distinguishing Compounds with Antibacterial Activity by Artificial Neural Networks.* International Journal of Molecular Sciences, 2005. **6**(1): p. 63-86.

84. Varnek, A., et al., *Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures.* J Comput Aided Mol Des, 2005. **19**(9-10): p. 693-703.

85. Varnek, A., et al., *ISIDA - Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors.* Current Computer Aided-Drug Design, 2008. **4**(3): p. 191-198.

86. Bonachera, F., et al., *Fuzzy tricentric pharmacophore fingerprints. 1. Topological fuzzy pharmacophore triplets and adapted molecular similarity scoring schemes.* J Chem Inf Model, 2006. **46**(6): p. 2457-77.

87. Ruggiu, F., et al., *Individual Hydrogen-Bond Strength QSPR Modelling with ISIDA Local Descriptors: a Step Towards Polyfunctional Molecules.* Mol Inform, 2014. **33**(6-7): p. 477-87.

88. Ruggiu, F., et al., *ISIDA Property-Labelled Fragment Descriptors.* Molecular Informatics, 2010. **29**(12): p. 855-868.

89. Skvortsova, M.I., et al., *Chemical graphs and their basis invariants.* Journal of Molecular Structure: THEOCHEM, 1999. **466**(1-3): p. 211-217.

90. Skvortsova, M.I., et al., *A New Technique for Coding Chemical Structures Based on Basis Fragments.* Doklady Chemistry, 2002. **382**(4/6): p. 33-36.

91. Oprea, T.I., *Chemoinformatics in Drug Discovery*. Methods and Principles in Medicinal Chemistry, ed. H.K. Raimund Mannhold, Gerd Folkers. 2005: Wiley-VCH Verlag GmbH & Co. KGaA.

92. Leach, A.R., *Molecular modelling : principles and applications*. 1996, Harlow, England: Longman.

93. Kukol, A., *Molecular modeling of proteins*. 2008, Totowa, NJ: Humana Press.

94. Khan, F.I., et al., *Thermostable chitinase II from Thermomyces lanuginosus SSBP: Cloning, structure prediction and molecular dynamics simulations.* J Theor Biol, 2015. **374**: p. 107-14.

95. Khan, F.I., et al., *Molecular mechanism of Ras-related protein Rab-5A and effect of mutations in the catalytically active phosphate-binding loop.* Journal of Biomolecular Structure and Dynamics, 2016: p. 1-14.

96. Gramany, V., et al., *Cloning, expression, and molecular dynamics simulations of a xylosidase obtained from Thermomyces lanuginosus.* J Biomol Struct Dyn, 2016. **34**(8): p. 1681-92.

97. Khan, F.I., et al., *Molecular mechanism of Ras-related protein Rab-5A and effect of mutations in the catalytically active phosphate-binding loop.* J Biomol Struct Dyn, 2017. **35**(1): p. 105-118.

98. Khan, F.I., et al., *Structure prediction and functional analyses of a thermostable lipase obtained from Shewanella putrefaciens.* J Biomol Struct Dyn, 2016: p. 1-13.

99. Zhang, J., et al., *A comprehensive review on the molecular dynamics simulation of the novel thermal properties of graphene.* RSC Adv., 2015. **5**(109): p. 89415-89426.

100. Khan, F.I., et al., *Current updates on computer aided protein modeling and designing.* International Journal of Biological Macromolecules, 2016. **85**: p. 48-62.

101. Alder, B.J. and T.E. Wainwright, *Phase Transition for a Hard Sphere System.* The Journal of Chemical Physics, 1957. **27**(5): p. 1208.

102. Alder, B.J. and T.E. Wainwright, *Studies in Molecular Dynamics. I. General Method.* The Journal of Chemical Physics, 1959. **31**(2): p. 459.

103. Rahman, A., *Correlations in the Motion of Atoms in Liquid Argon.* Physical Review, 1964. **136**(2A): p. A405-A411.

104. Stillinger, F.H.R., Aneesur. , *Improved simulation of liquid water by molecular dynamics.* The Journal of Chemical Physics, 1974. **60**(4): p. 1545.

