

Structural studies of HTLV-1 reverse transcriptase for antiviral drug design

Thesis submitted in partial fulfilment of
the requirements for the degree of
Master of Biomedical Sciences by

Brecht BOLLAERT

Supervisor: Prof. dr. Kalyan DAS
Instructor: MChem. Adéla NOVOTNÁ

Laboratory of Virology and Chemotherapy,
Rega Institute for Medical Research

Leuven, 2023-2024

This Master's Thesis is an exam document. Possibly assessed errors were not corrected after the defense. In publications, references to this thesis may only be made with written permission of the supervisor(s) mentioned on the title page.

Structural studies of HTLV-1 reverse transcriptase for antiviral drug design

Thesis submitted in partial fulfilment of
the requirements for the degree of
Master of Biomedical Sciences by

Brecht BOLLAERT

Supervisor: Prof. dr. Kalyan DAS
Instructor: MChem. Adéla NOVOTNÁ

Laboratory of Virology and Chemotherapy,
Rega Institute for Medical Research

Leuven, 2023-2024

Preface

I would like to thank **Professor Kalyan Das** for welcoming me and giving me the opportunity to work in his group on a very interesting and challenging project. Also, to have given me the opportunity and his trust to present our work in front of a group of established researchers in the field of virology and HTLV-1. In addition, I would like to extend a special thank you to **Adéla Novotná**, my daily instructor, for all the support and valuable feedback she has given. This extends to experimental design, proof-reading and advice on the next steps after graduation. Also, to have given me the opportunity to work independently and to learn from the mistakes I made. I am grateful towards the entire team: **Mizar Oliva (PhD)**, **Julien Olivet (PhD)**, **Brent De Wijngaert**, **Quinten Goovaerts**, **Neesha Rajesh Shewakramani**, **Ming Hao Chang** and, last but not least, **Ine Van Hileghem**. Together, these people make a warm and welcoming work place, where “no” was never an answer to any of my many questions. I am truly convinced that a group like KD lab is one-of-a-kind, and I am honoured to have been a part of it for almost one full year. I want to thank Professor Kalyan Das, Adéla Novotná, Dr. Mizar Oliva and Quinten Goovaerts for proof-reading my manuscript, completely or partially, and their many helpful suggestions.

I would like to mention two people who have been part of the group only briefly but have made an everlasting impression. **Calvin Gordon (PhD candidate at the University of Alberta)** and **Dimitrios Xynos**, I would like to thank both of you for the many laughs and discussions we had throughout your stay.

I want to say thank you to **Evelien Vanderlinden (PhD.)**, **Clément Heymann** and **Gorrit-Cor Lootsma**, as it is because of you that I got fascinated by HTLV-1 during my first lab rotation.

Finally, I want to thank my family. First, **my parents** and **my girlfriend** for supporting me throughout these five years. For always helping me get back up during the hard times and to celebrate with during the good times. Also, **my grandparents**, specifically grandpa Willy, for always attentively listening to me rattle on about my experiences and the science topics I find interesting.

Table of contents

Preface	iv
Table of contents	v
List of abbreviations	viii
Abstract.....	xi
1. Introduction	1
1.1 The HTLV-1 virus and its genome	1
1.2 HTLV-1 replication cycle	3
1.3 Reverse transcription	4
1.3.1 Pre-initiation	5
1.3.2 Reverse transcription initiation	6
1.3.3 Reverse transcription pausing	7
1.4 Reverse transcriptase	8
1.4.1 Structure and conserved motifs	9
1.4.2 The molecular mechanism of dNTP incorporation	11
1.5 Aims	13
2. Materials and methods	13
2.1 HTLV-1 constructs cloning and validation	13
2.1.1 HTLV-1 pET42a (+) RT & HTLV-1 pET42a (+) pol constructs	14
2.1.2 HTLV-1 pET-Duet-1 p62/p49 in-house	15
2.2 Recombinant protein expression in <i>E. coli</i>	15
2.2.1 Transformation.....	15
2.2.2 HTLV-1 RT expression in <i>E. coli</i>	16
2.2.3 Protein expression evaluation by SDS-PAGE	16
2.3 Protein purification	18
2.4 Western blot	18
2.5 Polymerase chain reaction	19
2.6 RT-qPCR	19
2.6.1 RNA extraction and reverse transcription	19
2.6.2 Quantitative PCR.....	19

2.7 Bioinformatics.....	19
2.8 Statistics.....	20
3. Results.....	21
3.1 Sequence analysis	21
3.1.1 General parameters and characteristics	21
3.1.2 Phylogeny	22
3.1.3 Domain identification	24
3.1.4 Pairwise sequence alignment.....	26
3.1.5 Secondary structure prediction	27
3.2 In silico HTLV-1 RT model	29
3.2.1 Homology-based structure prediction	29
3.2.2 Ab initio structure prediction	30
3.2.3 Comparative quality assessment.....	32
3.2.4 Spatial Model comparison	34
3.2.5 Mg ²⁺ docking	35
3.2.6 DNA docking.....	36
3.3 Antiviral compound docking.....	37
3.4 Design and validation of protein expression constructs	40
3.4.1 Plasmid design.....	40
3.4.2 Plasmid construction and quality control	41
3.5 Recombinant protein expression	43
3.5.1 IPTG protein induction	43
3.5.2 Western blot shows low level expression of HTLV-1 RT.....	44
3.5.3 The effect of IPTG concentration on HTLV-1 RT expression	45
3.6 HTLV-1 RT purification	46
3.6.1 Purification of HTLV-1 RT via affinity chromatography	46
3.7 HTLV-1 RT mRNA expression in <i>E. coli</i>	47
3.7.1 RT-qPCR	47
3.7.2 HTLV-1 5' mRNA shows highly stable secondary structure	49
4. Discussion	50
4.1 Great homology with HIV-1 RT and HERV-K RT	50

4.2 In silico HTLV-1 RT model	50
4.3 PoC: Antiviral compound docking	52
4.4 Plasmid design and protein expression	52
4.5 Successful protein purification	53
4.6 GC-rich hairpin loop likely at the basis of low protein expression.....	53
5. Conclusion.....	54
Bibliography	55
Supplementary material	I
SI. Overview of used informatics tools	I
SII. Multiple sequence alignment	III
SIII. Homology-based structure prediction per model scores	VI
SIV. Comparative alignment: Alignment scores	VII
SV. Comparative alignment: Pruned pairs	VII
SVI. Comparative alignment: RMSD for pruned pairs	VIII
SVII. Comparative alignment: All pairs	VIII
SVIII. Comparative alignment: RMSD for all pairs.....	IX
SIX. Ca-Ca in HIV-1 RT active site for important residues.....	IX
SX. Ca-Ca in HTLV-1 RT (AF2) active site for important residues	X
SXI. HTLV-1 RT expression in various bacterial strains	XI
SXII. Amplification plot: HTLV-1 pET15b	XII
SXIII. Amplification plot: HTLV-1 pET42a (+) RT.....	XII
SXIV. Amplification plot: HIV-1 pET-Duet-1 p66/p51	XIII
SXV. Amplification plot: HTLV-1 pET-Duet-1 p62/p49 in-house	XIII
SXVI. Amplification plot: HTLV-1 pET-Duet-1 p62/p49 Gs	XIV

List of abbreviations

+3Dis	+3 Displaced
+3Int	+3 Intermediary
+3Pre	+3 Pre-translocation state
3TC	Lamivudine
AAs	Amino acids
AF2	AlphaFold2
ALV	Avian leukosis virus
Amp	Ampicillin
AmpR	Beta-lactamase ampicillin resistance gene
ATL	Adult T-cell leukemia
BEV	Baboon endogenous virus
BIV	Bovine immunodeficiency virus
Ca	Capsid
CAEV	Caprine arthritis encephalitis virus
cPPT	Central polypurine tract
Cryo-EM	Cryogenic electron microscopy
CysG	Siroheme synthase
d4T	Stavudine
dNTP	Deoxyribonucleoside triphosphate
DP	Diphosphate
DTT	Dithiothreitol
<i>E. coli</i>	<i>Escherichia coli</i>
EIAV	Equine infectious anemia virus
Env	Envelope
FEV	Feline endogenous virus
FIV	Feline immunodeficiency virus
FLV	Feline leukemia virus
GHAP	Golden hamster intracisternal A-particle
Glut-1	Glucose transporter 1
GO	Gene ontology
gRNA	Genomic (+)ssRNA
GST	Glutathione-S-transferase

H1	Helix 1
H2	Helix 2
HAM/TSP	HTLV-1 associated myelopathy/tropical spastic paraparesis
HBZ	HTLV-1 bZip factor
HcaT	3-Phenylpropionate transporter
HIV-1	Human immunodeficiency virus type 1
HIV-2	Human immunodeficiency virus type 2
HRP	Horseradish peroxidase
HSPG	Heparan sulphate proteoglycans
HTLV-(1,2,3)	Human T-cell leukemia virus (type 1, 2, 3)
Int	Integrase
IPTG	Isopropyl β -D-1-thiogalactopyranoside
JDV	Jembrana disease virus
JSR	Jaagsiekte sheep retrovirus
Kan	Kanamycin
KanR	Aminoglycoside O-phosphotransferase kanamycin resistance gene
LTRs	Long terminal repeats
Ma	Matrix
MeOH	Methanol
MIAP	Mouse intracisternal A-particle
MMLV	Moloney murine leukemia virus
MMTV	Mouse mammary tumor virus
MPMV	Mason-Pfizer monkey virus
MSA	Multiple sequence alignment
Nc	Nucleocapsid
NNRTI	Non-nucleoside reverse transcriptase inhibitors
NRP-1	Neuropilin-1
NRTI	Nucleoside reverse transcriptase inhibitors
OL	Ovine lentivirus
ORF	Open reading frame
PAE	Predicted aligned error
PBS	Primer binding site
PCR	Polymerase chain reaction
PDB	Protein databank

PIC	Pre-integration complex
pLDDT	Predicted local distance difference test
MPMA	(9-[9(R)-2-(phosphonomethoxy)propyl]adenine; Tenofovir
PPT	Polypurine tracts
Pr	Protease
P _{α, β, γ}	(α, β, γ)-Phosphate
qPCR	Quantitative PCR
RBS	Ribosome binding site
RNase H	Ribonuclease H
rNTP	Ribonucleoside triphosphates
RSV	Rous sarcoma virus
RT	Reverse transcriptase
RTIC	Reverse transcription initiation complex
RTRV	Reticuloendotheliosis virus
SBDD	Structure based drug design
SDS-PAGE	Sodium dodecyl sulphate polyacrylamide gel electrophoresis
SIV	Simian immunodeficiency virus
SMR	Squirrel monkey retrovirus
SR_1	Simian retrovirus 1
SR_2	Simian retrovirus 2
STLV-(1,2,3)	Simian T-cell leukemia virus (type 1, 2, 3)
SU	Surface subunit
T7RNAP	T7 RNA polymerase
TM	Transmembrane subunit
TP	Triphosphate
tRNA	Transfer-RNA
Vdw	Van der Waals
VMV	Visna/maedi virus
WB	Western blot
WMSV	Woolly monkey sarcoma virus
XMLV	Xenotropic MuLV-related virus
ZDV	Zidovudine
β-ME	Beta-mercaptoethanol

Abstract

Human T-cell leukaemia virus type 1 (HTLV-1) is a major problem in highly endemic regions and is the causative agent for adult T-cell leukaemia and HTLV-1-associated myelopathy/tropical spastic paraparesis. To date, no cure is available for either of these diseases, yet prevention of HTLV-1 infection has proven to be a viable option. The HTLV-1 reverse transcriptase (RT) enzyme provides an interesting drug target due to its vital function in the viral replication. In this thesis, we aim to obtain the structure of HTLV-1 RT via single-particle cryogenic electron microscopy (cryo-EM) and gain insight into the mechanisms it employs to perform its DNA polymerase function. In the first part, an extensive analysis was conducted on phylogeny, gene ontology, and homology, where we found HTLV-1 RT to have great similarity to both human immunodeficiency virus type 1 RT and human endogenous retrovirus K RT. Following this, a physiologically relevant model of HTLV-1 RT in complex with a nucleic acid dimer was constructed and used in an antiviral compound docking study. In the second part, multiple plasmids were designed to express HTLV-1 RT in various forms and were transformed into multiple *Escherichia coli* (*E. coli*) strains for recombinant protein expression. After induction, HTLV-1 RT was isolated using nickel-histidine affinity chromatography. Here, the recombinant protein yield remained very low, which prompted us to perform an in-depth analysis of the HTLV-1 RT encoding mRNA. Via RT-qPCR, significant variation in mRNA expression levels between constructs was observed, but the presence of HTLV-1 RT encoding mRNA was confirmed. In silico RNA folding of the HTLV-1 RT mRNA suggested the occurrence of a highly stable GC-rich hairpin loop at the 5' end of the mRNA. In conclusion, we substantiated the link between HIV-1 RT and HTLV-1 RT necessary to validate the effort in screening anti-HIV-1 RT compounds against HTLV-1 RT. Furthermore, by docking these compounds into a physiologically relevant in silico model of HTLV-1 RT, we showed that some of those would favourably bind to HTLV-1 RT's polymerase active site with sub-nanomolar affinity. Additionally, we successfully expressed and isolated HTLV-1 RT in low quantities. This low yield is most likely explained by the presence of a highly stable GC-rich hairpin loop at the 5' end of the mRNA, reducing the ribosomes' binding capacity for successful translation. Although we succeeded in expressing HTLV-1 RT, higher quantities are required to perform structural studies. In the future, once sufficient protein is obtained, a high-resolution model of HTLV-1 RT can be constructed, which will allow for high-throughput drug screening and the development of HTLV-1 RT-specific drugs, for which there is a great need.

1. Introduction

1.1 The HTLV-1 virus and its genome

Human T-cell leukaemia virus type 1 (HTLV-1) is a human oncovirus from the *deltaretrovirus* genus of *orthoretrovirinae* subfamily¹. It is able to infect a great variety of host cells, such as B lymphocytes, dendritic cells, monocytes and endothelial cells, but has a main tropism for CD4⁺ and CD8⁺ T-cells². Following its discovery, HTLV-1's physiology and pathogenicity was characterized, showing that the virus is able to infect humans, however very inefficiently, and is associated with certain forms of T-cell leukaemia and lymphoma³. Recent data shows that 1-5% of HTLV-1 carriers in Japan (♂:4-6%, ♀:2-6%) develop adult T-cell leukaemia (ATL), an aggressive mature peripheral T-cell lymphoma for which the median age of onset of 56 years showcases HTLV-1's long latency⁴. However, this varies greatly per region⁴⁻⁶. Currently, there are four approved therapeutic approaches, namely antiviral-, chemo-, immunotherapy and curative allogeneic stem cell transplantation⁷. HTLV-1 is also the causative agent for HTLV-1 associated myelopathy/tropical spastic paraparesis (HAM/TSP), caused by inflammation and subsequent lethal degeneration of the spinal cord⁸. HAM/TSP is commonly seen in 30-50 years old HTLV-1 infected patients, but rarely seen in children⁸⁻¹¹. As of now, there is no cure accessible for HAM/TSP. Consequently, therapy typically revolves around managing symptoms or addressing causes. This often involves the use of immunomodulatory drugs such as corticosteroids and interferon alpha¹²⁻¹⁴. HTLV-1 has a hard time spreading in a cell-free manner and is therefore mainly transmitted by infected cell-containing fluids such as blood, sperm/vaginal fluids and breast milk, only rarely by virus-containing fluids¹². The three most important mechanisms responsible for cell-cell transmission are virological synapses, viral biofilms and cellular conduits.

Previously characterised as type C retroviral particle, HTLV-1 presents itself as an enveloped, icosahedral, (+)ssRNA virus with a size between 90-270 nm^{15,16} (Figure 1A-B). The envelope proteins (Env), located in the lipid envelope are composed by a gp46 surface subunit (SU) (oriented outward) and a gp21 transmembrane subunit (TM)^{1,17}. Underneath this envelope, the matrix protein (Ma) forms an inner lining. Within this structure lays the icosahedral capsid, made up of the capsid protein (Ca), containing two identical strands of the nucleocapsid (Nc)-coated genomic (+)ssRNA (gRNA) associated with the viral protease (Pr), integrase (Int) and reverse transcriptase (RT) proteins.

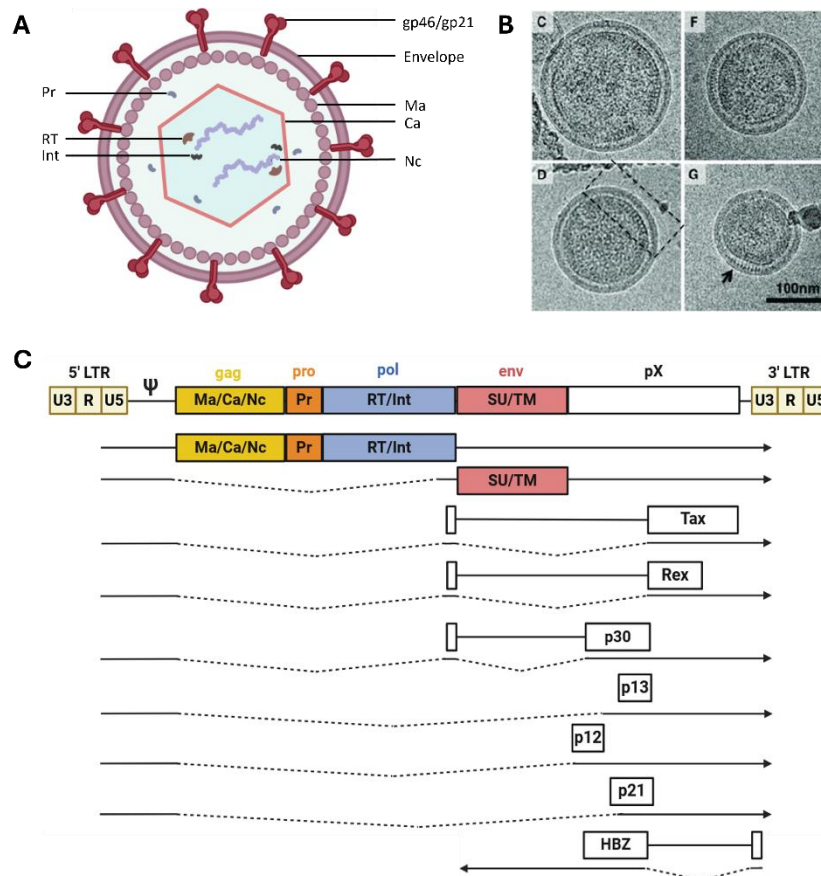


Figure 1. The HTLV-1 viral particle and the proviral DNA.

(A) Schematic overview of HTLV-1 viral particle. Pr: Protease, Int: Integrase, RT: Reverse transcriptase, Nc: Nucleocapsid, Ca: Capsid, Ma: Matrix, Gp46/Gp21: Envelope proteins. (B) Cryogenic transmission electron microscopy pictures of HTLV-1 particle¹⁶. Scale bar in G is applicable for C-G. (C) Overview of the proviral DNA organisation. The 5'/3' long terminal repeat (LTR) presenting its three subdomains (U3, R, U5), gag, pro, pol, Env, pX. (A, C) Made in Biorender.

After reverse transcription, the integrated HTLV-1 provirus is about 9 kb in length and can be divided into six domains^{1,18} (Figure 1C). At the 5' and 3' end, two long terminal repeats (LTRs) are present which are further subdivided into U3, R and U5 regions. These contain all necessary promoter and regulatory sequences needed for viral transcription and protein translation. Downstream of the 5'LTR, the *gag* domain encodes for Ma, Ca and Nc protein, followed by the Pr encoding *pro* domain. Pr is required to generate functionally active proteins from polyprotein precursors. Downstream of *pro* sits the *pol* domain which encodes the RT and Int proteins. These are important in proviral integration as RT converts the gRNA into proviral DNA, followed by integration into the host's chromosome by Int. The *env* domain is located further downstream of the *pol* domain, encoding the gp62 precursor protein. Gp62 will be cleaved into gp46 and gp21, but both remain connected through an intersubunit disulfide bond. Lastly, the *pX* domain encodes accessory proteins that are the result of alternative splicing events and alternate codon usage¹⁸. These mainly act in the regulation of viral replication (Tax, Rex, HTLV-1 bZip factor (HBZ)), viral latency (p8, p30, p13), immune evasion (p8, p12, p30) and oncogenesis (Tax, HBZ)¹⁹.

The structural and enzymatically active proteins are coded into the gag-pro-pol mRNA which contains three separate but overlapping open reading frames (ORF)^{1,20}. Translation of ORF1 with no ribosomal frameshift solely produces gag protein. ORF2 is translated through a -1 ribosomal frameshift, resulting in synthesis of both gag and pro into a preliminary gag-pro polyprotein²¹. ORF3 is translated via a -2 ribosomal frameshift, generating the gag-pro-pol polyprotein which uniquely encodes the RT and Int protein sequences¹. This rather complex but elegant system allows for the production of a great amount of structural proteins (gag) needed for the viral particle assembly, and only low quantities of enzymatically active proteins (pr, pol).

1.2 HTLV-1 replication cycle

Attachment of the viral particle is commenced by heparan sulphate proteoglycans (HSPG) on the cell surface that capture gp46's C-terminal domain^{22,23} (Figure 2, step 1). Quickly after, gp46 binds Neuropilin-1 (NRP-1) undergoing a conformational change leading to the exposure of gp46's Glut-1 binding domain²². Upon binding to Glut-1, disulfide isomerisation in gp46's C-terminal domain takes place, changing the interaction between gp46 and gp21. This allows gp21 to perform its fusogenic function resulting into a merging of HTLV-1's and the target cell's lipid membrane (Figure 2, step 2). After cell entry, HTLV-1's capsid starts to degrade and releases two copies of the 5' capped gRNA that will be used as template for dsDNA generation via reverse transcription¹ (Figure 2, step 3). This process will be thoroughly explained later on. Once produced, the proviral DNA is part of a nucleoprotein complex containing both viral and host proteins, called the pre-integration complex (PIC)²⁴. Int binds to the proviral DNA as a tetramer, forming the intasome, as part of the PIC that will interact with the host chromosomal DNA for proviral integration²⁵ (Figure 2, step 4). Transcriptional units, alphoid repetitive sequences and actively transcribed genes have shown to be preferential integration sites for HTLV-1/2²⁶. Interestingly, disease development is strongly correlated with the integration site type as integration into transcriptionally active regions increases viral gene expression necessary for development of HAM/TSP and ATL. It is from the integration step onward that HTLV-1 can commence its clonal expansion. This is responsible for persistence of infection and maintaining a high proviral load, predisposing the infected patient for development of HAM/TSP and ATL²⁷. After successful integration of the provirus, RNA polymerase II dependent transcription takes place, generating mRNA encoding any of the viral proteins, but also new gRNA¹ (Figure 2, step 5). In the cytoplasm, double spliced (Tax, Rex, p30, p12, p21, HBZ) and unspliced mRNAs (gag, gag-pro, gag-pro-pol precursor polyproteins) are translated by free ribosomes, whereas single spliced mRNAs (env) are translated by membrane-bound ribosomes (Figure 2, step 6). Yet, contradictory evidence is present on the mechanism of protein translation

initiation in HTLV-1 as it can be accomplished in a canonical cap-dependent and non-canonical cap-independent manner^{28,29}. The priorly generated gRNA is bound by the precursor polyproteins and is relocated towards the plasma membrane, where it will be packaged into the viral particle upon budding¹ (Figure 2, steps 7-8). Finally, all precursor polyproteins will be processed by the viral protease to form a mature viral particle (Figure 2, step 9).

