Computational Search for Nature-Derived Dual-Action Inhibitors of HIV-1 Reverse Transcriptase and Integrase: A Potential strategy to mitigate drug resistance progression

Presented in partial fulfillment of the requirements for the Degree of Master of Science in Bioinformatics at Rhodes University

By

Luyando Mwiinga

Department of Biochemistry and Microbiology

Faculty of Science

Rhodes University

2023

Supervisor: Professor Kevin Lobb

ABSTRACT

Human immunodeficiency virus Type 1 (HIV-1) is a devastating viral infection affecting millions worldwide and presents significant challenges in treatment and management. In 2022, approximately 39 million people were living with HIV with Sub-Saharan Africa having two thirds of these infections. Devastatingly, there were approximately 300 000 HIV/AIDS related deaths in Sub-Saharan Africa alone in 2022 alone. Antiretroviral therapy (ART) which is fundamental for HIV treatment, comprises of a combination of drugs such as nucleoside reverse transcriptase inhibitors (NRTIs), non-nucleoside reverse transcriptase inhibitors (NNRTs), protease inhibitors (PIs) and integrase strand transfer inhibitors (INSTIs). However, although 28.7 million people out of the estimated 38.4 million people living with HIV in 2021 were receiving ART, the emergence of drug-resistant strains further complicates treatment efforts, highlighting the need for novel therapeutic approaches. This study aimed to address the challenges raised by drug resistance and significant side effects by identifying potential dual inhibitors against HIV-1 Reverse Transcriptase (RT) and Integrase (IN) using in silico techniques. RT RNase H and IN were chosen as targets for their shared dependency on Mg^{2+} ions within their active sites, which are crucial for catalytic activity. The selection of dual inhibitors was motivated by the fact that the virus would need to replicate at two points simultaneously to develop resistance, making it less likely. The objectives of this study included the creation of a natural derivative compound library using RDKit with the aid of SciFinder, utilizing (-)-epigallocatechin-3-O-gallate (EGCG), because of its dual inhibitory effects against RT and IN, as indicated by a study conducted by Sanna et al. 2019. The natural derivatives were chosen to take advantage of their chemical diversity and to explore potential novel therapeutic options for combating HIV drug resistance. The compound library created comprised of 125 203 compounds. Then docking studies were conducted to assess protein-ligand binding. After the correlation of the RT and IN docking studies, 288 compounds were filtered to have potential dual inhibitory activity. Then quantitative estimation of druggability (QED) analysis identified three compounds with superior properties compared to EGCG and FDA-approved drug raltegravir (RAL). Molecular docking simulations revealed interactions between the inhibitors and the key active site residues of RT and IN, along with the chelation of at least one

Mg^{2+}, suggesting the potential for enzymatic disruption. Furthermore, molecular dynamic (MD) simulations were then conducted to assess protein-ligand system behavior, through RMSD and RMSF analysis. The RMSD analysis uncovered instability in the IN-Sci30703 complex, leading to its exclusion as a potential dual action inhibitor. RMSF analysis for IN showed that all the inhibitors had the ability to limit the flexibility of the catalytic loop which is essential for catalytic activity. Therefore, further in vitro studies are required to evaluate the effectiveness of the remaining two EGCG derivatives (Sci33211 and Sci48919) in inhibiting RT and IN through the chelation of at least one Mg^{2+} ion to determine if they have superior dual inhibitory effects compared to EGCG. This study adds to the ongoing efforts to develop effective strategies against HIV-1 drug resistance and emphasizes the importance of continued research in this field.

DECLARATION

I, Luyando Mwiinga , hereby declare that this thesis submitted to Rhodes University is my original work and has never been submitted to any institution for a degree or diploma.

.....L.Mwiinga..... **.....14/02/2024.....**

Signature **Date**

DEDICATION

I dedicate this thesis to my parents, your support,sacrifices and unwavering belief in my ability has been the foundation of my academic success. Being stubborn can also be a strength :)

I love you

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Professor Kevin Lobb, for his amazing patience and invaluable guidance throughout this study. Your guidance has helped develop my understanding and skills in bioinformatics, and I am really grateful for the opportunity to work under your supervision during my master's programme. I would also like to say thank you to Professor Özlem Tastan Bishop for giving me the opportunity to pursue my masters in bioinformatics. I'd like to express my sincere appreciation to all of the teachers who have helped me learn about bioinformatics, particularly Dr. Olivier. Your patience and commitment to teaching in our coursework has been critical to my development and learning.

To my classmates, I am grateful for the companionship and the support that we shared during our studies. Your encouragement and friendly competition pushed me to excel, and the laughter we shared during breaks made the journey more enjoyable. Special thanks to Curtis, whose willingness to help made the challenges more manageable. I would also like to say thank you to Khaya for your support and being a constant source of encouragement. Your listening ear and feedback on my practice presentations have been invaluable and I really appreciate you. Lastly, I want to extend my heartfelt gratitude to my parents for putting me through all my schooling years, I hope to pay you back ten-fold one day

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

1D	One-dimensional
2D	Two-dimensional
3D	Three-dimensional
3'-P	3'-Processing
β	Beta
γ	Gamma
AIDS	Acquired Immunodeficiency Syndrome
Ala	Alanine
ALOGP	Ghose-Crippen octanol-water partition coefficient
AMBER	Assisted Model Building with Energy Refinement
AROM	Aromatic rings
ART	Antiretroviral therapy
Asp	Aspartic Acid
CADD	Computer-Aided Drug Design
cART	Combination antiretroviral therapy
CAS	Chemical Abstracts Service

CTD	C-terminal domain
Cys	Cystine
D	Aspartic Acid
DFT	Density functional theory
DKAs	Diketoacid analogs
dsDNA	Double-stranded viral DNA
E	Glutamic Acid
EGCG	(-)-epigallocatechin-3-O-gallate
Glu	Glutamic Acid
Gly	Glycine
Gp120	Glycoprotein 120
HAART	Highly Active Antiretroviral Therapy
HBA	Hydrogen bond acceptors
HBD	hydrogen bond donors
HEPTs	1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine
His	Histidine
HIV-1	Human Immunodeficiency Virus Type 1
IN	Integrase

INSTIs	Integrase strand transfer inhibitors
K	Lysine
KCl	Potassium chloride
MD	Molecular Dynamic
Mg^{2+}	Magnesium ions
Mn^{2+}	Manganese ions
Mw	Molecular weight
NNRTIBP	Non-nucleoside reverse transcriptase inhibitor binding pocket
NNRTIs	Non-nucleoside reverse transcriptase inhibitors
NRTIs	Nucleoside/nucleotide reverse transcriptase inhibitors
PBC	Periodic boundary conditions
PDB	Protein Data Bank
PIC	Preintegration complex
PIs	Protease inhibitors
PSA	Polar Surface Area
Q	Glutamine
QED	Quantitative estimation of druggability
QM	Quantum mechanics

R	Arginine
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
RT	Reverse Transcriptase
SAR	Structure-activity relationship
WHO	World Health Organisation
XTB	Extended Tight Binding
Y	Tyrosine

CHAPTER ONE

LITERATURE REVIEW

1.1 INTRODUCTION

Human immunodeficiency Virus (HIV) is a widespread infectious disease that presents a significant global health threat and ranks amongst the leading causes of morbidity and mortality worldwide (Danforth et al., 2017). The virus is primarily spread through direct contact with certain physiological fluids, such as blood, sperm, vaginal fluids, rectal fluids and breast milk (CDC, 2021). HIV targets the immune system, specifically CD^{4+} T cells, therefore compromising the body's defense mechanism and potentially leading to the progression of Acquired Immunodeficiency Syndrome (AIDS). Early detection is critical for effective management, with symptoms ranging from flu-like symptoms to fever, tiredness and swollen lymph nodes during the early stages. As the infection progresses, more severe symptoms may arise (Balasubramaniam, Pandhare, & Dash, 2019). Comprehensive medical care, access to antiretroviral medication (ART) and continuous monitoring are critical for HIV management. However, the emergence of drug resistance emphasizes the significance of careful medication use and the continuous research to address this evolving challenge and ensure long term success of treatment regimens.

1.2 BACKGROUND

1.2.1 HIV Epidemiology in Sub-Saharan Africa

The global impact of HIV/AIDS has been monumental since their emergence in the early 1980s. As of 2022, approximately 39 million people globally were living with HIV, with 1.3 million new infections recorded that year (UNAIDS, 2023). The burden of the epidemic is unevenly spread, with Sub-Saharan Africa having more than two-thirds of new HIV infections, enduring the highest degree of the disease's impact. With 20.6 million people which is approximately 54% of all HIV cases. Although the number of new HIV cases remains high, there has been a 44% reduction from 2010 to 2021. Similarly, the number of AIDS-related deaths has decreased significantly by 58%, as illustrated in Figure 1.1 (UNAIDS, 2022).

The response to HIV/AIDS has been both successful and challenging with significant geographical variation. Sub-Saharan Africa should be a focal point for intervention given its significant impact. To fully comprehend the impact of HIV/AIDS and its prevalence factors such as age, gender and socioeconomic status should be accounted for.

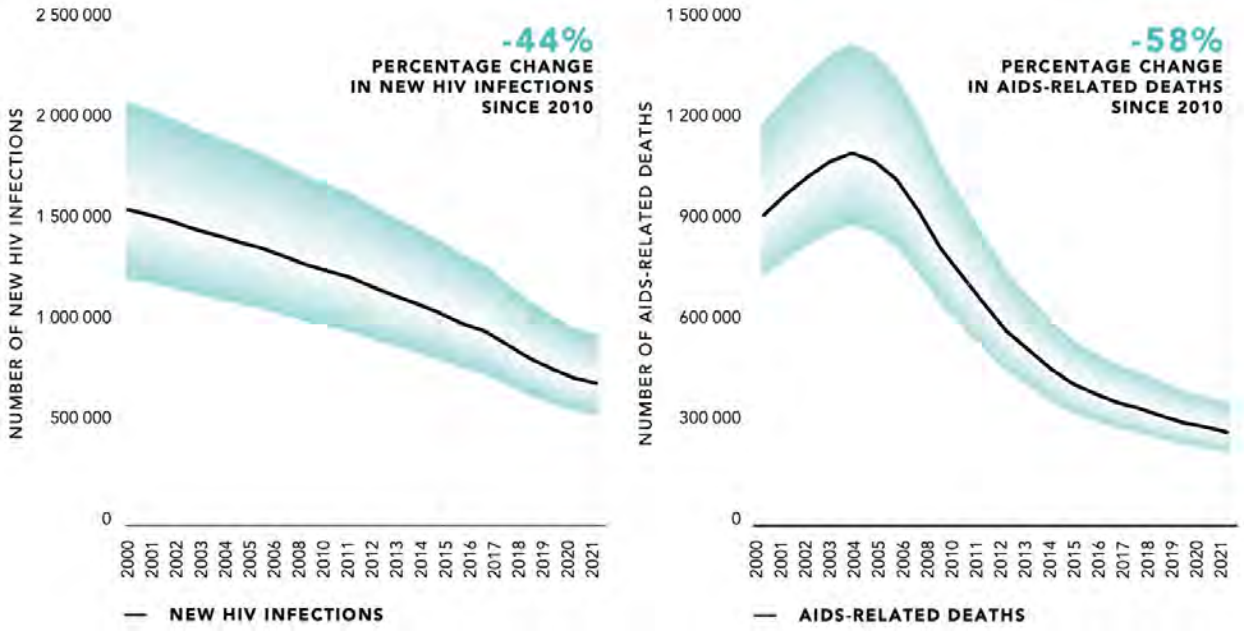

Figure 1.1: The number of new HIV infections and AIDS-related deaths in Sub-Saharan Africa from 2000–2021
(Directly sourced from the UNAIDS data report 2022)

Antiretroviral therapies have evolved, significantly improving the quality of life for individuals living with HIV. As well as preventative programs such as initiatives like Pre Exposure Prophylaxis (PrEP), are making strides in reducing new infections (Pyra et al., 2019). However, persistent challenges slow down progress towards global targets, such as the ambitious UNAIDS 90-90-90 goals, which aim for 90% of people with HIV to be aware of their status, 90% of those diagnosed to receive sustained antiretroviral therapy (ART) and 90% of those on ART to achieve viral suppression (Sohail et al., 2020). The management of HIV necessitates significant financial resources, a necessity worsened by recent occurrences like the COVID-19 pandemic. The reallocation of HIV resources owing to the pandemic has disrupted services, necessitating the implementation of adaptive methods to provide ongoing continuity of HIV/AIDS management

(UNAIDS, 2022). Furthermore, environmental stressors such as drought, flooding and civil instability also have noticeable effects on the dynamics of HIV statistics.

1.2.2 Understanding the HIV Life Cycle

By understanding the HIV life cycle and the important roles played by the enzymes such as reverse transcriptase (RT), integrase (IN) and protease (PR), researchers are able to develop focused therapeutics tailored to disrupt key stages of viral replication (Kirchhoff, 2013). HIV-1 is the most prevalent and virulent type of the virus and relies on these enzymes for its replication. The stages of the HIV-1 life cycle, from initial contact with a target cell to the production of new infectious virions, are outlined below and illustrated in Figure 1.2.

1. Attachment and Entry: HIV envelope glycoprotein (Env) initially attaches to the host cell (CD^{4+} T cells) by the interaction of the virus's glycoprotein 120 (gp120) with the host cell's CD4 receptor and a co-receptor, typically CCR5 or CXCR4 (Kirchhoff, 2013). Following attachment, the viral envelope glycoprotein gp41 undergoes conformational changes which causes the viral envelope to fuse with the host cell membrane. This fusion allows the viral core to enter the host cell and release its genetic material (single stranded viral RNA) (Verma et al., 2020).
2. Reverse Transcription: The viral enzyme RT then produces a complementary DNA (cDNA) strand from the RNA template. Following the breakdown of the initial viral RNA strand, RT creates a second DNA strand that is complementary to the first cDNA strand. This produces double-stranded viral DNA (dsDNA) within the host cell (German Advisory Committee Blood, Subgroup 'Assessment of Pathogens Transmissible by Blood', 2016).
3. Integration: The viral dsDNA is transported to the cell nucleus and integrated into the host's genomic DNA by the enzyme integrase. This is a critical phase in the HIV life cycle because it allows the virus to become part of the host cell, making it difficult for the immune system to eradicate the infection (Engelman and Cherepanov, 2012).
4. Transcription and Translation: The integrated viral DNA uses the host cell's machinery to produce new viral RNA and proteins. The process follows normal cellular transcription and translation processes, creating long chains of HIV proteins (Kirchhoff, 2013).

5. Assembly and Maturation: The newly produced viral RNA and proteins assemble at the cell membrane to form immature, noninfectious virions. The HIV-1 protease enzyme cleaves the long protein chains into individual proteins, which allows the virus to mature into its infectious form (Engelman and Cherepanov, 2012).

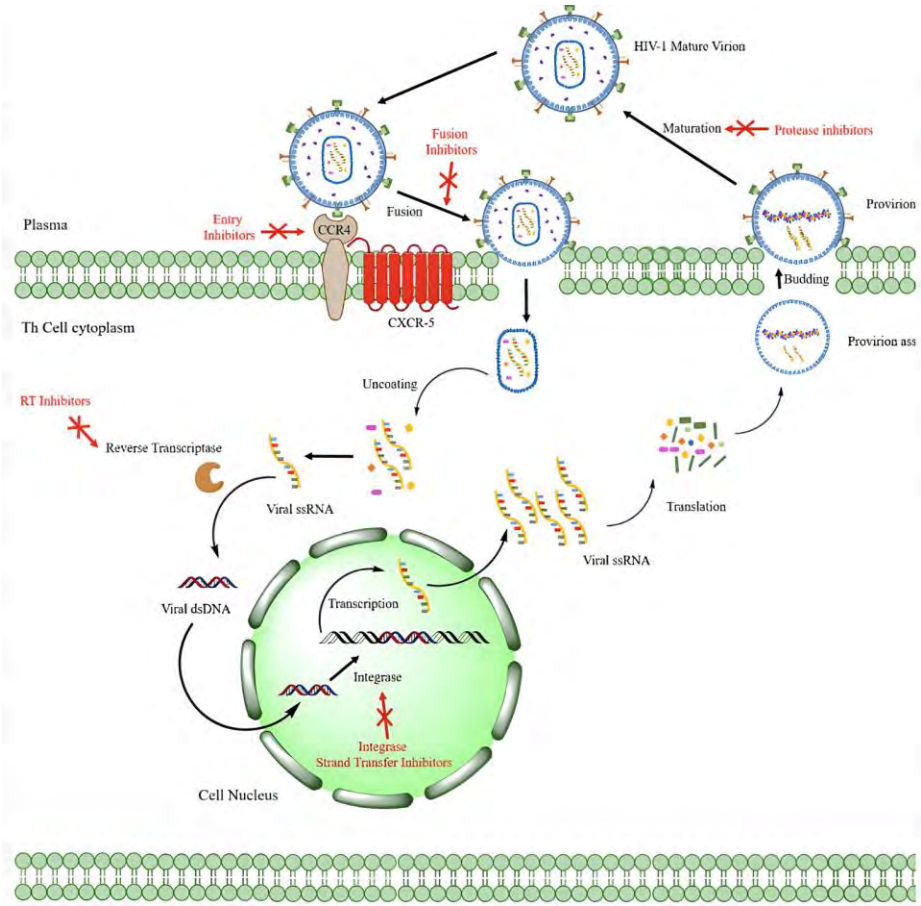

Figure 1.2: HIV life cycle (Mahboubi-rabbani et al., 2021)

Thus, these three enzymes; RT, IN and PR are targets for ART, which aim to suppress viral replication and prevent the progression to AIDS. However, for the purpose of this study, attention will be primarily directed towards RT and IN. The structures of these two enzymes will be explored in greater detail, specifically examining the potential for them to be targets for dual action inhibitors for RT and IN.

1.2.3 Structure of RT and IN

Understanding the structures of HIV-1 RT and IN is critical for the development of dual inhibitors. These enzymes play critical roles in HIV replication, with RT converting viral RNA into DNA and IN integrating the viral DNA into the host genome (Ferguson et al., 2002). Understanding their 3D structures enables the development of inhibitors that can specifically target areas that are functionally important, thereby inhibiting their function and preventing viral replication. The structural knowledge may provide insight into how these enzymes might change under drug pressure. This understanding can guide the development of next-generation drugs.

1.2.3.1 HIV-1 RT

HIV-1 RT is a heterodimer which consists of two subunits, p51 and p66. The p51 subunit consists of 440 residues and the larger p66 subunit consists of 560 residues (Ceccherini Silberstein et al., 2005). Both of these subunits originate from the viral Gag-Pol polyprotein, a process facilitated by the activity of the viral protease through proteolytic cleavage. The p51 subunit primarily provides structural support while the larger subunit, p66, contains the polymerase and RNase H active sites as illustrated in Figure 1.3 (Maga, 2013).

1.2.3.2 Polymerase Active Site

The polymerase domain of the p66 subunit is divided into four subdomains: fingers (residues 1-85 and 118-155), palm (residues 86-117 and 156-236), thumb (residues 237-318) and connection (residues 319-426). These, along with the p66 RNase H subdomain, constitute the nucleic-acid binding cleft, which facilitates the appropriate interaction with the nucleic acid substrate (Mahboubi-Rabbani et al., 2021).

The polymerase active site is found in the palm subdomain of the p66 subunit, containing three catalytic carboxylates, aspartic acids at position 110, 185, 186 (D110, D185, D186). These residues bind to two divalent ions, typically magnesium ions (Mg^{2+}) in vivo or manganese ions (Mn^{2+}) in vitro which are crucial for catalytic activity. The thumb domain contains αH and αI helices which help with the correct positioning of nucleic acids. The DNA primer grip is a highly conserved motif which consists of a β12-β13 hairpin which positions the 3'-OH end of the primer at the polymerase active site (Sarafianos et al., 2009).

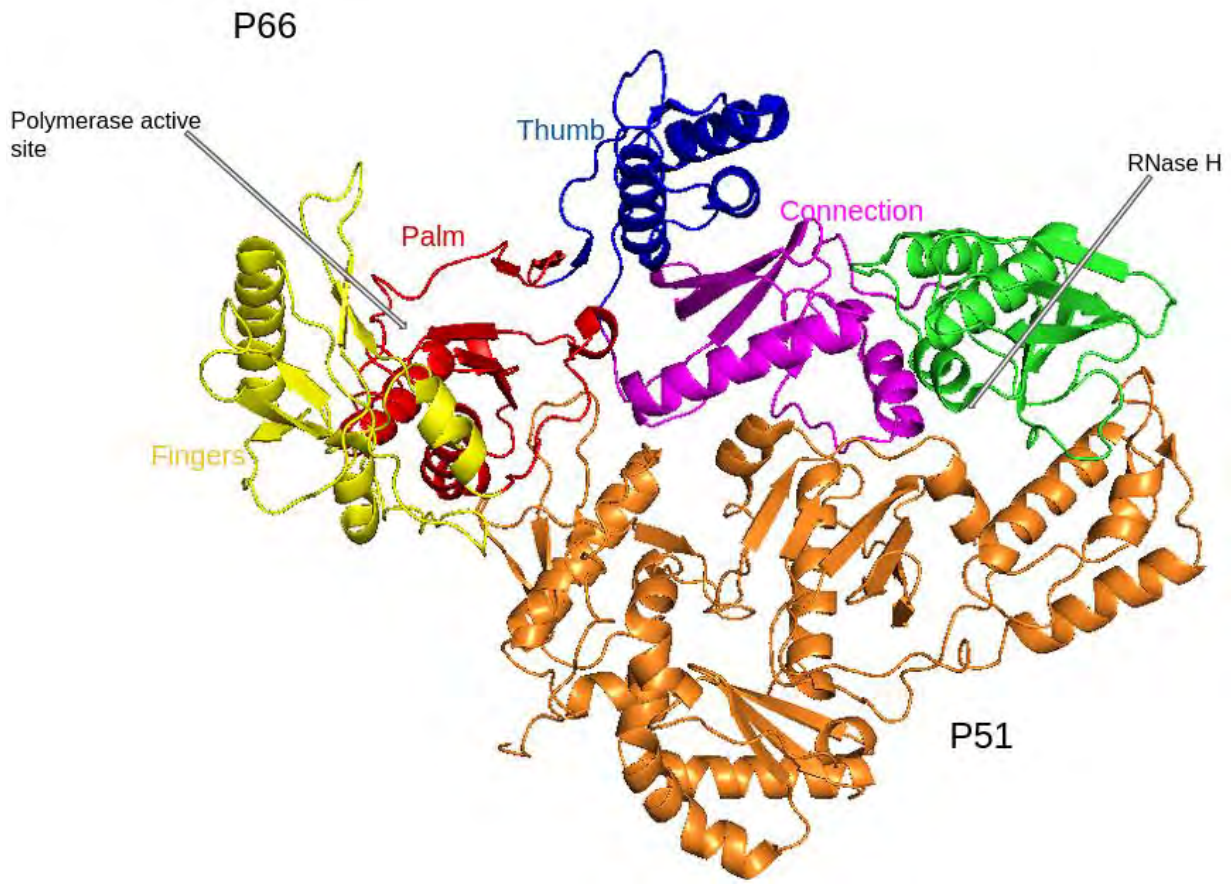

Figure 1.3:Ribbon representation of HIV-1 reverse transcriptase. The fingers, palm, thumb, connection, and RNase H subdomains of the p66 subunit are shown in yellow, red, blue, purple and green, respectively. The p51 subunit is shown in orange. Created in Chimera visualiser

Several key residues are involved in DNA synthesis by HIV-1 RT. These residues include Arginine 72 (R72) and Lysine 65 (K65) play critical roles by binding to the beta (β)- and gamma (γ)-phosphates of the incoming deoxyribonucleotide triphosphate (dNTP). Tyrosine 115 (Y115), contributes to the binding of the deoxyribose ring of the incoming dNTP. Additionally, Y115 acts as a steric gate distinguishing between deoxy and ribonucleoside triphosphates. Another crucial residue is Glutamine 151 (Q151), which directly interacts with the 3'-hydroxyl (3'-OH) of the incoming dNTP. Together, these key residues ensure the efficiency of DNA synthesis (Sarafianos et al., 2009).