105. Karplus, M. and G.A. Petsko, *Molecular dynamics simulations in biology.* Nature, 1990. **347**(6294): p. 631-9.

106. McQuarrie, D.A., *Statistical mechanics*. 2000, Sausalito, Calif.: University Science Books. xii, 641 p.

107. MacKerell, A.D., et al., *CHARMM: The Energy Function and Its Parameterization.* 2002.

108. Oostenbrink, C., et al., *A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6.* J Comput Chem, 2004. **25**(13): p. 1656-76.

109. Jorgensen, W.L., D.S. Maxwell, and J. Tirado-Rives, *Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids.* Journal of the American Chemical Society, 1996. **118**(45): p. 11225-11236.

110. Ponder, J.W. and D.A. Case, *Force fields for protein simulations.* Adv Protein Chem, 2003. **66**: p. 27-85.

111. Wang, J., et al., *Development and testing of a general amber force field.* J Comput Chem, 2004. **25**(9): p. 1157-74.

112. D.A. Case, T.A.D., T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R., et al., *AMBER 12.* University of California, San Francisco., 2012.

113. Van Der Spoel, D., et al., *GROMACS: fast, flexible, and free.* J Comput Chem, 2005. **26**(16): p. 1701-18.

114. Phillips, J.C., et al., *Scalable molecular dynamics with NAMD.* Journal of Computational Chemistry, 2005. **26**(16): p. 1781-1802.

115. Plimpton, S., *Fast Parallel Algorithms for Short-Range Molecular Dynamics.* Journal of Computational Physics, 1995. **117**(1): p. 1-19.

116. Bowers, K.J., et al., *Molecular dynamics---Scalable algorithms for molecular dynamics simulations on commodity clusters.* 2006: p. 84.

117. Trott, O. and A.J. Olson, *AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading.* J Comput Chem, 2010. **31**(2): p. 455-61.

118. Wang, R., L. Lai, and S. Wang, *Further development and validation of empirical scoring functions for structure-based binding affinity prediction.* J Comput Aided Mol Des, 2002. **16**(1): p. 11-26.

119. Velec, H.F.G., H. Gohlke, and G. Klebe, *DrugScoreCSDKnowledge-Based Scoring Function Derived from Small Molecule Crystal Data with Superior Recognition Rate of Near-Native Ligand Poses and Better Affinity Prediction.* Journal of Medicinal Chemistry, 2005. **48**(20): p. 6296-6303.

120. Eldridge, M.D., et al., *Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes.* Journal of Computer-Aided Molecular Design, 1997. **11**(5): p. 425-445.

121. Jones, G., et al., *Development and validation of a genetic algorithm for flexible docking.* Journal of Molecular Biology, 1997. **267**(3): p. 727-748.

122. Rarey, M., et al., *A Fast Flexible Docking Method using an Incremental Construction Algorithm.* Journal of Molecular Biology, 1996. **261**(3): p. 470-489.

123. Krammer, A., et al., *LigScore: a novel scoring function for predicting binding affinities.* Journal of Molecular Graphics and Modelling, 2005. **23**(5): p. 395-407.

124. B�hm, H.-J., *The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure.* Journal of Computer-Aided Molecular Design, 1994. **8**(3): p. 243-256.

125. Lewis, R.M. and V. Torczon, *A Globally Convergent Augmented Lagrangian Pattern Search Algorithm for Optimization with General Constraints and Simple Bounds.* SIAM Journal on Optimization, 2002. **12**(4): p. 1075-1089.

126. Kirkpatrick, S., C.D. Gelatt, and M.P. Vecchi, *Optimization by Simulated Annealing.* Science, 1983. **220**(4598): p. 671-680.

127. Goldberg, D.E., *Genetic algorithms in search, optimization, and machine learning*. 1989, Reading, Mass.: Addison-Wesley Pub. Co. xiii, 412 p.

128. Morris, G.M., et al., *Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function.* Journal of Computational Chemistry, 1998. **19**(14): p. 1639-1662.

129. Sitkoff, D., K.A. Sharp, and B. Honig, *Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models.* The Journal of Physical Chemistry, 1994. **98**(7): p. 1978-1988.

130. Kollman, P.A., et al., *Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models.* Accounts of Chemical Research, 2000. **33**(12): p. 889-897.

131. Dong, F., B. Olsen, and N.A. Baker, *Computational Methods for Biomolecular Electrostatics.* 2008. **84**: p. 843-870.

132. Gilson, M.K. and B.H. Honig, *The dielectric constant of a folded protein.* Biopolymers, 1986. **25**(11): p. 2097-2119.