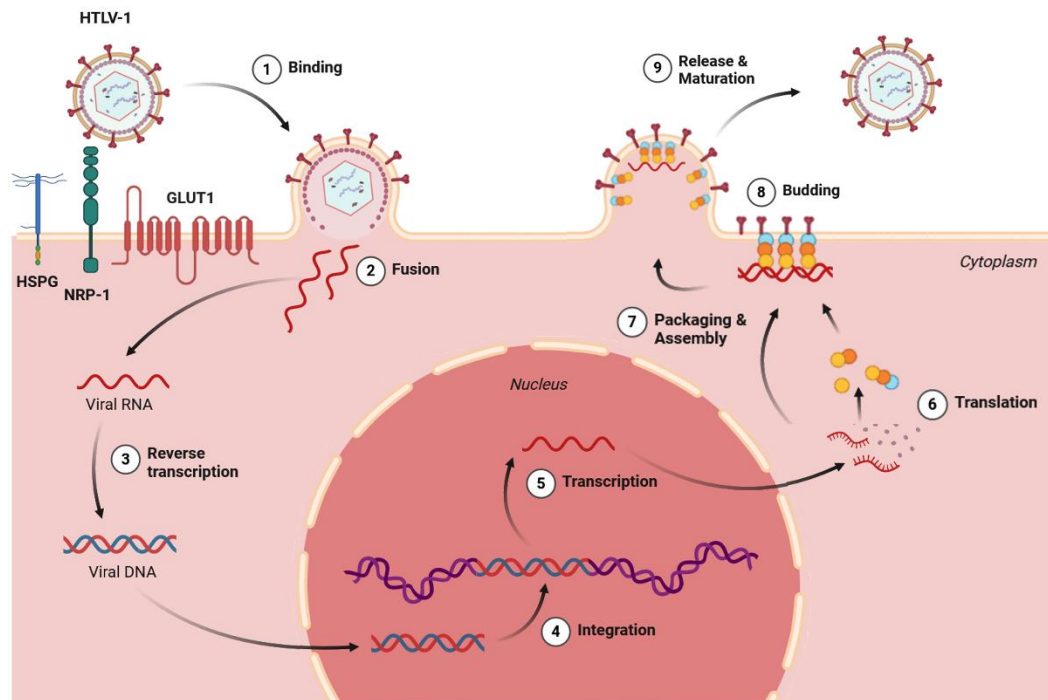


Figure 2. HTLV-1 viral replication cycle.

(1) HTLV-1 binds its entry receptors via the gp46 surface subunit, followed by (2) fusion via gp21. (3) Using its RT, the gRNA is converted to proviral DNA and translocated into the nucleus. (4) Int is employed to integrate the proviral DNA into the host chromosome. (5) Using cellular RNA-polymerase II transcription machinery, the proviral DNA is transcribed into mRNA and gRNA. The mRNA is processed by the spliceosome upon transcription. (6) The mRNA is translated by membrane-bound ribosomes (Env) and free ribosomes (gag, gag-pro, gag-pro-pol). (7) The polyprotein binds the gRNA and assembles at the cell membrane, which starts budding under influence of the ESCRT-complex. (9) After release, all polyproteins are processed by Pr and full assembly takes place to generate a fully active viral particle. Scheme made in Biorender.

1.3 Reverse transcription

Reverse transcription in context of HTLV-1 is still poorly understood, but likely follows a similar mechanism to other retroviruses (e.g. HIV-1) due to the high overall mechanistic conservation³⁰ (Figure 3A). After cell entry, HTLV-1's capsid starts to degrade releasing two copies of the 5' capped gRNA that will be used as template for dsDNA generation. RT requires an RNA oligonucleotide primer as it can only bind a double stranded nucleic acid substrate. For this, the 3' end of tRNA^{Pro} is used, that is packaged into the viral particles upon viral budding^{30,31}. The terminal 18 nucleotides of tRNA^{Pro} are complementary to the primer binding site (PBS, Figure 3B) in the 5' LTR and will be able to bind this upon reverse transcription initiation³⁰ (step 1). Then, RT binds the gRNA-tRNA^{Pro} hybrid and generates (-)ssDNA of the 5' R-U5 region, while ribonuclease H (RNase H) degrades the gRNA template (step 2). During (-) strand transfer, the (-)ssDNA will relocate to the

complementary 3' R region in the gRNA, acting as a new binding site for RT (step 3). Reverse transcription leads to generation of (-)ssDNA from the full gRNA. Meanwhile, RNase H has degraded the full template gRNA, except for some purine-rich regions (step 4). To produce the sense strand, a new primer is needed for the RT to be able to bind the (-)ssDNA. In HIV-1, RNase H-resistant polypurine tracts (PPT) are employed as primers, located at the edge of the 3' LTR U3 region and centrally in HIV-1's genome (cPPT)³² (step 5). This cPPT seems to be unique to HIV-1 as only a 3' PPT has been reported in HTLV-1 and other retroviruses. The produced (+)ssDNA of the 3' LTR region relocates towards the complementary 5' LTR region, where it will bind to the PBS (step 6). Finally, full length (+)ssDNA is synthesized, forming a complete proviral DNA ready for integration.

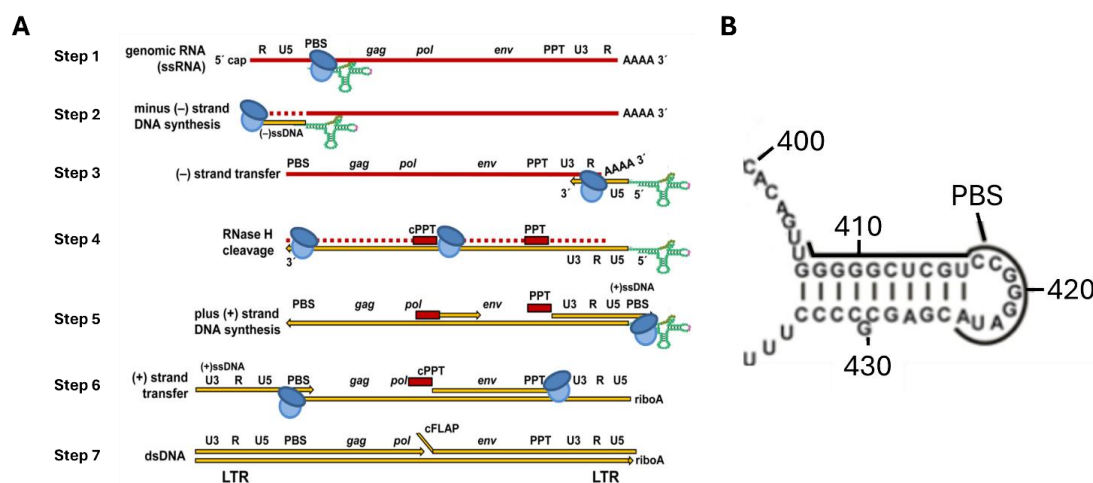


Figure 3. The HIV-1 reverse transcription process³⁰.

(A) Colour code: RNA, DNA, RT, tRNA. PBS: primer binding site, (-)ssDNA: minus-strand strong stop DNA, (c)PPT: (central) polypurine tract, (+)ssDNA: plus single strand DNA, (-)ssDNA: negative single strand DNA, LTR: long terminal repeat. (B) Focus on the secondary structure of the primer binding site in HTLV-1 gRNA³⁵.

1.3.1 Pre-initiation

The pre-initiation step takes place after entry of the gRNA into the cytoplasm when the tRNA^{Pro} primer binds onto the PBS, forming a nucleic acid dimer complex that can be bound by RT³⁰. The PBS is located in the 5'LTR downstream of the U5 region and is crucial during the start of the reverse transcription process. Commonly, the PBS holds no particular secondary structure as seen in HIV-1 and Rous sarcoma virus (RSV)^{33,34}. Nevertheless, HTLV-1's PBS is a highly structured self-enclosing hairpin loop that needs to be destabilised by a strong chaperone protein³⁵ (Figure 3B).

The first evidence of low molecular mass RNA being used as primer for reverse transcription was in a study on RSV, where its primer was identified as tRNA^{Trp}³⁶. Later, it was determined that this was very common among various retroviral species, thus also HTLV-1 and HIV-1 which use tRNA^{Pro}

and tRNA^{Lys,3}, respectively³¹. As the terminal 18 nucleotides of tRNA anneal to the complementary PBS, a gRNA/tRNA hybrid construct is generated^{30,31} (Figure 3A, step 1). The 3' end of the tRNA holds a free hydroxyl group, normally used by the aminoacyl-tRNA synthase to covalently bind an amino acid to its cognate tRNA³⁷. Here, RTs use the 3' hydroxyl group as transcription start site. Interestingly, an 18 nucleotide long 3' end fragment, called tRF-3019, is present in the HTLV-1 viral particles at levels greater or equal to tRNA^{Pro}, generated by specific cleavage events from tRNA^{Pro}³⁸. In vitro studies proved that tRF-3019 was able to fully activate reverse transcription, indicating the tRNA^{Pro} fragment's potential function as a primer during reverse transcription in vivo. Its size poses a lower energy barrier to overcome for primer annealing compared to tRNA^{Pro}, which might be useful due to the highly stable PBS (Figure 3B). In contrast, tRF-3019 could also act as a RT inhibitor, as seen in endogenous retrovirus where tRFs target the PBS and inhibit tRNA binding³⁹.

1.3.2 Reverse transcription initiation

Reverse transcription begins with the assembly of the reverse transcription initiation complex (RTIC), consisting of a tRNA primer, the gRNA template and RT³⁰. Recently, the initial HIV-1 RTIC structure (RTIC 0) and its conformation during the first dNTP incorporation (RTIC +1) were solved^{40,41} (Figure 4A). Together, these structures give an idea of the conformational changes that the RTIC undergoes during the early stages of reverse transcription initiation.

Great reorganisation in both gRNA and tRNA is observed as there is the generation of two self-annealing hairpin loop helices in the former, and a loss of the three-way junction in the latter (Figure 4A). The gRNA PBS neighbouring domains can be divided into three regions, namely helix 1 (H1), helix 2 (H2) and the PBS⁴⁰ (Figure 4B). H1 is formed by annealing of the 5' end of the gRNA template and a complementary gRNA sequence downstream of the PBS. H2 is formed by annealing of bases just upstream of the PBS, generating a complex and sterically hindering structure upstream of the polymerisation initiation site. The tRNA primer unfolds into a single self-annealing hairpin helical structure and binds to the PBS using its 3' end. Since these form an RNA-RNA duplex, it folds into a thermodynamically preferred A-type geometry making it a suboptimal substrate for an RNA-dependent-DNA-polymerase^{42,43}. This is also posed as the main cause of rapid RT-RNA dissociation during reverse transcription initiation⁴⁴. Dissociation rates during the elongation phase are much lower, most likely because of the more favourable intermediary A-B-type geometry adopted by the DNA/RNA-duplex^{42,43,45}.

1.3.3 Reverse transcription pausing

HIV-1 reverse transcription initiation is slow and non-progressive with quick dissociation of the RT from the RNA duplex from nucleotide incorporation +3 to +5⁴⁶. This is in stark contrast with what is observed during the elongation phase, suggesting that rate limiting factors are present during initiation causing HIV-1 RT dissociation and premature termination⁴⁵.

A molecular dynamics study performed on the HIV-1 RTIC complex proposes a flipping mechanism to explain how incorporation proceeds between nucleotides +3 till +5⁴⁷ (Figure 4C). Briefly, incorporation of nucleotides +1 till +3 goes rather smoothly, whereas from nucleotides +3 till +5 the rate severely drops⁴⁵. At position +3, RT encounters H2, which forces HIV-1 RT to “flip”, switching into a polymerase inactive state⁴⁷. In this conformation, the polymerase active site is at great distance from the 3' end of the primer strand. Once sufficient DNA is synthesised, H2 can

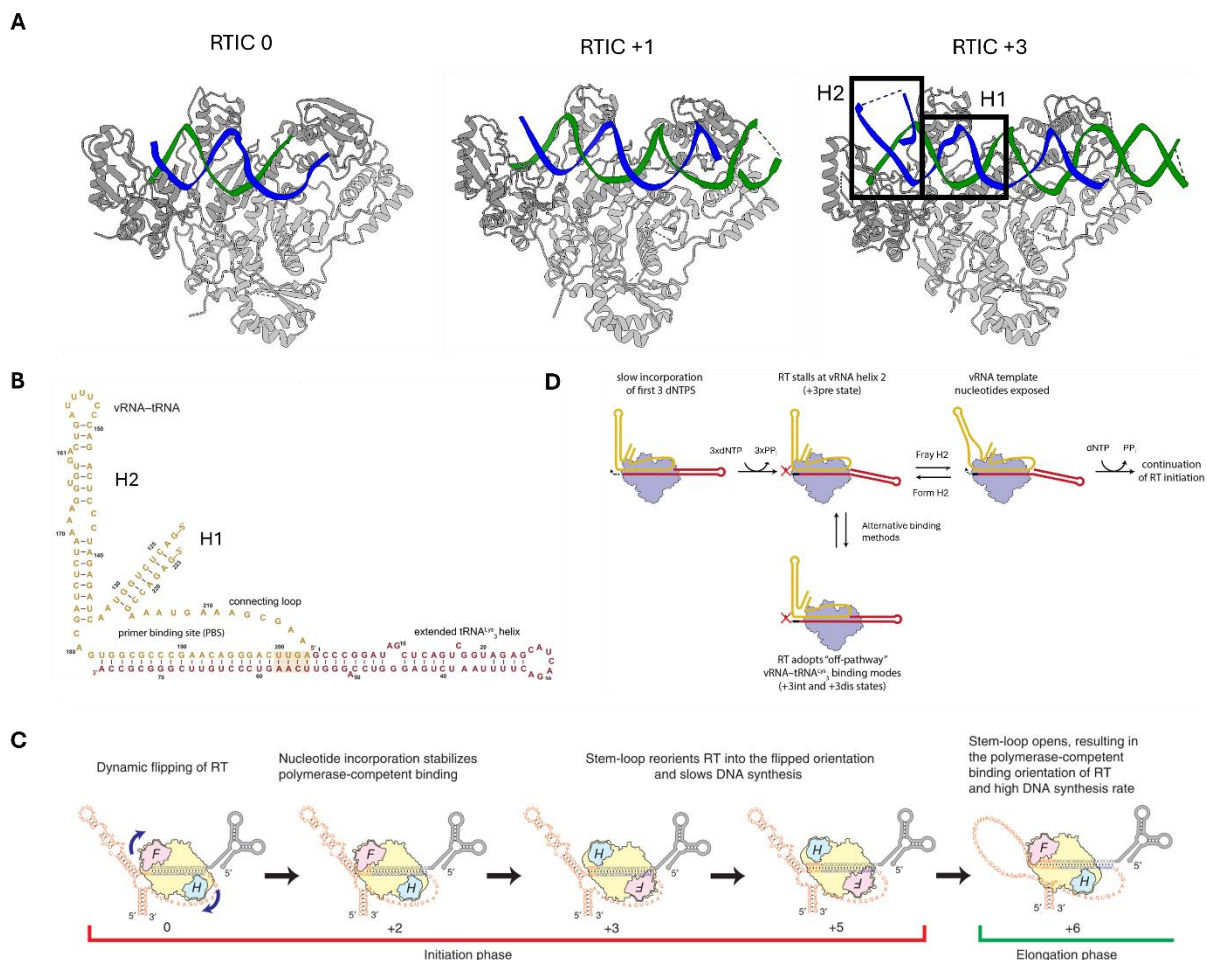


Figure 4. HIV-1 RT flipping and distinct conformational changes during reverse transcription initiation.

(A) Depiction of HIV-1 RT polymerase before dNTP incorporation (RTIC 0), after one dNTP incorporation (RTIC +1) and after three dNTP incorporations (RTIC +3, +3Pre). Black boxes on RTIC +3 structure indicate helix 1 (H1) and helix 2 (H2), also shown in (B). colour code: p66, p51, template strand, primer strand. Figure made in ChimeraX, PDB RTIC 0: 6HAK⁴¹, RTIC +1: 6WAZ⁴⁰, RTIC +3: 6WB0⁴⁸. (B) Schematic representation of the template:primer dimer secondary structures during transcription initiation⁴⁰. (C) RT flipping upon encountering H2 at the third dNTP incorporation, and reversion at the sixth incorporation⁴⁷. (D) Simplified model for pausing during the third dNTP incorporation, showing the three previously identified states (+3Pre, +3Int and +3Dis)⁴⁸.

be unwound with the help of the HIV-1 Nc chaperone protein, and HIV-1 RT can flip back to its polymerase active state conformation at position +6. This is followed by entry into the elongation phase.

Recently, the structural basis causing the +3 HIV-1 RTIC to pause were elucidated by Larsen, K. *et al.* (2020)⁴⁸. They observed that H2 indeed was present deeper into the HIV-1 RT polymerase active site during incorporation +3, but also obtained three different conformations of the same state. The RTIC +3 pre-translocation state (+3Pre) closely resembles RTIC +1, but holds the RNA-RNA duplex less firmly due to shifting at the RNase H domain^{40,48} (Figure 4A, RTIC +3). The two alternative states, named RTIC +3 Intermediary (+3Int) and RTIC +3 Displaced (+3Dis) showed conformational changes in the thumb domain as it moves towards the finger domain, pushing the 3' end of the tRNA primer out of the polymerase active site. It is in one of these two states that HIV-1 RT waits till the H2 template nucleotides are exposed for further polymerisation (Figure 4D). Once exposed, it causes a repositioning of the 3' end of the tRNA primer into the polymerase active site. In addition, further shifting of the RNA-RNA duplex along the binding cleft repositions the nucleic acid dimer to regain a close contact with multiple HIV-1 RT residues. This counteracts the strain obtained from the changing conformations of the finger and thumb domain. Finally, H2 is dismantled and +4 incorporation can take place, as from here on polymerisation rate increases again. It is important to mention that even though the existence of these intermediate states explains RT pausing very well, it does not mean that the RT-flipping theory is disproven (Figure 4C). Moreover Larsen, K. *et al.* (2020) mentions that the two theories are not mutually exclusive, but rather that these states are most likely accompanied by the flipping of RT as it all will lead to pausing.

1.4 Reverse transcriptase

It is well established that HTLV-1 RT is a heterodimer composed by a major p62 subunit, containing the polymerase active site and RNase H domain, and a minor p49 subunit that has a rather supportive function. This configuration is identical to HIV-1 RT made of a major p66 subunit and a smaller p51 subunit (Figure 5A). For both HIV-1 RT and HTLV-1 RT, the major subunits can be divided into five highly conserved subdomains, namely finger, palm, thumb, connection and RNase H (Figure 5B). Similarly to HIV-1 RT, it is also believed that the minor p49 subunit of HTLV-1 RT is the product of a proteolytic cleavage event of the major subunit with degradation of the RNase H domain. In addition, even though both subunits have identical sequences, HTLV-1 RT p49 presumably holds an alternative fold compared to its p62 subunit, as seen in HIV-1 RT's metamorphic p51 subunit⁴⁹. Here, p51 adopts a slightly different fold of the polymerase domain

into an inactive form with loss of polymerisation function. Yet, it does play an important role in stabilising the full HIV-1 RT complex, nucleic acid binding and overall protein function⁵⁰. Due to the high predicted grade of similarity between HIV-1 RT and HTLV-1 RT, an in-depth study on HIV-1 RT structure and mechanism of action was performed as this might allow for extrapolation towards our protein of interest.

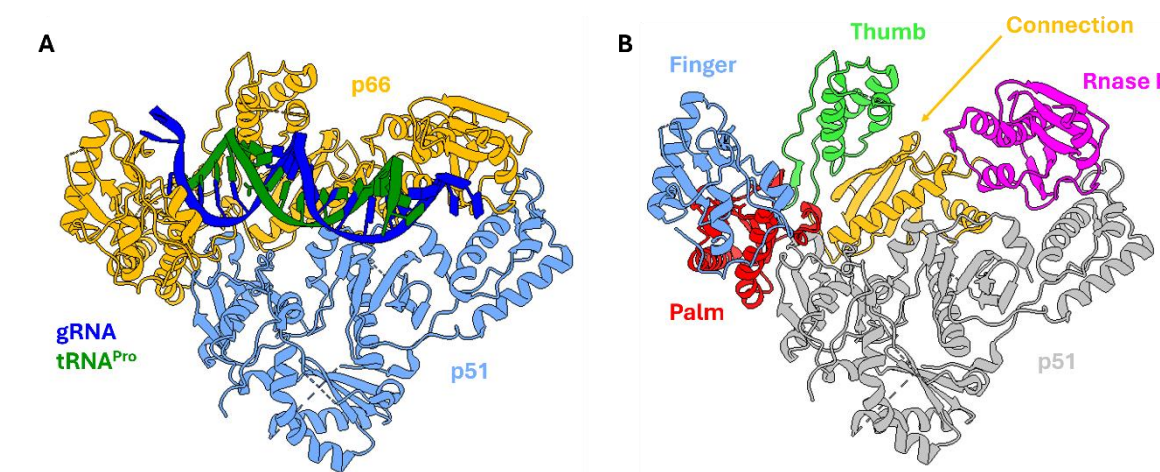


Figure 5. **Depiction of HIV-1 RT and its subdomains.**

(A) Front facing view of HIV-1 RT with coloured subunits and nucleic acid dimer. (B) Front facing view of HIV-1 RT with coloured subdomains in p66. Nucleic acid dimer was deleted for graphical reasons. Figures made in ChimeraX, PDB: 6UIT¹⁰⁶.

1.4.1 Structure and conserved motifs

The nucleic acid binding cleft in HIV-1 RT measures 17-18 bp in length and is formed by polar residues present in all of p66's subdomains in addition to p51's thumb and connection domain. It forms upon nucleic acid binding as the thumb domain of p66 moves away from its finger domain⁴⁹. Several highly conserved motifs are present within this cleft, such as the helix clamp (p66's αH-αI hairpin), primer grip (p66's β12-β13 hairpin) and template grip (p66's β4 and αB, p51's β8-αE connecting loop and β5a). In addition to residues in p66's and p51's connection domain, these motifs form a "translocation track" responsible for correctly positioning the nucleic acid complex in the polymerase and RNase H active site⁴⁹⁻⁵¹.

The helix clamp is part of the αH-αI helix-turn-helix motif in the thumb domain⁵² (Figure 6A-C, red; Table 1: helix clamp). In HIV-1 RT, it helps correctly positioning the nucleic acid dimer both into the

Table 1. **Involved residues in the highly conserved motifs in HIV-1 RT.**

Colour code follows domain colouring in Figure 6.

Motif	Subunit	Residues involved
Helix clamp	p66	V(X) ₄ K ²⁵⁹ LVGKL(X) ₁₆ KLLR ²⁸⁴ (X) ₄ L
Template grip	p66	V ⁷⁵ , R ⁷⁸ , N ⁸¹ , E ⁸⁹ , Q ⁹¹ , G ⁹³ , G ¹⁵² , W ¹⁵³ , K ¹⁵⁴ , P ¹⁵⁷
Primer grip	p66	F ²²⁷ L ²²⁸ W ²³⁰ M ²³¹ G ²³² Y ²³³ E ²³⁴ L ²³⁵
RNase H primer grip	p66	G ³⁵⁹ , A ³⁶⁰ , H ³⁶¹ , T ⁴⁷³ , Q ⁴⁷⁵ , K ⁴⁷⁶ , Y ⁵⁰¹ , I ⁵⁰⁵
	p51	K ³⁹⁵ , E ³⁹⁶

active site for polymerisation as well as into the RNase H active site for template RNA degradation. It was shown that this highly conserved motif could bind to any nucleic acid sequence, without sequence specificity⁵³. Helix α H (Q258, G262, W266, Q269) of the helix clamp and I94 in β 5b sheet in the palm domain were found to form a track interacting with the primer strand bases -2 till -6 forming the minor groove of the nucleic acid dimer⁵⁴. These residues are highly conserved among lentiviral RTs and much evidence suggests that they are crucial for adequate and correct reverse transcription^{54–56}.