Interestingly, an allosteric, hydrophobic pocket known as the non-nucleoside reverse transcriptase inhibitor binding pocket (NNRTIBP) lies near the polymerase's active site. This pocket is approximately 10 Å wide, comprises several residues and includes two additional amino acids from the p51 subunit. All mutations that cause NNRTI resistance occur within this binding pocket (Mahboubi-Rabbani et al., 2021).

1.2.3.3 RNase H Active Site

The RNase H active site is located approximately 17-18 base pairs away from the polymerase active site. It plays a significant role in the hydrolysis of the RNA strand in RNA-DNA hybrids during reverse transcription (Schultz and Champoux, 2008). HIV-1 RNase H active site has two Mg^{2+} which are approximately 4Å apart, as well as highly conserved residues: Aspartic Acid 442 (Asp442), Glutamic Acid 478 (Glu478), Aspartic Acid 498 (Asp498) and Aspartic Acid 549 (Asp549) (Mahboubi-Rabbani et al., 2021).

1.2.3.4 HIV IN

HIV-1 IN is a 32 kDa enzyme that often exists as a dimer. It consists of 288 residues and is stabilized by salt bridges and hydrogen bonds between beta strand 3 and alpha helices 1,3,5 and 6. In its dimeric state, the HIV integrase can carry out the 3'-Processing (3'-P) step, which prepares the viral DNA for integration into the host DNA (Chiu & Davies, 2004). Figure 1.4 is an illustration of the dimeric HIV-1 integrase.

1.2.3.4.1 N-terminal Domain

The N-terminal domain (residues 1-49) contains an HHCC zinc finger motif with four conserved residues Histidine 12 (His12), Histidine (His16), Cystine 40 (Cys40) and Cystine 43 (Cys43), essential for enzyme multimerization. This domain is made up of four alpha helices, with zinc playing a critical role in its correct folding (Cai et al., 1997).

1.2.3.4.2 Catalytic Core Domain (CCD)

The conserved catalytic core domain (residues 50-212) contains the catalytic triad motif D, D, 35E which comprises of the acidic residues Asp^{64} (D), Asp^{116}(D), and Glu^{152}(E) which coordinate with two Mg^{2+} ions which play a key role with the enzymes activity (Hajimahdi & Zarghi, 2016). If any

of these residues are changed, it would result in the loss of the three key functions of IN. These functions include 3' processing, strand transfer, and disintegration, all of which are critical for the viral life cycle (Chiu & Davies, 2004). Two additional residues, K156 and K159, are located near the DDE motif and are involved in viral DNA binding by interacting with the sugar-phosphate backbone of the DNA (Mahboubi-Rabbani et al., 2021).

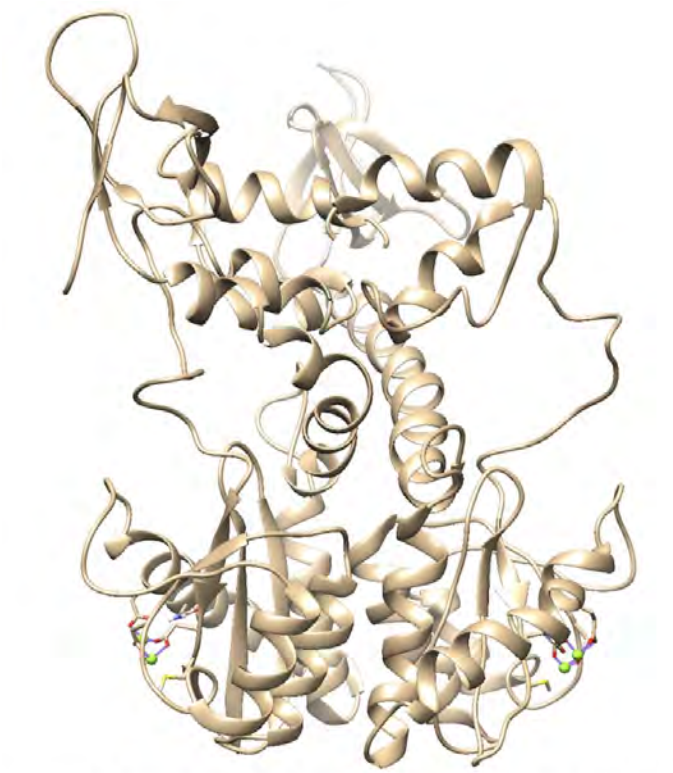

Figure 1.4: Homodimer of HIV-1 Integrase. Mg^{2+} ions (depicted as green balls) coordinated with active site residues ASP64, ASP116, and GLU152.

1.2.3.4.3 C-terminal Domain

The C-terminal domain (CTD) (residues 213-288) makes up a five-stranded beta-barrel structure. The CTD includes an SH3-like fold, which facilitates non-specific DNA binding. Additionally, the CTD has a saddle cleft, which plays a role in interacting and binding with the host cell's double-stranded DNA (Rocchi et al., 2022).

1.2.4 Active site similarities in HIV-1 RT RNase H and IN

The active sites of RT RNase H and IN are structurally similar, with both having five-stranded β sheets which are surrounded by α helices. Both sites include the DDE amino acid triad and require

two divalent metal cofactors (Mg^{2+}) for catalytic activity (Davies et al., 1991). Conserved acidic residues (D443, E478, D498, and D549) in the RNase H site neutralize the Mg^{2+} ions, resulting in an electrostatically neutral pocket. In contrast, the IN active site has only three acidic residues (D64, D116, and E152) that oppose the Mg^{2+} ions, allowing negatively charged carboxylate moieties to bind more favorably (Corona et al., 2014). This understanding of shared features and subtle differences in the active sites forms the basis for the rationale of designing more effective and specific enzyme inhibitors, which is a crucial step in advancing antiviral strategies against HIV.

1.2.5 Current RT and IN HIV-1 treatment

Understanding the modes of action of FDA-approved RT and IN medications provides critical insights into how these drugs interact with the virus. This understanding allows for informed decisions for HIV-1 therapy to effectively combat drug resistance and improve patient outcomes.

1.2.5.1 FDA approved HIV-1 RT and IN inhibitors

RT inhibitors are divided into two distinct subclasses. The first is known as nucleoside/nucleotide reverse transcriptase inhibitors (NRTIs), while the second is referred to as non-nucleoside reverse transcriptase inhibitors (NNRTIs). These classes have distinct modes of action, giving a wide range of antiretroviral therapy options (Patel and Zulfiqar, 2023).

1.2.5.1.1 NRTIs mode of action

NRTIs were the first antiretroviral drugs to be approved by the FDA. Before these drugs can be activated, they need to be taken into the host cell and then phosphorylated. They differ from other compounds because they lack a 3'-hydroxyl group at the 2'-deoxyribosyl moiety and instead have a nucleoside or nucleotide base (Patel and Zulfiqar, 2023). These inhibitors competitively bind to RT which leads to the termination of the formation of the DNA strand because of the lack of 3'-hydroxyl group which prevents the formation of a 3'-5'-phosphodiester bond (Eggleton & Nagalli, 2020). Some of the current FDA approved NRTIs are Zidovudine (Azidothymidine), Didanosine, Emtricitabine, Stavudine, Tenofovir, Zalcitabine and Abacavir (Mahboubi-Rabbani et al., 2021).

1.2.5.1.2 NNRTIs mode of action

The mode of action of NNRTIs is through being a non-competitive inhibitor by binding to RTs hydrophobic allosteric site. This binding creates a conformational change within the substrate-binding site which reduces the polymerase activity (Eggleton & Nagalli, 2020). Some of the FDA approved drugs are Nevirapine, Delavirdine, Doravirine, Efavirenz, Etravirine and Rilpivirine (Mahboubi-Rabbani et al., 2021).

1.2.5.1.3 FDA approved HIV-1 IN inhibitors

Integrase strand transfer inhibitors (INSTIs) target the IN active site, preventing the preintegration complex (PIC) from binding to the host DNA which prevents the integration of the viral DNA into the host's genome (Jóźwik et al., 2020). Some of the current FDA approved INSTIs include; Raltegravir, Elvitegravir and Dolutegravir (Mahboubi-Rabbani et al., 2021).

1.2.6 Pill regimen and drug resistance

Highly Active Antiretroviral Therapy (HAART) also called antiretroviral therapy (ART) or combination antiretroviral therapy (cART), is a combination treatment approach for HIV, typically involving three or more antiretroviral drugs (Eggleton & Nagalli, 2020). HAART is based on combining multiple drugs to limit viral replication through various mechanisms. This approach prevents the virus from thriving despite being resistant to individual drugs.

1.2.6.1 Understanding drug resistance

HAART has some downsides, despite its demonstrated effectiveness in suppressing viral proliferation and disease progression. With the rapid mutation rate of HIV, the virus can develop resistance to antiretroviral drugs. As these drugs apply selective pressure on the virus, mutations that cause resistance can become dominant in the viral population, therefore, making the drug less effective over time (Tang & Shafer, 2012).

However, resistance is not the sole issue, as adverse side effects and long-term toxicity of antiretroviral drugs can also cause substantial problems for people on treatment (Chawla et al., 2018). Some patients experience side effects ranging from mild to severe. Patients experience a

variety of symptoms, including diarrhea, lipodystrophy, cardiovascular problems, mitochondrial toxicity, peripheral neuropathy, osteoporosis just to name a few. Due to the demanding pill regimen and related side effects, many people abandon their medication, resulting in drug resistance and an increased viral load (Lu et al., 2018). Tables 1.1 - 1.3, provide a comprehensive summary of the most common clinically important resistance mutations linked with NRTIs, NNRTIs and INSTIs, respectively.

Table 1.1: Most prevalent clinically important NRTI resistance mutations.

	Discriminatory Mutations					Thymidine Analog Mutations (TAMs)						MDR Mutations	
Consensus	184 M	65 K	70 K	74 L	115 Y	41 M	67 D	70 K	210 L	215 T	219 K	69 T	151 Q
3TC	VI	R										Ins	M
FTC	VI	R										Ins	M
ABC	VI	R	E	VI	F	L			W	FY		Ins	M
TDF	---	R	E		F	L		R	W	FY		Ins	M
AZT	--	--		-		L	N	R	W	FY	QE	Ins	M

Source:(Stanford University, 2023)

Table 1.2: Most prevalent clinically important NNRTI resistance mutations.

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

Source:(Stanford University, 2023)

Table 1.3: Most prevalent clinically important INSTI resistance mutations

Consensus	66 T	92 E	118 G	138 E	140 G	143 Y	147 S	148 Q	155 N	263 R
Bictegravir (BIC)	K	Q	R	KAT	SAC			HRK	H	K
Cabotegravir (CAB)	K	Q	R	KAT	SACR			HRK	H	K
Dolutegravir (DTG)	K	Q	R	KAT	SAC			HRK	H	K
Elvitegravir (EVG)	AIK	Q	R	KAT	SAC		G	HRK	H	K
Raltegravir (RAL)	AIK	Q	R	KAT	SAC	RCH		HRK	H	K

Source:(Stanford University, 2023)

In the context of Tables 1.1 - 1.3, the mutations are categorized based on their impact on resistance susceptibility. Mutations highlighted in bold red indicate the highest levels of reduced viral response, suggesting a significant impact on treatment effectiveness. Mutations in bold also lead to a reduced viral response. Mutations present in plain text contribute to a decreased viral load, particularly when combined with other mutations associated with resistance (Stanford University, 2023).

While the primary focus of this work is not on mutations, it does highlight the ongoing difficulty presented by their emergence, leading to the development of resistance over time. This highlights the urgent need for continued research and development of new medication to effectively counteract evolving HIV-1 strains.

1.2.7 Dual action inhibitors

Given HIV's high mutation rate, drug resistance is a significant concern. Therefore, a strategic approach used to potentially combat this, is the exploration of dual action inhibitors. Dual-action inhibitors target two separate parts of the viral life cycle, reducing the likelihood of acquiring drug resistance because the virus would need to simultaneously replicate in multiple areas, which is less likely to happen (Strasfeld & Chou, 2010). The simplification of treatment regimens, which are made by dual action inhibitors is an added advantage. This benefit encourages better patient compliance as treatment regimens involve fewer medications and require less frequent dosing (Mahboubi-Rabbani et al., 2021). As mentioned in section 1.2.4, the similarities in the active sites of RT RNase H and IN make them suitable targets for dual action inhibitors.

1.2.7.1 Known dual inhibitors of RT and IN

The enzymatic function of both RT and IN is dependent on the presence of Mg^{2+} ions present within their active sites. Inhibitors capable of forming multiple bonds in a process known as chelation, with these ions have the potential to disrupt and inhibit the functions of these enzymes. By exploiting the shared Mg^{2+} ion dependency between these two enzymes, researchers have discovered several classes of RT/IN dual inhibitors (Gill et al., 2019).

HIV-1 RT and IN dual inhibitors can be categorized into various classes based on their chemical structure. These classes include Diketoacid analogs (DKAs), 1-[(2-hydroxyethoxy)methyl]-6-

(phenylthio)thymine (HEPTs) , ,Pyridopyrazine analogs, Quinazoline analogs, Isoquinoline analogs,Chromenone/Coumarin-based analogs, Hybridized analogs and Natural-derived analogs (Mahboubi-Rabbani et al., 2021). From these classes, this study will specifically focus on evaluating compounds sourced from natural origins.

1.2.8 Investigation of natural derived dual action inhibitors targeting RT and IN

Plants have been used for medicinal purposes for thousands of years. Many plant species produce secondary metabolites which have biological activities beneficial to treating human diseases (Barbieri et al., 2017). According to the World Health Organisation (WHO), 60% of the world's population utilizes herbal medicine and approximately 80% of the people in developing countries rely heavily on traditional medicine for their basic medicinal needs. The increasing demand for plant-based medications is a global trend that may be seen in both developing and developed countries. This increase can be due to an increasing understanding that natural products are safe, have few side effects, and are widely available (Ahmad & Ahmad, 2019).

Although the study of medicinal plants remains important in drug discovery, it also has its downsides, such as the acquisition of plant materials, careful selections and use of high-throughput screening assays and the scaling of the active compounds (Balunas and Kinghorn, 2005). Given these obstacles, the development of analogues became imperative. Analogues in respect to existing drugs are compounds that have similarities, either in structure or function. Generally, analogues are categorized into three main groups based on their characteristics and properties. Firstly there are Direct Analogues, which exhibit both chemical and pharmacological similarities to the original compound. This involves slight modifications of the structure but maintains the fundamental framework. Secondly, are Structural analogues, which are designed to closely mimic the chemical structure of a reference compound. Despite their structural similarity, these analogues have clear differences in their pharmacological properties when biological assays are conducted. The last category is Functional Analogues. These compounds do not share structural similarities with the reference molecule, but have similar biological properties (Wermuth, 2006). The categories are important for drug discovery which enable the refinement of existing drugs, the discovery of unexpected therapeutic properties and the design of compounds with similar biological effects.

1.2.8.1 Selection rationale of (-)-epigallocatechin-3-O-gallate

Natural compounds distinguish themselves from synthetic drugs in the drug discovery process due to their wide range of chemical diversity and biochemical specificity (Atanasov et al., 2015). However, isolation of natural medicinal extracts is time consuming and costly (Petersen and Amstutz, 2008). (-)-epigallocatechin-3-O-gallate (EGCG) extracted from the plant *Limonium morisianum*, demonstrated antiviral properties against both HIV-1 RT and IN in a study conducted by Sanna et al.(2019). With IC_{50} values that ranged from from 1.92 to 6.47 μM (Sanna et al., 2019). However, *L.morisianum* which belongs to the *Plumbaginaceae* is endemic to Sardinia (Italy), thus, scalability issues arise. As a result, this study will focus on the computational synthesis of structural analogues of EGCG. This approach allows for the exploration of potential novel compounds beyond their endemic origin,ensuring broader accessibility and capitalizing on the wide chemical diversity that natural compounds offer.

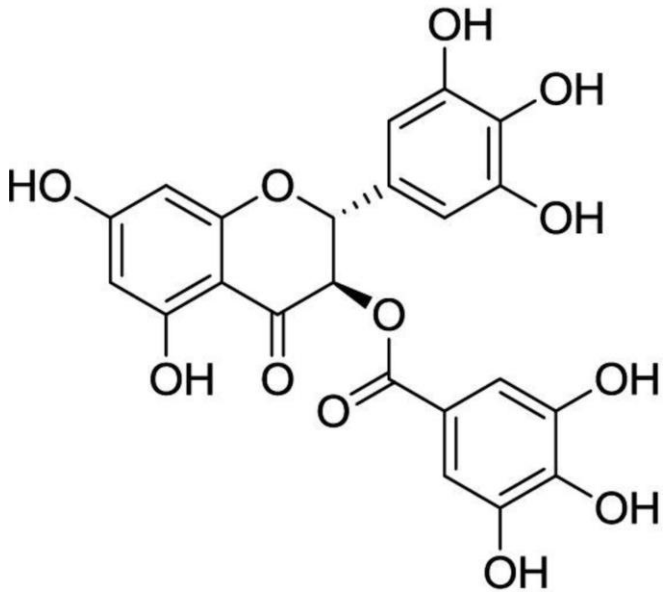

Figure 1.5 : Chemical structure of EGCG
(Mahboubi-Rabbani et al., 2021)

1.2.9 Rationale for selecting HIV-1 Subtype B in this study

HIV-1 subtype C is the most prevalent strain in South Africa. Its dominance in sub-Saharan Africa emphasizes the importance of understanding its molecular characteristics, particularly in relation to drug resistance and treatment response (Giovanetti et al., 2020). Given its epidemiological significance, research focused on HIV-1 subtype C is essential for improving therapeutic options and addressing the challenges of managing HIV in this population.

In this study, I initially aimed to focus on subtype C, however, despite numerous efforts to model the RT and IN enzymes of subtype C, I encountered significant difficulties. The models consistently failed validation , implying that the structural and sequencing data available for subtype C may be incomplete or unreliable compared to those of other subtypes. Due to the time constraints of this project, these challenges were intensified, which limited my ability to explore alternative modeling approaches.

Given these constraints, I made the practical decision to use HIV-1 subtype B RT and IN in my research. Subtype B is the most extensively studied subtype globally, and there is an abundance of validated structures available (Junqueira & Almeida, 2016). This makes subtype B a reliable model system for studying these enzymes. While there are genetic differences between subtypes B and C, many fundamental aspects of RT and IN function are conserved across subtypes (Santerre et al., 2019). Therefore, the insights gained from studying subtype B enzymes can still provide valuable information that may be applicable to subtype C, particularly in the absence of reliable subtype C models.

I acknowledge the limitations of using subtype B in a context where subtype C is more relevant. However, I believe that my current work lays an important foundation for future research. As more reliable data for subtype C becomes available, future studies can build on my findings to explore the unique characteristics of subtype C enzymes in greater detail. Ultimately, my decision to use subtype B was driven by the necessity to advance my research within the given timeframe while still contributing meaningful insights to the broader field of HIV-1 research.

1.3 PROBLEM STATEMENT

HIV is still a serious global public health concern that has claimed over 40 million lives to date and is still spreading across the world. An estimated 1.3 million people had contracted the virus in 2022, with two thirds of the cases observed in Sub-Saharan Africa. The death toll is still high, with approximately 300 000 people dying from HIV/AIDS related diseases in 2022 alone (UNAIDS, 2023).

HIV does not have a cure, which highlights the importance of effective disease management (Sued, Figueroa & Cahn, 2016). While Highly Active Antiretroviral Therapy (HAART) has achieved success in slowing down disease progression,it has some drawbacks which include a high pill burden, which increases drug to drug interactions which therefore results in adverse side effects which leads to reduced patient compliance which eventually leads to drug resistance (Lu et al., 2018). Drug resistance is a significant concern due to the high mutation rate of HIV, which allows the virus to swiftly adapt and develop resistance to antiviral medication (Iyidogan & Anderson, 2014). Therefore, to address the issues related to pill burden, side effects and drug resistance, the potential solution is the exploration of natural derived dual action inhibitors.