133. Baker, N.A., *Poisson–Boltzmann Methods for Biomolecular Electrostatics.* 2004. **383**: p. 94-118.

134. Fogolari, F., et al., *Biomolecular Electrostatics with the Linearized Poisson-Boltzmann Equation.* Biophysical Journal, 1999. **76**(1): p. 1-16.

135. *Amber 14 Reference Manual*. 2014.

136. Garg, R., et al., *Comparative Quantitative Structure−Activity Relationship Studies on Anti-HIV Drugs.* Chemical Reviews, 1999. **99**(12): p. 3525-3602.

137. Pungpo, P., et al., *Computer-aided molecular design of highly potent HIV-1 RT inhibitors: 3D QSAR and molecular docking studies of efavirenz derivatives.* SAR and QSAR in Environmental Research, 2006. **17**(4): p. 353-370.

138. Medina-Franco, J.L., et al., *Quantitative Structure–activity Relationship Analysis of Pyridinone HIV-1 Reverse Transcriptase Inhibitors using the k Nearest Neighbor Method and QSAR-*

based Database Mining. Journal of Computer-Aided Molecular Design, 2005. **19**(4): p. 229-242.

139. Carlsson, J., L. Boukharta, and J. Åqvist, *Combining Docking, Molecular Dynamics and the Linear Interaction Energy Method to Predict Binding Modes and Affinities for Non-nucleoside Inhibitors to HIV-1 Reverse Transcriptase.* Journal of Medicinal Chemistry, 2008. **51**(9): p. 2648-2656.

140. Rawal, R.K., V. Murugesan, and S.B. Katti, *Structure-activity relationship studies on clinically relevant HIV-1 NNRTIs.* Curr Med Chem, 2012. **19**(31): p. 5364-80.

141. Prabhakar, Y.S., et al., *CP-MLR/PLS Directed Structure-Activity Modeling of the HIV-1 RT Inhibitory Activity of 2,3-Diaryl-1,3-thiazolidin-4-ones.* QSAR & Combinatorial Science, 2004. **23**(4): p. 234-244.

142. Toropova, A.P., et al., *QSAR models for HEPT derivates as NNRTI inhibitors based on Monte Carlo method.* Eur J Med Chem, 2014. **77**: p. 298-305.

143. Douali, L., et al., *Artificial neural networks: non-linear QSAR studies of HEPT derivatives as HIV-1 reverse transcriptase inhibitors.* Mol Divers, 2004. **8**(1): p. 1-8.

144. Afantitis, A., et al., *A novel simple QSAR model for the prediction of anti-HIV activity using multiple linear regression analysis.* Mol Divers, 2006. **10**(3): p. 405-14.

145. Bazoui, H., et al., *QSAR for anti-HIV activity of HEPT derivatives.* SAR QSAR Environ Res, 2002. **13**(6): p. 567-77.

146. Rodriguez-Barrios, F. and F. Gago, *Understanding the basis of resistance in the irksome Lys103Asn HIV-1 reverse transcriptase mutant through targeted molecular dynamics simulations.* J Am Chem Soc, 2004. **126**(47): p. 15386-7.

147. Garner, J., et al., *A new methodology for the simulation of flexible protein–ligand interactions.* Journal of Molecular Graphics and Modelling, 2007. **26**(1): p. 187-197.

148. Shen, L., et al., *Steered Molecular Dynamics Simulation on the Binding of NNRTI to HIV-1 RT.* Biophysical Journal, 2003. **84**(6): p. 3547-3563.

149. Das, K., et al., *High-resolution structures of HIV-1 reverse transcriptase/TMC278 complexes: strategic flexibility explains potency against resistance mutations.* Proc Natl Acad Sci U S A, 2008. **105**(5): p. 1466-71.

150. Xu, H.T., et al., *Compensation by the E138K mutation in HIV-1 reverse transcriptase for deficits in viral replication capacity and enzyme processivity associated with the M184I/V mutations.* J Virol, 2011. **85**(21): p. 11300-8.

151. Cohen, C.J., et al., *Efficacy and safety of rilpivirine (TMC278) versus efavirenz at 48 weeks in treatment-naive HIV-1-infected patients: pooled results from the phase 3 double-blind randomized ECHO and THRIVE Trials.* J Acquir Immune Defic Syndr, 2012. **60**(1): p. 33-42.