The primer grip is a highly conserved motif located within p66's palm domain⁵⁷ (β 12- β 13, 227-235, Figure 6A-C, orange; Table 1, Primer grip). This structural motif helps in correctly positioning the 3'-hydroxyl group at the end of the primer strand into the active site, preparing it for nucleophilic attack on an incoming deoxyribonucleoside triphosphate (dNTP). It has been shown that mutations in this highly conserved motif cause severe increase in misinsertion rate and altered RNase H functionality^{58–60}.

The template grip is located in p66's finger domain and interacts with the 5' end of the RNA template strand (Figure 6A-C, yellow). Rather than a true motif it is made up by multiple conserved residues that tightly hold onto the RNA template⁶¹ (Table 1, Template grip). Mutations in these residues have been linked to altered processivity and resistance against both nucleoside reverse

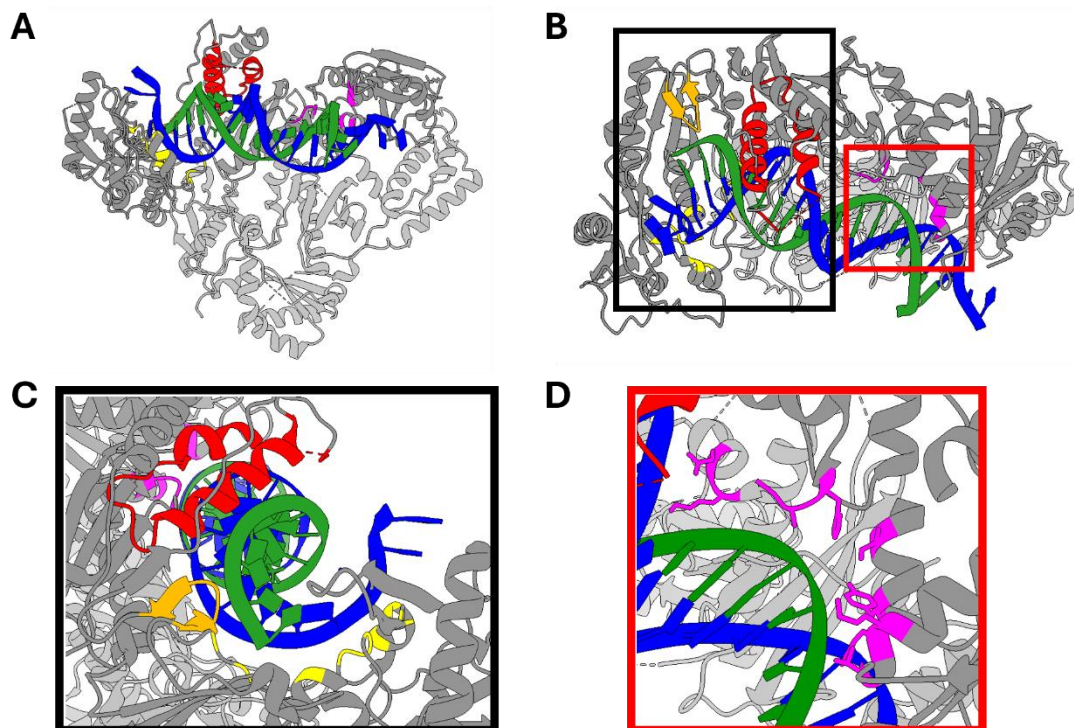


Figure 6. Depiction of HIV-1 RT with coloured nucleic acid dimer coordinating regions.

(A) Frontfacing view of HIV-1 RT. (B) Topview of HIV-1 RT. Black box corresponds to (C), red box corresponds to (D). (C) Focus on helix clamp, primer grip and template grip. (D) Focus on RNase H primer grip. Colour code for all Figures: p66, p51, template strand, primer strand, helix clamp, primer grip, template grip, RNase H primer grip. Figures made in ChimeraX, PDB: 6UIT¹⁰⁶.

transcriptase inhibitors (NRTIs) and non-nucleoside reverse transcriptase inhibitors (NNRTIs)^{51,61-66}. This NRTI-resistance is mainly due to the decreased catalytic rate in these mutants, that allows for better discrimination between a true nucleoside triphosphate and its antiretroviral analogue⁶⁴.

A fourth and final important domain for coordination of the nucleic acid dimer is present near the C-terminal end of the p66 subunit, called the RNase H primer grip^{51,57} (Figure 6A-B, D, magenta). It is composed of residues in both HIV-1 RT p66 and p51 subunits (Table 1, RNase H primer grip). All of these act to properly position the 5' end of the RNA template strand into the RNase H active site for cleavage during reverse transcription. Data shows that mutations in the RNase H primer grip enhance NRTI resistance and greatly affect substrate affinity^{67,68}.

The steric gate, Y115 in HIV-1 RT, acts in discriminating between ribonucleoside triphosphates (rNTPs) and dNTPs^{69,70} (Figure 7A, orange). The 2'-hydroxyl group of NTPs clashes with the bulky phenol sidechain of the tyrosine, which prevents it from entering the active site⁷⁰. Mutational studies showed that loss of the aromatic character at this position led to both decreased substrate specificity and overall RT activity^{69,71}. Additionally, a Y115F substitution did not significantly change polymerase activity as such, but slightly increased sensitivity to multiple NRTIs. Similar to HIV-1 RT, the steric gate is also found in Moloney Murine Leukemia virus RT (MMLV, F155), yeast DNA polymerase η (F18) and human DNA polymerase η (F35), however here phenylalanine is used instead of tyrosine^{70,72}. Here, the conserved aromatic nature of both residues allows for functional replacement.

1.4.2 The molecular mechanism of dNTP incorporation

Upon entry in the active site, the dNTP interacts with K65 and R72 in the highly conserved $\beta 3$ - $\beta 4$ hairpin loop in p66's finger domain through its γ -phosphate (P_γ) and β -phosphate (P_β), respectively⁷³ (Figure 7A-B). In addition, Q151 that is present in palm domain of p66 interacts with the 3' hydroxyl group of the incoming dNTP. Two Mg^{2+} ions are present in the active site, separated ~ 3.6 Å. Those are coordinated by residues D110, V111, D185 and D186. One Mg^{2+} interacts with the 3' hydroxyl group of the last base of the primer strand and α -phosphate (P_α) of the incoming dNTP^{74,75}. The other interacts with the P_β and P_γ of the incoming dNTP. In addition to metal-ion coordination bonds with the Mg^{2+} ions itself, the dNTP also interacts with the template base via Watson-Crick base pairing (Figure 7B). For DNA polymerases, both the correct hydrogen bonding as well as overall shape complementarity between template base and incoming dNTP are crucial for incorporation⁷⁶. First of all, residues from the polymerase itself form hydrogen bonds with the minor groove side of template base:incoming dNTP pair, acting as a ruler to measure whether there is proper base pair spacing. Also, entry of the dNTP into the active site triggers a

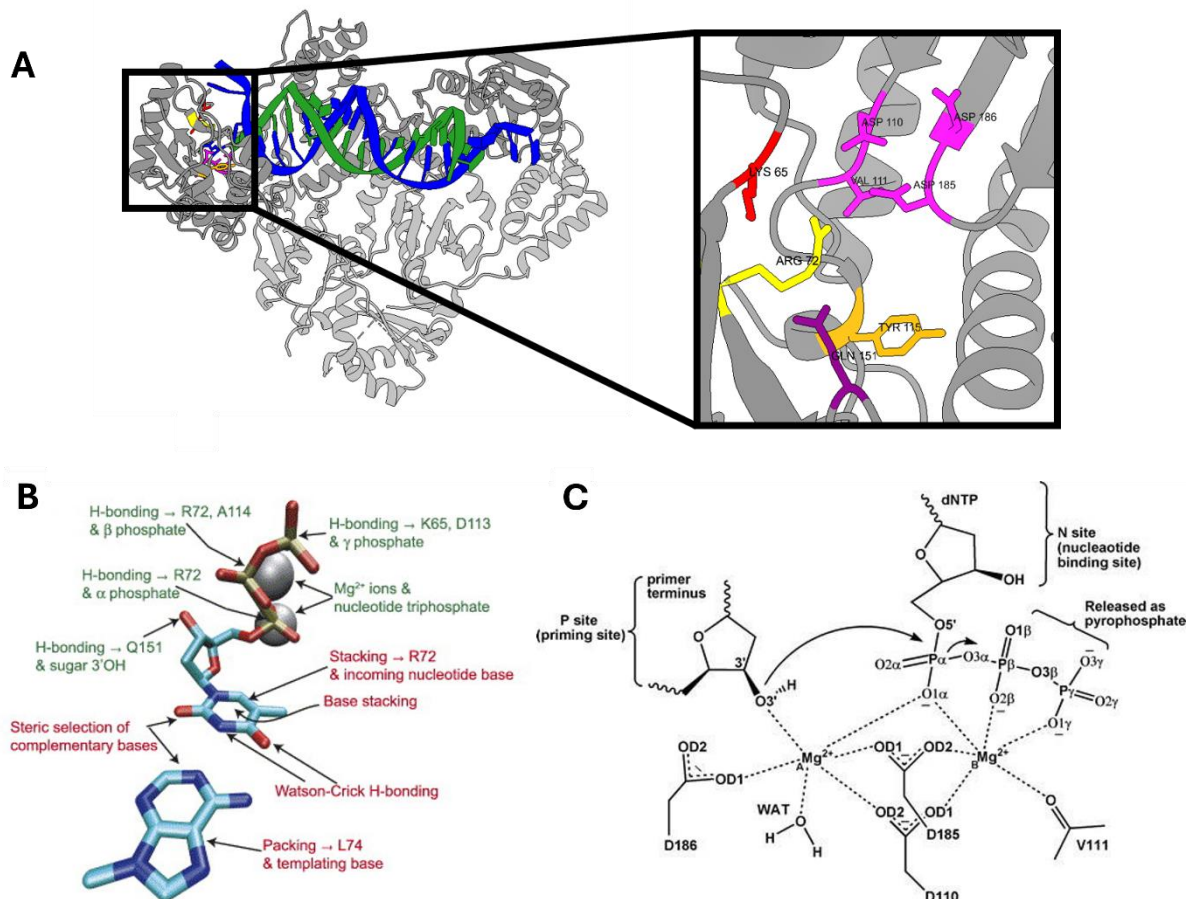


Figure 7. Structure of the HIV-1 RT polymerase active site, dNTP coordination and catalysis.

(A) Depiction of HIV-1 RT polymerase active site. Important residues in the polymerase active site coloured (K65, R72, Y115, Q151, D110, V111, D185, D186). Mg^{2+} ions not shown. Figure made in ChimeraX, PDB: 6UIT¹⁰⁶. (B) Overview of all coordinating interactions between HIV-1 RT active site residues and the dNTP⁷³. (C) Overview of catalysis reaction for primer elongation⁷⁷.

conformational change in the finger domain to form a pocket in which only a true Watson-Crick base pair will fit, namely A:T or G:C in DNA/DNA and U:A, A:T, G:C in RNA/DNA. Since wrong Watson-Crick base pairing will lead to incorrect base pair spacing and formation of a truncated dNTP binding pocket, both these mechanisms will ensure that there is correct dNTP incorporation. The second mechanism was already mentioned above where p66's finger domain changes conformation and encloses the active site after dNTP entry⁶¹. To date, no information on the presence of a similar ruler residues in HIV-1 RT is available.

Once the active site is completed, the catalysis process can take place. In summary, primer elongation happens by a type 2 nucleophilic substitution (S_N2) of the deprotonated 3' hydroxyl group of the terminal primer nucleotide on the dNTP's P_α , where P_β and P_γ are released⁷⁷ (Figure 7C). First, the dNTP needs to properly bind into the active site, where it is coordinated by residues of HIV-1 RT and two octahedrally coordinated Mg^{2+} ions. Both are crucial to the catalytic reaction, since Mg^{2+}_A is necessary to lower the pKa of the primer's 3' hydroxyl group easing deprotonation for the nucleophilic attack on P_α ^{75,78}. Additionally, Mg^{2+}_B is responsible to correctly position the

dNTP's triphosphate. In step 1, 3' OH of the primer releases its proton, which is accepted by a nearby aspartic acid residue⁷⁹. In step 2, the 3' O⁻ performs a nucleophilic attack on the P_α which generates a new phosphodiester bond, forming an intermediary negatively charged pentavalent complex stabilised by Mg²⁺_B⁷⁷. In step 3, P_β-P_γ is released and receives a proton from a still unknown donor. HIV-1 RT holds a preference for Mg²⁺, but is still functional with alternative bivalent ions such as Mn²⁺, Co²⁺ and Zn²⁺⁸⁰. Zn²⁺ increased HIV-1 RT's fidelity up to 3-fold, but held a severely decreased extension rate compared to Mg²⁺. Also Mn²⁺ and Co²⁺ showed a stark decrease in extension rate and a dose-dependent mutation rate. This shows that even though fidelity was relatively conserved or even increased in case of Zn²⁺, the slow extension rates indicate disfavour towards these bivalent cations by HIV-1 RT during viral replication.

1.5 Aims

Among the three enzymatically active proteins of HTLV-1, protease and integrase structures have been solved providing crucial information for structure-based drug design (SBDD), molecular docking and molecular dynamics for mechanistic studies⁸¹⁻⁹⁰. HTLV-1 RT, which is yet to be solved, is a crucial protein for HTLV-1 survival as it is in charge of one of the most important steps in the viral replication cycle, being the reverse transcription of the gRNA to proviral DNA^{1,30,44}. This makes it an attractive drug target in prophylactic treatment against HTLV-1 infection. However due to the little amount of data present on HTLV-1 RT function, mechanism and structure, no HTLV-1-specific drugs targeting its RT have been developed.

This thesis aims to obtain the structure of HTLV-1 RT through single particle cryogenic electron microscopy (cryo-EM) and to provide a platform for the first steps of structure based drug design targeting the HTLV-1 RT. To this end, a detailed literature study was performed, combined with experimental and in silico approaches to obtain a multifaceted understanding of HTLV-1 RT. Once the structure is solved, SBDD can be performed targeting this protein, increasing the amount of available options to counter HTLV-1 infection in highly endemic areas.

2. Materials and methods

2.1 HTLV-1 constructs cloning and validation

The HTLV-1 pET15b, encoding the Pr-Pol polyprotein, and HTLV-1 pET-Duet-1 p62/p49 Gs, encoding both p62 and p49 subunits of HTLV-1 RT, expression vectors were purchased from Genscript Europe (Netherlands). These were used for both protein expression and as a template for construction of derivate constructs (Figure 8).

2.1.1 HTLV-1 pET42a (+) RT & HTLV-1 pET42a (+) pol constructs

HTLV-1 RT p62 and HTLV-1 pol genes were amplified by polymerase chain reaction (PCR) from the HTLV-1 pET15b plasmid using sequence specific primers designed also to introduce *MfeI* and *XhoI* cutting sites at the 5' and 3', respectively (Table 2). Then, both PCR products and pET42a (+) plasmid were cut using the *MfeI* and *XhoI* restriction enzymes (10 U, 37°C, 1 h, New England Biolabs) followed by ligation using T4 ligase. For this, 100 ng vector DNA was mixed with insert DNA to a molar ratio of 1:3, supplemented with T4 ligase (1 U, Thermo Scientific) and T4 DNA ligase buffer (10x, Thermo Scientific), incubating at 22°C for 1 h. T4 ligase was then inactivated at 65°C for 10 min. After each step, the synthesised product was checked and purified on a 1% agarose (in 0.5% TAE) gel. For this, the sample was mixed with Trisack DNA loading buffer (6x, Thermo Scientific) and Midori green intercalating DNA dye (Nippon Genetics). For size comparison, 1 µg Generuler 1kB DNA ladder (Thermo Scientific) was used. Finally, visualisation was performed by using the Gel Doc™ EZ imager (Bio-Rad) with analysis in Image Lab™ software. The QIAquick® Gel Extraction Kit (Qiagen) was used to isolate the construct from the agarose gel. validation of the construct was done by Sanger sequencing (Macrogen Europe B.V.), PCR and restriction digest.

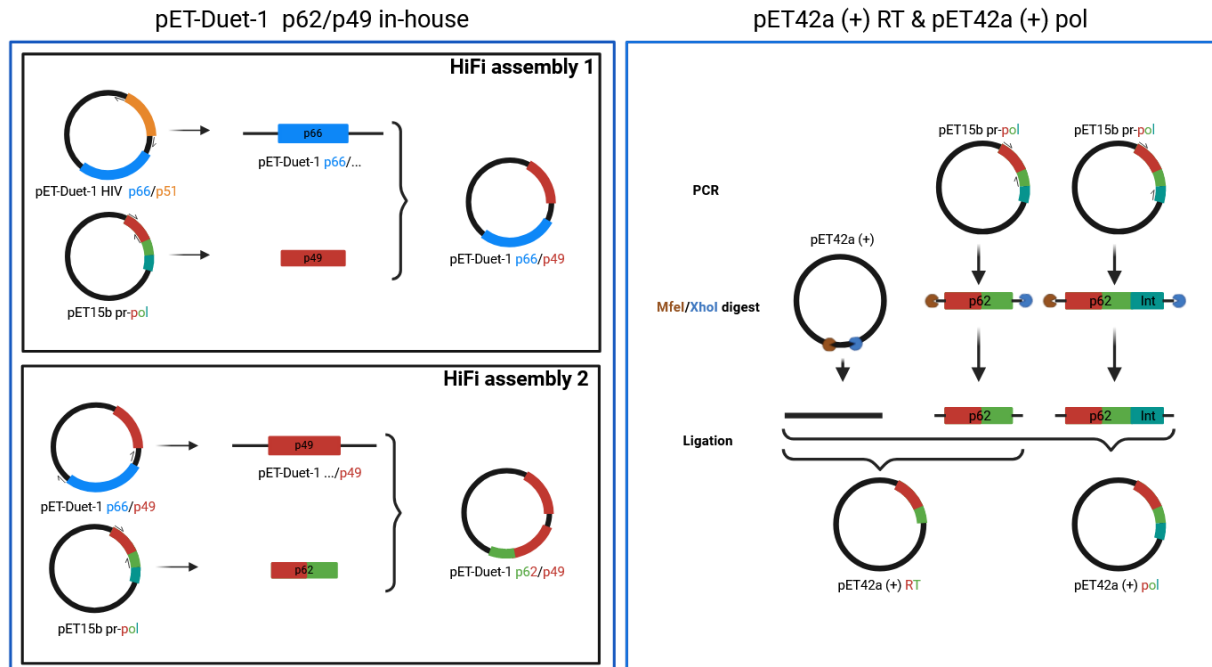


Figure 8. Schematic overview of plasmid construct design for recombinant HTLV-1 RT expression.

Table 2. Primers used for HTLV-1 pET-Duet-1 p62/p49 in-house and pET42a(+) RT/pol construction.

Final plasmid	Technique		Template	Product	Introduced cutsite	Primer (5'-3')	F/R		
HTLV-1 pET-Duet-1 p62/p49 in-house	HiFi assembly	1	HIV-1 pET-Duet-1 p66/p51	HIV-1 pET-Duet-1 .../p51		CCCGGTGCTGCAGCTGTAACGAGTCTGGTAAAGA ACCCGC	F		
						GGCGGCAGATGTTCCAGGCCCATATGTATATCTCCTT CTTATACTTAACATAATACTAA	R		
			HTLV-1 pET15b	HiFi ready p62		ATGGGCTGGAACATCTGCCGC	F		
						TTACAGCTGCAGACCCGGG	R		
		2				pET-Duet-1 intermediary p62/p51	pET-Duet-1 intermediary p62/...	GCGGCCGCATAATGCTTAAG	F
								AGGTCCCTGGAATAAGACCTCAAG	R
			HTLV-1 pET15b	HiFi ready p49		CTTGAGGTCTTATTCAGGGACCTGGCCT GGAACATCTGCCG	F		
						CTTAAGCATTATGCGGCCGCTTACATCAGCGCTTTCA CCGG	R		
HTLV-1 pET42 (+) RT	Restriction digest	HTLV-1 pET15b	P62 with MfeI/XhoI cutsites	MfeI	ATTCAATTGGCCTGGAACATCTGCCGC	F			
			XhoI	TTACCTCGAGCAGCTGCAGCACCGGGGTA	R				
HTLV-1 pET42 (+) pol			Pol with MfeI/XhoI cutsites	MfeI	ATTCAATTGGCCTGGAACATCTGCCGC	F			
			XhoI	GGATCCTCGAGCATATGTTAGCCAT GATGCTG	R				

2.1.2 HTLV-1 pET-Duet-1 p62/p49 in-house

HTLV-1 RT p62 and p49 genes were amplified by PCR from the HTLV-1 pET15b plasmid using sequence specific primers (Table 2). PCR products were run on a 1% agarose (in 0.5% TAE) gel and isolated using the QIAquick® Gel Extraction Kit (Qiagen). Then, final assembly was performed according the NEBuilder® HiFi DNA Assembly protocol (New England Biolabs). Validation of the construct was done by Sanger sequencing (Macrogen Europe B.V.), PCR and restriction digest.

2.2 Recombinant protein expression in *E. coli*

2.2.1 Transformation

Expression constructs were transformed in different *E. coli* strains following manufacturer's protocol. One Shot BL21 (DE3) star: Briefly, 1.7% (V/V) β-mercaptoethanol (β-ME) was added and cells were incubated for 10 min on ice, while shaking every 2 min. After, 20 ng plasmid DNA was added and the mix was left on ice for 30 min. This was followed by 30 s heat-shock at 42°C. Then, cells were left on ice for 2 min, diluted 1 in 10 with LB and incubated at 37°C, for 1 h at 250 rpm before plating on LB-agar. BL21-Codonplus-(DE3)-RIL & ArcticExpress (DE3): In summary, 2% (V/V) XL-10-gold (1:10) β-ME was added, and cells were left on ice for 10 min, while shaking every 2 min. After, 50 ng of plasmid DNA was added and cells were incubated on ice for 30 min, followed by heat-shock at 42°C for 20 s. Then, cells were left on ice for 2 min, diluted 1 in 10 with LB and incubated at 37°C, for 1 h at 250 rpm before plating on LB-agar. DH5α: 1-5 µL ligation mixture or 20 ng plasmid DNA was added and the mix was left on ice for 30 min. Then, heat shock was

performed for 45 s at 37°C, followed by a 2 min incubation on ice. After, cells were diluted 1:10 in LB medium and incubated at 37°C for 1 h at 250 rpm before plating on LB-agar. All strains were inoculated onto LB-agar plates supplemented with 100 µg/mL Ampicillin (Amp) (HTLV-1 pET-Duet-1 p62/p49 in-house/Gs, HTLV-1 pET15b) or 50 µg/mL Kanamycin (Kan) (HTLV-1 pET42a (+) RT, HTLV-1 pET42a (+) pol) and grown overnight at 37°C. QIAprep® Spin Miniprep Kit (Qiagen) was used for plasmid extraction for Sanger sequencing, PCR or restriction digest test. Concentration was measured using nanodrop (260 nm, Nanodrop oneC (Thermo scientific)).