There are three main enzymes that are essential to the survival of the virus: RT,IN and PR. RT is responsible for converting viral single stranded RNA into double stranded DNA. IN facilitates the integration of the viral DNA into the host's genome. Lastly, PR matures the virus, making it infectious (Engelman and Cherepanov, 2012). However, in this study the targets that will be considered for dual action inhibitors are RT and IN. The active sites of RT RNase H and IN in these enzymes are potential targets for dual inhibitors because they share a Mg^{2+} ion dependency. Therefore, inhibitors capable of chelating the Mg^{2+} ions can disrupt their function, making them a viable option for dual action inhibitors (Gill et al., 2019).

Lastly, natural compounds are used to synthesize natural derived analogues and exploit natural compounds chemical complexity in the hopes of discovering novel drugs. According to a study conducted by Sanna et al (2019), Epigallocatechin-3-O-gallate (EGCG), had dual inhibitory effects against RT and IN and will therefore serve as the reference compound to create the derivatives and assess them through in silico techniques. The approach to create and assess natural derived dual

action inhibitors is used to mitigate a complex pill regimen, adverse side effects and most importantly, drug resistance.

1.4 AIM OF THE STUDY

The aim of this study is to identify compounds with dual inhibitory activity against RT and IN which surpass the effectiveness of EGCG which serves as our reference compound through in silico techniques. Discovering dual inhibitors for HIV has the potential to provide a solution to the challenges of a high pill burden, significant side effects and most importantly, drug resistance.

1.5 OBJECTIVES

1. Acquisition of inhibitors
    a. Create a compound library of EGCG analogues with the aid of SciFinder
2. Protein acquisition
    a. Retrieval of RT and IN protein structures from PDB database

3. Protein preparation
    a. Addition of two Mg^{2+} ions onto both of the structures
    b. Quantum mechanics to ensure accurate placements of the ions
4. Molecular docking
    a. Employing molecular docking for ligand binding affinity studies
5. Druglikeness assessment
    a. QED and Lipinski's Rule of Five
6. Molecular Dynamic Simulations
    a. MD simulations and trajectory analysis to assess protein-ligand system behavior over time for both RT and IN.

CHAPTER TWO

COMPUTATIONAL COMPOUND LIBRARY CREATION

2.1 INTRODUCTION

The development of diverse and comprehensive compound libraries is a crucial step in the pursuit of novel therapeutic agents, particularly in the context of combating HIV. A computational compound library is a systematically curated collection of chemical structures that are created, organized and analyzed using computational methods. These libraries are created by searching for ligands that share similarities with known active ligands (Sliwoski et al., 2014). This chapter gets into the complex process of developing a compound library with the aid of SciFinder, with emphasis on analogues produced from EGCG, identified for its dual inhibitory effects against HIV-1 RT and IN, which serves as a reference compound for the library (Sanna et al., 2019). This chapter establishes the foundation for future research into potential dual-action inhibitors for HIV-1 RT and IN by the creation of an EGCG derivative compound library.

2.1.1 Utilizing SciFinder in compound library construction

SciFinder was first introduced in 1995 (Ridley, 2002). It is a resource taken from the Chemical Abstracts Service (CAS) which is a curated database that has a focus on chemistry. SciFinder is a valuable tool for acquiring background information on chemicals, drugs and other substances. SciFinder has three distinct search categories: references, substances and reactions. The reference records are obtained from the CAplus and MEDLINE databases. CAplus includes a wide range of content, such as books, journal articles, patents, clinical trials and more. Records retrieved from the SciFinders substances search is sourced from the CAS REGISTRY, which is a collection of over 142 million chemical substances dating back to the early 1800s. It is updated regularly and has important data such as molecular weight, chemical structure and boiling points. Lastly, records retrieved from the reaction search provides extensive information from the CASREACT database which contains millions of single and multistep reactions from 1840 and is updated daily (Gabrielson, 2018).

Additionally, Scifinder has a tool to generate full retrosynthesis plans of known and unknown compounds. Retrosynthesis analysis was developed by E.J Corey in the 1960s. It works backwards by taking the product molecule and finds suitable pathways to create reactants. This helps strategize the synthesis of compounds by proposing potential pathways for compound creation (Coley et al., 2018).

2.1.2 RDKit

RDKit is a cheminformatics toolkit with powerful API's. An illustration of its power is best served by looking at studies where RDKit has been used, and is a straightforward tool due to its wide use and applicability across the field (Bento et al., 2020; Lovrić et al., 2019; Duarte Ramos Matos et al., 2023). For example, RDKit has been used to flag chemical structures and confirm structure in CHEMBL (Bento et al., 2020), or it has been used together with other tools to work on molecules at an extremely large scale (Lovrić et al., 2019), and this synergy is also apparent when looking at de novo design (Duarte Ramos Matos et al., 2023). However there are tools within

RDKit that have been extensively used within this project, and some of these tools include working with SMARTS, Chem and AllChem, to create an extensive compound library (de la Vega de Leon *et al.*, 2017).

2.2 MATERIALS AND METHODS

2.2.1 SciFinder retrosynthesis plan

The SciFinder retrosynthesis planning tool improves chemical synthesis and accelerates the process and utilizes a database with reactions and structures. The retrosynthesis plan was used as the blueprint to create analogues of the nature derived compound EGCG as illustrated by Figure 1. A synthetic depth of three was selected. Common sets of rules supporting predicted reactions were selected. The starting materials cost limit was set to 1000 USD/mol (SciFinder 2023).

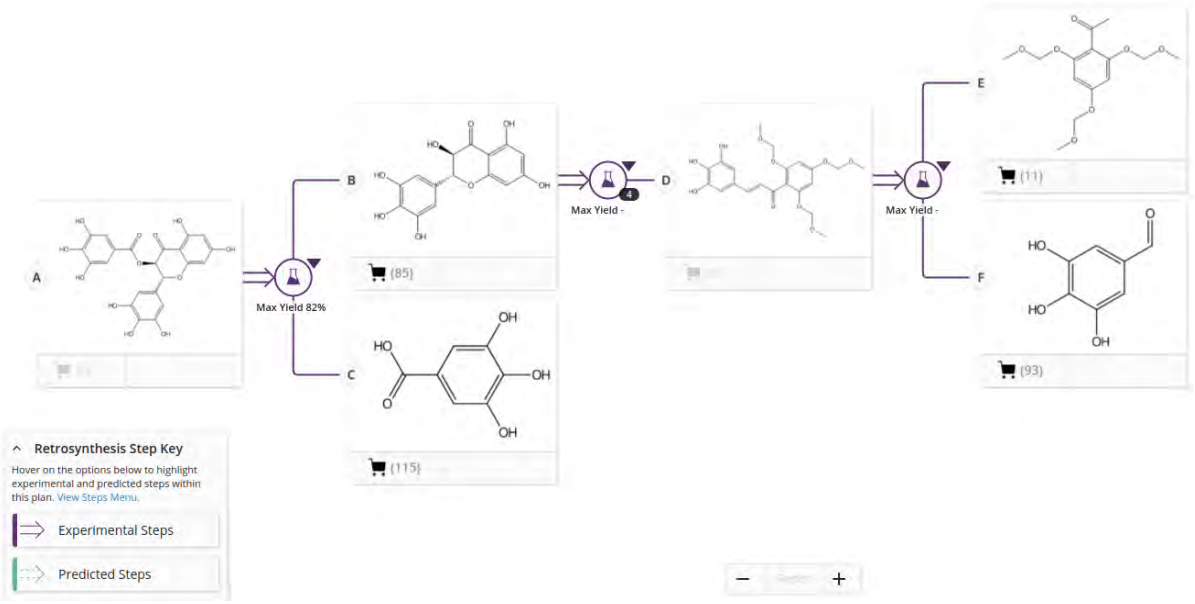

Figure 2.1: Retrosynthesis plan of EGCG. Created using Scifinder. Synthetic depth of three was used, Compound A is EGCG

2.2.2 Analogue Generation of reactants

Analogues for Compound F and C were first searched for using PubChem. In the case of Compound F, all hydroxyl groups were intentionally removed, while for Compound C, one hydroxyl group was retained to ensure no ester groups were formed on the analogues. The purpose of removing the hydroxyl groups was to enhance the binding specificity of the resulting compounds. To reduce the file size for both compounds, the molecular weight was restricted to 180 Da and below. Compound F and Compound C analogues' purpose was to create variation within the library.

2.2.3 Reactions

The RDKit tool was used to implement all of the reactions below:

Reaction E F in Figure 2.2 (A) illustrates Compound E, which has a carbonyl group bound to a carbon atom and an aromatic ring. This compound reacts with Compound F, which has a carbonyl group linked to a carbon atom and an aromatic ring. As a result, a product with two aromatic rings linked by a double bond is created. Then RDKit is used to perform a chemical reaction which uses Compound E and F and creates the product Compound D. The analogues of Compound F and Compound E are used as reactants in the reaction E F to create variations of Compound D using the RDKit's RunReactants function.

Similarly, Reaction D B in Figure 2.2 (B) illustrates the first reactant, which has an oxygen atom bonded to an aromatic ring and a carbon atom with a carbonyl group, undergoes a chemical reaction with the second reactant, which is an oxygen atom. This reaction produces a product in which the oxygen atom in the second reactant forms a new bond with the carbon atom next to the carbonyl group in the first reactant. The final product has a cyclic structure, with the additional oxygen atom inserted into the ring. The products created in the previous step above and an oxygen molecule were used as reactants in the reaction D B using the RDKit's RunReactants function.

Lastly, Reaction C B in Figure 2.2 (C) illustrates the first reactant, which is represented by a carbon atom attached to a carbonyl group and an oxygen atom bonded to a hydrogen atom, reacts chemically with the second reactant, which is represented by an oxygen atom bonded to a carbon atom and a hydrogen atom. This reaction produces a product in which the oxygen atom from the second reactant is incorporated into the structure of the first reactant, forming an oxygen-carbon bond. A hydrogen-oxygen bond was also formed in the product.

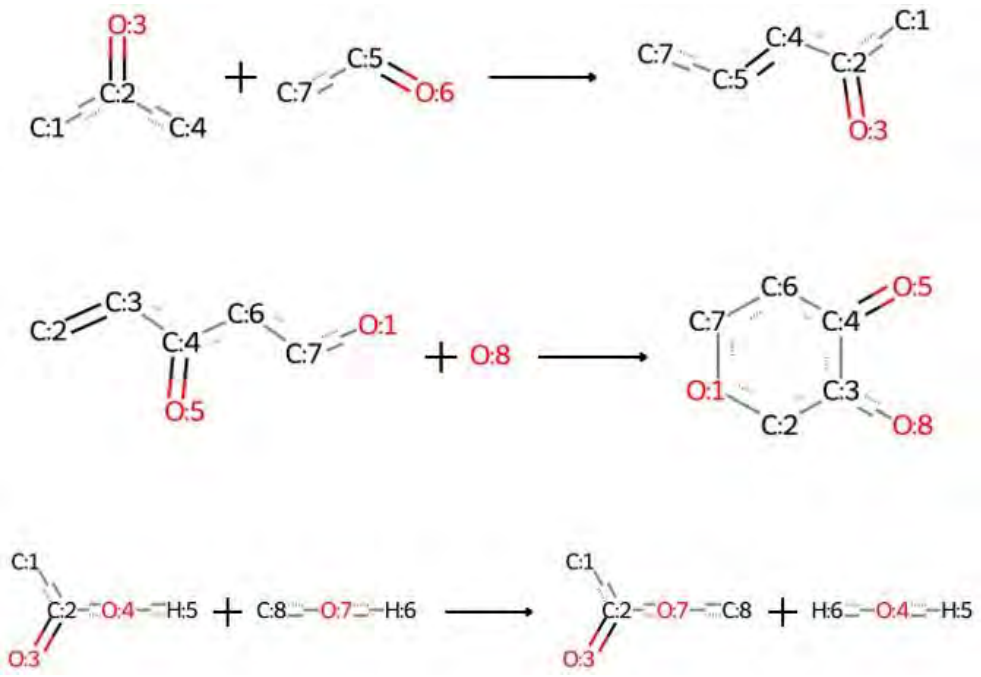

Figure 2.2: Reactions of the retrosynthesis plan. A) Reaction E F, B) Reaction D B and C) Reaction C B

The products created in the previous step above and the analogues of Compound C were used as reactants in the reaction C B using the RDKit's RunReactants function. This resulted in the creation of the complete library.

2.2.4 Rough filtering of compounds

Given the substantial analogue sizes for compound C and F, the complete resulting chemical library would also be large. Therefore, compound screening to assist in narrowing the focus to lead compounds using the ligand efficiencies of known reverse transcriptase compounds from the CHEMBL database was used as a guide therefore, the formulas below were used :

$$ligand\ efficiency\ = \frac{SEI}{BEI} = \frac{TPSA/100}{Mw/1000}$$

$$\text{LI}\ ratio\ = \frac{SEI}{BEI}$$

$$(ratio\ <\ 4)\ and\ (ratio\ >\ 0.25)\ \ and\ (SEI\ <\ 1.05)\ and\ (BEI\ <\ 1.05)$$

After filtering the compounds, the resulting compound library consisted of 125 203 EGCG derivatives.

2.2.5 3D structure conversion

The structures in the library were converted to 3D molecules using part of the AMADAR tool which finds the 3D coordinates and looks for conformations if knotting occurs (Isamura and Lobb, 2022).

2.2.6 Optimisation

All the analogues were optimized using extended tight binding (XTB) within the program xtb (Bannwarth et al., 2021). The method used was the GFN2-xTB method implemented within this program (Bannwarth et al., 2021). Optimization was performed using the command, where filename.xyz was the xyz formatted ligand file.

xtb filename.xyz --opt

2.3 RESULTS AND DISCUSSION

This section presents the results of the compound analogue library generated from the natural compound EGCG. The library consists of 125 203 compounds which were carried forward into molecular docking analysis.

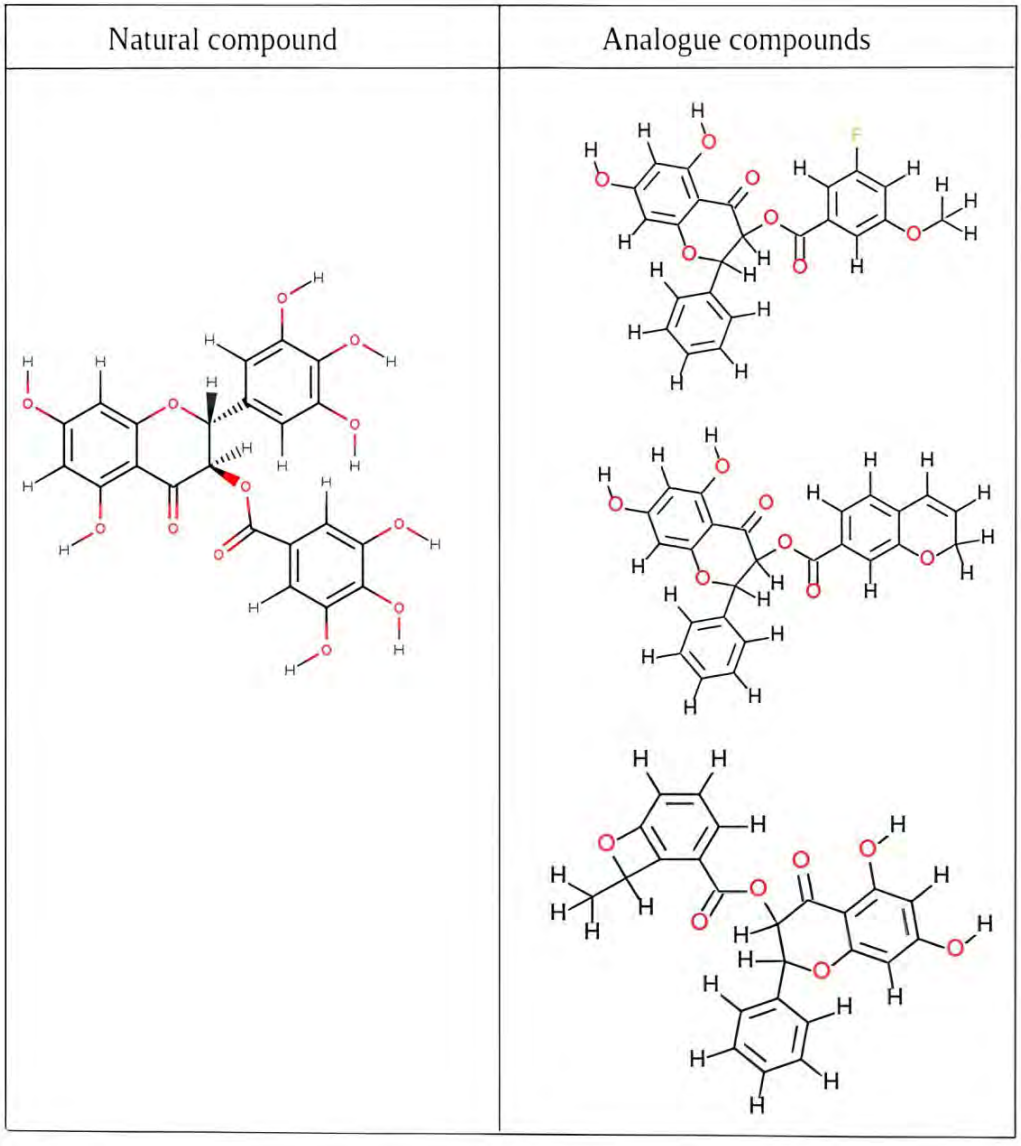

Figure 2.3: Natural compound EGCG and an illustration of the analogues in the library. The depicted analogues represent a subset of the structured analogues in the comprehensive library.

Referring to Figure 2.3, the natural compound used as a template EGCG was used to create analogues. All the analogues observed above have the same core structure, with some shared functional groups such as carboxyl groups on two aromatic rings.

Figure 2.4: Three-Dimensional Representation of Library Compounds. EGCG was used as a template to create this library. 2D structures were converted 3D structures using part of the AMADAR tool

Figure 2.4, is an illustration of the compounds in 3D. When studying chemical compounds, 1D (one-dimensional) and 2D (two-dimensional) approaches are frequently used because comparison algorithms can easily be created. However, because atomic interactions take place in the 3D space, 1D and 2D molecules do not take into consideration the binding affinity between molecules and

target proteins (Shin et al., 2015). Thus, the conversion from 2D to 3D is deemed necessary, especially for downstream analysis.

Additionally ligand optimisation is important to ensure accuracy of each atom. Quantum mechanics (QM) is the mathematical framework to accurately describe the behavior of electrons. Hence, It has the capability to precisely predict the properties of individual atoms or molecules (Yüksel, 2017).

Furthermore, the electronic structure of atoms, molecules and solids can be calculated using the density functional theory (DFT), which is a quantum-mechanical (QM) technique. DFT allows for the prediction of electronic structure theory of larger molecules with accuracy. As a result, DFT is the most used electronic structure method (Van Mourik et al., 2014).

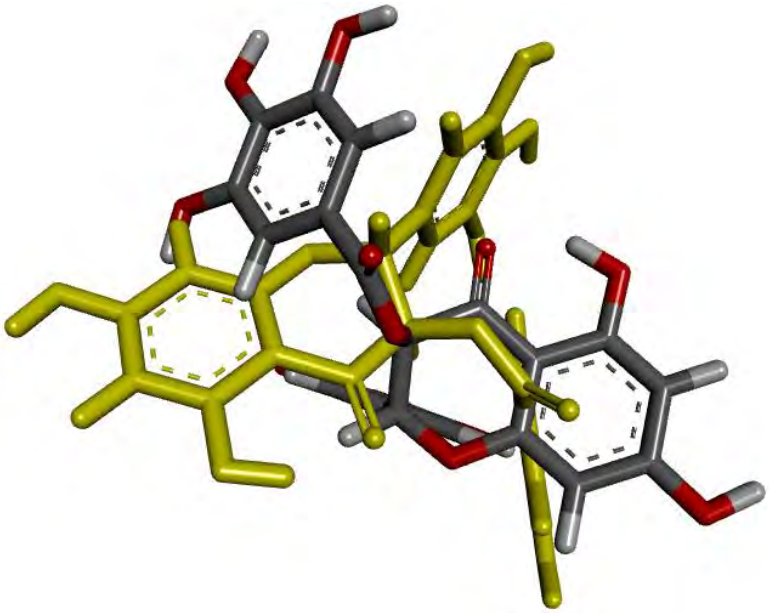

Figure 2.5: Superimposition of Density Function and XTB Formatted Compounds of EGCG. The yellow compound is the density function format, the grey is the xtb formated compound. Shows a clear difference in orientation between the optimised structures

However, due to the time constraint of this project, using DFT was not a viable option due its computational expense, therefore Extended Tight Binding (XTB) was used to optimize the ligands. The XTB program is a semi-empirical extended tight-binding package that predicts

molecule structures and characteristics accurately (Moore et al., 2022). However, XTB has its shortcomings as seen in Figure 2.5, the XTB ligand is in a less favorable conformation in comparison to DFT ligand. To mitigate this constraint a conformational search would need to be conducted. However, as stated above there is a time constraint. Due to the large size of the library it is possible that some of the ligands would be in the correct conformation. Therefore XTB is used without a conformational search. However, it is important to note that this assumption could be detrimental to the focused library after high throughput screening is conducted. If the screened compounds used in downstream analysis are in the wrong conformation, it could lead to inaccurate results.

2.4 CHAPTER SUMMARY

This chapter explores the process of constructing a comprehensive EGCG derivative compound library with the aid of SciFinder using RDKit. This compound library consisted of 125 203 compounds with similar core structures and functional groups. XTB was used for compound optimisation to enhance their accuracy. However, an assessment of conformational accuracy was conducted using DFT revealed that it was superior in comparison to XTB, but due to the time constraints of this project and the large compound library XTB was deemed to be acceptable. Although, caution was exercised in carrying forward with XTB optimisation and acknowledging potential limitations. The compounds were then converted into 3D structures using the AMADAR Tool, yielding an accurate 3D compound library ready for molecular docking in the following chapter.