152. Sushko, Y., et al., *Prediction-driven matched molecular pairs to interpret QSARs and aid the molecular optimization process.* J Cheminform, 2014. **6**(1): p. 48.

153. Cumming, J.G., et al., *Chemical predictive modelling to improve compound quality.* Nat Rev Drug Discov, 2013. **12**(12): p. 948-62.

154. Makatini, M.M., et al., *Pentacycloundecane-based inhibitors of wild-type C-South African HIV-protease.* Bioorg Med Chem Lett, 2011. **21**(8): p. 2274-7.

155. Makatini, M.M., et al., *Synthesis, screening and computational investigation of pentacycloundecane-peptoids as potent CSA-HIV PR inhibitors.* European Journal of Medicinal Chemistry, 2012. **57**: p. 459-467.

156. Chen, H., et al., *Design, microwave-assisted synthesis and HIV-RT inhibitory activity of 2-(2,6-dihalophenyl)-3-(4,6-dimethyl-5-(un)substituted-pyrimidin-2-yl)thiazolidin -4-ones.* Bioorg Med Chem, 2009. **17**(11): p. 3980-6.

157. Costi, R., et al., *Structure-activity relationship studies on potential non-nucleoside DABO-like inhibitors of HIV-1 reverse transcriptase.* Antivir Chem Chemother, 2000. **11**(2): p. 117-33.

158. Kertesz, D.J., et al., *Discovery of piperidin-4-yl-aminopyrimidines as HIV-1 reverse transcriptase inhibitors. N-benzyl derivatives with broad potency against resistant mutant viruses.* Bioorg Med Chem Lett, 2010. **20**(14): p. 4215-8.

159. Mai, A., et al., *5-Alkyl-2-alkylamino-6-(2,6-difluorophenylalkyl)-3,4-dihydropyrimidin-4(3H)-ones, a new series of potent, broad-spectrum non-nucleoside reverse transcriptase inhibitors belonging to the DABO family.* Bioorg Med Chem, 2005. **13**(6): p. 2065-77.

160. Mai, A., et al., *Synthesis and anti-HIV-1 activity of thio analogues of dihydroalkoxybenzyloxopyrimidines.* J Med Chem, 1995. **38**(17): p. 3258-63.

161. Mai, A., et al., *5-Alkyl-2-(alkylthio)-6-(2,6-dihalophenylmethyl)-3, 4-dihydropyrimidin-4(3H)-ones: novel potent and selective dihydro-alkoxy-benzyl-oxopyrimidine derivatives.* J Med Chem, 1999. **42**(4): p. 619-27.

162. Mai, A., et al., *Dihydro(alkylthio)(naphthylmethyl)oxopyrimidines: novel non-nucleoside reverse transcriptase inhibitors of the S-DABO series.* J Med Chem, 1997. **40**(10): p. 1447-54.

163. Mai, A., et al., *Structure-based design, synthesis, and biological evaluation of conformationally restricted novel 2-alkylthio-6-[1-(2,6-difluorophenyl)alkyl]-3,4-dihydro-5-alkylpyrimidin-4(3H)-on es as non-nucleoside inhibitors of HIV-1 reverse transcriptase.* J Med Chem, 2001. **44**(16): p. 2544-54.

164. Nugent, R.A., et al., *Pyrimidine thioethers: a novel class of HIV-1 reverse transcriptase inhibitors with activity against BHAP-resistant HIV.* J Med Chem, 1998. **41**(20): p. 3793-803.

165. Cao, D.S., et al., *A new strategy of outlier detection for QSAR/QSPR.* J Comput Chem, 2010. **31**(3): p. 592-602.

166. Qin, H., et al., *Synthesis and biological evaluation of novel 2-arylalkylthio-4-amino-6-benzyl pyrimidines as potent HIV-1 non-nucleoside reverse transcriptase inhibitors.* Bioorg Med Chem Lett, 2010. **20**(9): p. 3003-5.

167. Ragno, R., et al., *Computer-aided design, synthesis, and anti-HIV-1 activity in vitro of 2-alkylamino-6-[1-(2,6-difluorophenyl)alkyl]-3,4-dihydro-5-alkylpyrimidin-4(3H)-o nes as novel potent non-nucleoside reverse transcriptase inhibitors, also active against the Y181C variant.* J Med Chem, 2004. **47**(4): p. 928-34.