2.2.2 HTLV-1 RT expression in *E. coli*

One shot BL21 (DE3) star: Cells were grown in presence of 100 µg/mL Amp (HTLV-1 pET-Duet-1 p62/p49 in-house/Gs, HTLV-1 pET15b) or 50 µg/mL Kan (HTLV-1 pET42a (+) RT, HTLV-1 pET42a(+) pol) till the OD₆₀₀ equals 0.6. For induction, both temperature and IPTG concentration were varied. BL21-CodonPlus-(DE3)-RIL: Cells were grown overnight in presence of 100 µg/mL Amp (HTLV-1 pET-Duet-1 p62/p49 in-house/Gs, HTLV-1 pET15b) or 50 µg/mL Kan (HTLV-1 pET42a (+) RT, HTLV-1 pET42a(+) pol) in addition to 50 µg/mL Chloramphenicol. Then, cells were diluted 1 in 20 and grown for 2 h at 37°C, in absence of antibiotics. Induction was performed by adding 1 mM IPTG for 2 h at 37°C. ArcticExpress (DE3): Cells were grown overnight at 37°C in presence of 100 µg/mL Amp (HTLV-1 pET-Duet-1 p62/p49 in-house/Gs, HTLV-1 pET15b) or 50 µg/mL Kan (HTLV-1 pET42a (+) RT, HTLV-1 pET42a(+) pol) and 20 µg/mL Gentamycin. After, the culture was equilibrated to 12°C at 220 rpm for 10 min, followed by induction for 24h upon addition of 1 mM IPTG. A summary of all the conditions tested can be found in Table 3.

2.2.3 Protein expression evaluation by SDS-PAGE

A 2 mL sample was taken before induction, and an OD₆₀₀-equivalent culture volume was taken after induction. Cells were spun down at max speed in a table-top centrifuge for 5 min, and the pellet was resuspended in lysis buffer (50 mM Tris pH 8.0, 600 mM NaCl, 5 mM MgCl₂, 10% glycerol, 5 mM Dithiothreitol (DTT), 0.1% Triton-X, 0.2 mg/mL lysozyme, 0.05% benzonase). Optionally, samples were sonicated (amplitude: 0.7, 10 seconds, 3 times) and centrifuged at max speed for 5 min. The supernatant (soluble fraction) was moved into a clean tube and the pellet (insoluble fraction) was resuspended in lysis buffer. Loading buffer (5x) was added to the samples, boiled at 95°C for 5 min and loaded onto 4-12% gradient or 12% Sodium Dodecyl sulphate Polyacrylamide gel electrophoresis (SDS-PAGE, (MiniProtean TGX stain free (4-20%, 12%), Bio-Rad)) gel, which was run for 30-40 min at 180 V. As marker, Precision Plus Protein™ Unstained Standard was used. Visualization was performed using the Gel Doc™ EZ imager (Bio-Rad) with analysis in Image Lab™ software.

Table 3. Overview of the tested conditions during recombinant protein expression.

Construct	Cell line	IPTG (mM)	Time (h)	Temperature (°C)
HTLV-1 pET15b	One Shot BL21 (DE3) star	0.05	2	28
	One Shot BL21 (DE3) star	0.2	5	17
	One Shot BL21 (DE3) star	0.2	5	30
	One Shot BL21 (DE3) star	0.5	3	37
	One Shot BL21 (DE3) star	0.5	18	20
	One Shot BL21 (DE3) star	0.6	5	17
	One Shot BL21 (DE3) star	0.6	5	30
	One Shot BL21 (DE3) star	2	2	28
HTLV-1 pET42a (+) RT	One Shot BL21 (DE3) star	0.05	2	28
	One Shot BL21 (DE3) star	0.2	3	20
	One Shot BL21 (DE3) star	0.2	3	30
	One Shot BL21 (DE3) star	0.5	3	37
	One Shot BL21 (DE3) star	0.5	18	20
	One Shot BL21 (DE3) star	0.5	18	30
	One Shot BL21 (DE3) star	0.6	3	20
	One Shot BL21 (DE3) star	0.6	3	30
	BL21-CodonPlus-(DE3)-RIL	1	2	37
	One Shot BL21 (DE3) star	2	2	28
HTLV-1 pET42a(+) pol	One Shot BL21 (DE3) star	0.05	2	28
	One Shot BL21 (DE3) star	0.5	3	37
	One Shot BL21 (DE3) star	0.5	3.5	37
	One Shot BL21 (DE3) star	0.5	6	20
	One Shot BL21 (DE3) star	0.5	6	30
	One Shot BL21 (DE3) star	0.5	6	37
	One Shot BL21 (DE3) star	0.5	18	20
	BL21-CodonPlus-(DE3)-RIL	1	2	37
	ArcticExpress (DE3)	1	24	12
	One Shot BL21 (DE3) star	2	2	28
HTLV-1 pET-Duet-1 p62/p49 Gs	One Shot BL21 (DE3) star	0.05	18	20
	One Shot BL21 (DE3) star	0.5	6	37
	One Shot BL21 (DE3) star	0.5	18	20
	One Shot BL21 (DE3) star	2	18	20
HTLV-1 pETDUET-1 p62/p49 in-house	One Shot BL21 (DE3) star	0.05	2	28
	One Shot BL21 (DE3) star	0.5	3	37
	One Shot BL21 (DE3) star	0.5	6.5	37
	One Shot BL21 (DE3) star	0.5	14	20
	One Shot BL21 (DE3) star	0.5	14	30
	One Shot BL21 (DE3) star	0.5	16	20
	One Shot BL21 (DE3) star	0.5	18	20
	BL21-CodonPlus-(DE3)-RIL	1	2	37
	ArcticExpress (DE3)	1	24	12
	One Shot BL21 (DE3) star	2	2	28

2.3 Protein purification

Larger cell cultures (6-10 L) were prepared using the best expressing conditions identified as previously described. Cells were harvested by centrifugation at 7,000g for 15 min and the pellet resuspended in lysis buffer. Per liter of LB culture, 20 mL lysis buffer (50 mM Tris pH 8.0, 600 mM NaCl, 5 mM DTT, 5 mM MgCl₂, 10% glycerol, 0.1% Triton-X, 1 mM PMSF, 5 mM imidazole) was used. The resuspended cells were freeze-thawed three times using liquid nitrogen and then sonicated on ice for 10 min in intervals of 10 s ON followed by 20 s OFF at 50% power (Branson 250 Sonifier). Lysed sample was centrifuged for 45 min at 13,000 rpm (Sigma 6-16KS) and the supernatant was separated from the pellet. Finally, the supernatant was filtered to get rid of all impurities that could interfere with subsequent purification steps.

Ni-NTA beads (HisPur™ Ni-NTA resin, Thermo scientific) were equilibrated with lysis buffer and added to the supernatant at approximately 0.5 mL per liter of culture, incubating overnight at 4°C. The next day, the beads were put into gravity columns and drained from the supernatant. The flowthrough was re-passed onto the beads and the bound protein was eluted using elution buffer (50 mM Tris pH 8.0, 200 mM NaCl, 5 mM MgCl₂, 1 mM DTT, 10% glycerol) with step-wise increasing imidazole concentration (0 mM, 5 mM, 10 mM, 50 mM, 100 mM, 250 mM). The collected fractions were analysed using Nanodrop (280 nm, Nanodrop oneC (Thermo scientific)) and SDS-PAGE to evaluate sample purity and yield.

2.4 Western blot

Samples obtained from recombinant protein expression and isolation were evaluated for the presence of the N-terminal 6xHis-tag on the expressed protein by western blot (WB). The polyvinylidene fluoride membrane (0.2 µm, Immobilon®-PSQ) was first soaked in methanol (MeOH) and then washed in transfer buffer (25 mM Tris, 190 mM glycine, 20% MeOH, 3.4 mM SDS). All components required for the western blot assembly were soaked in transfer buffer during setup. Blotting was performed at 300 mA for 90 min. Following this, the membrane was washed with Milli-Q water for 5 min while shaking at 4°C. The blot was then blocked using blocking buffer (3% milk powder in TBST, containing 0.1% Tween 20, 150 mM NaCl, 20 mM Tris, pH 7.5) for 30 min while shaking at 4°C. After blocking, the blot was incubated overnight at 4°C with mouse anti-6xHis primary antibody (1:3 000, Invitrogen) while shaking. The next day, the blot was rinsed five times with TBST for 5 min each, while rocking. It was then incubated for 1 h with horseradish peroxidase (HRP)-linked goat anti-mouse secondary antibody (1:5 000, Invitrogen) at room temperature. The blot was again rinsed five times with TBST for 5 min each, while rocking. After a final rinse for 10 min, the blot was incubated with a 1:1 mixture of SuperSignal™ West Pico PLUS

Luminol/Enhancer and SuperSignal™ West Pico PLUS Stable Peroxide (Thermo Scientific) for 2 min and visualized both chemiluminescently and colorimetrically. Final analysis was performed using Image Lab™ software (Bio-Rad).

2.5 Polymerase chain reaction

To perform PCR, the standard protocol described by New England Biolabs for Q5® High-Fidelity DNA Polymerase was used. Each reaction mix consisted of 190 µM dNTPs, Q5 Reaction buffer (5x), 0.5 U Q5 High-Fidelity DNA polymerase, 1 µM forward/reverse primer mix, 1 µL cDNA sample and 0.6 µL Midori Green intercalating dye (Nippon Genetics). As setup, initial denaturation (1 min, 98°C, 1 cycle) was followed by 25 cycles of denaturation (98°C, 10 s), annealing (30 s) and elongation (30 s/kB, 72°C). Final elongation (2 min, 72°C) was followed by an infinite hold at 4°C to preserve the sample.

2.6 RT-qPCR

2.6.1 RNA extraction and reverse transcription

RNA extraction was performed using the RiboPure-yeast kit (Invitrogen), and accompanying protocol. For reverse transcription, the final reaction mix consisted of 400 ng purified RNA mixed with 0.2 µg oligo-dT primer, AffinityScript RT-buffer (10x, Aligent technologies), 10 mM DTT, 1 mM dNTPs and 1.25% AffinityScript Multiple Temperature Reverse Transcriptase (Aligent technologies) all brought to final volume using MilliQ water (RNase- and DNase free). The mix was incubated at 42°C for 2 h, after which the reverse transcriptase was denatured by heating the mix to 70°C for 15 min. The resulting cDNA was stored at -20°C until usage.

2.6.2 Quantitative PCR

Based the One Shot BL21 (DE3) star genome (CP053602.1), primers were designed for detection of T7 RNA polymerase (T7RNAP), Siroheme synthase (CysG) and 3-phenylpropionate transporter (HcaT) (Table 4). To perform qPCR, the basic PCR protocol by New England Biolabs for Q5® High-Fidelity DNA Polymerase was used (see *Polymerase chain reaction*, Table 5).

2.7 Bioinformatics

A tabular overview of all tools used is shown in Table SI.

Table 4. Overview of the target genes, primer sequences and amplicon sizes for RT-qPCR.

Gene	Abbreviation	Genbank accession	Gene location in CP053602.1	Orientation	primers (5'-3')	Amplicon length (bp)
T7 RNA polymerase	T7RNAP	FJ881694.1	751 719-754 370	F	CGTGGCATAAGGAAGACTCTATT	108
				R	GACCTACTACGCCAGCATTT	
Siroheme synthase	CysG	VWQ03902.1	3 358 727-3 360 100	F	CCGTCTGGTTTCTGACGATATT	120
				R	CGCAGCAGGATCTGGTTAAT	
3-phenylpropionate transporter	HcaT	NP_417031.1	c(2 534 949-2 533 811)	F	GCTGGTGATGATTGGCTTTAAC	111
				R	AGTCGCACITTTGCCGTAAT	
HTLV-1 RT p62 subunit	p62			F	GTGCTGCTGCTGGATAAGTA	110
				R	TGGTCGCTGGTTTGGATAAA	
HIV-1 RT p51 subunit	p51			F	TCTATTCGGTACCTCTTGACG	109
				R	CCCCATAGGCAATACGTTGTATTGG	

Table 5. Overview of RT-qPCR experimental setup.

Target	Phase 1: initial denaturation		Phase 2:PCR			Phase 3: melting curve		
			Denaturation	Annealing	Elongation	Heating	Cooldown	Heating
HTLV-1 p62	50°C, 2 min	95°C, 10 min	95°C, 12 s	55°C, 30 s	72°C, 20 s	95°C, 15 s	60°C, 1 min	95°C, 1 s
HIV-1 p51				52.5°C, 25 s	72°C, 25 s			
	1 cycle		40 cycles			1 cycle		

2.8 Statistics

Experimental data are expressed as mean $\pm$ standard deviation (SD). Differences between two groups were analysed using parametric two-sample t-test. Differences between multiple ($n > 2$) groups were analysed using the nonparametric Kruskal-Wallis test. If a statistical difference was detected ($P < 0.05$), post hoc pairwise group comparisons were performed by two-sample Wilcoxon test with Bonferroni's multiple-testing correction. R (v4.3.1) was used for statistical analysis, Microsoft Excel for graphing. Differences were considered statistically significant at $P < 0.05$.

3. Results

To date, the structure of the active HTLV-1 RT/tRNA^{Pro}:ssRNA complex remains unknown and therefore analysing the mechanism has not been feasible. Fortunately, the amino acid sequences of both subunits, p49 and p62, are known allowing for identification of homologous proteins, and prediction of characteristics related to the behaviour and mechanism of both secondary and tertiary structure. For this, many bioinformatics tools have been developed that aid with HTLV-1 RT's sequence analysis.

3.1 Sequence analysis

3.1.1 General parameters and characteristics

Sequence derived physicochemical characteristics can give interesting insights into the behaviour of a protein in vitro and in cellulo. To that end, amino acid sequences of HTLV-1 p62 (P03362; 583-1167 - from now referred to as p62) and p49 (P03362; 583-1021 - from now referred to as p49), both as monomers and as HTLV-1 RT heterodimer, were given as input to *ProtParam* and *PepStats* sequence analysis algorithms^{91,92} (Table 6).

Table 6. General parameters and characteristics of HTLV-1 RT by *ProtParam* and *PepStats*.

^astable: < 40, unstable: > 40. ^bHigher values indicates higher thermostability. ^chydrophilic: < 0, hydrophobic: > 0.

Parameter	P62			P49			P62/p49 complex		
Length (AA)	585			439			1024		
Molecular weight (kDa)	65.5			49.1			114.67		
Theoretical isoelectric point	9.50			8.98			9.34		
Amino acid composition	Ala	39	6.7%	Ala	28	6.4%	Ala	67	6.5%
	Arg	28	4.8%	Arg	17	3.9%	Arg	45	4.4%
	Asn	19	3.2%	Asn	15	3.4%	Asn	34	3.3%
	Asp	18	3.1%	Asp	13	3.0%	Asp	31	3.0%
	Cys	8	1.4%	Cys	6	1.4%	Cys	14	1.4%
	Gln	47	8.0%	Gln	40	9.1%	Gln	87	8.5%
	Glu	13	2.2%	Glu	12	2.7%	Glu	25	2.4%
	Gly	25	4.3%	Gly	20	4.6%	Gly	45	4.4%
	His	25	4.3%	His	19	4.3%	His	44	4.3%
	Ile	29	5.0%	Ile	22	5.0%	Ile	51	5.0%
	Leu	95	16.2%	Leu	67	15.3%	Leu	162	15.8%
	Lys	18	3.1%	Lys	14	3.2%	Lys	32	3.1%
	Met	5	0.9%	Met	5	1.1%	Met	10	1.0%
	Phe	23	3.9%	Phe	17	3.9%	Phe	40	3.9%
	Pro	54	9.2%	Pro	42	9.6%	Pro	96	9.4%
	Ser	47	8.0%	Ser	31	7.1%	Ser	78	7.6%
	Thr	43	7.4%	Thr	35	8.0%	Thr	78	7.6%
	Trp	10	1.7%	Trp	8	1.8%	Trp	18	1.8%
	Tyr	15	2.6%	Tyr	10	2.3%	Tyr	25	2.4%
	Val	24	4.1%	Val	18	4.1%	Val	42	4.1%
[negatively charged residues]	31			25			56		
[positively charged residues]	46			31			77		
Formula	C ₂₉₈₉ H ₄₆₉₁ N ₈₁₃ O ₈₁₉ S ₁₃			C ₂₂₃₇ H ₃₄₉₉ N ₆₀₅ O ₆₂₁ S ₁₁			C ₅₂₂₆ H ₈₁₈₈ N ₁₄₁₈ O ₁₄₃₉ S ₂₄		

$\epsilon_{280\text{nm}}$ ($\text{M}^{-1}\text{cm}^{-1}$) (in water)	Naïve : 77850 Reduced : 77350	Naïve : 59275 Reduced : 58900	Naïve : 137125 Reduced : 136250
Estimated half-life (hours)	30	30	30
Instability index [‡]	58.84	55.39	57.37
Aliphatic index [‡]	101.23	97.33	99.56
Grand average of hydropathy [†]	-0.089	-0.148	-0.114
Average residue weight (Da)	112.029	111.977	111.989
Overall charge	27.5	15.5	43
Improbability of expression in inclusion bodies	0.887	0.835	0.866

3.1.2 Phylogeny

Looking for evolutionary related species that share great RT homology, multiple sequence alignment (MSA, Figure SII) was performed on 34 retroviral RT sequences from the *orthoretrovirinae* (lentiviruses, alpha-, beta-, gamma-, deltaretroviruses and intracisternal A-particle) subfamily. The phylogenetic tree revealed a clustering of the viruses into their respective genus and shows that HTLV-1 RT (p62) is more closely related to simian T-lymphotropic virus type 1 (STLV-1) RT than HTLV-2 RT, which is its human counterpart (Figure 9). In addition, this holds true for all other *deltaretroviruses*. These results prove that HTLV-1 RT is not very closely related to HIV-1 RT as they diverged into separate branches very early in the tree. In conclusion,

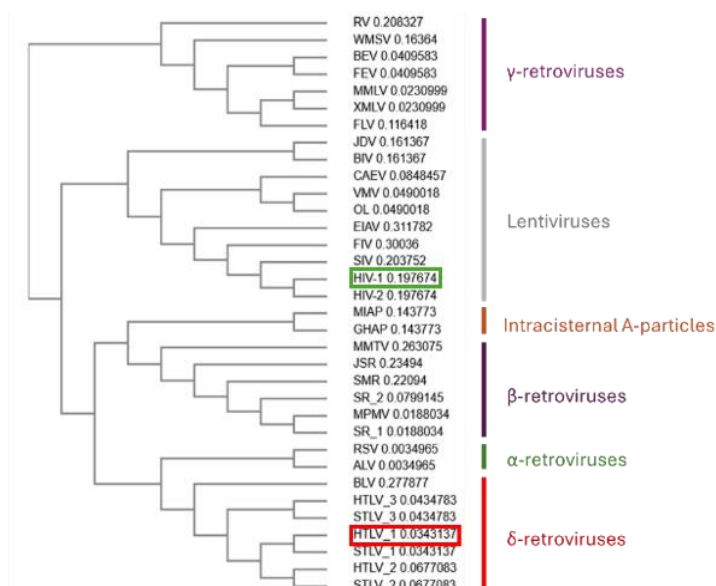


Figure 9. HTLV-1 RT is only distantly linked to HIV-1 RT.

Phylogenetic tree using amino acid sequences from 34 retroviral RTs. Numerical values correspond to the node support ([0,1]). Red box highlights HTLV-1 RT, green box highlights HIV-1 RT. RV: Reticuloendotheliosis virus, WMSV: Woolly money sarcoma virus, BEV: Baboon endogenous virus, FEV: Feline endogenous virus, MMLV: Moloney murine leukemia virus, XMLV: Xenotropic MuLV-related virus, FLV: Feline leukemia virus, JDV: Jembrana disease virus, BIV: Bovine immunodeficiency virus, CAEV: Caprine arthritis encephalitis virus, VMV: Visna/maedi virus, OL: Ovine lentivirus, EIAV: Equine infectious anemia virus, FIV: Feline immunodeficiency virus, SIV: Simian immunodeficiency virus, HIV-1: Human immunodeficiency virus type 1, HIV-2: Human immunodeficiency virus type 2, MIAP: Mouse intracisternal A-particle, GHAP: Golden hamster intracisternal A-particle, MMTV: Mouse mammary tumor virus, JSR: Jaagsiekte sheep retrovirus, SMR: Squirrel monkey retrovirus, SR_2: Simian retrovirus 2, MPMV: Mason-Pfizer monkey virus, SR_1: Simian retrovirus 1, RSV: Rous sarcoma virus, ALV: Avian leukosis virus, HTLV: Human T-cell lymphotropic virus (type 1, 2, 3), STLV: Simian T-cell lymphotropic virus (type 1, 2, 3).

HTLV-1 RT's evolutionary link to HIV-1 RT is rather weak but holds a great homology to other *deltaretroviral* RTs.

As seen before, many highly conserved residues in the HIV-1 RT active site are important for correct incorporation of an incoming dNTP. The first residue encountered upon active site entry is Y115, which acts as a steric gate. In the MSA, this tyrosine is 100% conserved within the lentiviruses (Figure SII). However, in HTLV-1 RT and 70% of the RTs, this tyrosine is mutated to a phenylalanine. Yet, an aromatic amino acid is present at this specific position in all retroviral RTs displayed, and hence serve as a functional steric gate. Next, the dNTP enters the polymerase active site and is coordinated by multiple residues. First, K65 and R72 of HIV-1 RT are responsible for binding P_β and P_γ of the incoming dNTP. These two residues are 100% conserved in all discussed retroviral RTs. Second, three catalytic carboxylates (D110, D185, D186) are present in the HIV-1 RT active site and are responsible for coordinating two Mg^{2+} ions. These ensure the correct orientation of the incoming dNTP's P_α for polymerisation. D185 and D186 are part of the highly conserved YxDD sequence in all retroviral RTs and is maintained in 97% of mentioned RT sequences. Q151 in the xLPQy HIV-1 RT motif interacts directly with the 3' hydroxyl group of the incoming dNTP and is conserved in 97% of all aligned RT sequences. An additional function of the dNTP coordinating residues, is correctly positioning the DNA into the polymerase active site. Upon comparing all residues involved in HIV-1 RT/DNA contact to HTLV-1 RT, only 49% of residues are fully conserved. However, when considering physicochemical properties instead of residue identity, 66% of residues are conserved. This is expected since not all amino acids are suitable for DNA interaction and thus similar amino acids can occur. A closer look was taken at the primer grip responsible for correctly positioning the 3' hydroxyl group of the primer strand into the polymerase active site. The HIV-1 RT primer grip sequence is very poorly conserved in all 34 RT sequences, but this increased when assessing the sequences only considering the physicochemical characteristics (Table 7). In conclusion, HTLV-1 RT holds great sequence homology to all *deltaretroviral* discussed RTs, but also to HIV-1 RT. This is in contrast with what was observed in the phylogenetic analysis where HTLV-1 and HIV-1 were found to be distantly related. This indicates that even though HTLV-1 and HIV-1 hold a weak evolutionary link, their RTs might have a lot in common structurally, functionally and mechanistically.