CHAPTER THREE

MOLECULAR DOCKING

3.1 INTRODUCTION

Molecular docking is an important computational method for drug discovery and protein-ligand interaction investigations. It predicts how small molecules (ligands), such as drugs, bind to a target protein by simulating atomic-level interactions (Morris and Lim-Wilby, 2008). This chapter explores molecular docking techniques, emphasizing their importance in how compounds, particularly EGCG analogues, interact with the RT RNase H and IN active sites, specifically looking at Mg^{2+} ion chelation in both RT and IN.

3.2 NOTE ON PROTEIN-LIGAND BINDING

Protein-ligand binding refers to the interaction between a protein molecule and a smaller molecule known as a ligand (In this case the inhibitors). The study of protein-ligand interactions is a fundamental aspect of biochemistry, molecular biology, and pharmacology (Du et al., 2016).
Currently, there are four conceptual models that explain the interaction between proteins and ligands which include: lock-and-key model, induced-fit model, selected-fit model, and finally, keyhole-lock-key model (Kokkonen et al., 2019). The Lock-and-Key model was proposed by Emil Fischer in 1894, this model describes enzyme-substrate interaction as a rigid, complementary fit. It assumes that the active site of an enzyme has a specific shape that matches the substrate (Navanietha Krishnaraj et al., 2017). Expanding on the lock-and-key model the induced fit model was created. It allows for some flexibility in both the enzyme and the substrate. The principle is that the binding of the ligand induces conformational changes in the enzyme. Additionally, the selected-fit model was created, which is also known as conformational selection. This model implies that the enzyme exists in multiple conformational states, some of these states can bind to

the ligand and stabilize the reactive conformation. Lastly, the Keyhole-Lock-Key was proposed to address enzymes with buried active sites and access tunnels. It introduces the concept of tunnels (keyholes) that allow ligands to reach the catalytic site and it describes the passage of ligands through tunnels for catalysis (Kokkonen et al., 2019).

3.3 MOLECULAR DOCKING STUDIES

Computer-Aided Drug Design (CADD) is a computational method widely used in drug discovery. It complements wet-lab approaches which allows researchers to investigate drug resistance mechanisms, as well as potentially discovering new antiviral targets as well as design new antiviral compounds. A notable benefit of CADD is its capability to produce an atomic level structure-activity relationship (SAR) which is efficient and cost effective (Yu and MacKerell, 2017). Molecular Docking is part of CADD, the aim of molecular docking is to predict and understand molecular recognition by finding energetically favorable binding modes and predicting binding affinity (Morris and Lim-Wilby, 2008). It involves two main steps, the first is sampling conformations of the ligand in the receptor's active site and the second is scoring the conformations and giving them a ranking (Meng et al., 2011). Molecular docking usually involves the interaction between small molecules and larger target macromolecules, which is known as ligand-protein docking. Protein-protein docking is also common; however, this chapter focuses on ligand-protein docking and refers to the ligand as the compound (Hernández-Santoyo et al., 2013). Molecular docking has many applications in drug discovery such as binding prediction for mutagenesis studies, structural activity studies and finding potential leads through virtual screening. The latter will be employed in this study (Morris and Lim-Wilby, 2008).

3.3.1 AUTO DOCK VINA

In the context of computational molecular docking, the selection of the appropriate software tool is crucial for accurate predictions due to the complex characteristics of ligand-receptor interactions. In this study, AutoDock Vina was utilized, it is a well-established and recognised

software for performing molecular docking analysis. The rationale behind the choice of this software is that AutoDock Vina has a simple scoring function and it quickly explores potential ligand-receptor conformations through gradient optimisation (Forli et al., 2016).

3.3.1.1 Sampling conformations

Due to the conformational degrees of freedom and the six degrees of rotational and translational freedom for both ligand and protein, there are many binding modes between the two molecules, therefore, it would be too computationally expensive to generate all the possible conformations (Meng et al., 2011). As a result of many conformational degrees of freedom, different sampling algorithms are created. AutoDock Vina uses a quasi-Newton method called Broyden-Fletcher-Goldfarb-Shanno (BFGS) for local optimisation during molecular docking and Monte Carlo sampling for global optimisation (Ng et al., 2015). Local optimization in this case refers to the refinement of the ligands position and orientation in the binding site in order to find the arrangement that maximizes the binding affinity. This method is favorable because it iteratively adjusts the ligands position and orientation based on the scoring function and the gradient with respect to its various parameters. The parameters could be the orientation and position of the ligand, as well as the values of the torsions for the rotatable bonds in the ligand and flexible residues (Trott and Olson, 2010; Wang et al., 2023). This approach allows for a detailed exploration of the ligand-receptor conformational space, which therefore enhances the accuracy and the efficiency of the optimisation process.

3.3.1.2 Scoring function

The scoring function in molecular docking plays a crucial role in distinguishing between accurate binding poses from incorrect ones, enabling the identification of ligands that effectively bind to the protein within a reasonable computational time frame. However, the scoring function is only an estimation of the binding affinity rather than an exact calculation. Scoring functions are usually

classified as force field based, empirical or knowledge based scoring functions (Meng et al., 2011). AutoDock Vina uses an empirical scoring function derived from experimental data as expressed in the formula below:

$$c = \sum_{i<j} f_{t_i t_j}(r_{ij}) \quad (1)$$

The expression above represents the conformation dependent part of the scoring function in AutoDock Vina. It calculates the cumulative score (c) by summing all of the pairs of atoms (i,j) that are not rigid in the ligand-receptor complex. $f_{titj}\ (r_{ij})$ represents the interaction between the atom types t_i *and* t_j at a specific distance (r_{ij}). This equation, therefore, accounts for pairwise interactions, to assess the overall energy of a particular ligand-receptor conformation (Trott and Olson, 2010). The scoring function is a valuable tool for the prediction and understanding of ligand-receptor interactions.

3.4 MATERIALS AND METHODS

To explore potential interactions between EGCG and its derivatives, RT and IN were selected due to their structurally similar active sites. Both enzymes share a Mg^{2+} ion dependency for functionality. This selection was motivated by the prospect for potentially uncovering dual action inhibitors.

3.4.1 Structure Preparation

3.4.1.1 RT

The crystal structure of RT was obtained from the Protein Data Bank (PDB ID: 4G1Q) and was determined using X-ray diffraction with a high resolution of 1.51 Å. It also had R-Value Free and R-Value Work of 0.193 and 0.154 respectively. To focus on the catalytic activity of the enzyme, the p51 domain was excluded because its primary role is to provide structure. In the p66 domain,

all non-standard residues except for the Mg^{2+} ion were removed. These modifications were done using UCSF Chimera (version 1.17.3).

3.4.1.1.1 Reverting mutations

The engineered mutations in the RT structure were reverted back to their original residues and the expression tags were deleted using Pymol as tabulated in Table 3.1 (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC).

Table 3.1: Discrepancies between the modeled and the reference sequence UNP P03366

Residue	Modeled	Actual	Comment
-1	MET	-	Expression tag
0	VAL	-	Expression tag
172	ALA	LYS	Engineered mutation
173	ALA	LYS	Engineered mutation
280	SER	CYS	Engineered mutation

3.4.1.1.2 Correct placements of Mg^{2+} ions

RT (4G1Q) contains one magnesium ion, however, from literature we know that RT contains two Mg^{2+} ions (Tian et al., 2018). To add the second Mg^{2+} ion onto the 4G1Q structure, another RT structure (8JYI) with the same conformation which contained two Mn^{2+} ions (Mg^{2+} and Mn^{2+} are interchangeably used in the crystal structures of RT) were superimposed onto each other with an RMSD value of 0.367 and the second Mn^{2+} ion was copied onto the 4G1Q RT structure using Pymol (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC). The Mn^{2+} ion

was then changed to Mg^{2+} in Discovery Studio Visualizer (v21.1.0.20298) as depicted in Figure 3.1.

3.4.1.1.3 Optimisation using Quantum mechanics

Although the superimposition of the structures gave an educated estimation for the position of the Mg^{2+} ions, to be precise, further optimization using quantum mechanics was necessary. Before proceeding with quantum mechanics, the charge of the whole protein as well as the active sites were determined. The whole protein's charge was determined at a pH of 7.4 to mimic the physiological environment of the proteins using H^{++} (Hopkins, Sanvictores and Sharma, 2022). The active site's charge was calculated to be zero (ASP443, ASP498, ASP549, GLU478, two Mg^{2+} ions) and the entire protein charge was calculated to be nine. The gaussian optimization was combined quantum mechanics/molecular mechanics (QM/MM) at the ONIOM PM6:UFF mixed level (PM6 semiempirical for the Magnesium and active site residues, universal force-field for the MM) using Gaussian 09 (version g09.E01).

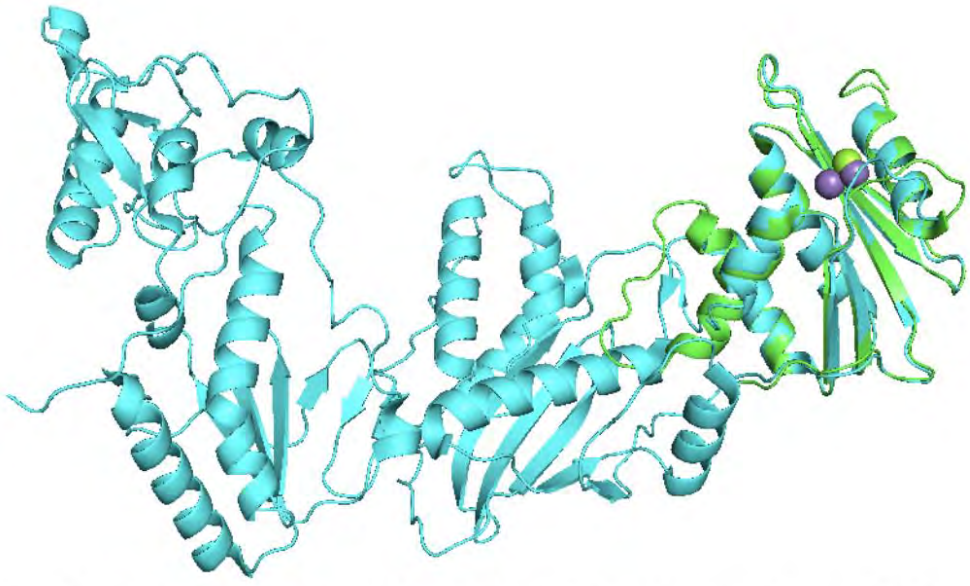

Figure 3.1: Structural Superimposition of RT Rnase H Domain (PDB ID: 8JYI) onto RT Structure (PDB ID: 4G1Q). The green ball represents the Mg^{2+} ion belong to the 4G1Q structure and the purple balls represent the Mn^{2+} ions belonging to the 8JYI structure

3.4.1.1.4 Correction of atoms in PDB

Due to the mixed level optimization, the active site residues D498 and D549 were mislabeled as "UNK" with incorrect atoms and were therefore corrected.

3.4.1.2 IN protein

The integrase structure utilized in this report had the PDB ID 1WKN, this structure is a plausible full length structure modeled by De Luca et al., (2003). This structure took precedence over other crystal structures because it contained the correct placement of the two Mg^{2+} ions in coordination with the DDE amino acid triad. Although this Integrase structure contained both magnesium ions , it also underwent protonation and optimisation to ensure uniform treatment of both enzymes

3.4.2 Docking Parameters

3.4.2.1 Initial docking with RT

All 125 203 compounds were docked to RT using AutoDock Vina (Trott & Olson, 2010; Eberhardt et al., 2021). All phenyl hydrogens were removed from the ligands to ensure coordination with the Mg^{2+} ions (Case, Zubieta and Doyle, 2020). Due to the large compound library an exhaustiveness value of 8 was selected because it is the default setting in AutoDock and has been used in studies to evaluate scalability and general performance (Minibaeva, Ivanova and Polishchuk, 2023). The active site coordinates were determined based on the average position of the two Mg^{2+} ions, resulting in coordinates of (-42.29, 13.99, -9.20) for the x,y and z axis respectively. The docking box size was 22 Å. Additional parameters included an energy range of 4 and the utilization of 8 CPUs for computational efficiency were selected.

3.4.2.2 Redocking validation

To assess the reliability of the molecular docking protocol, validation was performed by redocking the compound EGCG into the active site of RT. Since the crystal structure lacked a bound ligand, redocking of EGCG served as an internal validation method. The root mean square deviation (RMSD) value for the redocked EGCG compound was 0.0722 (Figure 3.2), indicating a close alignment with the experimental binding pose and validating the accuracy of the docking approach.

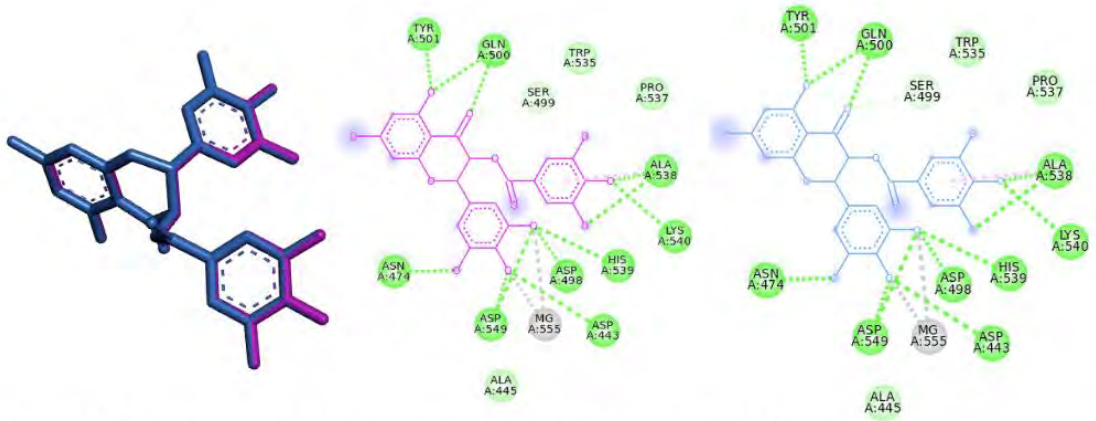

Figure 3.2: Redocking of RT at an exhaustiveness of 8 with RMSD value of 0.0722. Receptor ligand interactions of the redocked compound EGCG shows that the compound interacts with all of the same residues

3.4.2.3 Initial docking with IN

All 125 203 compounds were docked to IN using AutoDock Vina (Trott & Olson, 2010; Eberhardt et al., 2021). All phenyl hydrogens were removed from the ligands to ensure coordination with the Mg^{2+} ions. Due to the large compound library an exhaustiveness value of 8 was selected to ensure computational efficiency. The active site coordinates were determined based on the average position of the two Mg^{2+} ions, resulting in coordinates of (16.95, -11.82, -18.58) for the x,y and z axis respectively. The docking box size was 22 Å. Additional parameters included an energy range of 4 and the utilization of 8 CPUs for computational efficiency were selected.

3.4.2.4 Redocking validation

I utilized the same validation method for IN as those employed for RT. The RMSD value for the redocked EGCG compound was 0.0279 (Figure 3.3)

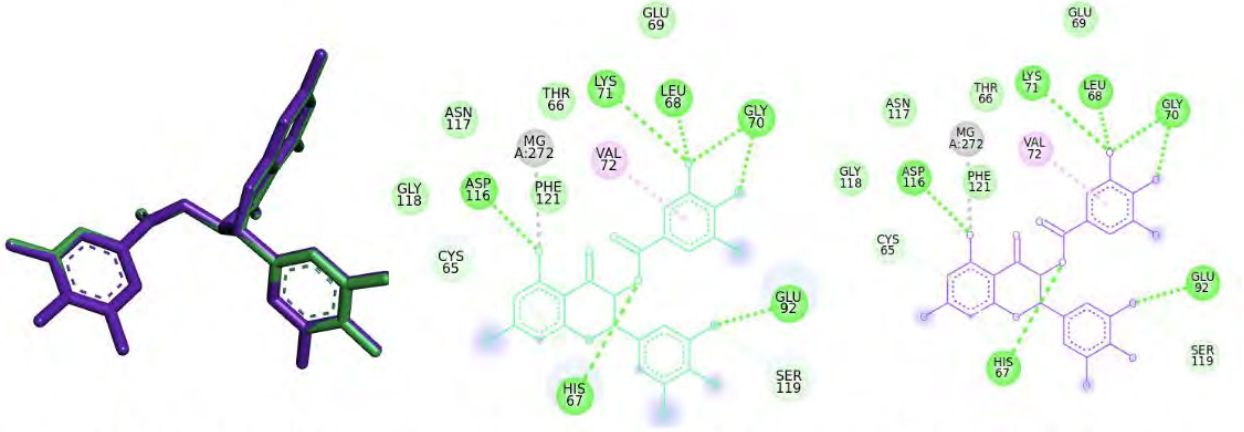

Figure 3.3: Redocking of integrase at an exhaustiveness of 8 with RMSD value of 0.0279. Receptor ligand interactions of the redocked compound EGCG shows that the compound interacts with all of the same residues

3.4.3 RT and IN docking correlation

To assess the correlation between the docking results obtained for RT and IN, the average binding affinities were calculated. Simultaneously, the average distance of the oxygen atoms on the compounds to the Mg^{2+} ions on the enzymes were plotted using Matplotlib. The aim was to identify compounds with high binding affinities for both enzymes and a short distance between oxygen atoms and the Mg^{2+} ions to ensure effective Mg^{2+} coordination. Then, a cutoff threshold of 7.5 for the average binding affinity and 4 Å for the average distance of the Mg^{2+} ions was set. Subsequently, the compounds were filtered resulting in 288 compounds (Figure 3.8).

3.4.4 Enhanced docking precision (Exhaustiveness: 256)

The filtered 288 compounds as well as EGCG were docked to RT and IN using AutoDock Vina (Trott & Olson, 2010; Eberhardt et al., 2021) at an exhaustiveness of 256 to enhance docking precision (Le and Do, 2022). The active site coordinates and docking parameters are identical as above.

3.4.4.1 Redocking validation (Exhaustiveness: 256)

The docking validation at an exhaustiveness of 256 was done as stated earlier (Figure 3.4 and 3.5).

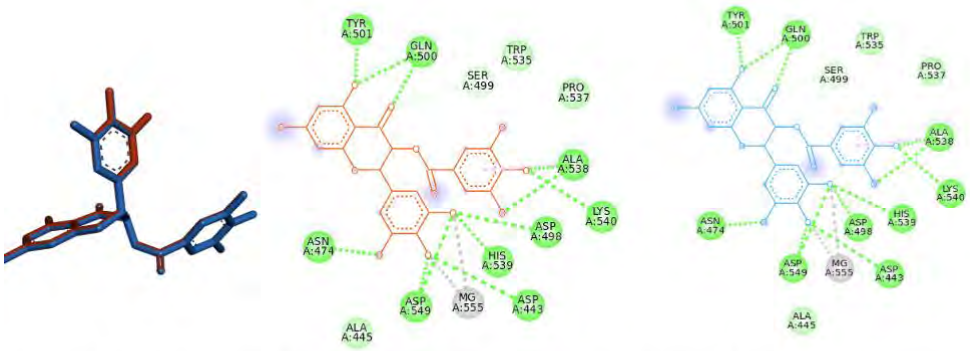

Figure 3.4: Redocking of Reverse Transcriptase at an exhaustiveness of 256. With RMSD value of 0.0113 Receptor ligand interactions of the redocked compound EGCG shows that the compound interacts with all of the same residues

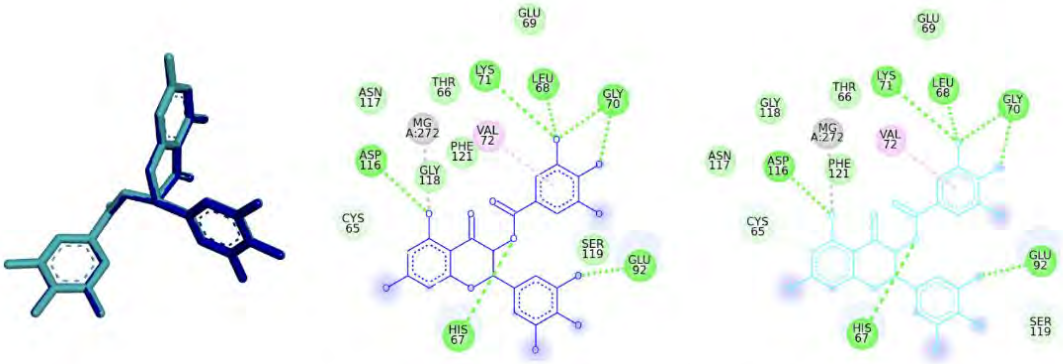

Figure 3.5: Redocking of Integrase at an exhaustiveness of 256 with RMSD value of 0.0109. Receptor ligand interactions of the redocked compound EGCG shows that the compound interacts with all of the same residues

3.4.5 Calculation of Drug-Like Properties

To evaluate the potential oral bioavailability of the filtered compounds, calculations on pharmaceutical properties were calculated using RDKit (2023.09.4). The parameters considered were molecular weight (Mw), Ghose-Crippen octanol-water partition coefficient (ALOGP), hydrogen bond acceptors (HBA), hydrogen bond donors (HBD), polar surface area (PSA), number of rotatable bonds (ROTB), aromatic rings (AROM), alerts (Alert Flags for undesirable chemical features), and the overall quantitative estimation of Drug-likeness (QED) score (Table 3.2).

3.5 RESULTS AND DISCUSSION

The accurate placement of Mg^{2+} within the proteins is crucial for enzymatic stability and function (Miri et al., 2014). Both of these proteins have a shared dependency on magnesium therefore inhibitors that are able to chelate the Mg^{2+} ions, have the potential to disrupt the enzymatic function of RT and IN (Gill et al., 2019). Therefore, ensuring the correct coordination of Mg^{2+} ions within these protein structures is essential for our research goals.