168. Rao, A., et al., *2-(2,6-Dihalophenyl)-3-(pyrimidin-2-yl)-1,3-thiazolidin-4-ones as non-nucleoside HIV-1 reverse transcriptase inhibitors.* Antiviral Res, 2004. **63**(2): p. 79-84.

169. Tang, G., et al., *Exploration of piperidine-4-yl-aminopyrimidines as HIV-1 reverse transcriptase inhibitors. N-Phenyl derivatives with broad potency against resistant mutant viruses.* Bioorg Med Chem Lett, 2010. **20**(20): p. 6020-3.

170. Zhang, L., et al., *Synthesis and biological evaluation of novel 2-arylalkylthio-5-iodine-6-substituted-benzyl-pyrimidine-4(3H)-ones as potent HIV-1 non-nucleoside reverse transcriptase inhibitors.* Molecules, 2014. **19**(6): p. 7104-21.

171. Ludovici, D.W., et al., *Evolution of anti-HIV drug candidates. Part 3: Diarylpyrimidine (DAPY) analogues.* Bioorg Med Chem Lett, 2001. **11**(17): p. 2235-9.

172. Tramontano, E. and Y.C. Cheng, *HIV-1 reverse transcriptase inhibition by a dipyridodiazepinone derivative: BI-RG-587.* Biochem Pharmacol, 1992. **43**(6): p. 1371-6.

173. Balzarini, J., et al., *Kinetics of inhibition of human immunodeficiency virus type 1 (HIV-1) reverse transcriptase by the novel HIV-1-specific nucleoside analogue [2',5'-bis-O-(tert-butyldimethylsilyl)-beta-D-ribofuranosyl]-3'-spiro-5 "- (4"-amino-1",2"-oxathiole-2",2"-dioxide)thymine (TSAO-T).* J Biol Chem, 1992. **267**(17): p. 11831-8.

174. Kalliokoski, T., et al., *Comparability of mixed IC(5)(0) data - a statistical analysis.* PLoS One, 2013. **8**(4): p. e61007.

175. Meng, G., et al., *Design and synthesis of a new series of modified CH-diarylpyrimidines as drug-resistant HIV non-nucleoside reverse transcriptase inhibitors.* Eur J Med Chem, 2014. **82**: p. 600-11.