Table 7. **MSA results of the HIV-1 RT primer grip showing both identity and similarity.**

^ψF, Y, H and W. [†]A, V, L, P, M, G, I and C. [‡]D, E, K, N, Q, R, S and T.

HIV-1 RT primer grip		F	L	W	M	G	Y	E	L	H
Identity		3%	3%	27%	9%	79%	39%	30%	58%	18%
Aromatic ^ψ		24%	3%	85%	21%	0%	70%	3%	0%	24%
Aliphatic	Non-polar [†]	58%	24%	0%	79%	100%	21%	12%	94%	30%
	Polar [‡]	18%	73%	15%	0%	0%	9%	85%	6%	45%

3.1.3 Domain identification

To further substantiate previous findings, domain identification and gene ontology (GO) analysis was performed on the HTLV-1 RT (p62). Hence, algorithms capable of finding homologous protein sequences to a specific input query sequence, such as ScanProSite, InterProScan and PfamScan, were used⁹³. Based on the sequences found, a query sequence can be subdivided into functional domains that help clarify the function of the protein of interest. All above mentioned algorithms were used, as this allows for identification of proteins in multiple databases (PROSITE, InterPro, Pfam, PANTHER, SUPERFAMILY and CATH-Gene3D) at once. All of these databases contain signatures of conserved regions that are linked to detailed documentation on annotation of the protein family and conserved domains. Each search algorithm reports that both a RT and RNase H domain is present within the p62 sequence (Table 8). Uniquely, InterProScan reports an association between HTLV-1RT and HIV-1 RT in the polymerase domain, validating previous results. In addition, ScanProSite was able to detect three D-residues (D98, D173 and D174) that might be implicated in Mg²⁺ ion coordination in the polymerase active site, and four D-residues (D458, D492, D514 and D575) that perform the same function in the RNase H active site. Also, InterProScan identified great homology between HTLV-1 RT and the human endogenous retrovirus group K polymerase protein (HERV-K RT), a retrovirus family that is highly conserved within Old World monkeys and apes, associated various types of cancer and neurological diseases. Lastly, GO analysis indicated that HTLV-1 RT is functional in nucleic acid binding and RNA-DNA hybrid ribonuclease activity. In conclusion, and in addition to the phylogenetic and MSA analysis, it was proven that HTLV-1 RT should be able to perform RT-like functions. Also, these data show that HTLV-1 RT is linked evolutionary and functionally, to both HIV-1 RT and HERV-K RT.

Table 8. Results from domain identification using ScanProSite, InterProScan and PfamScan.

Algorithm	Region	Identified domain	Name	Predicted feature	Accession number	Database
ScanProSite	32-222	RT_POL	Reverse transcriptase (RNA-dependent DNA polymerase)	Mg2+ binding: D98, D173, D174	PS50878	ProSite
	449-583	RNASE_H_1	RNase H 1 domain	Mg2+ binding: D458, D492, D514, D575	PS50879	ProSite
InterProScan	32-222	RT_POL	Reverse transcriptase (RNA-dependent DNA polymerase)		PS50878, IPR000477	ProSite, InterPro
	449-583	RNASE_H_1	RNase H 1 domain		PS50879, IPR002156	ProSite, InterPro
	32-222	RT_dom	Reverse transcriptase (RNA-dependent DNA polymerase)		IPR000477	InterPro
	32-222	RT_POL	Reverse transcriptase (RNA-dependent DNA polymerase)		PS50878	ProSite
	51-222	RVT_1	Reverse transcriptase (RNA-dependent DNA polymerase)		PF00078	Pfam
	454-564	RNase_H	RNase H 1 domain		PF00075	Pfam
	449-583	RnaseH_domain	RNase H 1 domain		IPR002156	InterPro
	449-583	RNASE_H_1	RNase H 1 domain		PS50879	Pfam
	78-186, 215-308	Rev_trsase/Diguanyl_cyclase	Reverse transcriptase (RNA-dependent DNA polymerase)		IPR043128	InterPro
	78-103, 145-186, 215-308	Rev_trsase/Diguanyl_cyclase	Reverse transcriptase (RNA-dependent DNA polymerase)		G3DSA:3.30.70.270	CATH-Gene3D
	448-579	RNaseH_sf	RNase H 1 domain		IPR036397	InterPro
	448-579	RNaseH_sf	RNase H 1 domain		G3DSA:3.30.420.10	CATH-Gene3D

	6-562	DNA/RNA_pol_sf	Reverse transcriptase (RNA-dependent DNA polymerase)		IPR043502	InterPro
	6-345, 477-562	DNA/RNA polymerases	Reverse transcriptase (RNA-dependent DNA polymerase)		SSF56672	SUPERFAMILY
	11-580	ENDOGENOUS RETROVIRUS GROUP K MEMBER POL PROTEIN			PTHR41694	PANTHER
	7-77, 104-144	HIV Type 1 Reverse transcriptase subunit A			G3DSA:3.10.10.10	CATH-Gene3D
PfamScan	51-222	RVT_1	Reverse transcriptase (RNA-dependent DNA polymerase)		PF00078	Pfam
	454-564	Rnase_H	RNase H 1 domain		PF00075	Pfam

3.1.4 Pairwise sequence alignment

Since HIV-1 RT is the most studied RT to date and was previously found to be homologous by domain identification, it is important to look how these proteins compare one-to-one. For this, Dotlet was used, a tool that provides a two-dimensional graphical view of a pairwise sequence alignment⁹⁴. As the tool was parameterised to a window size of 50 AAs, a single highly conserved region is seen (Figure 10A). This corresponds to N137 – D186 in HIV-1 p66, where a major part of the polymerase active site is located, and N125 – D174 in HTLV-1 p62. On sequence level, great conservation is seen at the YMDD motif which contains two metal ion coordinating aspartic acids (Figure 10B, red), but also at the overall highly conserved VLPQG amino acid motif (Figure 10B, green). Similarly, high local conservation between the two sequences is seen when matching the predicted D98 of HTLV-1 p62 to D110 of HIV-1 p66, known to be important in metal ion coordination (Figure 10C, Table 8). Since HERV-K RT matched the whole sequence of HTLV-1 p62 by InterProScan domain identification (Table 8), it is interesting to perform a similar analysis comparing HTLV-1 RT's p62 to HERV-K RT (sequence PDB: 7SR6). Previous research found that D121, D196 and D197 in the polymerase active site of HERV-K RT correspond to D110, D184 and D184 in HIV-1 RT's polymerase active site⁹⁵. Alike previous comparison, high conservation is seen after aligning the VLPQG and Y[I/M]DD motif in both sequences, in addition to alignment of HTLV-1 RT D98 to HERV-K RT D121 (Figure 10D-F). Taken together, these and previous data provide a first glimpse of the possibility that HTLV-1 RT performs its function in a very similar way to HERV-K RT and HIV-1 RT. In addition, it shows that general mechanics about HERV-K RT and HIV-1 RT could be largely extrapolated towards HTLV-1 RT.

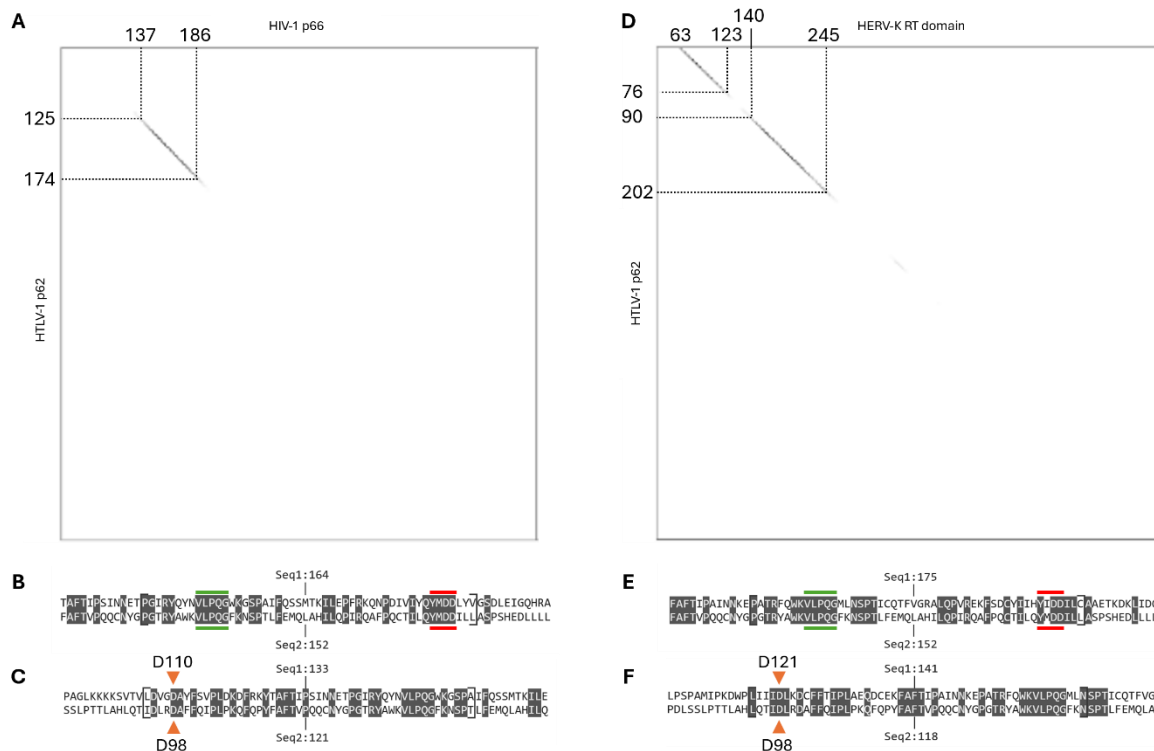


Figure 10. DotPlot shows sequential conservation at the polymerase active site between HTLV-1 RT, HIV-1 RT and HERV-K RT.

(A) Dotplot of pairwise sequence alignment of HTLV-1 RT and HIV-1 RT. (B) Pairwise sequence alignment of HTLV-1 RT and HIV-1 RT, focus on the YMD and VLPQG motif. (C) Pairwise sequence alignment of HTLV-1 RT and HIV-1 RT, focus on HTLV-1 RT D98 and HIV-1 RT D110 alignment. (D) Dotplot of pairwise sequence alignment of HTLV-1 RT and HERV-K RT. (E) Pairwise sequence alignment of HTLV-1 RT and HERV-K RT, focus on the YMD and VLPQG motif. (F) Pairwise sequence alignment of HTLV-1 RT and HERV-K RT, focus on HTLV-1 RT D98 and HERV-K RT D121 alignment. (A-F) Figures made using Dotlet tool, window size = 50, scores were restricted to [0, 149]. (B, C, E, F) Seq1 refers to HIV-1 RT (B, C) and HERV-K RT (E, F), seq2 refers to HTLV-1 RT.

3.1.5 Secondary structure prediction

The secondary and tertiary structure of a protein are highly dependent on the amino acid sequence. In absence of a structure of HTLV-1 RT, predictions can be made purely based on the combined physicochemical properties of the amino acids. By this, the presence of certain highly conserved region such as the primer grip, helix clamp or YxDD motif can be confirmed without having to rely on structure solving. ProtScale allows for visualisation of several predicted properties solely based on amino acid sequence, however here only α -helix and β -sheet propensity were analysed⁹². At first sight, six regions of very high α -helix propensity are distinguishable on the graph (Figure 11A, green arrows). The location of the polymerase active site, RNase H active site, primer and template grip were identified previously by sequence alignment of HIV-1 RT, HERV-K RT and HTLV-1 RT (Figure SII, Figure 10, Table 8). Previously, it was determined that the active site sequence of both HIV-1 RT and HERV-K RT was highly conserved compared to HTLV-1 RT (Figure 10). The YxDD motif of HIV-1 RT and HERV-K RT lays within a turn connecting two β -sheets (HIV-1 RT: β 9- β 10, HERV-K RT: β 7- β 8). Region 171-174 in HTLV-1 RT suggests a similar motif as it is preceded and followed by increased β -sheet propensity (Figure

11B). In addition, both D110 in HIV-1 RT and D121 in HERV-K RT are located directly downstream of a β -sheet (HIV-1 RT & HERV-K RT: β 4), a similar pattern is distinguishable at D98 in HTLV-1 RT as β -sheet propensity strongly decreases in this region (Figure 11C). The primer grip in HIV-1 RT and HERV-K RT is located downstream of the polymerase active site and composes of a linker region between two β -sheets (HIV-1 RT: β 12- β 13, HERV-K RT: β 9- β 10). MSA proposes that region 215-223 is responsible for the formation of the primer grip in HTLV-1 RT, which indeed shows a slight increase in β -sheet propensity (Figure SII, Figure 11D). In addition, a severe increase in β -sheet propensity is observed in the region 220-230, which is possibly a third β -sheet. The helix clamp is

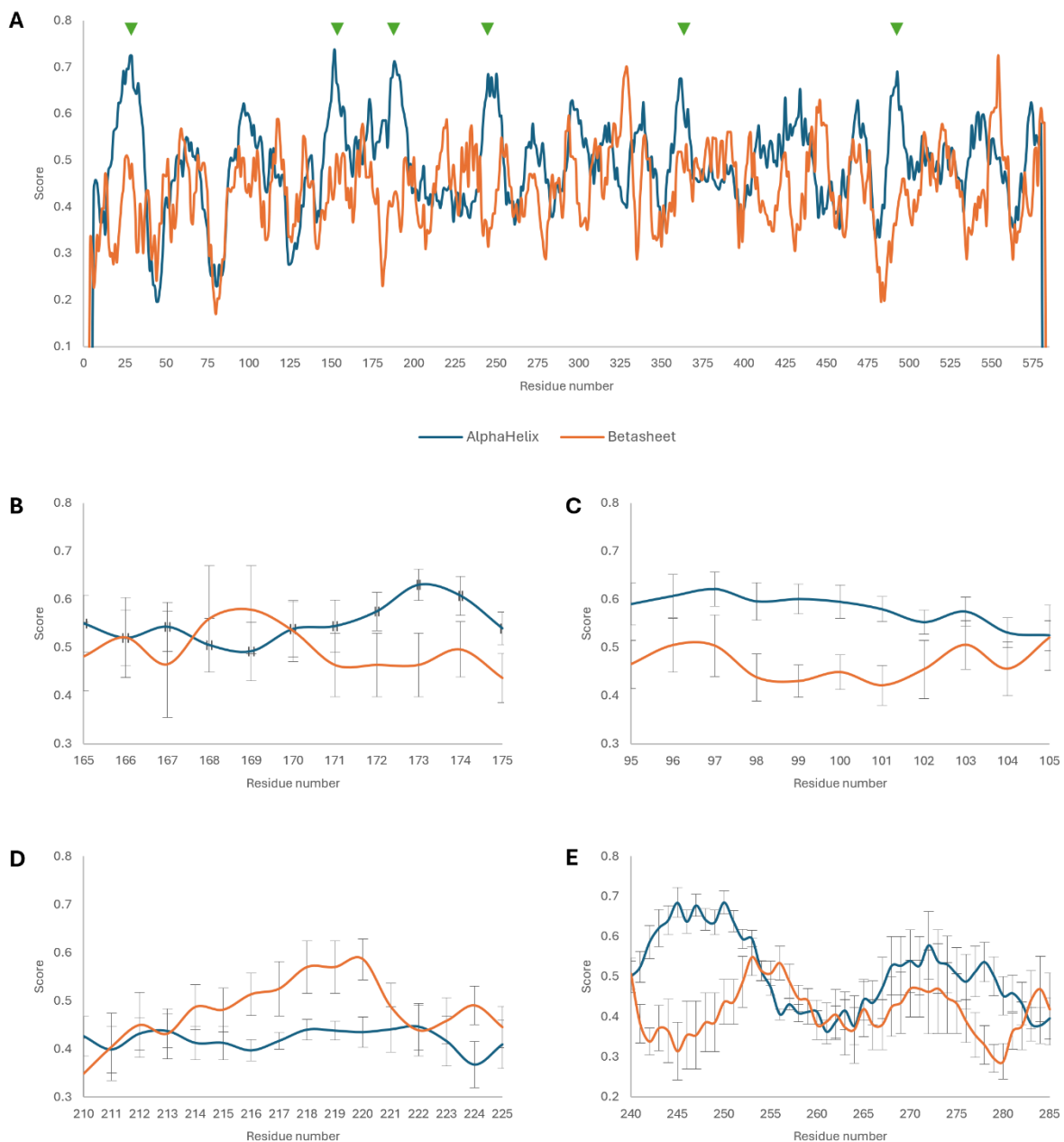


Figure 11. **Secondary structure prediction based on sequence via ProtScale.**

(A) Whole sequence overview. Green arrows indicate regions of high α -helix propensity. (B) Focus on polymerase active site of p62. (C) Focus on D98 residue of p62. (D) Focus on primer grip region in p62. (E) Focus on template grip region in p62. (B-E) data is represented as mean $\pm$ SD (n = 3).

a third highly conserved region in HIV-1 RT and HERV-K RT, and seems to correspond to region 243-280 in HTLV-1 RT. This conserved structure is composed of two α -helices linked by a small disordered linker oligopeptide. Indeed, a hill-valley-hill pattern in the α -helix propensity of p62 is seen, indicating a motif consisting of two high α -helix propensity regions connected by a low α -helix propensity region (Figure 11E). In addition, this linker region also does not show a severe increase in beta sheet propensity, further proving that this is most likely a disordered linker oligopeptide between two α -helices. In conclusion, this approach allows for identification of conserved regions within HTLV-1 RT that largely resemble structures seen in RTs from HIV-1 and HERV-K. The presence of these regions further substantiates previous data that HTLV-1 RT structure will greatly resemble HIV-1 RT.

3.2 In silico HTLV-1 RT model

Great progress has been made in sequence-based protein structure prediction, especially with the arrival of AlphaFold2 (AF2). As multiple approaches have proven their value over the years, models were generated by six different structure prediction algorithms, where AF2, OmegaFold and ESMFold were employed for ab initio protein structure prediction and Swiss-Model, Phyre² and Modeller for homology-based structure prediction⁹⁶⁻¹⁰¹.

3.2.1 Homology-based structure prediction

In homology modelling, five stages can be distinguished. To start, all possible templates are identified through sequence similarity tools such as Basic Local Alignment Search Tool (BLAST) against a curated database of experimentally determined structure¹⁰². Then, the query sequence is aligned against the selected templates to look for homologous regions, commonly done by Clustal Omega or MUSCLE algorithms. Following this, a preliminary model is constructed based on the structural information of the template(s) used, after which the model is refined using energy minimization (GROMACS or AMBER), molecular dynamics simulations (MD) and loop modelling. Finally, the model is evaluated to assess reliability and quality using the algorithms unique scoring functions.

After multiple sequence alignment, all three structure prediction algorithms found HERV-K RT (PDB: 7SR6) to be the most optimal template to use in HTLV-1 RT model construction (Table SIII). Phyre² structure shows the conserved domains within RTs (finger, palm, thumb, connection and RNase H), which can also be distinguished in the models by SWISS-MODEL and Modeller (Figure 12A-B, D). Both Modeller and SWISS-MODEL return a per-residue predicted Local Distance Difference Test score (pLDDT, [0,1]), indicating the local accuracy and stereochemical validity.

Most of the scores of the SWISS-MODEL model vary between 0.6 and 0.8, suggesting that these are of adequate accuracy (Figure 12C). Yet, some regions seem to have a low to very low score (<0.5). Modeller, however, shows great variation in pLDDT scores and does not seem to have high scoring regions (Figure 12E). These results indicate that the models should be used with caution as they could contain great inaccuracies in spatial configuration or folding.

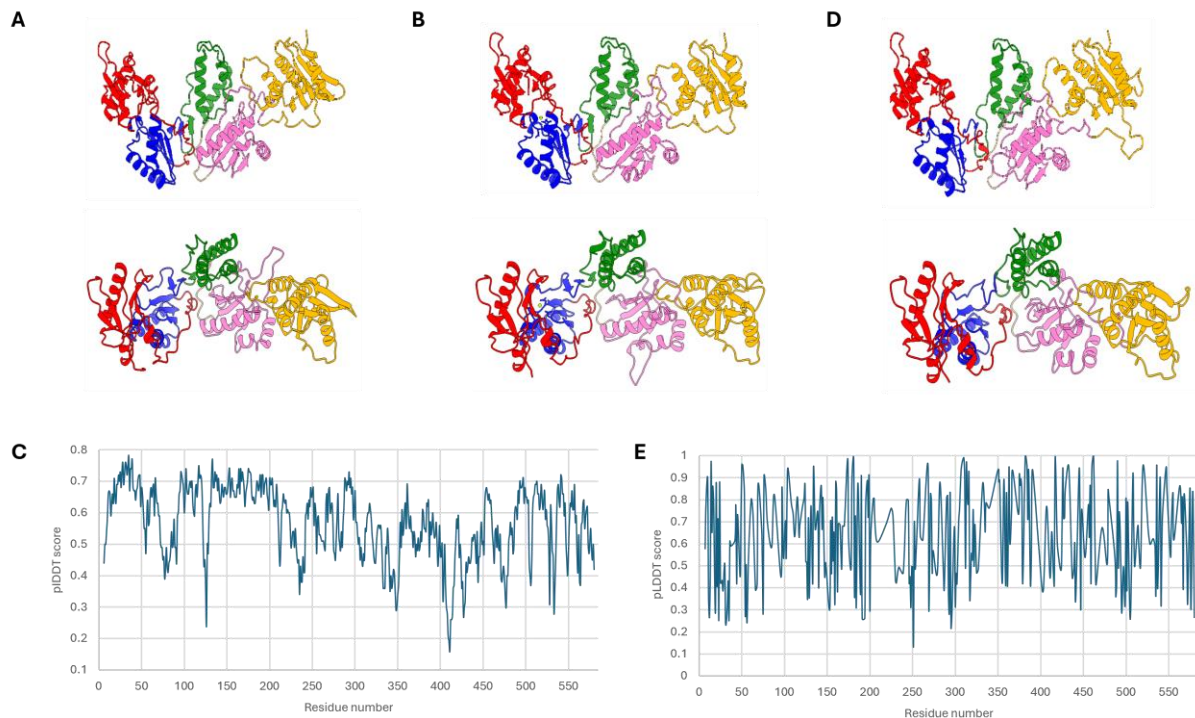


Figure 12 . **Quality assessment of HTLV-1 RT model by homology modelling.**

(A) Front view and top view of Phyre² predicted model. (B) Front view and top view of SWISS-MODEL predicted model. (C) Per-residue pLDDT plot for the SWISS-MODEL predicted model. (D) Front view and top view of MODELLER predicted model. (E) Per-residue pLDDT plot for the Modeller predicted model. (A-B, D) colour code: Finger, Palm, Thumb, Connection, RNaseH domain.

3.2.2 Ab initio structure prediction

Ab initio structure prediction uses solely computational methods based on the fundamental principles of protein folding to predict a query protein's structure from its sequence. For this, Monte Carlo sampling or MD simulations are employed to generate and evaluate a large number of candidate structures, iteratively adjusting the conformations of both backbone and sidechains to minimize the energy of the system. Once a structure is obtained, it is refined using energy minimization methods, MD simulations, after which the quality is assessed using predefined validation criteria and scoring functions. It is important to mention that AF2 is not a true ab initio structure prediction model but combines elements from homology modelling (e.g. MSA), ab initio modelling and deep learning. In contrast, OmegaFold and ESMFold are true ab initio modelling algorithms since these do not perform MSA to find homologues sequences or templates.