The placement of the Mg^{2+} ions within the enzymes RT RNase H and IN active sites is depicted in Figure 3.6, which illustrates the highly conserved residues which coordinate with Mg^{2+} ions. In the RT RNase H active site, the crucial residues ASP443, GLU478, ASP498, and ASP549 are in coordination with the Mg^{2+} ions, which are at a distance of 4.176 Å apart. This distance between Mg^{2+} ions is consistent with the findings in literature, which suggest a distance of approximately 4 Å between the ions (Goldschmidt et al., 2006). Similarly, in the IN active site, the highly conserved residues ASP64, ASP116, and GLU152 coordinate with the Mg^{2+} ions. The distance between the ions is 3.459 Å. This value slightly deviates from the literature, specifically reported by Liao and Nicklaus (2010), which expressed an Mg^{2+} ion distance of 3.70 Å. However, it's important to note that our IN structure was obtained from a computational model, which is in contrast to the experimental X-ray cocrystal structure utilized in the literature.

Given that EGCG is our reference compound, particularly due to its demonstrated dual inhibitory effects as shown in the study conducted by Sanna et al. (2019), it is imperative to assess whether its mechanism of action involves metal chelation. Therefore,it is crucial to evaluate whether EGCG possesses the ability to chelate the Mg^{2+} ions in both RT and IN.

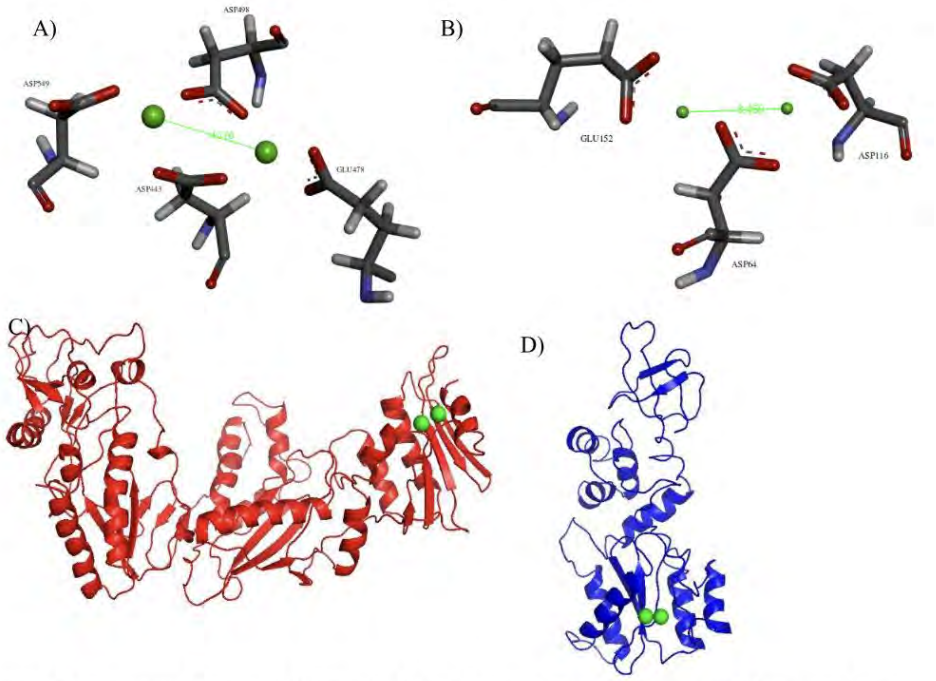

Figure 3.6: Structural Analysis of RT and IN Active Sites and final proteins with Magnesium Ions. A)Geometry of RT active site, Contains acidic aspartic acid (ASP) and glutamic acid (GLU), with two magnesium ions with a distance of 4.176 Å. B) Geometry of IN active site, Contains acidic aspartic acid (ASP) and glutamic acid (GLU), with two magnesium ions with a distance of 3.459 Å. C)Final RT enzyme with both magnesium ions. D) Final IN enzyme with both magnesium ions.

In Figure 3.7, the protein-ligand interactions between EGCG and the enzymes RT and IN are depicted. Within the RT-EGCG complex, hydrogen bonds are formed between two of the four highly conserved amino acid residues which make up the DDE motif (ASP443 and ASP549) (Mahboubi-Rabbani et al., 2021). This could potentially mean that the hydrogen bonds between EGCG and these residues could potentially interfere with enzymatic activity by interacting with key catalytic residues. Additionally, EGCG chelates both Mg^{2+} ions and would therefore theoretically inhibit RT's enzymatic function (Gill et al., 2019). However, experimental validation would need to be conducted to confirm EGCG's inhibitory mechanism.

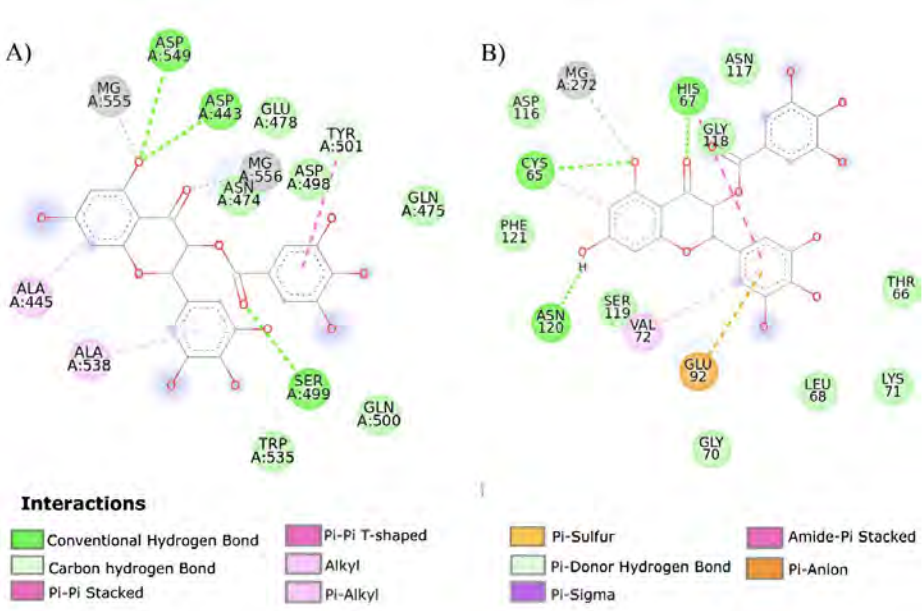

Figure 3.7: Receptor-ligand interaction of EGCG docked to RT and IN. A) EGCG docked to RT. B) EGCG docked to IN and visualised using Discovery studio visualiser. With the grey bonds depicting metal chelation.

In the context of the IN-EGCG complex, the lack of interactions between EGCG and DDE motif residues implies that EGCG's inhibitory mechanism does not include the direct interference with catalytic residues. However, EGCG's ability to chelate one Mg^{2+} within the active site of IN indicates that it could potentially impact its enzymatic function. Raltegravir (RAL), which is an FDA approved HIV-1 IN drug, chelates both Mg^{2+} ions and disrupts enzymatic function (Agrawal et al., 2012). Therefore, it is unclear if the chelation of a single magnesium ion by EGCG is sufficient to disrupt its function. Thus, experimental validation would need to be conducted to assess if the chelation of one Mg^{2+} ion is sufficient to disrupt IN's enzymatic function.

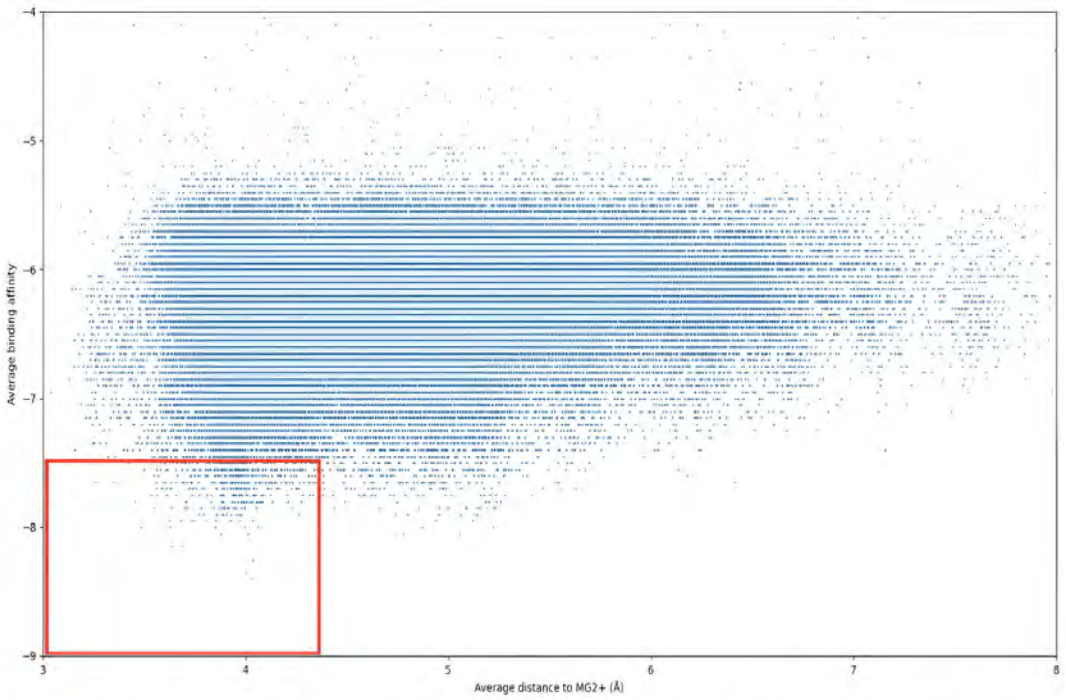

Figure 3.8: A scatter plot illustrating the relationship between the average binding affinity for both RT and IN and the average distance of oxygen atoms from magnesium ions. The x-axis represents the average binding distance to the magnesium ions, while the y-axis depicts the average binding affinity for both RT and IN. The red box represents the cut off threshold chosen to be 7.5 and 4 A which resulted in 288 compounds

Furthermore, it is essential to analyze the docking results of both RT and IN together, forming a correlation between the two, in order to comprehensively evaluate their potential for dual inhibitory effects and identify compounds that could potentially have better efficacy to EGCG. Referring to Figure 3.8, it illustrates the average binding affinity for both RT and IN in correlation with the average distance of oxygen (O^2) atoms to the Mg^{2+} ions. Specifically, the O^2 atoms within the inhibitors were selected to ensure successful coordination with the Mg^{2+}. The objective of this analysis is to identify inhibitors with low binding affinities and shorter distances between O^2 atoms and Mg^{2+}. This is because the binding affinity represents the strength of interaction between two molecules that bind reversibly (Kastritis and Bonvin, 2013) and the closer distance of O^2 atoms to Mg^{2+} ions would be more favorable because it indicates a stronger potential for metal chelation.

Additionally, the red box on the graph indicates the threshold we set to filter the compound library, with an average binding affinity of -7.5 and an average distance of O^2 atoms to Mg^{2+} ions of 4 Å. This threshold was aimed to narrow down the extensive compound library, resulting in the filtration of only 288 compounds out of the initial 125 203 EGCG derivative compounds.

However, the relatively small number of compounds filtered in comparison to the big compound library, suggests that there is a lack of focus in our library for our research objectives. A more focused library would ideally have more compounds that have lower binding affinities and shorter distances to the Mg^{2+} ions. However, it's important to note that the docking was conducted at an exhaustiveness of 8, which is relatively low. Due to the large size of the compound library, this exhaustiveness was necessary to manage computational resources while still producing useful results. Therefore, when analyzing the docking results, it's important to consider that higher exhaustiveness values might provide higher exhaustive sampling of the conformational space, which could potentially result in more accurate predictions of protein-ligand interactions. Nonetheless, the chosen exhaustiveness was suitable considering the computational limitations and well as time constraints caused by the size of the compound library.

The 288 filtered compounds were docked again, at a higher exhaustiveness of 256 to increase the accuracy of protein-ligand interactions. However, to filter them down further to ensure low numbers for downstream analysis, calculation of QED was conducted. QED is a straightforward way to rank compounds based on their physicochemical properties. QED values range from 0 to 1, with 0 indicating unfavorable properties and 1 indicating favorable properties in all aspects. Therefore, QED values that are closer to one are considered to be more drug-like (Bickerton et al., 2012). In addition to QED, some physicochemical properties were assessed, including molecular weight (Mw), lipophilicity (ALOGP), hydrogen bond acceptors (HBA), hydrogen bond donors (HBD), polar surface area (PSA), rotational bonds (ROTB), and aromatic rings (AROM). By considering these properties, alongside QED, a comprehensive evaluation of compound drug-likeness can be achieved, which directs the selection of potential hit compounds for further investigation.

Table 3.2 : Quantitative estimation of druggability (QED) of the filtered ligands

Compound	Mw	ALOGP	HBA	HBD	PSA	ROTB	AROM	ALERTS	QED
Raltegravir	444.4	0.9	8	3	152.2	6	3	0	0.51
EGCG	472.4	1.9	12	8	214.4	3	3	2	0.20
Sci30703	486.5	4.6	7	2	102.3	3	2	1	0.61
Sci33211	482.5	4.8	7	2	102.3	3	2	1	0.60
Sci48919	482.5	4.8	7	2	102.3	3	2	1	0.60

Table 3.2 presents the quantitative estimation of druggability for the top three EGCG derivative compounds with the highest QED values (Sci30703,Sci33211 and Sci48919), along with EGCG and Raltegravir (RAL). RAL, being an FDA-approved drug, serves as a benchmark for comparison. This comparison allows for the identification of potential drug candidates among the EGCG derivatives by evaluating their druggability against a known HIV-1 metal chelating FDA-approved drug (Mahboubi-Rabbani et al., 2021). As seen in the table, all EGCG derivatives have higher QED values than RAL and EGCG, with the values being 0.61, 0.60, 0.60, in relation to 0.51 and 0.20 respectively. This comparison implies that, theoretically, EGCG derivatives have better drug-like properties than Raltegravir and EGCG.

In addition to QED, Lipinski's Rule of Five is important in drug discovery for predicting the oral bioavailability of compounds. Lipinskies rule of five states that compounds are more likely to have good oral bioavailability if they meet a specific criteria, such as molecular weight ≤ 500, ALOGP ≤ 5, HBA ≤ 10, HBD ≤ 5 and PSA ≤ 140 Å² (Mahgoub et al., 2022). It's important to note that a compound can deviate from one of these criteria and still be considered to be compliant with Lipinski's Rule as long as it only has one violation. In this case RAL violates one of Lipinski's Rules due to its PSA value exceeding 140 Å², yet it still passes because it only has one violation.

Based on this rule the Sci30703, Sci33211 and Sci48919 do not violate this rule. On the other hand, EGCG violates Lipinski's Rule by exceeding the limits for HBA, HBD and PSA. However, Lipinski's Rule can not be applied to natural products like EGCG (Müller-Kuhrt, 2003). Lipinski's Rule is important in assessing the drug-like qualities of compounds and accelerating the drug discovery process. However, it's important to note that Lipinski's Rule of Five cannot be used as

a direct comparison in this case because EGCG is a natural compound and the derivatives would be synthetic. Thus, we depend on the QED values as a metric to evaluate their suitability for drug development.

Table 3.3:Binding affinities of compounds to RT and IN with average docking energies

Compound	RT Energy (kcal/mol)	IN Energy (kcal/mol)	Average energy (kcal/mol)
sci30703	-8.5	-6.9	-7.7
sci33211	-8.7	-6.4	-7.55
sci48919	-8.6	-6.5	-7.55

To supplement Table 3.2, Table 3.3 shows the three compounds with the best QED scores and their binding energies were selected for further analysis. These compounds had average binding energies of -7.7, -7.55, and -7.55 kcal/mol, respectively. As shown in Table 3.3, the binding energies to RT were consistently more negative than those to IN. Specifically, RT binding energies ranged from -8.5 to -8.7 kcal/mol, whereas IN binding energies ranged from -6.4 to -6.9 kcal/mol. This indicates a stronger predicted interaction between the ligands and RT compared to IN. One possible reason for this difference is the composition of the active sites. The RT active site consists of negatively charged residues Asp443, Asp498, Asp548, and Glu478 and the two magnesium ions (Mg^{2+}), the net charge of the RT active site becomes neutral (0). In contrast, the IN active site, composed of Asp64, Asp116, and Glu152 and two magnesium ions, results in a net positive charge (+1) due to the fewer negatively charged residues.

The charge difference between the two active sites plays a significant role in their respective binding affinities. However, the inhibitors studied at a pH of 7.4 are all neutral but if the inhibitors studied were negatively charged, they might bind more favorably to IN due to electrostatic interactions. However, because the inhibitors in this case are neutral, such interactions are not as prevalent, limiting the variety of favorable interactions and contributing to the overall weaker binding affinities observed for IN (Corona et al., 2014).

The selection of these top three compounds based on their average binding energies highlights their potential as lead candidates for further development. These findings suggest that targeting RT may yield more effective inhibition due to the stronger ligand binding observed, while IN, with its more selective active site, may require more specific ligands to achieve similar binding efficacy. Further experimental validation and structure-activity relationship studies will be essential to confirm these computational predictions and to refine these compounds for potential therapeutic use.

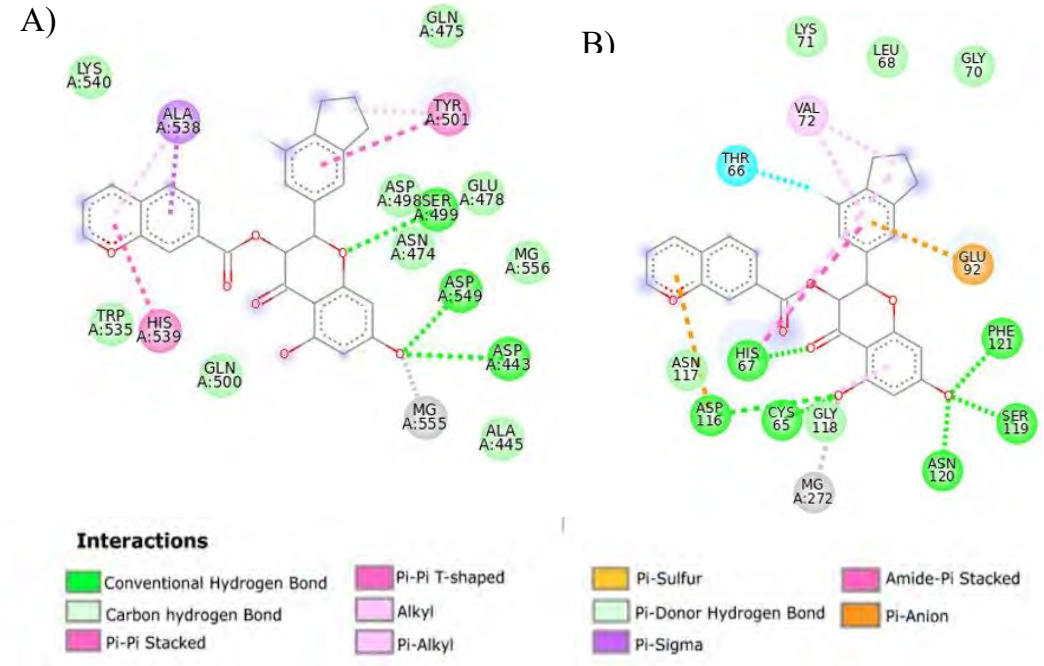

Figure 3.9: Ligand-receptor interacts of the docked filtered Scifinder compound (Sci30703). At an exhaustiveness of 256. A) RT B) IN. With the grey bonds depicting metal chelation

Lastly, compounds Sci30703, Sci33211 and Sci48919 were docked at an exhaustiveness of 256 to improve the accuracy of protein-ligand interactions and to evaluate potential metal chelation. Figure 3.9 illustrates the interaction of the inhibitor Sci30703 with RT and IN, it shows that the

RT-Sci30703 complex demonstrates the compound's ability to chelate one Mg^{2+} ion and interact with two key active site residues (ASP443 and ASP549). Similarly, in the IN-Sci30703 complex, the inhibitor also has the ability to chelate one Mg^{2+} and interacts with the key active site residue ASP116 through conventional hydrogen bonds. Figure 3.10 are the protein-ligand interactions of the inhibitor Sci33211 in complex with RT and IN. The RT-Sci33211 shows the inhibitor's ability to chelate both Mg^{2+} ions as well as the formation of hydrogen bonds with the key active site residues ASP443 and ASP549. This is in contrast to the IN-Sci33211 complex, in which the inhibitor only has the ability to chelate one Mg^{2+} ion and interacts with one key active site residue ASP116.