176. Liu, Z., et al., *Design, synthesis and anti-HIV evaluation of novel diarylnicotinamide derivatives (DANAs) targeting the entrance channel of the NNRTI binding pocket through structure-guided molecular hybridization.* Eur J Med Chem, 2014. **87**: p. 52-62.
177. Yang, S., et al., *Molecular design, synthesis and biological evaluation of BP-O-DAPY and O-DAPY derivatives as non-nucleoside HIV-1 reverse transcriptase inhibitors.* Eur J Med Chem, 2013. **65**: p. 134-43.
178. Tetko, I.V., *Associative neural network.* Methods Mol Biol, 2008. **458**: p. 185-202.
179. Tetko, I.V., *Neural network studies. 4. Introduction to associative neural networks.* J Chem Inf Comput Sci, 2002. **42**(3): p. 717-28.
180. Tollenaere, T., *Supersab - Fast Adaptive Back Propagation with Good Scaling Properties.* Neural Networks, 1990. **3**(5): p. 561-573.
181. Matthias Dehmer, K.V., Danail Bonchev, , *Statistical Modelling of Molecular Descriptors in QSAR/QSPR*, ed. F. Emmert-Streib. Wiley-Blackwell. 456.
182. Jens Sadowski , J.G., Gerhard Klebe, *Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures.* journal of chemical information and modeling 1994. **34** (4): p. 701-1028.
183. Whitley, D.C., M.G. Ford, and D.J. Livingstone, *Unsupervised forward selection: a method for eliminating redundant variables.* J Chem Inf Comput Sci, 2000. **40**(5): p. 1160-8.
184. Breiman, L., *Bagging predictors.* Machine Learning, 1996. **24**(2): p. 123-140.
185. Tetko, I.V., et al., *Development of dimethyl sulfoxide solubility models using 163,000 molecules: using a domain applicability metric to select more reliable predictions.* J Chem Inf Model, 2013. **53**(8): p. 1990-2000.
186. Hopkins, A.L., et al., *Complexes of HIV-1 reverse transcriptase with inhibitors of the HEPT series reveal conformational changes relevant to the design of potent non-nucleoside inhibitors.* J Med Chem, 1996. **39**(8): p. 1589-600.
187. Bernstein, F.C., et al., *The Protein Data Bank: a computer-based archival file for macromolecular structures.* Arch Biochem Biophys, 1978. **185**(2): p. 584-91.
188. Frisch, M.J., et al., *Gaussian 09*. 2009, Gaussian, Inc.: Wallingford, CT, USA.
189. Stewart, J.J., *Optimization of parameters for semiempirical methods V: modification of NDDO approximations and application to 70 elements.* J Mol Model, 2007. **13**(12): p. 1173-213.
190. Forli, S., *Raccoon|AutoDock VS: an automated tool for preparing AutoDock virtual screenings",* . (accessed 01/12/2014).
191. Sanner, M.F., *Python: a programming language for software integration and development.* J Mol Graph Model, 1999. **17**(1): p. 57-61.
192. Hussain, J. and C. Rea, *Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets.* J Chem Inf Model, 2010. **50**(3): p. 339-48.
193. Holm, S., *A Simple Sequentially Rejective Multiple Test Procedure.* Scandinavian Journal of Statistics, 1979. **6**(2): p. 65-70.
194. Vorberg, S. and I.V. Tetko, *Modeling the Biodegradability of Chemical Compounds Using the Online CHEmical Modeling Environment (OCHEM) Molecular Informatics Volume 33, Issue 1.* Molecular Informatics, 2014. **33**(1): p. 73-85.
195. Zhu, H., et al., *Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis.* J Chem Inf Model, 2008. **48**(4): p. 766-84.
196. Tetko, I.V., et al., *Can we estimate the accuracy of ADME-Tox predictions?* Drug Discov Today, 2006. **11**(15-16): p. 700-7.
197. Tetko, I.V., et al., *Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection.* J Chem Inf Model, 2008. **48**(9): p. 1733-46.
198. Sushko, I., et al., *Applicability domains for classification problems: Benchmarking of distance to models for Ames mutagenicity set.* J Chem Inf Model, 2010. **50**(12): p. 2094-111.

199. Das, K., et al., *Crystallography and the design of anti-AIDS drugs: conformational flexibility and positional adaptability are important in the design of non-nucleoside HIV-1 reverse transcriptase inhibitors.* Prog Biophys Mol Biol, 2005. **88**(2): p. 209-31.

200. Carta, A., et al., *Activity and molecular modeling of a new small molecule active against NNRTI-resistant HIV-1 mutants.* Eur J Med Chem, 2009. **44**(12): p. 5117-22.

201. Verma, R.P. and C. Hansch, *An approach toward the problem of outliers in QSAR.* Bioorg Med Chem, 2005. **13**(15): p. 4597-621.

202. Kim, K.H., *Outliers in SAR and QSAR: 2. Is a flexible binding site a possible source of outliers?* J Comput Aided Mol Des, 2007. **21**(8): p. 421-35.

203. Kim, K.H., *Outliers in SAR and QSAR: is unusual binding mode a possible source of outliers?* J Comput Aided Mol Des, 2007. **21**(1-3): p. 63-86.

204. Tetko, I.V., *The perspectives of computational chemistry modeling.* J Comput Aided Mol Des, 2012. **26**(1): p. 135-6.

205. Herbst, A.J., et al., *Adult mortality and antiretroviral treatment roll-out in rural KwaZulu-Natal, South Africa.* Bull World Health Organ, 2009. **87**(10): p. 754-62.

206. Bor, J., et al., *Increases in adult life expectancy in rural South Africa: valuing the scale-up of HIV treatment.* Science, 2013. **339**(6122): p. 961-5.

207. Manasa, J., et al., *High-levels of acquired drug resistance in adult patients failing first-line antiretroviral therapy in a rural HIV treatment programme in KwaZulu-Natal, South Africa.* PLoS One, 2013. **8**(8): p. e72152.

208. Bacheler, L.T., et al., *Human immunodeficiency virus type 1 mutations selected in patients failing efavirenz combination therapy.* Antimicrob Agents Chemother, 2000. **44**(9): p. 2475-84.

209. Mobley, D.L. and K.A. Dill, *Binding of small-molecule ligands to proteins: "what you see" is not always "what you get".* Structure, 2009. **17**(4): p. 489-98.