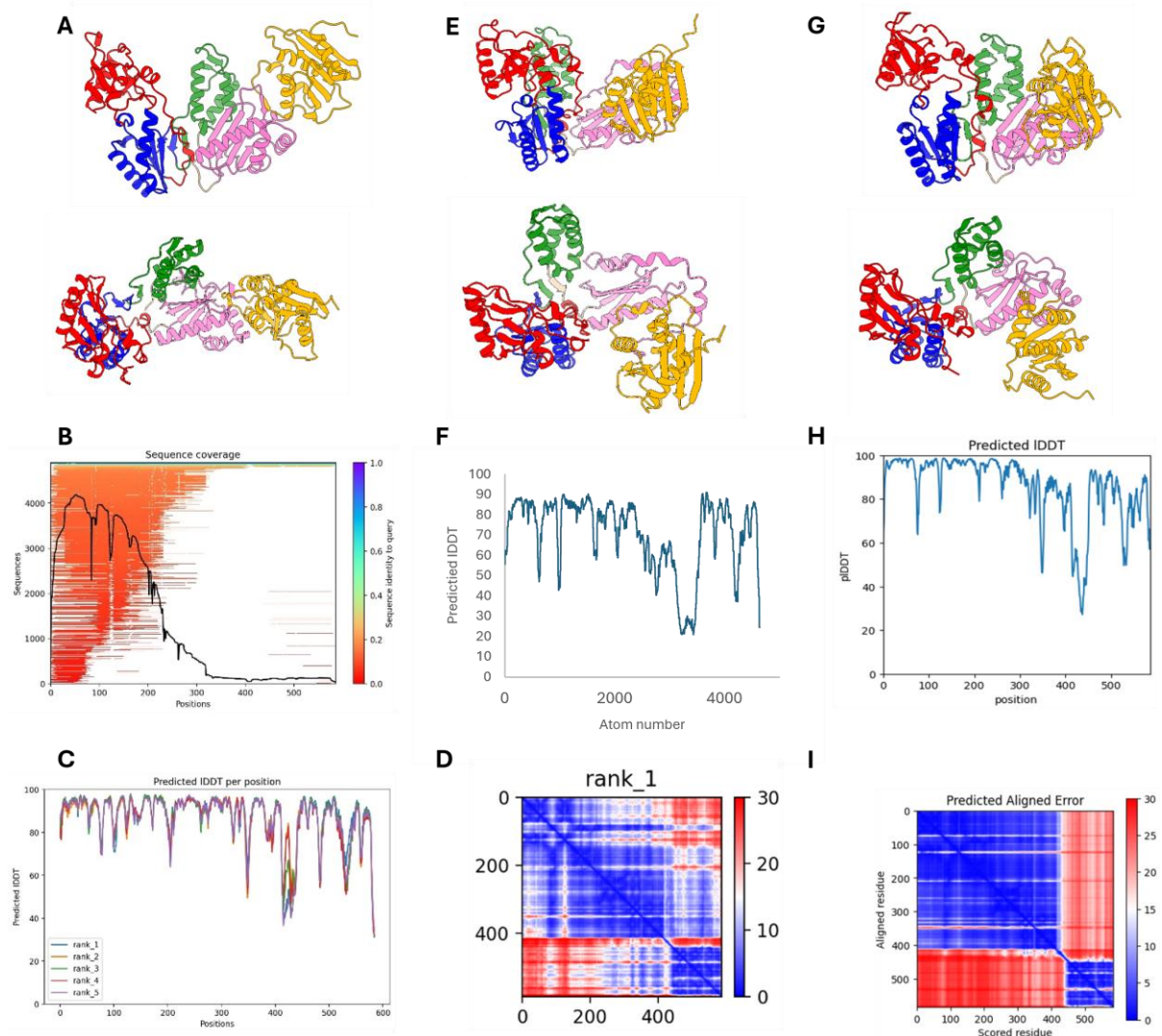


Figure 13. **Quality assessment of HTLV-1 RT model by ab initio modelling.**

(A) Front view and top view of AF2 predicted model. (B) AF2 sequence coverage plot. (C) AF2 pLDDT plot. (D) AF2 PAE plot. (E) Front view and top view of OmegaFold predicted model. (F) OmegaFold pLDDT plot. (G) Front view and top view of ESMFold predicted model. (H) ESMFold pLDDT plot. (I) ESMFold PAE plot. (A, E, G) colour code: Finger, Palm, Thumb, Connection, RNaseH domain.

AF2 provided a predicted model for HTLV-1 RT (p62) where 92,1% of all residues held a pLDDT value larger than 70, and 58,9% greater than 90 (Figure 13A & C). In addition, 46 residues held a pLDDT score below 70 in a region where low sequence coverage is present (Figure 13B). The predicted alignment error (PAE) of the model distinguishes three domains, namely polymerase domain (G1-N413, A433-K436), disordered linker domain (L414-L432) and RNase H domain (A437-L585) (Figure 13D). Both OmegaFold and ESMFold returned less extensive output information than AF2, and are therefore harder to analyse correctly. OmegaFold returned a model of HTLV-1 RT (p62) in which only 63.1% of residues holds a pLDDT score larger than 70, and only 0.97% larger than 90 (Figure 13E-F). A big cluster of sub 30 scores is present in the region spanning N399-L437. This represents a highly disordered linker region between the connection domain and RNase H domain, similar to what was seen previously in AF2 (Figure 13C & F). Fortunately,

ESMFold returns more quality parameters as part of the output, such as pLDDT and PAE data (Figure 13G-I). First of all, 83.5% of all residues in the model hold a pLDDT larger than 70, and 43.6% larger than 90 (Figure 13H). Lowest pLDDT values were present in the region A416-V447 where pLDDT was mainly below 40, again similar to the OmegaFold and AF2 models (Figure 13C, F & H). The PAE of the model distinguishes three distinct regions, namely polymerase domain (G1-N413), disordered linker domain (L414-I449) and RNase H domain (N450-L585) (Figure 13I). In conclusion, OmegaFold and ESMFold return low resolution structures compared to AF2. On first sight, AF2's predicted model seems to be the most accurate and reliable out of all generated models.

3.2.3 Comparative quality assessment

Since each model performs quality control using its own unique scoring functions, it is difficult to truly compare each model to determine which is the best. For this, SAVES was used, a webserver which allows to run Verify3D and PROCHECK simultaneously on each of the models¹⁰³⁻¹⁰⁵. As a published experimental control structure, the p66 subunit of HIV-1 RT (PDB: 6UIT) solved by X-ray crystallography to a resolution of 2.81 Å was used¹⁰⁶.

Verify3D determines the compatibility of an atomic model with its respective amino acid sequence and compares it with experimentally determined high-resolution structures^{104,105}. If 80% of the protein structure has a 3D-1D compatibility score equal or greater than 0.1, the structure passes the test. In this case, only AF2 and ESMFold passed as they obtained a score of 82.05% and 81.37% respectively. Phyre² (69.67%), SWISS-MODEL (69.04%), Modeller (66.38%) and OmegaFold (69.40%) hardly came close to passing the test. Surprisingly, also the experimentally determined structure of HIV-1 RT p66 failed this test with a score of 77.80%. In addition, per residues scores show a similar pattern in all models (Figure 14A). Interestingly, the AF2 and ESMFold predicted models are consistently part of the best scoring subset. Also, the severe decrease in average score in the region 400-450 greatly resembles the low pLDDT region seen in AF2 (Figure 13C). This again could be an issue of low sequence coverage by MSA for AF2, Modeller, SWISS-MODEL and Phyre². This low scoring region is also present in ESMFold and OmegaFold models and could indicate that this region is truly intrinsically disordered.

Several stereochemical parameters were assessed by PROCHECK, for which each respective Z-score was calculated¹⁰³ (Figure 14B). AF2 is reported to have the highest Ramachandran plot validity ($z = 0.49$), also outscoring the experimentally determined HIV-1 RT p66 structure ($z = 0.25$). No models show any unexpected planarity values, proving that the peptide bonds in the structure's backbone hold a physiologically accurate and relevant ω angle. Furthermore, slight C_{α}

tetrahedral distortion is present in all predicted model backbones, as measured by the ζ notional torsion angle that is related to the O3'-P bond. Finally, AF2's model holds an overall G-factor value (0.1, $z = 1$) that is close to the experimentally determined HIV-1 RT p66 structure (0, $z = 0.66$), whereas all others hold low to very low values. This means that the AF2 model holds the most normal stereochemical properties out of all predicted models.

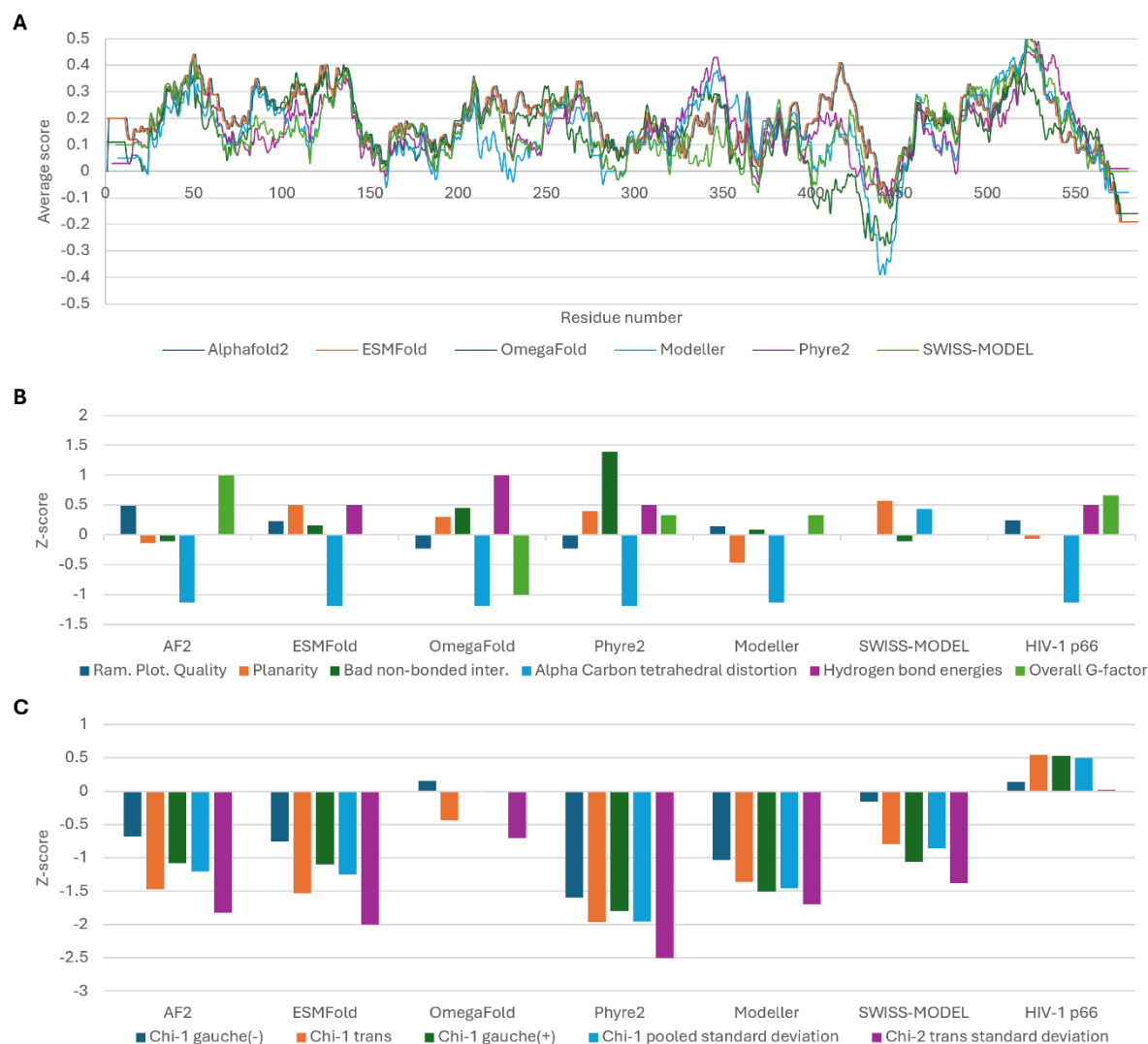


Figure 14 . Comparative quality assessment through VERIFY3D and PROCHECK.

(A) Per-residue average score by VERIFY3D. (B, C) Normalised PROCHECK output per model.

The chi-1 (χ_1) angle, defined as the angle between C_α and C_β , is restricted by the steric hindrance between the main chain and the gamma side chain atoms. Its conformations are referred to as *gauche(+)* ($64.1 \pm 15.7^\circ$), *trans* ($183.6 \pm 16.8^\circ$) and *gauche(-)* ($-66.7 \pm 15.0^\circ$). In addition, chi-2 (χ_2) refers to the angle between C_β - C_γ in the side chain, with a value of $177.4 \pm 18.5^\circ$. Compared to the experimentally determined HIV-1 RT p66 structure, all models showed acceptable χ_1 - χ_2 plots with no detected outliers (Figure 14C). Modeller and Phyre² are shown to have a highly divergent

χ_1 *gauche*(+) angles from the reference structures ($z = -1.6$ and $z = -1.03$, respectively), whereas the χ_1 *trans* angles are worst in the structures of AF2 ($z = -1.47$), ESMFold ($z = -1.52$), Phyre² ($z = -1.96$) and Modeller ($z = -1.35$). All structures except ESMFold and the HIV-1 RT p66 show highly diverging χ_1 *gauche*(-) angles. In conclusion, all models, including the HIV-1 RT p66 control, show deviations in χ_1 and χ_2 but not in ω indicating that while the backbone of the structure is reliable, some of the residue side chains are not. Yet, by various scoring parameters it was determined that the model by AF2 is the most reliable.

3.2.4 Spatial Model comparison

Since homology modelling and ab initio modelling was performed, it is important to check whether these produce similar protein models. If true, it increases the reliability of the predicted structure used in further analysis and model building. The greatest differences are present upon matching the AF2 predicted structure onto any of the other structures, resulting in a root mean square deviation (RMSD) of 1.139 ± 0.030 Å based on 248 ± 8 pruned pairs and 5.742 ± 0.693 Å based on 575.5 ± 1.5 pairs out of 585 AAs (Table SIV-SVIII). In contrast, mutually matching the other non-AF2 structures resulted in RMSD equal to 0.606 ± 0.29 Å for 375.5 ± 89.5 pruned pairs and 5.212 ± 0.724 Å for 574 ± 0.5 pairs out of 585 AAs. It is worth mentioning that SWISS-MODEL and Phyre² showed great similarity with RMSD of 0.316 Å for 468 pruned pairs and 4.988 Å for 575 pairs out of 585 AAs. Overall, none of the models tested showed greatly diverting structures. In addition, all models were aligned to wild-type HIV-1 RT where the reported overall RMSD values are 1.063 ± 0.132 Å for 188.5 ± 32.5 pruned pairs and 5.987 ± 0.745 Å for 516 ± 6 pairs out of 585 AAs. The models generated by OmegaFold and ESMFold were kept from further analysis as overlap showed great structural differences with all other models, but also with the HIV-1 RT p66 reference structure. Interestingly, a RMSD of 0.981 Å for 237 pruned pairs and 16.540 Å for all pairs was reported after aligning the OmegaFold and ESMFold models, suggesting that are vastly different from the others, but also from each other. For both, the connection and RNase H domain align very well, whereas finger, palm and thumb domain vary significantly.

Residues K65, R72, D110, V111, Q151, D185, D186 are necessary for catalysis in HIV-1 RT, with Y115 acting as a steric gate. To look for similarly positioned residues in the predicted model of HTLV-1 RT, residue G1-L315 (finger-palm-thumb) were isolated from the HTLV-1 RT AF2 model and superimposed with P1-L315 (finger-palm-thumb) of HIV-1 RT p66. Following residues in HTLV-1 RT AF2 model were found to spatially match the required residues in HIV-1 RT, namely K65:K53, R72:R59, D110:D98, V111:L99, D185:D173, D186:D174 and Y115:F103 (Table SIX-SX). It is important to note that the necessity of D98, D173 and D174 in metal ion binding was reported

previously by ScanProSite (Table 8) and that these spatially overlapping residues also overlap sequentially as seen in pairwise sequence alignment between HTLV-1 RT p62 and HIV-1 RT p66 (Figure 10A-C). Furthermore, the distance comparison between the two models shows that the active site of HTLV-1 RT is predicted to be less compact than the HIV-1 RT active site (Table SIX-SX). In conclusion, it seems that the AF2 model does not differ much in overall structure compared to the models obtained by homology modelling. This in stark contrast with the OmegaFold and ESMFold structures that show distinct folding differences compared to the others.

3.2.5 Mg²⁺ docking

RTs are only functional in the presence of two octahedrally coordinated metal ions, mainly Mg²⁺, as these are required for catalysis. Based on structural conservation of the polymerase active site between HIV-1 RT and the predicted HTLV-1 RT model, in addition to the prediction by ScanProSite (Table 8), residues D98, L99, D173 and D174 are most likely responsible for metal ion coordination. Unfortunately, none of the used protein structure prediction algorithms have the ability to predict and bind ligands into the model. To overcome this, the AF2 model was given as input to the MIB2 server which is able to identify metal ion binding sites as the input protein structure is compared to all templates in its database¹⁰⁷. A final model was constructed where Mg²⁺_A is coordinated by D98, D173 and D174 and Mg²⁺_B by D98, L99 carboxyl group and D173. As a control, MIB2 was run once for each previously generated model, including ESMFold and OmegaFold. All models showed a predicted Mg²⁺ coordination for D98, L99, D173 and D174, plus each logical combination of these residues as Mg²⁺_A: (D98, D173, D174) and

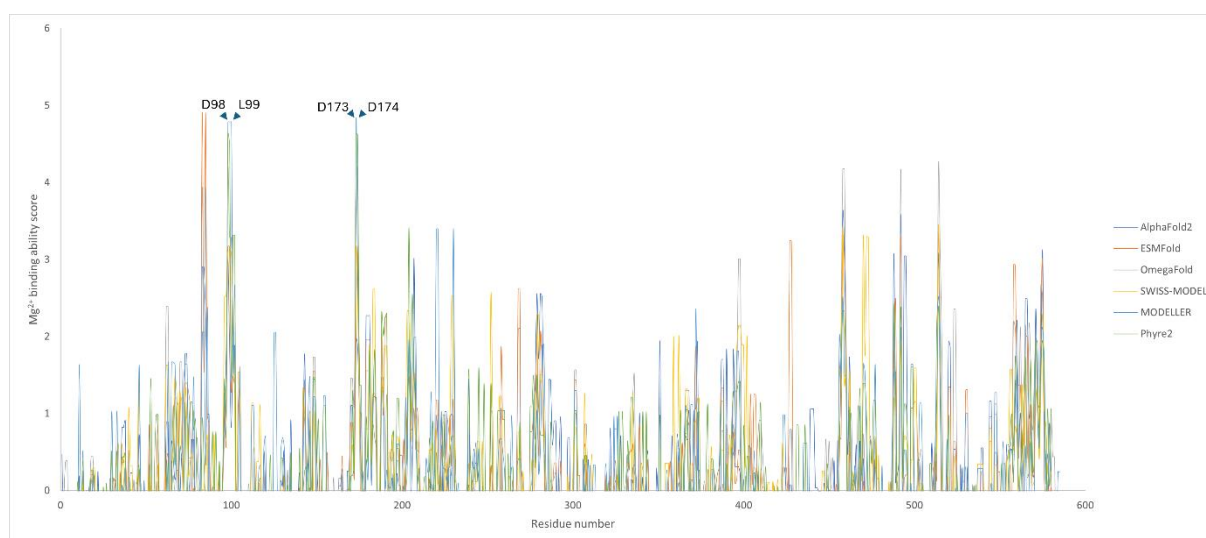


Figure 15. MIB2 Mg²⁺ docking on all models showed increased Mg²⁺ binding capacity at previously proposed residues D98, L99, D173 and D174.

Mg²⁺_B: (D98, L99, D173) (Figure 15). These data show that the incorporated Mg²⁺ are present in physiologically relevant positions, each coordinated in a similar way as seen in HIV-1 RT.

3.2.6 DNA docking

Since HTLV-1 RT's predicted model is highly superimposable with HIV-1 RT, a lot of the DNA-coordinating residue are likely to be very conserved. For this, HTLV-1 RT was superimposed with HIV-1 RT and spatially equivalent amino acids were identified (Table 9).

Table 9. Spatially equivalent DNA binding residues after HIV-1 RT/HTLV-1 RT superimposition¹²³.

HIV-1 RT	HTLV-1 RT		HIV-1 RT	HTLV-1 RT		HIV-1 RT	HTLV-1 RT		HIV-1 RT	HTLV-1 RT
W24	F14		D113	D101		Q255	E245		A355	P345
F61	F49		A114	A102		C258	Q247		G359	T349
K65	K53		Y115	F103		K259	A248		H361	S350
R72	R59		Q151	Q139		G262	G251		K374	L364
L74	I61		G152	G140		K263	E252		R448	T461
D76	D63		K154	K142		N265	Q254		K451	R463
R78	R65		P157	P145		W266	W255		T473	A488
N81	N68		Y183	Y171		C280	Y271		N474	Q489
E89	S75		M184	M172		R284	R276		K476	R490
Q91	S78		D185	D173		G285	H277		Q500	K516
L92	P79		D186	D174		T286	T278		Y501	Y517
G93	G80		M230	L218		L289	P280		H539	H560
I94	P81		G231	G219		K353	H343			

PyDockDNA was employed, an algorithm that is able to predict protein-DNA interactions between a given model and substrate¹⁰⁸. This algorithm scores over 10,000 DNA conformations based on electrostatic, desolvation and vdW energy, in combination with the percentage satisfied distance restraints to the input residues (Table 9). HIV-1 RT p66 was used as a reference structure to

validate the DNA positioning in the nucleic acid binding cleft and polymerase active site. As in Mg^{2+} docking, the DNA docking was repeated for each previously generated protein model to ensure reproducibility. In the end, the AF2 model was used to construct the final model. First, the reported mean electrostatic energy over all ranks is -98.67 ± 17.53 kcal/mol, while rank one holds an electrostatic energy of -97.303 kcal/mol (Figure 16A). Second, the mean desolvation energy estimated over all ranks is 21.59 ± 9.21 kcal/mol, whereas rank one holds a value of 6.60 kcal/mol (Figure 16B). Third, the mean vdW energy for all ranks is 8.41 ± 28.51 kcal/mol, which is much larger than rank one's vdW energy, namely -36.22 kcal/mol (Figure 16C). Lastly, rank one showed the and lowest PyDockDNARst scoring energy (-163.67 kcal/mol) out of all predicted conformations (Figure 16D). The post-DNA docking models of SWISS-MODEL, Modeller and Phyre² were overlayed with the AF2 DNA model, using HIV-1 RT p66 as a reference (Figure 16E-F). It can be seen that the DNA dimer in the AF2 model greatly represents the DNA binding state of HIV-1 RT, in both overall conformation and in the active site. Furthermore, the model of SWISS-MODEL and Modeller differ slightly in general DNA binding, but remain consistent with both AF2 and HIV-1 RT p66 at the active site. ESMFold and OmegaFold were again excluded from the comparison as DNA docking results showed the inability of adequate DNA docking because of the differently oriented RNase H domain. Together, these results show that the final model of HTLV-1 RT p62 by AF2 with DNA is reliable and can be used in subsequent docking studies using anti-HIV-1 antiviral compounds.