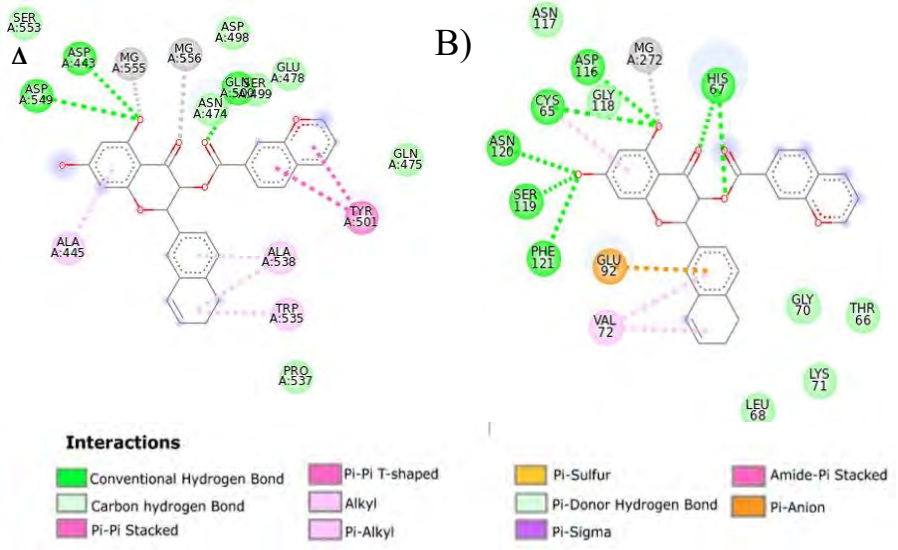

Figure 3.10: Ligand-receptor interacts of the docked filtered Scifinder compound (Sci33211). At an exhaustiveness of 256 A) RT B) IN. Where the grey interactions depict metal chelation

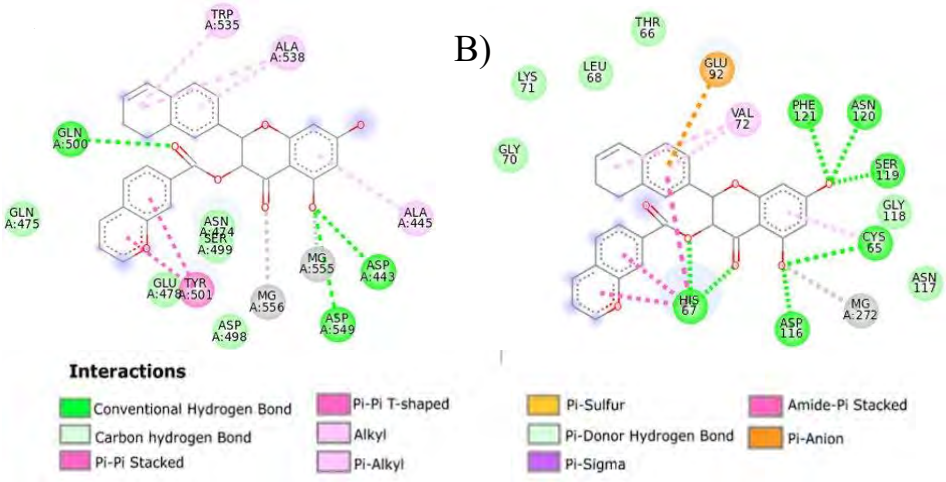

Figure 3.11: Ligand-receptor interacts of the docked filtered Scifinder compound (Sci48919). At an exhaustiveness of 256. A) RT B) IN. The grey interactions represent metal chelation

Finally, Figure 3.11 shows the protein-ligand interactions of the inhibitor Sci48919 in complex with RT and IN. Referring to the RT-Sci48919 complex, this inhibitor has the ability to chelate both Mg^{2+} ions and interact with two key active site residues ASP443 and ASP549 through conventional hydrogen bonding. Whereas the IN-Sci48919 has the ability to chelate one Mg^{2+} ion and interact with one key active site residue ASP116 through conventional hydrogen bonding. Regarding Figures 3.9 to 3.11 it is evident that none of these inhibitors have the ability to chelate both Mg^{2+} ions within both enzymes. This raises uncertainty regarding their potential as dual-action inhibitors through metal chelation. However, an interesting observation is that all inhibitors in complex with RT interact consistently with the same two active site residues, ASP443 and ASP549. Similarly, all inhibitors in complex with IN only interact with the active site residue ASP116, with each inhibitor chelating one Mg^{2+} ion. Notably, compounds Sci33211 and Sci48919 in complex with RT demonstrate the ability to chelate both Mg^{2+} ions. As a result, it remains uncertain if the interaction with select critical active site residues in both enzymes, together with the chelation of a single Mg^{2+} ion is sufficient for total enzymatic disruption. Further in vitro studies are necessary to determine if a single metal chelation is sufficient for enzymatic disruption.

3.6 CHAPTER SUMMARY

This chapter explores drug discovery methodologies, with specific focus on EGCG and its derivatives as potential dual action inhibitors of HIV-1 RT and IN through metal chelation. The selection of these enzymes comes from their shared dependency on Mg^{2+} ions, which indicates that inhibitors that are capable of chelating these ions could disrupt enzymatic function (Gill et al., 2019). Given the crucial role of the active site residues in interacting with Mg^{2+} ions, accurate placement of these residues and ions was ensured through QM/MM optimisation.

Additionally, docking experiments were then conducted on the entire library of EGCG derivative compounds (consisting of 125 203 compounds), separately to RT and IN. Correlation between the docking results aimed to identify potential dual-action inhibitors was done by assessing the average docking binding affinities for both RT and IN were graphed against the average distance of oxygen atoms to the Mg^{2+} ions. This graph allowed for the filtering of the compounds and was performed by setting a threshold average binding affinity of 7.5 and an average distance of O^2 atoms to Mg^{2+} ions of 4 Å, resulting in 288 compounds.

Since 288 compounds was still too high for downstream analysis, quantitative estimation of druggability was employed to further filter the compounds, focusing on those with the highest QED values. Assessment of the QED values and physicochemical properties revealed that three EGCG derivative compounds exhibited better drug-like properties compared to the FDA-approved drug RAL and EGCG itself.

Subsequently, the three selected EGCG derivatives were docked again at an exhaustiveness of 256 to assess metal chelation within both enzymes. The results indicated that RT in complex with the inhibitors consistently interacted with the same active site residues (ASP443 and ASP549), with Sci33211 and Sci48919 chelating two Mg^{2+} ions, while Sci30703 chelated one. In contrast, IN in complex with the inhibitors interacted with ASP116, with only one metal ion chelated in all cases. Therefore, further in vitro studies are needed to determine if the chelation of a single metal ion is sufficient for enzymatic inhibition. However, computationally, these three EGCG

derivatives will undergo MD simulations in order to get more insights into the potential mechanism of action.

CHAPTER FOUR

MOLECULAR DYNAMICS SIMULATIONS

4.1 INTRODUCTION

In this chapter, we explore the use of Molecular Dynamic (MD) simulations as a key computational technique to understand the dynamic interactions of protein-ligand systems. The main goal of this chapter is to understand the structural dynamics and behavior of these complexes over time. In the context of our study, the proteins are RT and IN, and the ligands under investigation include EGCG and its derivatives. By utilizing MD simulations we may be able to identify features in protein-ligand interactions that may not be readily visible in lab settings and therefore bridge experimental gaps. In addition to being helpful in enhancing our understanding of molecular dynamics, MD simulations are a crucial initial stage in the drug discovery process (Dewaker & Prabhakar, 2023). By simulating the behavior of biomolecular systems with a range of compounds, MD simulations allow us to efficiently assess potential treatment possibilities. This technique acts as a preliminary filter, screening possible compounds prior to investing funds in the labor-intensive wet lab studies. This method not only serves to streamline and focus the drug development process, but it also appears to be a time saving strategy in the search for novel therapeutics.

MD simulations are recognised as a powerful computational tool within the scientific community, by offering the unique capability to observe the physical interactions within a biophysical environment (Ghahremanian et al., 2022). They predict how every single atom in a molecule or in this case a protein, will move over time based on the physics of interatomic interactions. These simulations can capture biomolecular processes such as protein folding, conformational change and ligand binding. As well as how the biomolecules respond at an atomic level to changes such as protonation, mutations and phosphorylation (Hollingsworth and Dror, 2018).

4.2 FUNDAMENTAL PRINCIPLES OF MD SIMULATION

Essentially, MD simulations involve a force field, a mathematical model explaining atomic interactions in the system. The force field guides the dynamic study of a biomolecular system, such as a protein immersed in a solvent like water (Salo-Ahen et al., 2020). The simulation begins with a precise determination of all atoms' initial positions in the system. The forces exerted by each atom are then calculated, taking into account the complex interactions within the molecular system. Newton's equations of motion are then used to estimate how each atom's spatial position will change over time. To ensure a comprehensive study,periodic boundary conditions extend the system outside the simulations box, resulting in a continuous molecular environment. The simulation is carried out in incremental time steps, with forces repeatedly recalculated and positions and velocities updated accordingly. The end result is a detailed three-dimensional trajectory that depicts the system's atomic level configuration at each point in time during the simulation. MD simulations are thus an effective tool for unraveling the complicated dynamics of biomolecular systems and understanding their behaviors at an atomic level (Hollingsworth and Dror, 2018).

4.2.1 Periodic boundary conditions

As previously stated, periodic boundary conditions (PBC) are critical in generating a continuous molecular environment during simulations. This is especially imported when simulating systems where bulk solvents, such as water is crucial. Without PBC, simulations in droplets of water would simply evaporate. As a result, the simulation must include barriers (periodic boundaries) to contain the molecules in order to maintain thermodynamic parameters such as temperature,density and pressure (Bruininks et al., 2023).

The use of PBC in simulations allows for the approximation of an infinite system using a small part known as a unit cell, also known as a periodic box in molecular dynamics. To implement PBC, the unit cell is surrounded by translated copies in all directions, resulting in an environment that resembles a big system (Bekker, 1996). As a result, when a molecule crosses the simulation box’s boundary, it reappears on the other side. This novel approach ensures that each molecule consistently interacts with its neighbors, even if they are on opposing sides of the simulation box.

4.2.2 AMBER force fields

As stated earlier, a force field is a mathematical expression that describes how the energy of a system depends on the positions of its atoms (Salo-Ahen et al., 2020). A force field consists of two main components. Firstly, the force field requires an analytical form of the interatomic potential energy, represented by a mathematical function denoted as U(r). This function represents the system's potential energy in relation to the spatial coordinates of its atoms. Secondly, the force field includes a set of parameters that determine the shape and properties of the potential energy function (Konovalov et al., 2021). Parameters in a force field are usually obtained through two distinct methods: 1. Ab Initio or Semi-Empirical Quantum Mechanical calculations. Which involve quantum mechanical calculations based on first principles. 2. Aligning to Experimental Data, the parameters are adjusted to fit the force field to experimental data obtained from various techniques like electron diffraction, NMR and so many more (González, 2011).

Assisted Model Building with Energy Refinement (AMBER) was originally designed with a primary focus on protein and nucleic acid systems (Wang et al., 2004). It focuses on structures and nonbonded energies within molecules. The total energy of a molecule in the Amber force field is the sum of different energy components (Hornak et al., 2006). These components include components include stretching (Stretching), angle bending (Vangle bending), angle torsion (Vangle torsion), and electrostatic and Van der Waals forces as seen in the equation below:

$$V_{total} = V_{stretching} + V_{angle\ bending} + V_{angle\ torsion} + V_{electrostatics\ and\ Van\ der\ Waals\ forces} \quad (2)$$

This equation provides a holistic view of the energy contributions from different interactions within the molecule. However, a simplified functional form of the Amber force field is called the "Class I" model which was developed based on the equation above. The Class I model utilizes the equation below:

$$E_{pair} = \sum_{bonds} k_r(r - r_{eq})^2 + \sum_{angles} k_\theta(\theta - \theta_{eq})^2 \times [1 + cos(n\phi - \gamma)] + \sum_{i<j} \quad (3)$$

Where, the equilibration structural parameters (req and θ_{eq}) define the equilibrium values for bond lengths and angles. Force constants (Kr, Kθ, Vn) determine the strength of bonded interactions. Torsional angle parameters (n and γ) control the periodicity and phase of dihedral angle contributions. Nonbonded potentials are characterized by parameters A,B, and q. The dihedral angle (φ) represents the rotation about a bond and lastly, the phase angle γ for torsional angles can take values of either 0° or 180° (Guo, 2019).

In summary, the understanding of these parameters not only broadens our theoretical understanding, but also provides the way for practical applications in our search to determine whether specific inhibitors can chelate the Mg^{2+} ions present in RT and IN.

4.2.3 Trajectory Analysis

In MD simulations, trajectory analysis is pivotal in understanding the dynamic behavior of biomolecular systems over time (Hollingsworth and Dror, 2018). By tracking the atomic movement and interactions within protein-ligand systems, trajectory analysis provides valuable insights into understanding the mechanisms of action for inhibitors. In this study, the trajectory analysis conducted was Root Mean Square Deviation (RMSD) and Root Mean Square Fluctuation (RMSF) to assess protein-ligand system behavior.

4.2.3.1 RMSD

RMSD analysis is a fundamental method used in trajectory analysis to measure how much a biomolecular structure changes over time during molecular dynamic simulations (Schreiner et al., 2012). It works by aligning different structures at each time step with a reference structure and then calculating the average distance between corresponding atoms. This allows researchers to understand the extent of structural variation within protein-ligand complexes, providing valuable insights into their stability and dynamic behavior (Arnittali et al., 2019).

RMSD has two purposes: firstly, it acts as an indicator for evaluating the reliability of our simulation and secondly, it helps to ensure that our system archives equilibrium and remains structurally stable. Notably, all RMSD calculations are conducted using the Cα atoms, which represent the backbone carbon atoms (Arnittali et al., 2019). The RMSD equations is shown below:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N} m_i(r_i - r_{ref})^2}{\sum_{i=1}^{N} m_i}} \quad (4)$$

Where, the variable m_i represents the mass of atom *i*, while ri represents the coordinates of atom *i* at a specific moment in time. The term r_{ref} refers to the coordinates of atom *i* in its original or reference state.

4.2.3.2 RMSF

The RMSF analysis serves as a valuable tool for assessing the flexibility of a protein when interacting with a compound, as well as for identifying localized variations within the protein chain. RMSF shows the variation of flexibility and can be used to investigate inhibitors binding to their targets and which specific residues the inhibitor disrupts. The RMSF plot highlights residues that undergo significant changes during the simulation, with peaks indicating the residues greatest fluctuations (Ghahremanian et al., 2022). Therefore, higher RMSF values indicate greater flexibility within the protein's domain. RMSF is calculated using the equation below:

$$RMSF = \sqrt{\frac{1}{T}\sum_{i=1}^{T} (x_i - \underline{x})^2} \quad (5)$$

Where T is the trajectory trajectory frame numbers and $\underline{X}$ average position over time (Ghahremanian et al., 2022).

4.3 MATERIALS AND METHODS

MD simulations were conducted using the AMBER03 force field. The purpose of the MD simulations was to evaluate the protein-ligand system behavior and to assess the stability of the three compounds with the highest QED values as stated in chapter three, with EGCG serving as the reference compound. The primary objective was to compare the stability of the three compounds to EGCG for both RT and IN.

4.3.1 Structure preparation

RT and IN were both prepared according to the procedures outlined in Chapter Three, involving protonation at a pH of 7.4 and subsequent quantum mechanics calculations. To prepare the system for MD simulations, the proteins were subjected to a series of preprocessing steps using the GROMACS software package (2020.1)

4.3.2 Force field selection

The choice of the AMBER03 force field was deliberate because it was specifically designed for the accurate representation of protein and nucleic acids (Hornak et al., 2006). The parameterization of Mg^{2+} ions within RT and IN was redundant. The conclusion was based on the dynamic nature of the Mg^{2+} ions observed in current crystal structures of both proteins. The presence of Mg^{2+} ions varies from none to one or two, demonstrating their natural mobility. Given this dynamic behavior, our strategy leaned towards using non-bonded parameters, specifically relying on Van Der Waals forces. This choice effectively eliminated the need for explicit parameterisation, aligning with the Mg^{2+} ions natural flexibility inside these proteins.

4.3.3 Production of molecular dynamics (MD) simulation

The simulation box was defined as cubic, the proteins both maintained a distance of 1.0 nm from the edge of the box. Solvation of the system was done using the TIP3P water model. Ionization was performed by adding Potassium chloride (KCl) ions to the system at a concentration of 0.15 M ensuring system neutrality. To ensure system stability and readiness energy minimisation was conducted using a step size of 0.01 and underwent 50000 steps. The MD simulations were conducted utilizing the leap-frog integrator to numerically integrate the equations of motion at a

temperature of 310 K following the Maxwell distribution, and the Parrinello-Rahman method for pressure coupling. Lastly, the simulations were run for a total of 50 nanoseconds, utilizing the Particle Mesh Ewald method for long range electrostatics.

4.3.4 Analysis of trajectory

Trajectory analysis was conducted using RMSD and RMSF metrics to assess structural deviations of RT and IN inhibitor complexes using the GROMACS software package (2020.1).

4.4 RESULTS AND DISCUSSION

The MD simulations were conducted to explore the interactions between RT and IN in complex with EGCG and its derivatives. The objective was to evaluate the behavior of the protein ligand systems and determine whether the derivatives exhibited improved performance compared to the reference compound EGCG.

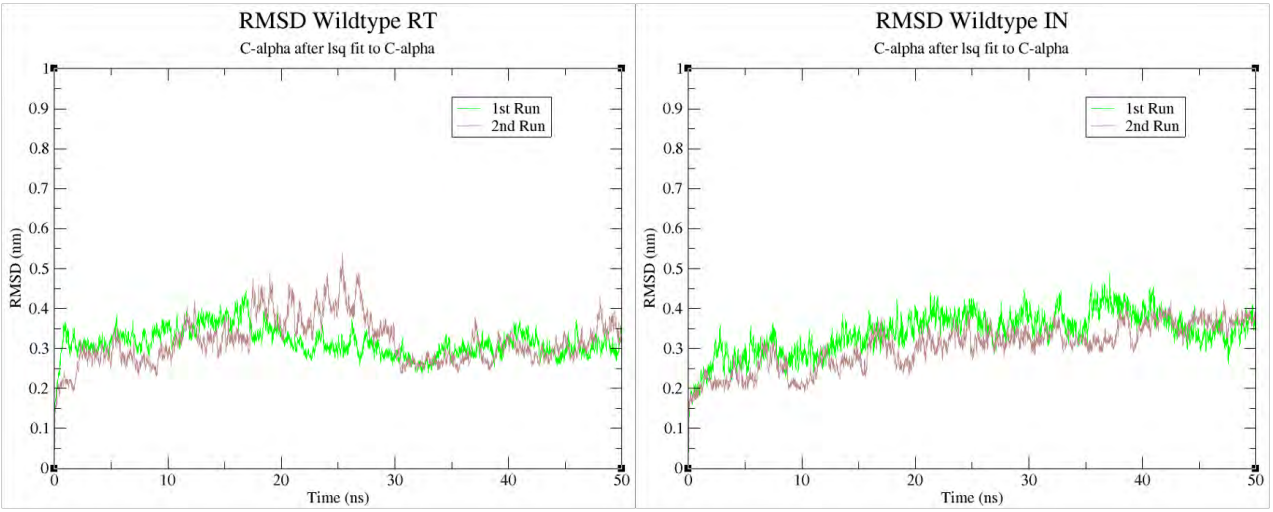

Figure 4.1: RMSD plots depicting the conformational stability of RT and IN wiltypes. Two independent runs were conducted for each enzyme, illustrating the RMSD values calculated for the C-alpha atoms over time

Figure 4.1 illustrates the two independent runs conducted for the RT and IN wild types. The wild type runs serve to ensure the reliability of all other runs involving complexes with inhibitors. Both runs for both proteins fluctuate between 0.15nm but do not exceed 0.5 nm, which is equivalent to 1.5 Å and 5 Å respectively. While there is no absolute rule for acceptable ranges of RMSD values, the general accepted value is < 3 Å (Manandhar et al., 2022). However, a study conducted by Aier et al. (2016) investigated the structural insights into the conformational stability of both wild type and mutant EZH2 receptor. Although this study used the OPLS force field, the RMSD C-alpha ranges calculated on the wild type protein are similar to mine, with a stabilization RMSD value of approximately 4.3 Å in a 50 ns MD run.

There is a distinct difference between run one and run two between 15 and 30 ns of the RT RMSD. However, both runs stabilize thereafter. In contrast, the IN wild type shows better consistency

between the two runs, with similar fluctuations. While the MD simulations conducted for both the RT and IN began with the same parameters and starting conditions, there are noticeable variations. These differences result from the stochastic nature of MD simulations. Therefore, while starting with the same initial conditions, the randomness of atomic motions causes diverse trajectories over time (Collette and King, 2021). As a result, it is critical to recognise that, although striving for reproducibility, MD simulations may produce variations, emphasizing the complexity and, to a certain extent, unpredictability in molecular dynamics.

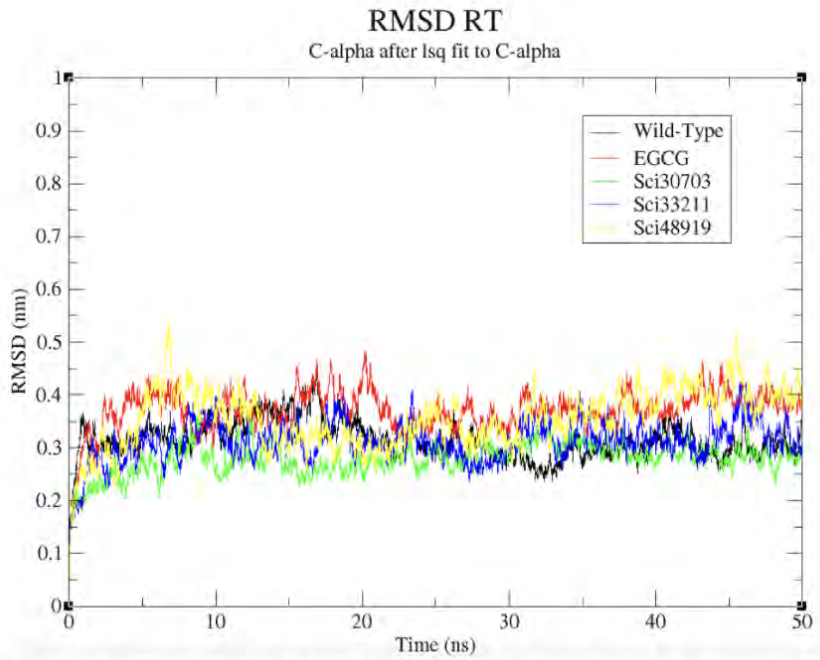

Figure 4.2:Depicts the RMSD graphs detailing the fluctuations in the C-alpha backbone of the RT wild type and EGCG, along with its derivatives. The color legend indicates the representation for each graph: Black) RT Wild Type, Red) EGCG, Green) EGCG derivative Sci30703, Blue) EGCG derivative Sci33211, Yellow) EGCG derivative Sci48919. All runs exhibit RMSD values that fluctuate below 0.5 nm.

Figure 4.2 represents the RMSD values illustrating the C-alpha variation from the wild type to EGCG and its derivatives. While all the graphs exhibit a consistent trend of fluctuating below 0.5 nm, RT in complex with Sci30703 shows minimal fluctuation throughout the 50 ns run and stabilizes below 0.3 nm (3 Å) which is an acceptable RMSD value (Manandhar et al., 2022). As EGCG serves as our reference compound, it shows higher RMSD fluctuation values compared to most of the protein-ligand complexes with Sci48919 being the exception. Nonetheless, the

variations across all runs remains below 2 Å, which suggests minimal structural differences between the wild type and the protein-ligand complexes (Li et al., 2018).