210. Hajduk, P.J., J.R. Huth, and S.W. Fesik, *Druggability indices for protein targets derived from NMR-based screening data.* J Med Chem, 2005. **48**(7): p. 2518-25.

211. Pettit, F.K. and J.U. Bowie, *Protein surface roughness and small molecular binding sites.* J Mol Biol, 1999. **285**(4): p. 1377-82.

212. Kortemme, T. and D. Baker, *A simple physical model for binding energy hot spots in protein-protein complexes.* Proc Natl Acad Sci U S A, 2002. **99**(22): p. 14116-21.

213. Lansdon, E.B., et al., *Crystal structures of HIV-1 reverse transcriptase with etravirine (TMC125) and rilpivirine (TMC278): implications for drug design.* J Med Chem, 2010. **53**(10): p. 4295-9.

214. Olsson, M.H., et al., *PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions.* J Chem Theory Comput, 2011. **7**(2): p. 525-37.

215. Sondergaard, C.R., et al., *Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values.* J Chem Theory Comput, 2011. **7**(7): p. 2284-95.

216. Advanced Chemistry Development, I.T., ON, Canada, *ACD/ChemSketch*. 2013.

217. Cornell, W.D., et al., *Application of Resp Charges to Calculate Conformational Energies, Hydrogen-Bond Energies, and Free-Energies of Solvation.* Journal of the American Chemical Society, 1993. **115**(21): p. 9620-9631.

218. Bayly, C.I., et al., *A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model.* The Journal of Physical Chemistry, 1993. **97**(40): p. 10269-10280.

219. Wang, J., et al., *Automatic atom type and bond type perception in molecular mechanical calculations.* J Mol Graph Model, 2006. **25**(2): p. 247-60.

220. Lindorff-Larsen, K., et al., *Improved side-chain torsion potentials for the Amber ff99SB protein force field.* Proteins, 2010. **78**(8): p. 1950-8.

221. William L Jorgensen, J.C., Jeffry D Madura, Roger W Impey, Michael L Klein, *Comparison of simple potential functions for simulating liquid water.* The Journal of chemical physics, 1983. **79**(2): p. 926-935.

222. Harvey, M.J. and G. De Fabritiis, *An Implementation of the Smooth Particle Mesh Ewald Method on GPU Hardware.* J Chem Theory Comput, 2009. **5**(9): p. 2371-7.

223. Ryckaert, J.-P., G. Ciccotti, and H.J.C. Berendsen, *Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes.* Journal of Computational Physics, 1977. **23**(3): p. 327-341.

224. Gotz, A.W., et al., *Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born.* J Chem Theory Comput, 2012. **8**(5): p. 1542-1555.

225. Roe, D.R. and T.E. Cheatham, 3rd, *PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data.* J Chem Theory Comput, 2013. **9**(7): p. 3084-95.

226. Schmidtke, P., et al., *MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories.* Bioinformatics, 2011. **27**(23): p. 3276-85.

227. Amadei, A., A.B. Linssen, and H.J. Berendsen, *Essential dynamics of proteins.* Proteins, 1993. **17**(4): p. 412-25.

228. Bakan, A., L.M. Meireles, and I. Bahar, *ProDy: protein dynamics inferred from theory and experiments.* Bioinformatics, 2011. **27**(11): p. 1575-7.

229. Humphrey, W., A. Dalke, and K. Schulten, *VMD: visual molecular dynamics.* J Mol Graph, 1996. **14**(1): p. 33-8, 27-8.

230. Wermuth, C.G., et al., *Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998).* Pure Appl. Chem., 1998. **70**: p. 1129-1143.

231. Sydow, D., *Dynophores: Novel Dynamic Pharmacophores - Implementation of Pharmacophore Generation Based on Molecular Dynamics Trajectories and Their Graphical Representation*. 2015, master thesis, Humboldt-Universität zu Berlin, Lebenswissenschaftliche Fakultät.

232. Wolber, G. and R. Kosara, *Pharmacophores from Macromolecular Complexes with LigandScout*, in *Pharmacophores and Pharmacophore Searches*, L. Thierry and R.D. Hoffmann, Editors. 2006, Wiley-VCH. p. 131-50.

233. Wolber, G. and T. Langer, *LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters.* Journal of Chemical Information and Modeling, 2005. **45**: p. 160-169.