3.3 Antiviral compound docking

The final model was used for docking of antiviral compounds using AutoDock Vina^{109,110}. As a proof of principle, the diphosphate form (DP) of Tenofovir (PMPA), and triphosphate form (TP) of Zidovudine (ZDV), Stavudine (d4T) and Lamivudine (3TC) were tested for their ability to bind into the active site of the final HTLV-1 RT model (Figure 17A-B). All are examples of phosphorylated NRTIs as they bind into the polymerase active site and prematurely terminate chain elongation. A grid of size [x: 40 Å, y: 36 Å, z: 34 Å] in the polymerase active site was defined prior to docking, giving the compound the ability to bind into a volume of $48\,960\text{ Å}^3$ ($= 4.896\text{ μm}^3$). This gives it ample space to try multiple conformations, diminishing operator bias. For each compound, AutoDock Vina was able to dock it into the HTLV-1 RT active site with varying affinity (Figure 17C).

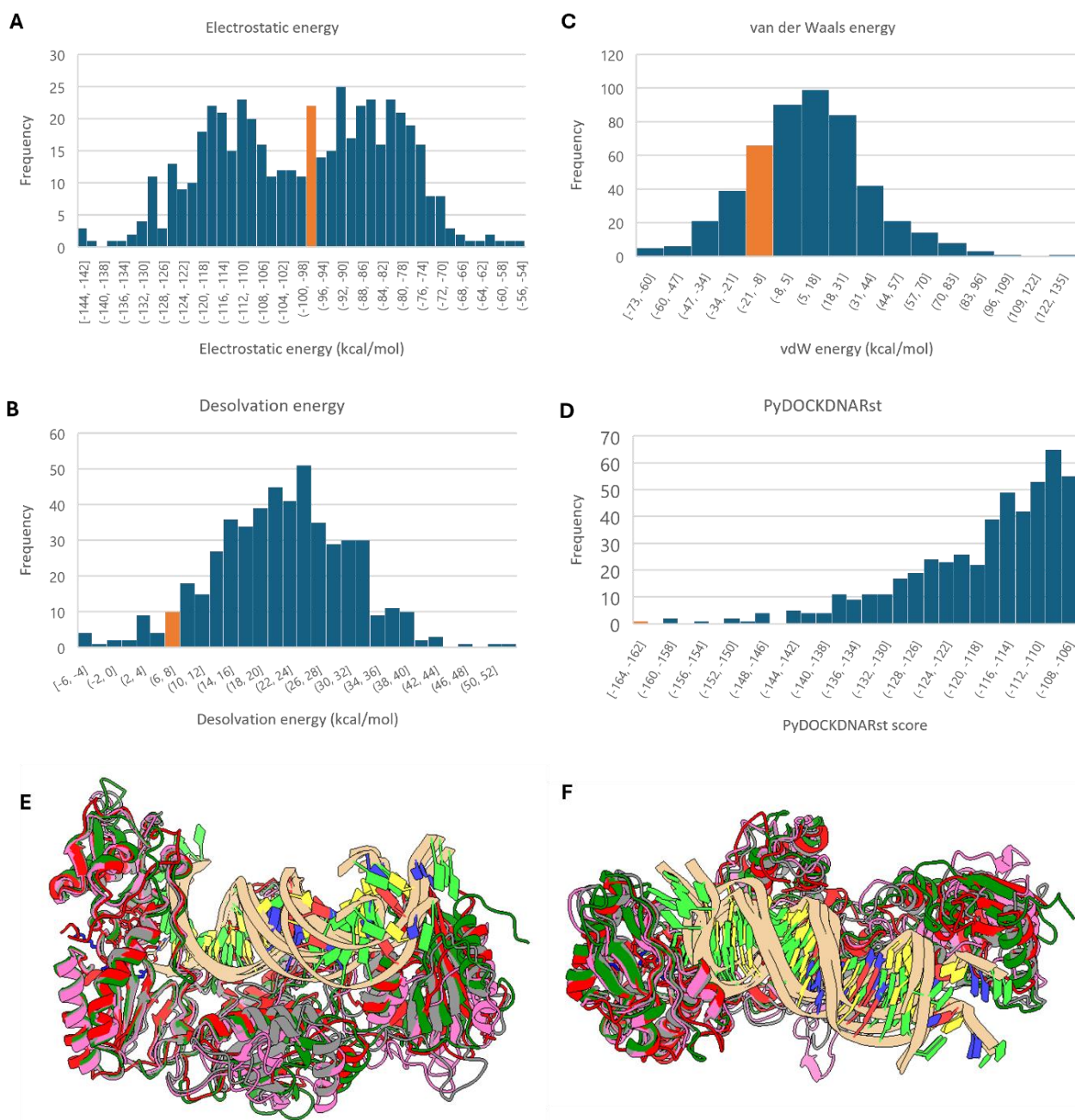


Figure 16. PyDOCKDNA rank one validates the use of the AlphaFold2 model for DNA docking.

For all parameters, histograms are shown representing the distribution of (A) electrostatic energy, (B) desolvation energy, (C) van der Waals energy and (D) PyDOCKDNAst scores amongst $n = 500$ conformations. The orange coloured bin represents rank one's position within the dataset. (E, F) Overlap structure of DNA bound HIV-1 RT (PDB: 6UIT) with AF2, Modeller, SWISS-MODEL and Phyre² PyDOCKDNA output structures.

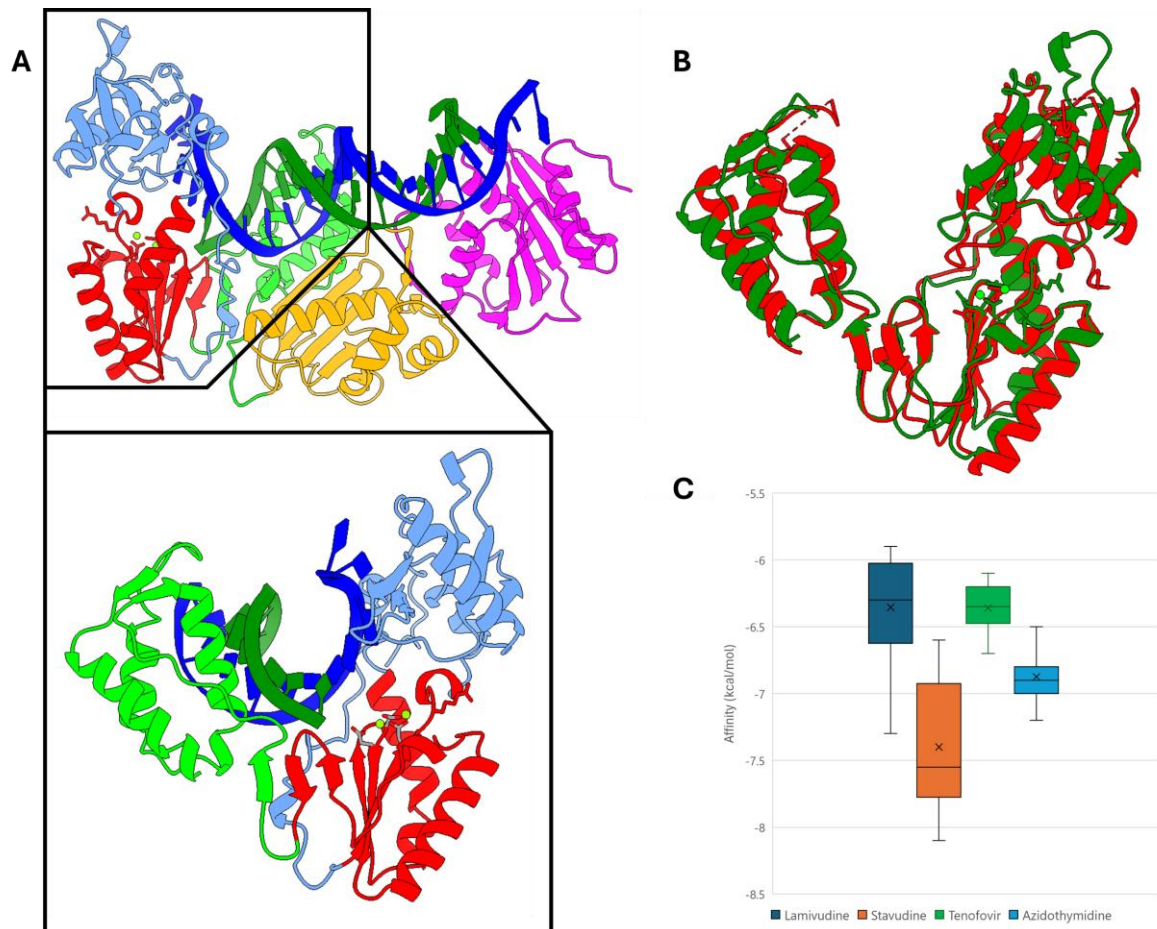


Figure 17. The predicted HTLV-1 RT binding pocket shows high structural conservation with HIV-1 RT and is able to bind NRTI-TP.

(A) The final model of HTLV-1 RT bound to a nucleic acid dimer (Template, Primer) and two Mg^{2+} ions. Black box highlights the dNTP binding pocket formed by the finger, palm and thumb domain showcasing D98, D173 and D174 coordinating two Mg^{2+} ions. (B) HTLV-1 RT predicted model superimposed with HIV-1 RT (PDB: 6UIT). Mg^{2+} ions belong the HTLV-1 RT model. Colour code: HIV-1 RT, HTLV-1 RT. (C) Boxplot presenting binding affinity for each compound for $n = 20$ conformations.

3.4 Design and validation of protein expression constructs

3.4.1 Plasmid design

HTLV-1 pET15b: As stated before, the gRNA of HTLV-1 produces mRNA for the gag-pro-pol polyprotein by a -2 ribosomal shift. In addition, the Pr protein is important in maturation of all HTLV-1 proteins, thus also in cleaving p62 to produce p49. To this end, the HTLV-1 pET15b plasmid encodes the terminal pro-pol part of the gag-pro-pol polyprotein. The expressed protein will hold an N-terminal 6xHis-tag to be used in Nickle-Histidine affinity chromatography (Figure 18A). **HTLV-1 pET42a (+) RT & HTLV-1 pET42a (+) pol:** In contrast to HTLV-1 pET15b, these plasmids hold the coding sequence for either RT or pol (Figure 18B-C). This allows for analysis of the RT or pol polyprotein without having to rely on Pr cleavage during maturation. The expressed protein is equipped with a N-terminal glutathione-S-transferase (GST)-tag that can be used in Glutathione-GST affinity chromatography in addition to a N-terminal 6xHis-tag. **HTLV-1 pETDuet-1 p62/p49 in-**

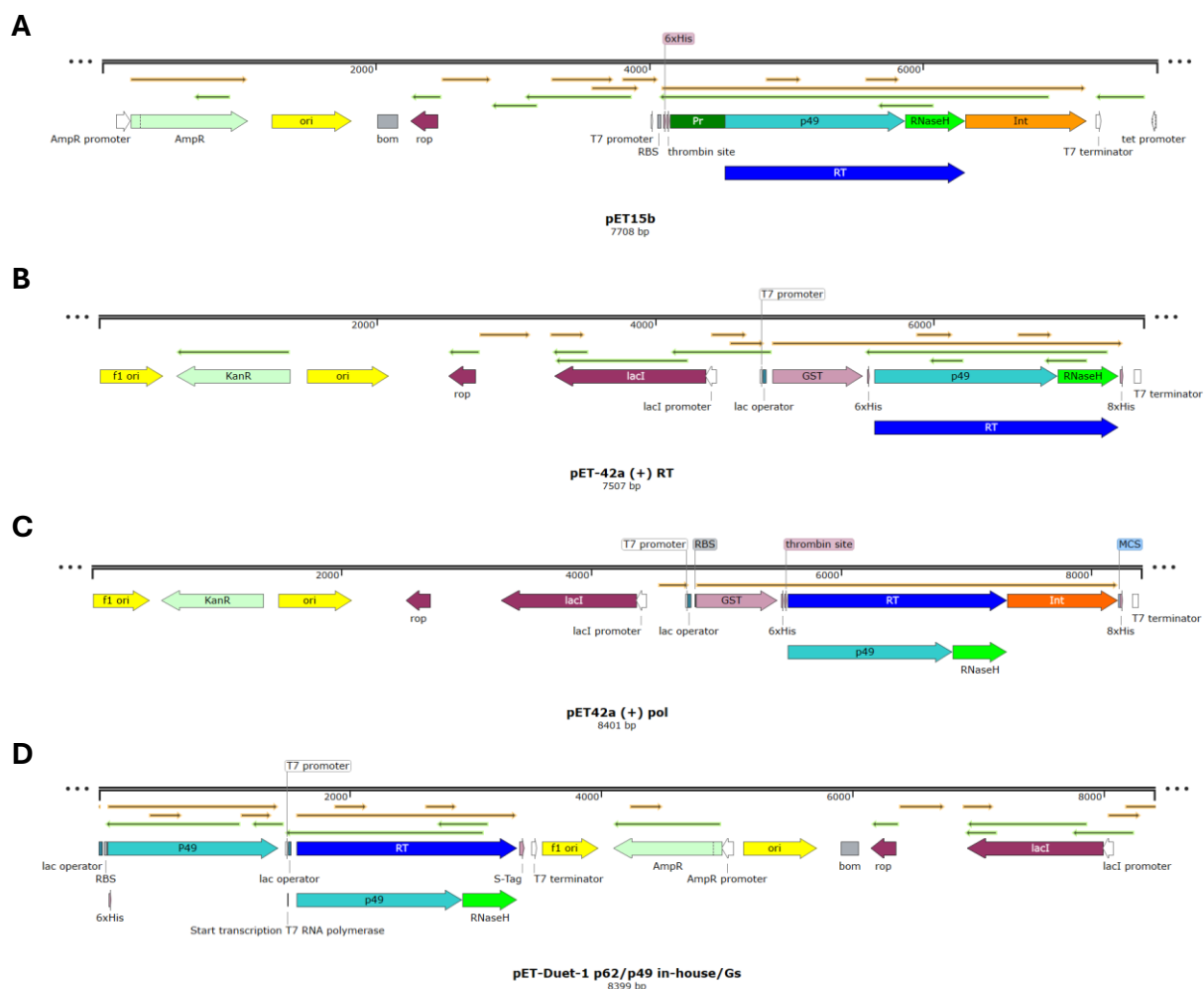


Figure 18. Design of the protein expression constructs.

(A) HTLV-1 pET15b, (B) HTLV-1 pET42a (+) RT, (C) HTLV-1 pET42a(+) pol, (D) HTLV-1 pET-Duet-1 p62/p49 in-house/Gs.

house/Gs: Due to the high grade of similarity between HTLV-1 RT and HIV-1 RT, and the successes of the pET-Duet-1 expression system in HIV-1 RT, a similar construct was developed for HTLV-1 RT. Upon expression, the p49 subunit will hold 6xHis-tag (Figure 18D). It is important to note that two HTLV-1 pET-Duet p62/p49 constructs were used. HTLV-1 pET-Duet-1 p62/p49 in-house refers to the pET-Duet-1 plasmid that was generated in the lab. HTLV-1 pET-Duet-1 p62/p49 Gs refers the pET-Duet-1 plasmid that was ordered directly from GenScript Europe. In general, the encoded beta-lactamase ampicillin resistance gene (AmpR) or aminoglycoside O-phosphotransferase kanamycin resistance gene (KanR) allow post-transformation antibiotic selection. Together, all of these constructs allow for the analysis of HTLV-1 RT in different states during maturation e.g. as a polyprotein in HTLV-1 pET15b and HTLV-1 pET42a (+) pol, but also as the final HTLV-1 RT heterodimer in HTLV-1 pET42a (+) RT and HTLV-1 pET-Duet-1 p62/p49 in-house/Gs.

3.4.2 Plasmid construction and quality control

Prior to induction, all plasmids were subjected to extensive validation by PCR, restriction digest or sequencing. HTLV-1 pET15b: Using *Bam*HI/*Nde*I, *Hind*III/*Xba*I and *Eco*RI restriction enzymes, the topology of the plasmid was checked (Figure 19A, lane 1-7). Here, however, incorrectly sized DNA fragments were observed. In addition, no DNA fragments were detected in the HIV-1 pET-Duet-1 p66/p51 positive control construct. In addition, a PCR was performed with sequence specific primers targeting the HTLV-1 RT encoding domain (Figure 18A, blue), which resulted in the presence of the correctly sized amplicons (Figure 19A, lane 8-16). The additional incorrect amplicons can be explained by the low annealing temperature used. Lastly, Sanger sequencing was performed on HTLV-1 pET15b for the both insert and upstream region. Results showed no deviations from coding sequence within the insert but also confirmed an intact upstream region. HTLV-1 pET42a (+) RT: Sequencing on the final product of HTLV-1 pET42a (+) RT construction for both insert and upstream region proved the presence of the insert, but also that it remained intact during assembly. Furthermore, the T7 promotor, GST-tag, RBS and 6xHis-tag showed no defects. HTLV-1 pET42a (+) pol: After transformation into One Shot BL21 (DE3) star cells, colony PCR was performed to check for the presence of the plasmid by using pol sequence specific primers. This showed the presence of an amplicon at size ~2.7 kb, as expected (Figure 19B). To confirm, Sanger sequencing was performed which proved that the insert showed no sign of sequential deviation, but did show the presence of an intact T7 promotor, RBS, GST- and 6xHis-tag. HTLV-1 pET-Duet-1 p62/p49 in-house: After transformation into DH5α cells, the plasmid was isolated and checked by restriction digest using the *Xba*I/*Not*I restriction enzyme pair to cut out p49 (1.4 kb), and *Nde*I/*Xho*I restriction enzyme pair to cut out p62 (1.7 kb) (Figure 19C). Both DNA fragments, in addition to full length HTLV-1 pET-Duet-1 p62/p49 in-house plasmid (8.3 kb), were present. The

full sequence was further confirmed by Sanger sequencing. HTLV-1 pET-Duet-1 p62/p49 Gs: Genscript delivered an extensive quality assessment report, noting both construction of the plasmid and quality control. In here, results from a restriction digest test showing the correct sizes after cleavage, namely 3349 bp and 5050 bp (Figure 19D). In addition, sequencing was performed proving an intact coding region for both inserts.

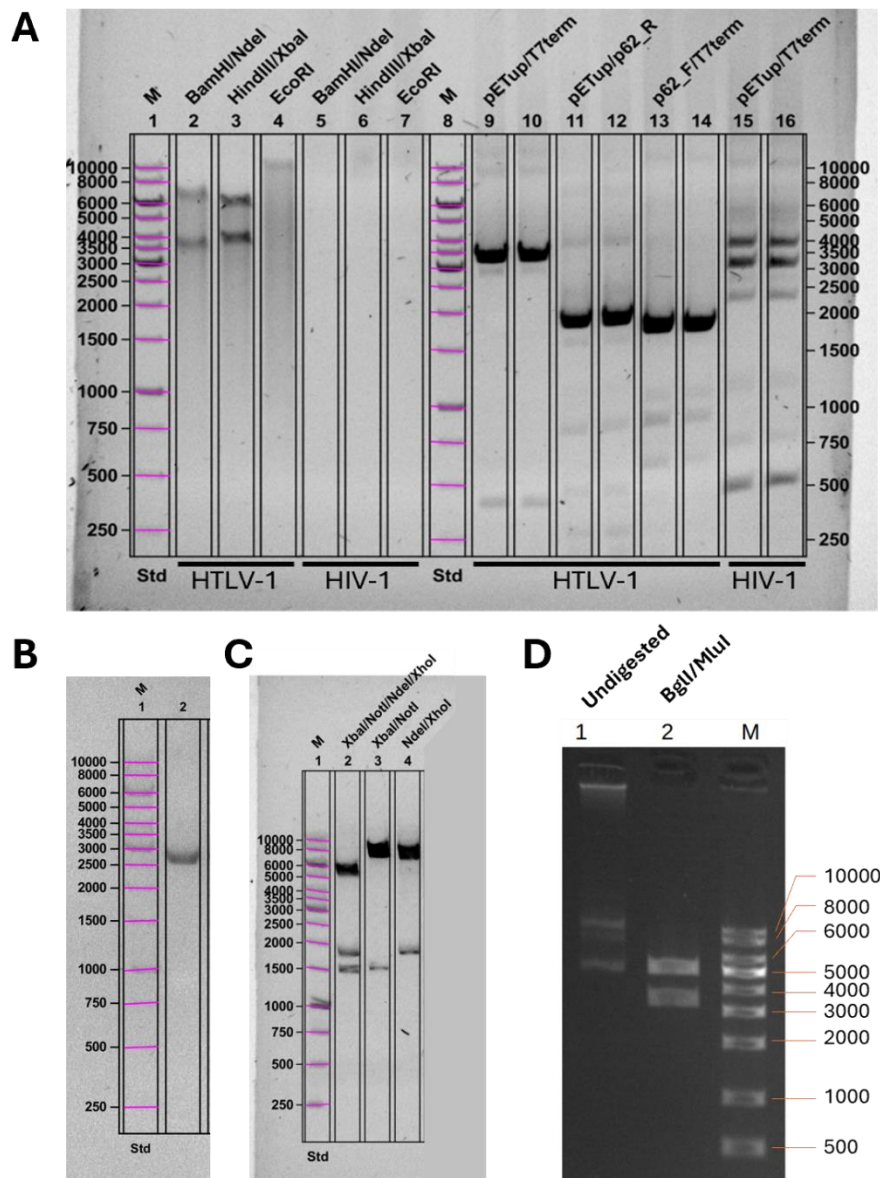


Figure 19. Quality control of the protein expression constructs.

(A) Restriction digest (lane 2-7) and PCR control (lane 9-16) on HTLV-1 pET15b (lane 2-4, 9-14) and HIV-1 RT pET-Duet p66/p51 (lane 5-7,15,16) HTLV-1 pET42a (+) pol PCR check. (C) Restriction digest on HTLV-1 pET-Duet-1 p62/p49 in-house. (D) Restriction digest on HTLV-1 pET-Duet-1 p62/p49 Gs. (A, B, C) M: Generuler 1 kb (1 µg, Thermo Scientific), (D) M: KB ladder.