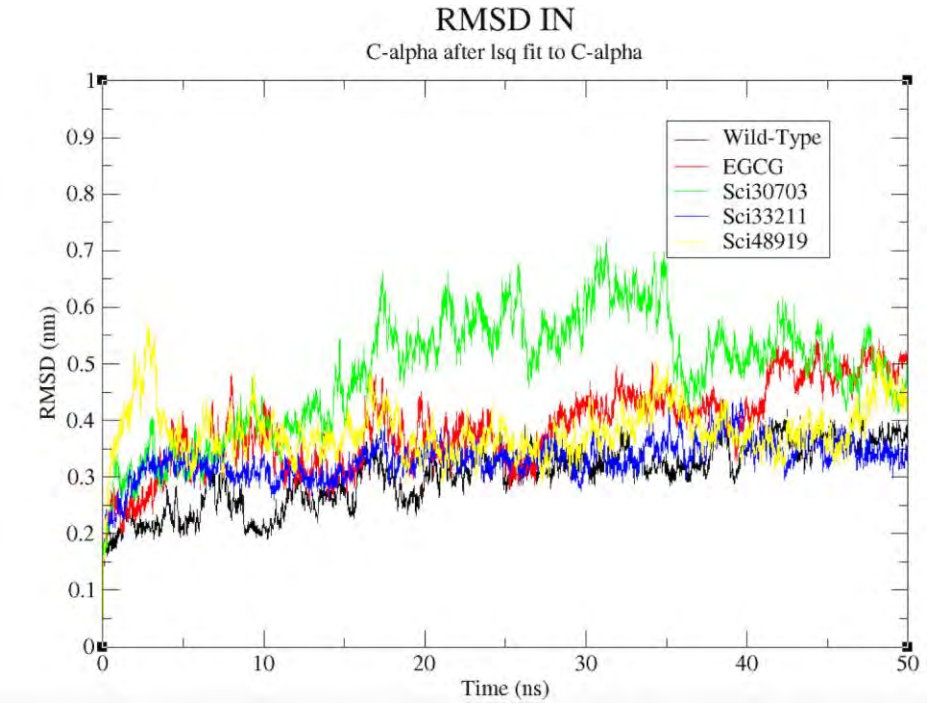

Figure 4.3:Depicts the RMSD graphs detailing the fluctuations in the C-alpha backbone of the IN wild type and EGCG, along with its derivatives. The color legend indicates the representation for each graph: Black) RT Wild Type, Red) EGCG, Green) EGCG derivative Sci30703, Blue) EGCG derivative Sci33211, Yellow) EGCG derivative Sci48919.

The RMSD values observed in Figure 4.3, show higher fluctuations than those observed in Figure 4.2. The wild type and protein complex with Sci33211 show similar minimal fluctuations throughout the 50 ns, with values that are slightly higher than 0.3 nm. IN in complex with Sci30703 shows the highest RMSD values which exceed 0.6 nm. This deviation is created because of the instability of the IN-Sci30703 complex. Instability of this complex is confirmed through visualization of the Sci30703 detaching from IN. Therefore, although the RT-Sci30703 complex shows the most stable RMSD values, the IN-Sci30703 complex is unstable. Thus because of this obvious disparity, the inhibitor Sci30703 is disqualified for being a possible dual action inhibitor.

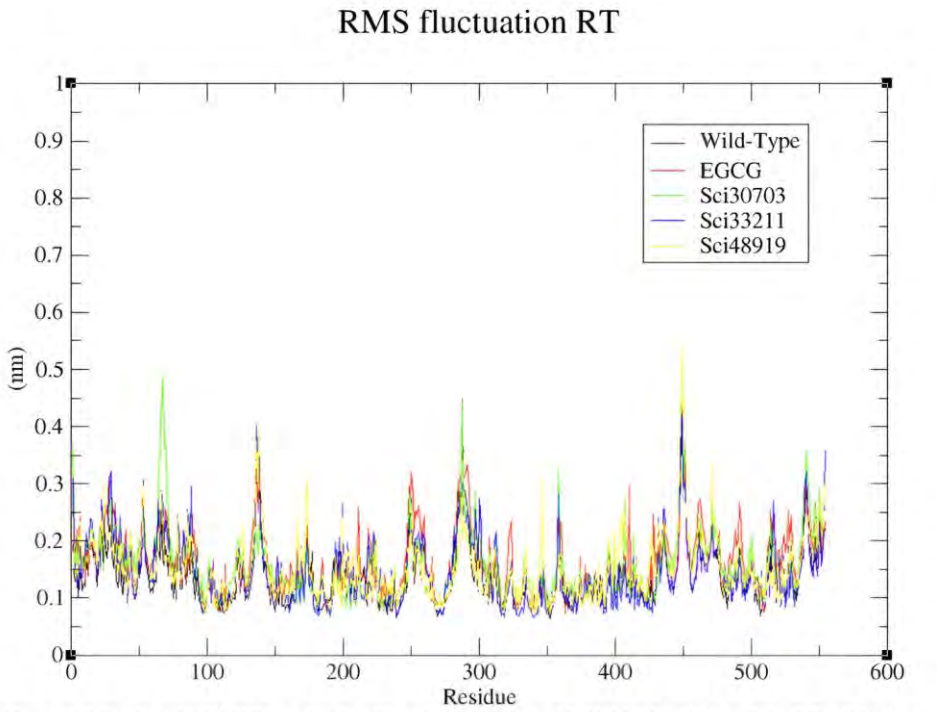

Figure 4.4:Depicts the RMSF graphs detailing the flexibility of individual residues of the RT wild type, RT in complex with EGCG, along with its derivatives. The color legend indicates the representation for each graph: Black) RT Wild Type, Red) EGCG, Green) EGCG derivative Sci30703, Blue) EGCG derivative Sci33211, Yellow) EGCG derivative Sci48919.

Moreover, RMSF is important because it illustrates the flexibility of individual residues. This information can offer insight into potentially important residues that undergo changes over time (Martínez, 2015). Most of the RT complexes with inhibitors share the same trend of residue fluctuation in relation to the wild type, with the exception of the peak that is created by the RT-Sci30703 complex, observed at the residue range 60 - 80. These residues form part of the β3-β4 hairpin loop. This loop is within the fingers subdomain of HIV-1 RT and regulates different aspects of enzyme activity. This includes the enzyme's ability to discriminate between the correct and incorrect nucleotides during DNA synthesis, its capacity for pyrophosphorolysis (the reversal of DNA synthesis) and its ability to remove existing DNA strands during replication. The β3-β4 loop is also targeted for mutations that give rise to resistance to nucleoside analogue medication, highlighting its role in both normal RT function and antiretroviral therapy response (Garforth et al., 2007).

The fluctuation observed in the β3-β4 loop of the RT-Sci30703 complex could possibly indicate allosteric effects of the inhibitor. Allosteric regulation occurs when the binding of a ligand (in this case the inhibitor Sci30703) at one site on the enzyme (RNase H residue 440-560) induces conformational changes at a distant site (β3-β4 loop residue 60-80) and potentially affects its function (Wenthur et al., 2014). These fluctuations may be a consequence of changes in the overall dynamics of the enzyme induced by inhibitor binding. Regarding the RNase H domain where the ligand is bound, all RMSF fluctuations are similar in this region including the wild type, with a sharp peak at approximately residues 445-460. Some of these residues are part of the metal binding pocket. This region plays a vital role in coordinating with the Mg^{2+} which are essential cofactors for the catalytic activity of the RNase H domain. Therefore, the observed fluctuations suggest that the ligand binding does not significantly alter the dynamic behavior of the metal-binding pocket which emphasizes its fundamental role in facilitating the catalytic activity within the RNase H domain (Mahboubi-Rabbani et al., 2021).

Additionally, Figure 4.5 shows the RMSF values of IN and IN complexed with EGCG and its derivatives. The IN-Sci30703 complex exhibits much higher fluctuations in comparison to the other complexes. This is consistent with the RMSD values observed in figure 4.3 which shows the overall instability of the ligand. Another noticeable disparity is the high peak (0.7 nm) created by the wildtype at a residue range of 130 - 155. This residue fluctuation is consistent with the location of a catalytic loop which is located at residues 139-152 (Lee et al., 2005). The catalytic loop is within the catalytic core domain of HIV-1 and it plays a crucial role in catalysis. Mutagenesis studies were conducted in which the flexible Glycine (Gly) within the loop was replaced with Alanine (Ala) and the flexibility of the loop decreased and affected catalytic function. This indicates the dynamics of the catalytic loop are essential for catalytic activity. The flexibility of the catalytic loop is necessary for accommodating substrate binding and facilitating the catalytic reactions carried out by IN, which includes the 3’ processing and the strand transfer reactions (Fitzkee et al., 2010).

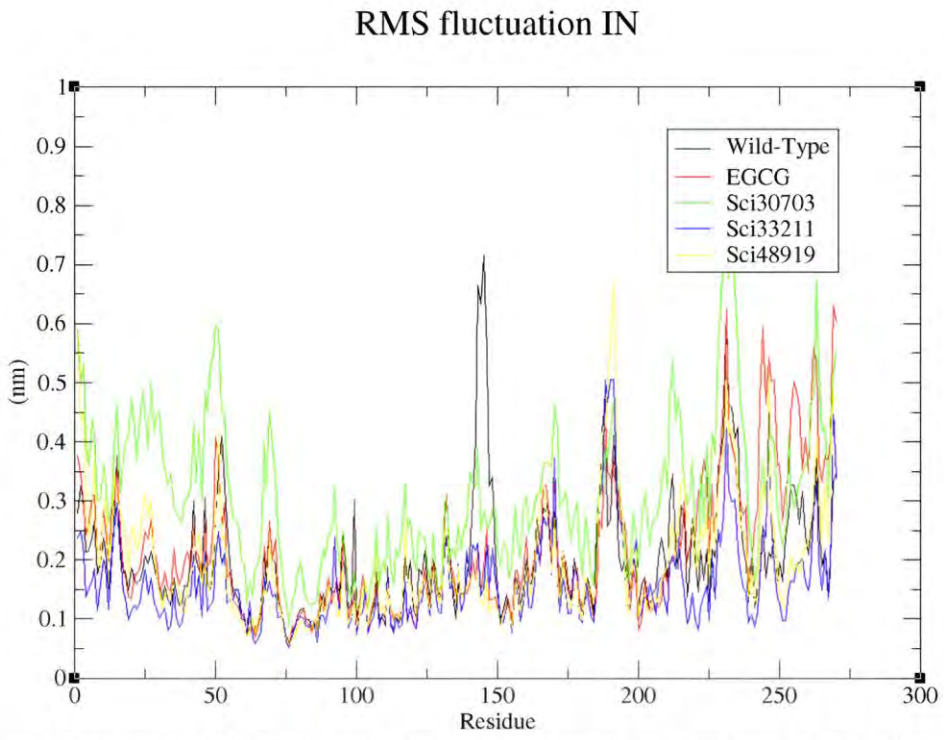

Figure 4.5:Depicts the RMSF graphs detailing the flexibility of individual residues of the IN wild type, IN in complex with EGCG, along with its derivatives. The color legend indicates the representation for each graph: Black) RT Wild Type, Red) EGCG, Green) EGCG derivative Sci30703, Blue) EGCG derivative Sci33211, Yellow) EGCG derivative Sci48919.

Therefore, the presence of the peak solely in the wild type and absent in the protein-ligand complexes, implies that the inhibitors effectively restrict the flexibility of the loop. The decreased flexibility of the loop would therefore reduce catalytic activity. Therefore, in vivo experimentation is needed to determine whether Sci33211 and Sci48919 inhibitors have the ability to act as dual action inhibitors and surpass the efficacy of EGCG.

4.5 CHAPTER SUMMARY

This chapter investigates the dynamics of HIV-1 RT and IN in complex with EGCG and its derivatives through MD simulations. The goal is to assess the performance of these derivatives compared to the reference compound EGCG in inhibiting both RT and IN.

The analysis utilized RMSD values to evaluate structural changes in proteins upon inhibitor docking. These findings revealed stable complexes for RT and mostly stable complexes of IN inhibitors except for Sci30703, which exhibited detachment from the protein. RMSF analysis showed elevated fluctuations in the RT-Sci30703 complex, at the residues forming β3-β4 hairpin loop, suggesting potential allosteric effects. While IN complexes effectively restrict the catalytic loop, which is crucial for strand transfer and 3' processing, the Sci30703 inhibitor was disqualified as a potential dual action inhibitor due to its instability in the IN complex. However, Sci33211 and Sci48919 in IN showed potential for reducing catalytic function, though further investigation is needed to confirm their dual inhibitory effects.

CONCLUDING REMARKS AND FUTURE WORK

The overall aim of this project was to identify compounds with dual inhibitory activity against HIV-1 RT and IN which surpass the effectiveness of EGCG which served as our reference compound through in silico techniques.

The compound library created with the aid of Scifinder resulted in the creation of 125 203 EGCG derivatives. However, once docking analysis was done and the correlation between the RT and IN docking was conducted, only 288 compounds were deemed suitable after a filtering threshold was set. This is less than 1% of the entire compound library which shows that there is a lack of focus in the compound library. Therefore reconstruction of the compound library would need to be conducted and the reactions created in order to form the EGCG derivatives would need to be reassessed. By doing so, this could potentially increase the number of compounds in down stream analysis and therefore increase the possibility of finding potential hit compounds. Additionally, conformational analysis was conducted on the ligands in relation to the DFT conformation (which is considered more accurate), revealing shortcomings in the XTB conformations that were selected for our study. Therefore, for further studies, DFT should be used to ensure more precise ligand conformations.

The 288 filtered compounds underwent QED analysis to evaluate their suitability as drug candidates. The three compounds with the highest QED values were selected and had superior drug-like properties compared to EGCG and RAL which is an FDA-approved drug that is used as a reference. These three compounds were docked to RT and IN and protein-ligand interactions were assessed. This assessment showed that the IN in complex with all the inhibitors had the ability to chelate one Mg^{2+} ion and interact with the active site residue ASP116. With regards to RT, all inhibitors interacted with the active site residues ASP443 and ASP549. The two inhibitors Sci33211 and Sci48919 demonstrated the ability to chelate two Mg^{2+} ions, indicating their potential to disrupt enzymatic function. However, there were no noticeable differences observed in RMSF fluctuations within the RNase H domain to confirm this.

Lastly, MD simulations were conducted to evaluate the protein-ligand system behavior, focusing on RMSD and RMSF. The RMSD analysis indicated stable complexes for RT, the IN-Sci30703

complex showed notably high RMSD values, indicating instability and confirmed by ligand detachment when observed in a visualizer. Consequently, Sci30703 was deemed unsuitable as a potential dual action inhibitor. However, RMSF analysis for RT showed no obvious fluctuations in the graphs in comparison to the wild type except for Sci30703, which is not considered for further analysis. RMSF of IN shows that all inhibitors have the ability to restrict the catalytic loop which is important for enzymatic function.Therefore, further in vitro studies are needed to be conducted on Sci33211 and Sci48919 to assess if the chelation of one metal ion and the interaction of select active site residues is sufficient for enzymatic disruption, potentially surpassing the efficacy of EGCG.

The discovery of dual inhibitors for HIV-1 holds promise in addressing challenges such as high pill burden, significant side effects and drug resistance. Therefore highlighting the importance of the continued research in this area to advance treatment strategies for HIV-1.

REFERENCES

Agrawal, A., DeSoto, J., Fullagar, J.L., Maddali, K., Rostami, S., Richman, D.D., Pommier, Y. and Cohen, S.M., 2012. Probing chelation motifs in HIV integrase inhibitors. *Proceedings of the National Academy of Sciences*, *109*(7), pp.2251-2256.

Ahmad, K.M. and Ahmad, I., 2019. Chapter 1-Herbal medicine: current trends and future prospects

Aier, I., Varadwaj, P.K. and Raj, U., 2016. Structural insights into conformational stability of both wild-type and mutant EZH2 receptor. *Scientific reports*, *6*(1), p.34984.

Arnittali, M., Rissanou, A.N. and Harmandaris, V., 2019. Structure of biomolecules through molecular dynamics simulations. *Procedia Computer Science*, *156*, pp.69-78.

Atanasov AG, Waltenberger B, Pferschy-Wenzig EM, et al. Discovery and resupply of pharmacologically active plant-derived natural products: A review. Biotechnol Adv. 2015;33(8):1582-1614.

Balasubramaniam, M., Pandhare, J. and Dash, C., 2019. Immune control of HIV. *Journal of life sciences (Westlake Village, Calif.)*, *1*(1), p.4.

Balunas, M.J. and Kinghorn, A.D., 2005. Drug discovery from medicinal plants. Life sciences, 78(5), pp.431-441.

Bannwarth, C., Caldeweyher, E., Ehlert, S., Hansen, A., Pracht, P., Seibert, J., Spicher, S. and Grimme, S., 2021. Extended tight-binding quantum chemistry methods. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *11*(2), p.e1493.

Barbieri, R., Coppo, E., Marchese, A., Daglia, M., Sobarzo-Sánchez, E., Nabavi, S.F. and Nabavi, S.M., 2017. Phytochemicals for human disease: An update on plant-derived compounds antibacterial activity. Microbiological research, 196, pp.44-68.

Bekker, H., 1996. *Molecular dynamics simulation methods revised*. Groningen: Rijksuniversiteit Groningen.

Bento, A.P., Hersey, A., Félix, E., Landrum, G., Gaulton, A., Atkinson, F., Bellis, L.J., De Veij, M. and Leach, A.R., 2020. An open source chemical structure curation pipeline using RDKit. *Journal of Cheminformatics*, *12*, pp.1-16.

Bickerton, G.R., Paolini, G.V., Besnard, J., Muresan, S. and Hopkins, A.L., 2012. Quantifying the chemical beauty of drugs. *Nature chemistry*, *4*(2), pp.90-98.

Bruininks, B.M., Wassenaar, T.A. and Vattulainen, I., 2023. Unbreaking Assemblies in Molecular Simulations with Periodic Boundaries. *Journal of chemical information and modeling*.

Cai, M., Zheng, R., Caffrey, M., Craigie, R., Clore, G.M. and Gronenborn, A.M., 1997. Solution structure of the N-terminal zinc binding domain of HIV-1 integrase. *Nature structural biology*, *4*(7), pp.567-577.

Case, D.R., Zubieta, J.P. and Doyle, R., 2020. The coordination chemistry of bio-relevant ligands and their magnesium complexes. *Molecules*, 25(14), p.3172. doi: 10.3390/molecules25143172.

Ceccherini Silberstein, F., Gago, F., Santoro, M., Gori, C., Svicher, V., Rodriguez Barrios, F., D'Arrigo, R., Ciccozzi, M., Bertoli, A., d'Arminio Monforte, A. and Balzarini, J., 2005. High sequence conservation of HIV-1 reverse transcriptase under drug pressure despite a continuous appearance of mutations. *JOURNAL OF VIROLOGY*, *79*(16), pp.10718-10729.

Centers for Disease Control and Prevention. (2021). HIV Transmission. [Online] Available at: https://www.cdc.gov/hiv/basics/hiv-transmission/body-fluids.html (Accessed: 13/08/2023)

Chawla, A., Wang, C., Patton, C., Murray, M., Punekar, Y., de Ruiter, A. and Steinhart, C., 2018. A review of long-term toxicity of antiretroviral treatment regimens and implications for an aging population. *Infectious diseases and therapy*, *7*, pp.183-195.

Chiu, T.K. and Davies, D.R., 2004. Structure and function of HIV-1 integrase. Current topics in medicinal chemistry, 4(9), pp.965-977.

Coley, C.W., Green, W.H. and Jensen, K.F., 2018. Machine learning in computer-aided synthesis planning. *Accounts of chemical research*, *51*(5), pp.1281-1289.

Collette, R. and King, J., 2021. Molecular dynamics simulations of radiation cascade evolution near cellular dislocation structures in additively manufactured stainless steels. *Journal of Nuclear Materials*, *549*, p.152872.

Corona, A., Di Leva, F.S., Thierry, S., Pescatori, L., Cuzzucoli Crucitti, G., Subra, F., Delelis, O., Esposito, F., Rigogliuso, G., Costi, R. and Cosconati, S., 2014. Identification of highly conserved residues involved in inhibition of HIV-1 RNase H function by diketo acid derivatives. *Antimicrobial agents and chemotherapy*, *58*(10), pp.6101-6110.

Danforth, K., Granich, R., Wiedeman, D., Baxi, S. and Padian, N., 2017. Global mortality and morbidity of HIV/AIDS. *Major infectious diseases*, *6*.

Davies, J.F., Hostomska, Z., Hostomsky, Z., Jordan, S.R. and Matthews, D.A., 1991. Crystal structure of the ribonuclease H domain of HIV-1 reverse transcriptase. *Science*, *252*(5002), pp.88-95.

de la Vega de Leon, A., Lounkine, E., Vogt, M. and Bajorath, J., 2017. Design of diverse and focused compound libraries. *Tutorials in Chemoinformatics*, pp.83-101.

De Luca, L., Pedretti, A., Vistoli, G., Barreca, M.L., Villa, L., Monforte, P. and Chimirri, A., 2003. Analysis of the full-length integrase–DNA complex by a modified approach for DNA docking. *Biochemical and Biophysical Research Communications*, *310*(4), pp.1083-1088.

Dewaker, V. and Prabhakar, Y.S., 2023. Molecular Dynamics Simulations of HDAC-ligand Complexes Towards the Design of New Anticancer Compounds. *Current Topics in Medicinal Chemistry*, *23*(29), pp.2743-2764.

Du, X., Li, Y., Xia, Y.L., Ai, S.M., Liang, J., Sang, P., Ji, X.L. and Liu, S.Q., 2016. Insights into protein–ligand interactions: mechanisms, models, and methods. *International journal of molecular sciences*, *17*(2), p.144.

Duarte Ramos Matos, G., Pak, S. and Rizzo, R.C., 2023. Descriptor-Driven de Novo Design Algorithms for DOCK6 Using RDKit. *Journal of Chemical Information and Modeling*, *63*(18), pp.5803-5822.

Eberhardt, J., Santos-Martins, D., Tillack, A.F., Forli, S. (2021). AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. Journal of Chemical Information and Modeling.

Engelman, A. and Cherepanov, P., 2012. The structural biology of HIV-1: mechanistic and therapeutic insights. Nature Reviews Microbiology, 10(4), pp.279-290.