234. Bermudez, M., C. Rakers, and G. Wolber, *Structural Characteristics of the Allosteric Binding Site Represent a Key to Subtype Selective Modulators of Muscarinic Acetylcholine Receptors.* Molecular Informatics, 2015. **34**(8): p. 526-530.

235. Das, K., et al., *Crystal structures of clinically relevant Lys103Asn/Tyr181Cys double mutant HIV-1 reverse transcriptase in complexes with ATP and non-nucleoside inhibitor HBY 097.* J Mol Biol, 2007. **365**(1): p. 77-89.

236. !!! INVALID CITATION !!!

237. Vijayan, R.S., E. Arnold, and K. Das, *Molecular dynamics study of HIV-1 RT-DNA-nevirapine complexes explains NNRTI inhibition and resistance by connection mutations.* Proteins, 2014. **82**(5): p. 815-29.

238. Chung, Y.T. and C.I. Huang, *Ion condensation behavior and dynamics of water molecules surrounding the sodium poly(methacrylic acid) chain in water: a molecular dynamics study.* J Chem Phys, 2012. **136**(12): p. 124903.

239. Wright, D.W., et al., *Thumbs down for HIV: domain level rearrangements do occur in the NNRTI-bound HIV-1 reverse transcriptase.* J Am Chem Soc, 2012. **134**(31): p. 12885-8.

240. Janssen, P.A., et al., *In search of a novel anti-HIV drug: multidisciplinary coordination in the discovery of 4-[[4-[[4-[(1E)-2-cyanoethenyl]-2,6-dimethylphenyl]amino]-2-pyrimidinyl]amino]benzonitrile (R278474, rilpivirine).* J Med Chem, 2005. **48**(6): p. 1901-9.

241. Das, K., et al., *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.* J Med Chem, 2004. **47**(10): p. 2550-60.

242. Rhee, S.Y., et al., *Distribution of human immunodeficiency virus type 1 protease and reverse transcriptase mutation patterns in 4,183 persons undergoing genotypic resistance testing.* Antimicrob Agents Chemother, 2004. **48**(8): p. 3122-6.

243. Bacheler, L., et al., *Genotypic correlates of phenotypic resistance to efavirenz in virus isolates from patients failing nonnucleoside reverse transcriptase inhibitor therapy.* J Virol, 2001. **75**(11): p. 4999-5008.

244. Eshleman, S.H., et al., *Phenotypic drug resistance patterns in subtype A HIV-1 clones with nonnucleoside reverse transcriptase resistance mutations.* AIDS Res Hum Retroviruses, 2006. **22**(3): p. 289-93.

245. Rimsky, L., et al., *Genotypic and phenotypic characterization of HIV-1 isolates obtained from patients on rilpivirine therapy experiencing virologic failure in the phase 3 ECHO and THRIVE studies: 48-week analysis.* J Acquir Immune Defic Syndr, 2012. **59**(1): p. 39-46.

246. Nizami, B., et al., *Molecular insight on the binding of NNRTI to K103N mutated HIV-1 RT: Molecular dynamics simulations and dynamic pharmacophore analysis.* Molecular BioSystems, 2016.

247. Ren, J., et al., *Structural insights into mechanisms of non-nucleoside drug resistance for HIV-1 reverse transcriptases mutated at codons 101 or 138.* FEBS J, 2006. **273**(16): p. 3850-60.

248. Gordon, J.C., et al., *H++: a server for estimating pKas and adding missing hydrogens to macromolecules.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W368-71.

249. Case, D., et al., *Amber 14.* 2014.

250. Maier, J.A., et al., *ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB.* J Chem Theory Comput, 2015. **11**(8): p. 3696-713.

251. Nizami, B., et al., *QSAR models and scaffold-based analysis of non-nucleoside HIV RT inhibitors.* Chemometrics and Intelligent Laboratory Systems, 2015. **148**: p. 134-144.

252. Jensen, F., *Introduction to computational chemistry*. 2nd ed. 2007, Chichester, England ; Hoboken, NJ: John Wiley & Sons. xx, 599 p.

253. Brändén, C.-I. and J. Tooze, *Introduction to protein structure*. 1991, New York: Garland Pub.

254. Chary, K.V.R. and G. Govil, *NMR in Biological Systems*. Vol. 6. 2008.