Eggleton, J.S. and Nagalli, S., 2020. Highly active antiretroviral therapy (HAART).

Ferguson, M.R., Rojo, D.R., Von Lindern, J.J. and O'Brien, W.A., 2002. HIV-1 replication cycle. *Clinics in laboratory medicine*, *22*(3), pp.611-635.

Fitzkee, N.C., Masse, J.E., Shen, Y., Davies, D.R. and Bax, A., 2010. Solution conformation and dynamics of the HIV-1 integrase core domain. *Journal of Biological Chemistry*, *285*(23), pp.18072-18084.

Forli, S., Huey, R., Pique, M.E., Sanner, M.F., Goodsell, D.S. and Olson, A.J., 2016. Computational protein–ligand docking and virtual drug screening with the AutoDock suite. *Nature protocols*, *11*(5), pp.905-919.

Gabrielson, S.W., 2018. SciFinder. *Journal of the Medical Library Association: JMLA*, *106*(4), p.588.

Garforth, S.J., Kim, T.W., Parniak, M.A., Kool, E.T. and Prasad, V.R., 2007. Site-directed mutagenesis in the fingers subdomain of HIV-1 reverse transcriptase reveals a specific role for the β3–β4 hairpin loop in dNTP selection. *Journal of molecular biology*, *365*(1), pp.38-49.

German Advisory Committee Blood (Arbeitskreis Blut), Subgroup ‘Assessment of Pathogens Transmissible by Blood', 2016. Human immunodeficiency virus (HIV). *Transfusion medicine and hemotherapy*, *43*(3), pp.203-222.

Ghahremanian, S., Rashidi, M.M., Raeisi, K. and Toghraie, D., 2022. Molecular dynamics simulation approach for discovering potential inhibitors against SARS-CoV-2: A structural review. *Journal of Molecular Liquids*, *354*, p.118901.

Gill, M.S.A., Hassan, S.S. and Ahemad, N., 2019. Evolution of HIV-1 reverse transcriptase and integrase dual inhibitors: Recent advances and developments. *European journal of medicinal chemistry*, *179*, pp.423-448

Giovanetti, M., Ciccozzi, M., Parolin, C. & Borsetti, A., 2020. Molecular epidemiology of HIV-1 in African countries: A comprehensive overview. *Pathogens*, 9(12), p.1072. Available at: https://doi.org/10.3390/pathogens9121072

Goldschmidt, V., Didierjean, J., Ehresmann, B., Ehresmann, C., Isel, C. and Marquet, R., 2006. Mg 2+ dependency of HIV-1 reverse transcription, inhibition by nucleoside analogues and resistance. *Nucleic acids research*, *34*(1), pp.42-52

González, M.A., 2011. Force fields and molecular dynamics simulations. *École thématique de la Société Française de la Neutronique*, *12*, pp.169-200.

Guo, Y. (2019). "Introduction of the Amber force field".

Hajimahdi, Z. and Zarghi, A., 2016. Progress in HIV-1 integrase inhibitors: A review of their chemical structure diversity. *Iranian journal of pharmaceutical research: IJPR*, *15*(4), p.595.

Hernández-Santoyo, A., Tenorio-Barajas, A.Y., Altuzar, V., Vivanco-Cid, H. and Mendoza-Barrera, C., 2013. Protein-protein and protein-ligand docking. *Protein engineering-technology and application*, pp.63-81.

Hollingsworth, S.A. and Dror, R.O., 2018. Molecular dynamics simulation for all. *Neuron*, *99*(6), pp.1129-1143.

Hopkins, E., Sanvictores, T. and Sharma, S., 2022. Physiology, acid base balance. In *StatPearls [Internet]*. StatPearls Publishing.

Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A. and Simmerling, C., 2006. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics*, *65*(3), pp.712-725.

Isamura, B.K. and Lobb, K.A., 2022. AMADAR: a python-based package for large scale prediction of Diels–Alder transition state geometries and IRC path analysis. *Journal of Cheminformatics*, *14*(1), p.39.

Iyidogan, P. and Anderson, K.S., 2014. Current perspectives on HIV-1 antiretroviral drug resistance. *Viruses*, *6*(10), pp.4095-4139.

Jóźwik, I.K., Passos, D.O. and Lyumkis, D., 2020. Structural biology of HIV integrase strand transfer inhibitors. *Trends in pharmacological sciences*, *41*(9), pp.611-626.

Junqueira, D.M. & Almeida, S.E.M., 2016. HIV-1 subtype B: Traces of a pandemic. *Virology*, 495, pp.173-184. Available at: https://doi.org/10.1016/j.virol.2016.05.003

Kastritis, P.L. and Bonvin, A.M., 2013. On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *Journal of The Royal Society Interface*, *10*(79), p.20120835.

Kirchhoff, F., 2013. HIV life cycle: overview. *Encyclopedia of AIDS*, pp.1-9.

Kokkonen, P., Bednar, D., Pinto, G., Prokop, Z. and Damborsky, J., 2019. Engineering enzyme access tunnels. *Biotechnology advances*, *37*(6), p.107386.

Konovalov, A., Symons, B.C. and Popelier, P.L., 2021. On the many-body nature of intramolecular forces in FFLUX and its implications. *Journal of Computational Chemistry*, *42*(2), pp.107-116.

Lee, M.C., Deng, J., Briggs, J.M. and Duan, Y., 2005. Large-scale conformational dynamics of the HIV-1 integrase core domain and its catalytic loop mutants. *Biophysical journal*, *88*(5), pp.3133-3146

Le, T.T.V. and Do, P.C., 2022. Molecular docking study of various Enterovirus-A71 3C protease proteins and their potential inhibitors. *Frontiers in Microbiology*, 13, p.987801

Li, Y., Cong, Y., Feng, G., Zhong, S., Zhang, J.Z., Sun, H. and Duan, L., 2018. The impact of interior dielectric constant and entropic change on HIV-1 complex binding free energy prediction. *Structural Dynamics*, *5*(6).

Liao, C. and Nicklaus, M.C., 2010. Tautomerism and Magnesium Chelation of HIV-1 Integrase Inhibitors: A Theoretical Study. *ChemMedChem*, *5*(7), pp.1053-1066.

Lovrić, M., Molero, J.M. and Kern, R., 2019. PySpark and RDKit: moving towards big data in cheminformatics. *Molecular informatics*, *38*(6), p.1800082.

Lu, D.Y., Wu, H.Y., Yarla, N.S., Xu, B., Ding, J. and Lu, T.R., 2018. HAART in HIV/AIDS treatments: future trends. *Infectious Disorders-Drug Targets (Formerly Current Drug Targets-Infectious Disorders)*, *18*(1), pp.15-22.

Maga, G., 2013. Reverse Transcriptase,pp.222-223

Mahboubi-Rabbani, M., Abbasi, M., Hajimahdi, Z. and Zarghi, A., 2021. HIV-1 Reverse Transcriptase/Integrase dual inhibitors: A review of recent advances and structure-activity relationship studies. Iranian Journal of Pharmaceutical Research: IJPR, 20(2), p.333.

Mahgoub, R.E., Atatreh, N. and Ghattas, M.A., 2022. Using filters in virtual screening: A comprehensive guide to minimize errors and maximize efficiency.

Manandhar, S., Sankhe, R., Priya, K., Hari, G., Kumar B, H., Mehta, C.H., Nayak, U.Y. and Pai, K.S.R., 2022. Molecular dynamics and structure-based virtual screening and identification of natural compounds as Wnt signaling modulators: possible therapeutics for Alzheimer's disease. *Molecular Diversity*, *26*(5), pp.2793-2811.

Martínez, L., 2015. Automatic identification of mobile and rigid substructures in molecular dynamics simulations and fractional structural fluctuation analysis. *PloS one*, *10*(3), p.e0119264.

Meng, X.Y., Zhang, H.X., Mezei, M. and Cui, M., 2011. Molecular docking: a powerful approach for structure-based drug discovery. *Current computer-aided drug design*, *7*(2), pp.146-157.

Minibaeva, G., Ivanova, A. and Polishchuk, P., 2023. EasyDock: customizable and scalable docking tool. *Journal of Cheminformatics*, 15, p.102.

Miri, L., Bouvier, G., Kettani, A., Mikou, A., Wakrim, L., Nilges, M. and Malliavin, T.E., 2014. Stabilization of the integrase-DNA complex by Mg2+ ions and prediction of key residues for binding HIV-1 integrase inhibitors. *Proteins: Structure, Function, and Bioinformatics*, *82*(3), pp.466-478.

Moore, C., Zhang, D., Rousseau, R., Glezakou, V.A. and McEwen, J.S., 2022. Determining the Adsorption Energetics of 2, 3-Butanediol on RuO2 (110): Coupling First-Principles Calculations With Global Optimizers. *Frontiers in Energy Research*, *9*, p.781001

Morris, G.M. and Lim-Wilby, M., 2008. Molecular docking. *Molecular modeling of proteins*, pp.365-382.

Müller-Kuhrt, L., 2003. Putting nature back into drug discovery. *Nature biotechnology*, *21*(6), pp.602-602.

Navanietha Krishnaraj, R., David, A. and Sani, R.K., 2017. Fundamentals of enzymatic processes. *Extremophilic enzymatic processing of lignocellulosic feedstocks to bioenergy*, pp.5-29.

Ng, M.C., Fong, S. and Siu, S.W., 2015. PSOVina: The hybrid particle swarm optimization algorithm for protein–ligand docking. *Journal of bioinformatics and computational biology*, *13*(03), p.1541007.

Patel, P.H. and Zulfiqar, H., 2023. Reverse Transcriptase Inhibitors. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing. Available from: https://www.ncbi.nlm.nih.gov/books/NBK551504/ [Accessed 14 July 2023]

Petersen, F. and Amstutz, R. eds., 2008. *Natural Compounds as Drugs, Volume II* (Vol. 66). Springer Science & Business Media.

Pyra, M.N., Haberer, J.E., Hasen, N., Reed, J., Mugo, N.R. and Baeten, J.M., 2019. Global implementation of PrEP for HIV prevention: setting expectations for impact. *Journal of the International AIDS Society*, *22*(8), p.e25370.

Ridley, D.D., 2002. *Information Retrieval: SciFinder and SciFinder Scholar*. John Wiley & Sons.

Rocchi, C., Gouet, P., Parissi, V. and Fiorini, F., 2022. The C-Terminal Domain of HIV-1 Integrase: A Swiss Army Knife for the Virus?. *Viruses*, *14*(7), p.1397

Salo-Ahen, O.M., Alanko, I., Bhadane, R., Bonvin, A.M., Honorato, R.V., Hossain, S., Juffer, A.H., Kabedev, A., Lahtela-Kakkonen, M., Larsen, A.S. and Lescrinier, E., 2020. Molecular dynamics simulations in drug discovery and pharmaceutical development. *Processes*, *9*(1), p.71.

Santerre, M., Wang, Y., Arjona, S., Allen, C. & Sawaya, B.E., 2019. Differential contribution of HIV-1 subtypes B and C to neurological disorders: Mechanisms and possible treatments. *AIDS Reviews*, 21(2), pp.76-83. Available at: https://doi.org/10.24875/AIDSRev.19000051

Sanna, C., Rigano, D., Corona, A., Piano, D., Formisano, C., Farci, D., Franzini, G., Ballero, M., Chianese, G., Tramontano, E. and Taglialatela-Scafati, O., 2019. Dual HIV-1 reverse transcriptase and integrase inhibitors from Limonium morisianum Arrigoni, an endemic species of Sardinia (Italy). Natural product research, 33(12), pp.1798-1803.

Sarafianos, S.G., Marchand, B., Das, K., Himmel, D.M., Parniak, M.A., Hughes, S.H., Arnold, E. (2009). Structure and Function of HIV-1 Reverse Transcriptase: Molecular Mechanisms of Polymerization and Inhibition. Journal of Molecular Biology, 385(3), 693-713.

Shin, W.H., Zhu, X., Bures, M.G. and Kihara, D., 2015. Three-dimensional compound comparison methods and their application in drug discovery. *Molecules*, *20*(7), pp.12841-12862.

Schreiner, W., Karch, R., Knapp, B. and Ilieva, N., 2012. Relaxation estimation of RMSD in molecular dynamics immunosimulations. *Computational and mathematical methods in medicine*, *2012*.

Schultz, S.J. and Champoux, J.J., 2008. RNase H activity: structure, specificity, and function in reverse transcription. *Virus research*, *134*(1-2), pp.86-103.

SciFinder®, https://scifinder-n.cas.org, accessed 2023-06-24.
Sliwoski, G., Kothiwale, S., Meiler, J. and Lowe, E.W., 2014. Computational methods in drug discovery. *Pharmacological reviews*, *66*(1), pp.334-395.

Sohail, M., Levitan, E.B., Rana, A.I., Heath, S.L., Rastegar, J., Kempf, M.C. and Long, D.M., 2020. Estimating the first 90 of the UNAIDS 90-90-90 Goal: a review. *Journal of the International Association of Providers of AIDS Care (JIAPAC)*, *19*, p.2325958220919290.

Stanford University. (2023). HIV Drug Resistance Database. Stanford University. Retrieved July 16, 2023, from [https://hivdb.stanford.edu/dr-summary/resistance-notes/NNRTI/].

Stanford University. (2023). Major integrase inhibitor (INSTI) resistance mutations. Stanford University. Retrieved from [https://hivdb.stanford.edu/dr-summary/resistance-notes/INSTI/]

Strasfeld, L. and Chou, S., 2010. Antiviral drug resistance: mechanisms and clinical implications. *Infectious Disease Clinics*, *24*(2), pp.413-437.

Sued, O., Figueroa, M.I. and Cahn, P., 2016. Clinical challenges in HIV/AIDS: Hints for advancing prevention and patient management strategies. *Advanced Drug Delivery Reviews*, *103*, pp.5-19.

Tang, M.W. and Shafer, R.W., 2012. HIV-1 antiretroviral resistance: scientific principles and clinical applications. Drugs, 72, pp.e1-e25.

Tian, L., Kim, M.S., Li, H., Wang, J. and Yang, W., 2018. Structure of HIV-1 reverse transcriptase cleaving RNA in an RNA/DNA hybrid. *Proceedings of the National Academy of Sciences*, *115*(3), pp.507-512.

Trott, O. and Olson, A.J., 2010. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, *31*(2), pp.455-461.

UNAIDS, 2023. Global HIV & AIDS statistics — 2022 fact sheet. [online] Available at: https://www.unaids.org/sites/default/files/media_asset/UNAIDS_FactSheet_en.pdf [Accessed 8 July 2023].

UNAIDS. (2022). Data Book 2022. [Online] Available at: https://www.unaids.org/sites/default/files/media_asset/data-book-2022_en.pdf (Accessed: 15 August 2023).

User Guide to Semiempirical Tight Binding — xtb doc 2023 documentation (no date). Available at: https://xtb-docs.readthedocs.io/.

Van Mourik, T., Bühl, M. and Gaigeot, M.P., 2014. Density functional theory across chemistry, physics and biology. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *372*(2011), p.20120488.

Verma, J., Subbarao, N. and Rajala, M.S., 2020. Envelope proteins as antiviral drug target. *Journal of Drug Targeting*, *28*(10), pp.1046-1052.

Wang, J., Wolf, R.M., Caldwell, J.W., Kollman, P.A. and Case, D.A., 2004. Development and testing of a general amber force field. *Journal of computational chemistry*, *25*(9), pp.1157-1174

Wang, Z., Zheng, L., Wang, S., Lin, M., Wang, Z., Kong, A.W.K., Mu, Y., Wei, Y. and Li, W., 2023. A fully differentiable ligand pose optimization framework guided by deep learning and a traditional scoring function. *Briefings in Bioinformatics*, *24*(1), p.bbac520.

Wenthur, C.J., Gentry, P.R., Mathews, T.P. and Lindsley, C.W., 2014. Drugs for allosteric sites on receptors. *Annual review of pharmacology and toxicology*, *54*, pp.165-184.

Wermuth, C.G., 2006. Similarity in drugs: reflections on analogue design. *Drug Discovery Today*, *11*(7-8), pp.348-354.

Yu, W. and MacKerell, A.D., 2017. Computer-aided drug design methods. *Antibiotics: methods and protocols*, pp.85-106.

Yüksel, T., 2017. *Undergraduate Students' Understanding of Concepts Foundational to the Learning of Quantum Mechanics (QM) via Models-Based Inquiry* (Doctoral dissertation, Purdue University).

SUPPLEMENTARY MATERIAL

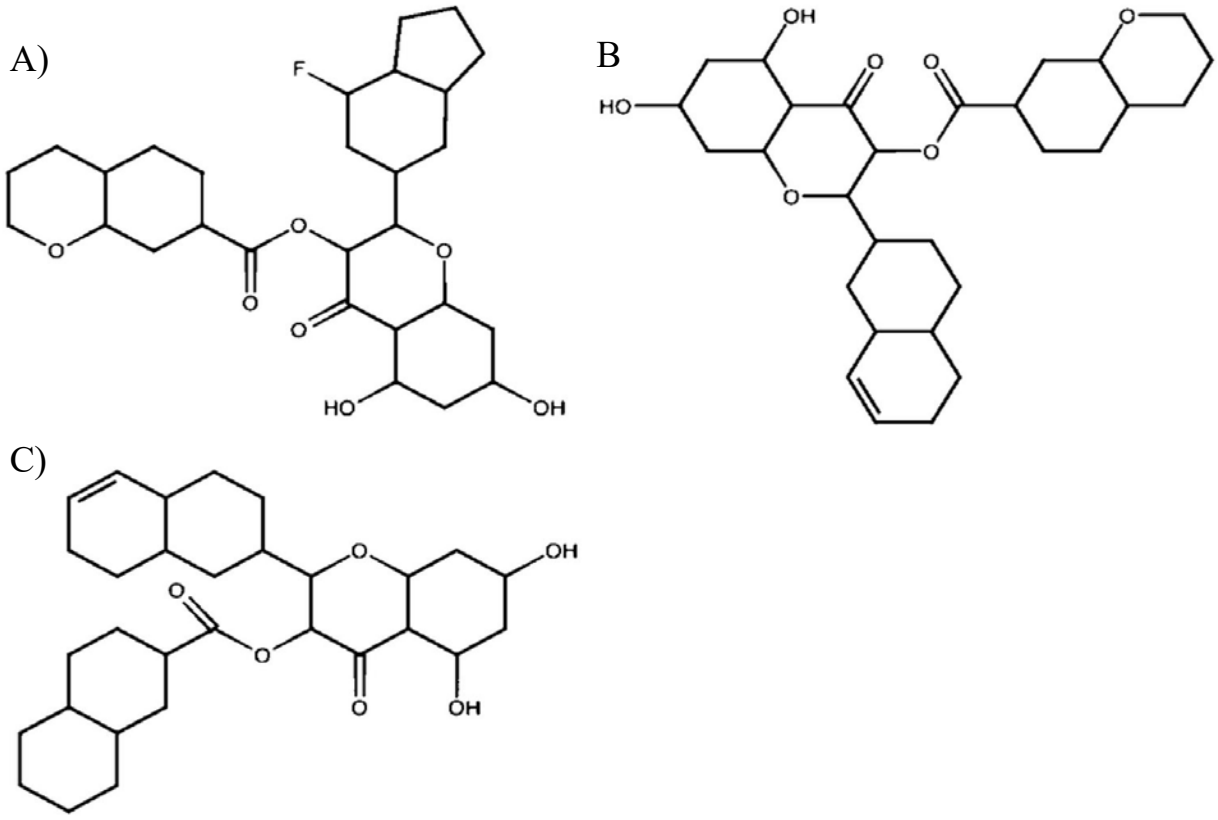

Figure S1:Chemical structures of the studied inhibitors. A)Sci30703, B) Sci33211, C) Sci48919. All structures were obtained from the SciFinder compound library

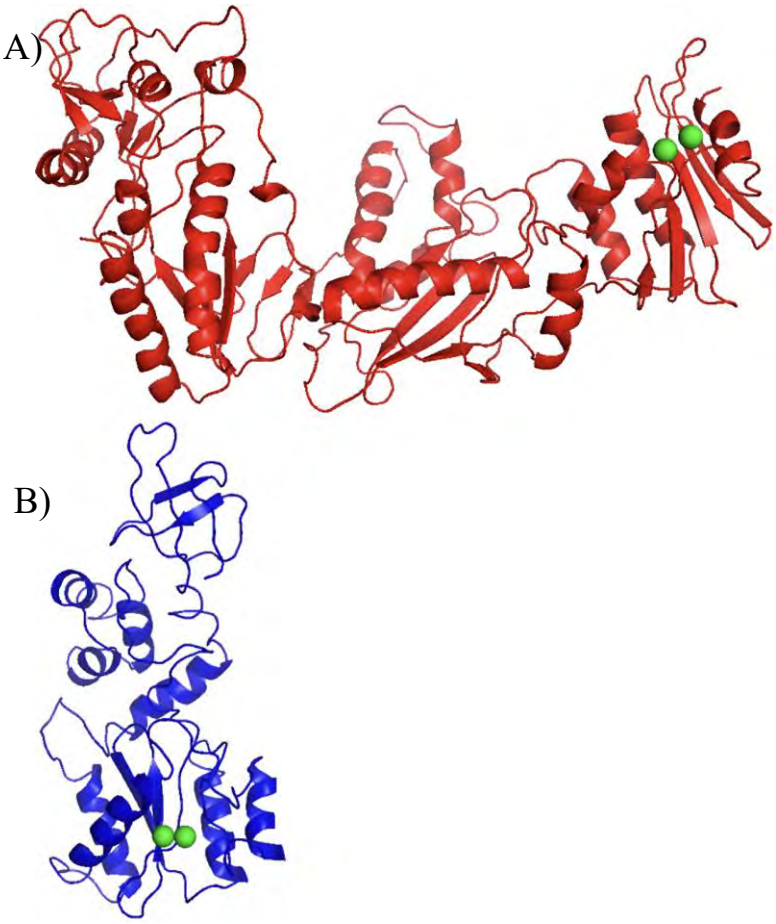

Figure S2: 3D structures of RT and IN. A)Reverse transcriptase with magnesium ions depicted in green. B) Integrase with the magnesium ions depicted in green