3.5 Recombinant protein expression

3.5.1 IPTG protein induction

The plasmids were transformed into various T7 RNA polymerase (T7RNAP, 99 kDa) expressing bacterial strains and induction was tested at a wide range of conditions (Table 3). Each time, HIV-1 pET-Duet-1 p66/p51 (products: 64.3 kDa (p66) & 52.7 kDa (6xHis-p51)) was induced alongside in the same bacterial strains as a positive control. BL21-CodonPlus-(DE3)-RIL: HTLV-1 pET-Duet-1 p62/p49 in-house (products: 65.7 kDa (p62) & 50.7 kDa (6xHis-p49)) and HTLV-1 pET42a (+) RT (product: 95 kDa (GST-6xHis-p62-8xHis)) were transformed into BL21-CodonPlus-(DE3)-RIL cells as the use of this cell strain for expression of a pET-Duet-1 construct was proven successful earlier¹¹¹. Here however, no expression of HTLV-1 RT could be detected, even though HIV-1 pET-Duet-1 p66/p51 presented with very strong induction (Figure 20A, yellow arrows). For all plasmids, T7RNAP (Figure 20A, red arrows) seems present after induction, which validates the cell's ability to transcribe the plasmid as also validated by HIV-1 RT expression. ArcticExpress (DE3): HTLV-1

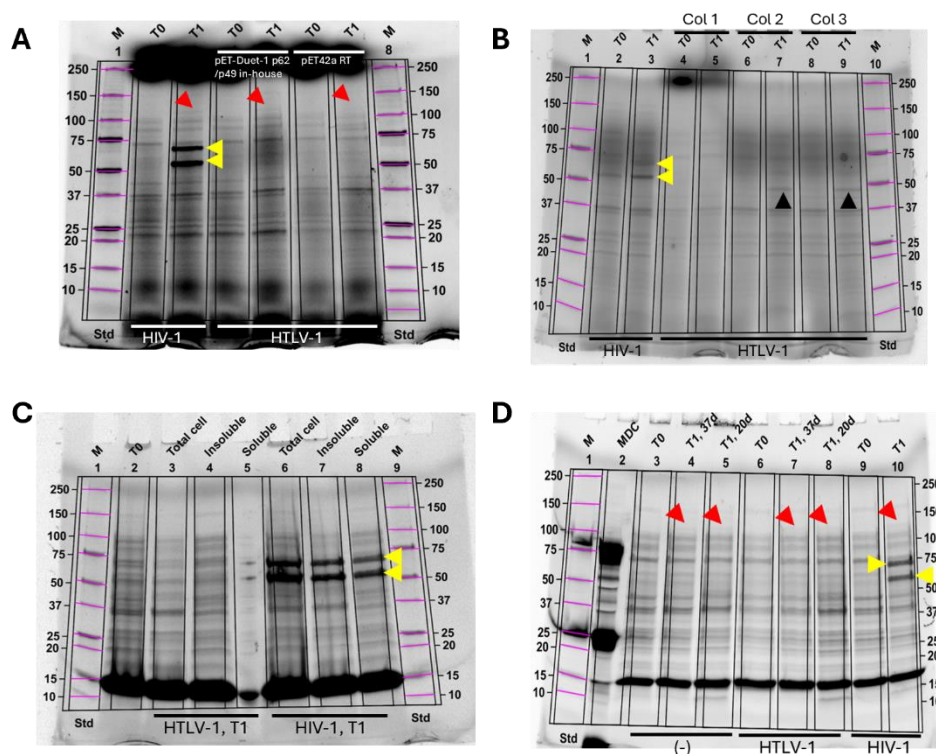


Figure 20. HTLV-1 RT remains undetectable on SDS-PAGE after induction in different cell strains at various induction conditions.

(A) Induction of pET-Duet-1 p62/p49 in-house [lane 4 - 5] and pET42a (+) RT [lane 6 - 7] in BL21-CodonPlus-(DE3)-RIL cells at 37°C for 2 hours with 1 mM IPTG. (B) Induction of HTLV-1 pET-Duet-1 p62/p49 in-house and HIV-1 pET-Duet-1 p66/p51 at 37°C for 3 hours with 1 mM IPTG in ArcticExpress (DE3) cell at 13°C for 18 hours at 1 mM IPTG. (C) Induction of HTLV-1 pET-Duet-1 p62/p49 in-house and HIV-1 pET-Duet-1 p66/p51 at 37°C for 3 hours with 1 mM IPTG in One Shot BL21 (DE3) star cells. (D) Induction of HTLV-1 pET-Duet-1 p62/p49 in-house and HIV-1 pET-Duet-1 p66/p51 at 37°C for 6 hours and 20°C overnight at 0.5 mM IPTG in One Shot BL21 (DE3) star cells. Negative control (-) is HTLV-1 pET-Duet-1 p62/p49 in-house plasmid with 6xHis out-of-frame with p49. T0: start point, T1: end point, M: Precision Plus protein unstained standards (Biorad), MDC: Precision plus protein dual color standards (Bio-Rad). Colour code: yellow arrows: HIV-1 RT, red arrows: T7RNAP, black arrow: possibly 6xHis-p49.

pET-Duet-1 p62/p49 in-house (product: 65.7 kDa (p62) & 50.7 kDa (6xHis-p49)) was transformed into ArcticExpress (DE3) cells since the presence of chaperone proteins (Cpn10 (10 kDa) & cpn60 (57 kDa)) might help protein folding. After induction, 6xHis-p51 and p66 are clearly expressed (Figure 20B, yellow arrows). Interestingly, a band at ~50 kDa is seen in colony 2 and 3, which could point to 6xHis-p49 (Figure 20B, black arrow). Yet, induction was tried in a bigger batch where HTLV-1 pET-Duet-1 p62/p49 in-house expression remained absent (Figure SXIC). In contrast, HIV-1 pET-Duet-1 p66/p51 showed clear expression of both protein products (Figure SXID, yellow arrows). One Shot BL21 (DE3) star: All constructs were transformed into One Shot BL21 (DE3) star cells as these are optimised for high-level expression of any non-toxic recombinant protein. HTLV-1 RT was induced in a broad range of conditions (3-16 h, 0.05-2mM IPTG, 20-37°C) however none of the construct was able to provide detectable protein expression (Figure 20C-D, Figure SXIE-F). In contrast, both HIV-1 RT subunits and T7RNAP are clearly expressed (Figure 20C-D). In conclusion, it seems that none of the plasmids are able to be used for protein expression in any of the *E. coli* strains, or at least not to the extent that it is detectable on SDS-PAGE.

3.5.2 Western blot shows low level expression of HTLV-1 RT

SDS-PAGE visualisation by ultraviolet light illumination allows for rapid processing, but provides lower sensitivity compared to Coomassie or silver stain, or western blot (WB). Due to this, low expression may be misinterpreted as absence of expression. Therefore, WB targeting the 6xHis tag was performed after induction of HTLV-1 pET15b and HTLV-1 pET-Duet-1p62/p49 Gs in One Shot BL21 (DE3) star cells. In contrary to what was seen on SDS-PAGE, it seems that both HTLV-1 pET15b (Figure 21A-B) and HTLV-1 pET-Duet-1 p62/p49 Gs (Figure 21C-D) express the encoded proteins, as both 6xHis-Pr-Pol (Figure 21B, red arrow) and 6xHis-p49 (Figure 21D, red arrow) were detected using mouse anti-6xHis monoclonal antibodies. Furthermore, specificity of the antibody is validated as no product was detected in the uninduced samples (Figure 21B-D) or negative control (Figure 21D). The presence of 6xHis-p51 in the uninduced sample can be explained by promotor leakage (Figure 21D, yellow arrow). In addition, a great amount of degradation product can be seen in both blots between 20 kDa and 37 kDa for HTLV-1 pET15b (Figure 21B, green arrow) and between 15 kDa and 20 kDa for HTLV-1 pET-Duet-1 p62/p49 Gs (Figure 21D, green arrow). Smaller amounts are visible between 50 kDa and 100 kDa for HTLV-1 pET15b (Figure 21B) and 25 kDa and 50 kDa for HTLV-1 pET-Duet-1 p62/p49 Gs (Figure 21D). In conclusion, both plasmids are capable of expressing recombinant protein at very low level. The yield from expression might be increased by inhibiting degradation of the protein product.

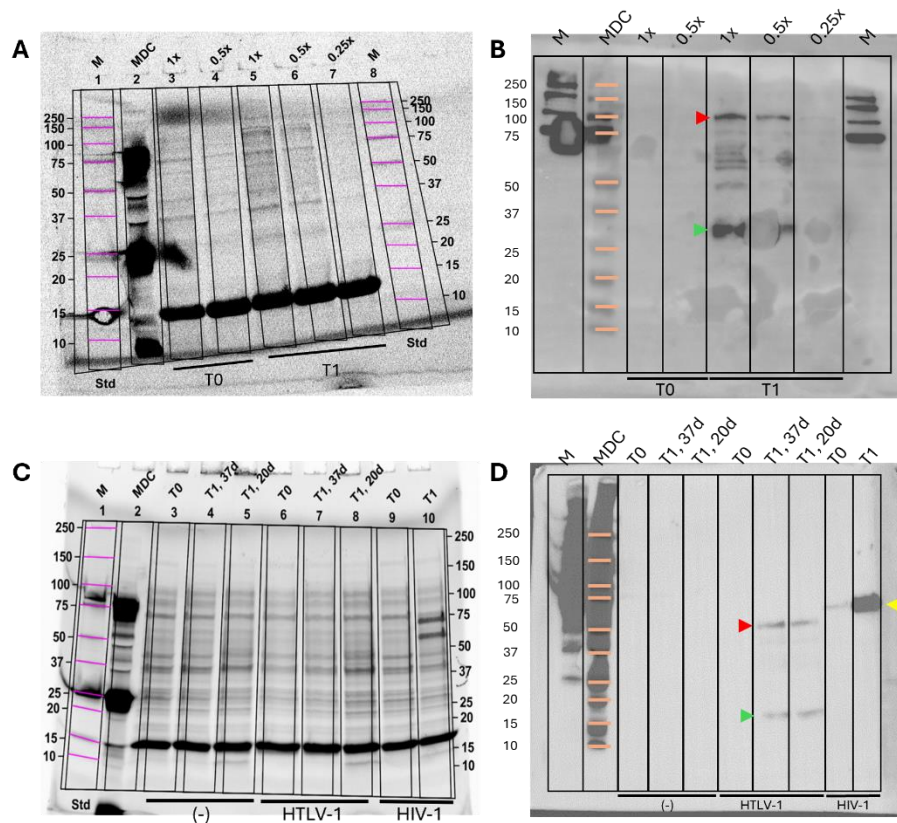


Figure 21. **Protein expression on all constructs.**

(A, B) Analysis of HTLV-1 pET15b expression on SDS-PAGE (A) and western blot (B) after induction at 37°C for 3 hours using 0.5 mM IPTG in One Shot BL21 (DE3) star cells. (A) SDS-PAGE with serial dilution of cellular lysate post-induction for HTLV-1 pET15b. (B) Western blot from SDS-PAGE in (A) using primary mouse anti-6xHis monoclonal antibody and secondary HRP-linked goat anti-mouse monoclonal antibody. (C, D) Analysis of HTLV-1 pET-Duet-1 p62/p49 Gs expression on SDS-PAGE (C) and western blot (D) after induction at 37°C for 6 hours and at 20°C overnight using 0.5 mM IPTG in One Shot BL21 (DE3) star cells. (C) SDS-PAGE with uninduced and induced samples for HTLV-1 pET-Duet-1 p62/p49 Gs. (D) Western blot from SDS-PAGE in (C) using primary mouse anti-6xHis monoclonal antibody and secondary HRP-linked goat anti-mouse monoclonal ab. T0: start point, T1: end point M: Precision protein plus unstained ladder (BioRad), MDC: Precision protein plus Dual Color ladder (BioRad). Red arrow: HTLV-1 6xHis-pr-pol (B, 115.3 kDa) or HTLV-1 6xHis-p49 (D, 50.7 kDa), yellow arrow: 6xHis-p51 (52.7 kDa), green arrow: 6xHis-positive degradation product.

3.5.3 The effect of IPTG concentration on HTLV-1 RT expression

Even though IPTG concentration shows no significant effect on the final recombinant protein yield, various concentrations for induction have been reported in literature. To test the effect of IPTG concentration on HTLV-1 RT production, One Shot BL21 (DE3) star cells transformed with HTLV-1 pET-Duet-1 p62/p49 Gs were induced at both 20°C overnight and 37°C for 6h using a range of IPTG concentrations. Again, HIV-1 pET-Duet-1 p66/p51 was used as positive control. All IPTG concentrations were able to induce T7RNAP expression in similar amount (Figure 22A & C, red arrows), suggesting the that IPTG concentration does not affect the protein expression. Yet, both HIV-1 RT and HTLV-1 RT are not detectable on SDS-PAGE. Due to this, WB was performed targeting the N-terminal 6xHis-tag on HIV-1 p51 and HTLV-1 p49 (Figure 22B & D). For both, higher RT expression is seen at 37°C compared to 20°C. (Figure 22B: black arrows (HTLV-1 RT), Figure 22D:

yellow arrows (HIV-1 RT)). The pre-induction sample (T0) of HIV-1 also shows the presence of 6xHis-p51, but can be attributed to promotor leakage. The optimal conditions for HTLV-1 RT expression were established (6h, 37°C, 0.5 mM IPTG) proving the IPTG concentration to be inconsequential.

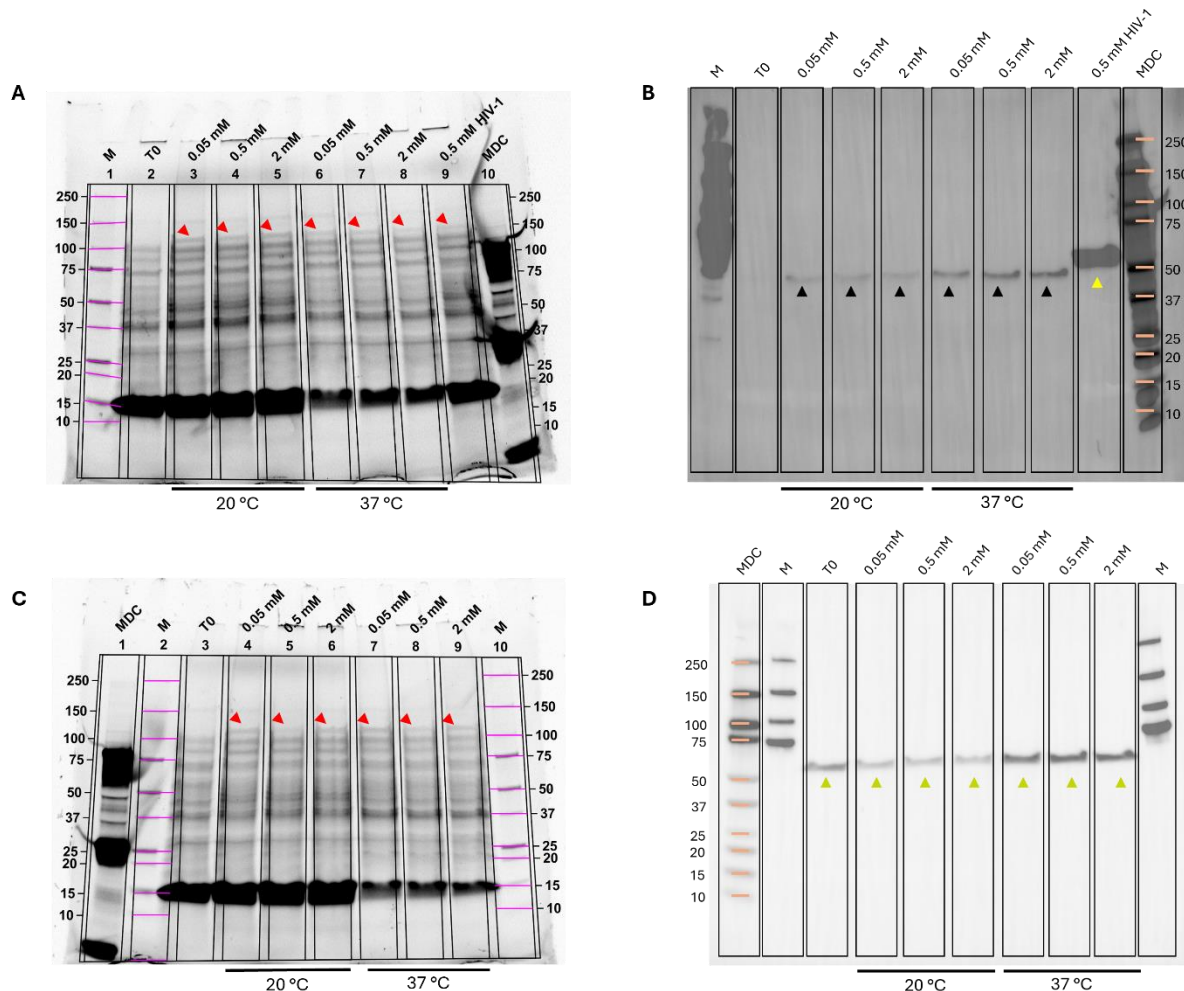


Figure 22. Culture conditions rather than IPTG concentration influence HTLV-1 RT expression.

(A, B) Analysis of HTLV-1 pET-Duet-1 p62/p49 Gs expression on SDS-PAGE (A) and western blot (B) after overnight induction at 20°C or 6 hours induction at 37°C using 0.05 mM, 0.5 mM or 2 mM IPTG in One Shot BL21 (DE3) star cells. (B) Western blot from SDS-PAGE in (A) using primary mouse anti-6xHis monoclonal antibody (1:3000) and secondary HRP-linked goat anti-mouse monoclonal antibody (1:5000). (C, D) Analysis of HIV-1 pET-Duet-1 p66/p51 expression on SDS-PAGE (C) and western blot (D) after overnight induction at 20°C or 6 hours induction at 37°C using 0.05 mM, 0.5 mM or 2 mM IPTG. (D) Western blot from SDS-PAGE in (C) using primary mouse anti-6xHis monoclonal antibody (1:3 000) and secondary HRP-linked goat anti-mouse monoclonal antibody (1:5 000). T0: start point, M: Precision protein plus unstained ladder (BioRad), MDC: Precision protein plus Dual Color ladder (BioRad). Red arrow: T7 RNA polymerase (99 kDa), yellow arrow: 6xHis-p51 (52.7 kDa), black arrow: 6xHis-p49 (50.7 kDa).

3.6 HTLV-1 RT purification

3.6.1 Purification of HTLV-1 RT via affinity chromatography

The presence of the N-terminal 6xHis-tag on the p49 subunit in the pET-Duet-1 p62/p49 Gs plasmid allows for isolation of HTLV-1 RT using nickle-chelated nitrilotriacetic acid (Ni-NTA) coupled to agarose beads. After induction, cells were lysed by sonication, centrifuged at high speed and the supernatant was incubated overnight with the Ni-NTA beads. HTLV-1 RT was eluted

with an imidazole gradient. Both p62 and 6xHis-p49 are visible on SDS-PAGE (Figure 23A, black arrows) in addition to both p66 and 6xHis-p51 (Figure 23A, yellow arrows). This was validated by performing WB targeting the N-terminal 6xHis-tag on HIV-1 p51 and HTLV-1 p49 (Figure 23B, HIV-1 6xHis-p51: yellow arrow, HTLV-1 6xHis-p49: black arrows). Yet, some common *E. coli* contaminants accompany HTLV-1 RT upon elution at high imidazole concentration, of which bifunctional polymyxin resistance protein ArnA (ArnA, 74.2 kDa, Figure 23A: red arrows) is the most notorious. These results indicate that upon expression of HTLV-1 RT 1) the 6xHis-tag on p49 is intact and functional, 2) p62 is able to coelute with 6xHis-p49 and that 3) the 6xHis-Ni-NTA affinity is strong enough to separate HTLV-1 RT from most of the *E. coli* contaminants, though some remain.

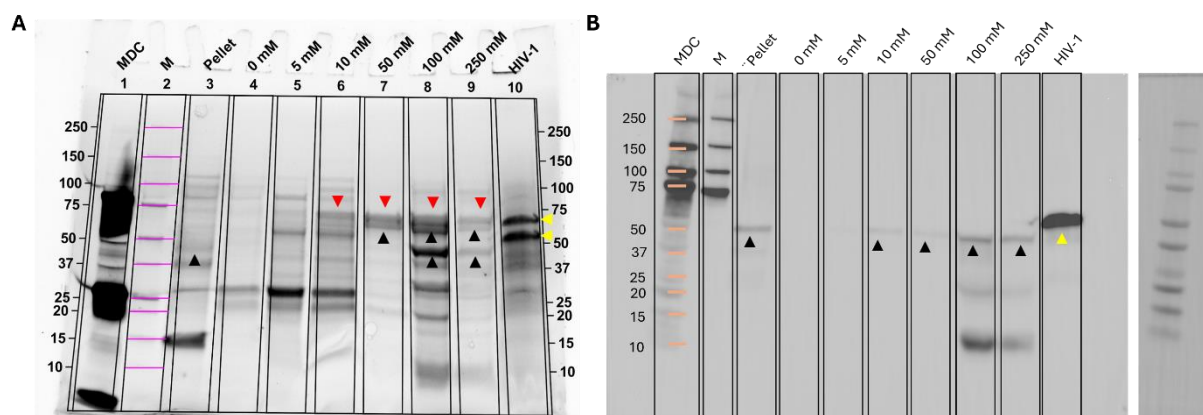


Figure 23. **Successful isolation of HTLV-1 RT via affinity chromatography.**

(A) SDS-PAGE analysis of different fractions after elution of HTLV-1 RT from Ni-NTA beads using a 0 mM to 250 mM imidazole gradient. (B) Western blot from SDS-PAGE in (A) using primary mouse anti-6xHis monoclonal antibody (1:3 000) and secondary HRP-linked goat anti-mouse monoclonal antibody (1:5 000). M: Precision protein plus unstained ladder (BioRad), MDC: Precision protein plus Dual Color ladder (BioRad). **Red arrow:** bifunctional polymyxin resistance protein ArnA (ArnA, 74.2 kDa), **yellow arrow:** 6xHis-p51 (52.7 kDa), black arrow: 6xHis-p49 (A - B, 50.7 kDa) and p62 (A, 65.7 kDa).

3.7 HTLV-1 RT mRNA expression in *E. coli*

3.7.1 RT-qPCR

Since low protein expression was observed on WB (Figure 21-23), it is important to check whether sufficient protein encoding mRNA is present in the sample. Primers were developed targeting HTLV-1 RT p62, HIV-1 RT p51 and T7RNAP. HcaT and CysG were used as reference genes to which all expression levels were normalised, a p62 PCR amplicon was used as positive control and a reaction master mix without template was considered as negative control. The final results can be seen in Figure 24 and the qPCR amplification plots are presented in Figure SXII-SXVI. T7RNAP seems to be highly expressed (fold change: 8.05 ± 5.33 , $n = 30$) in all samples. Interestingly, HIV-1 RT's p51 subunit shows only a minor (2.55 ± 0.03 , $W = 0$, $p = 0.666$, $n = 6$) increase in expression, even though T7RNAP (5.02 ± 0.05 , $W = 0$, $p = 0.666$, $n = 6$) is present in great abundance. HTLV-1 protein encoding mRNA shows varying expression levels depending on the construct used. The

greatest change in expression is seen in HTLV-1 pET-Duet-1 p62/p49 Gs where the synthesis of p62 mRNA increases considerably, (55.73 ± 24.45 , $W = 0$, $p = 0.666$) more than HTLV-1 pET15b (2.83 ± 0.31 , $W = 4$, $p = 0.333$), HTLV-1 pET42a (+) RT (5.60 ± 0.67 , $W = 4$, $p = 0.333$) and HTLV-1 pET-Duet-1 p62/p49 in-house (17.3 ± 1.76 , $W = 4$, $p = 0.333$). Here, T7RNAP also seems adequately upregulated (3.62 ± 1.59 , $W = 0$, $p = 0.666$). HTLV-1 pET15b seems to present expression of pr-pol mRNA (2.83 ± 0.31 , $W = 0$, $p = 0.666$) at equal levels to HIV-1 p51 (2.55 ± 0.03 , $W = 2$, $p = 1$), with non-significant differences in T7RNAP expression ($W = 0$, $p = 0.333$). Other than this, T7RNAP is very much expressed (8.78 ± 0.96 , $W = 0$, $p = 0.666$). The greatest increase in T7RNAP expression was seen in cell transformed with the pET-Duet-1 p62/p49 in-house plasmid which present an up to 19-fold (17.3 ± 1.76 , $W = 0$, $p = 0.666$) increase after induction. However, this is not accompanied by a great increase in p62 mRNA expression (1.75 ± 0.179 , $W = 0$, $p = 0.666$). Lastly, pET42a (+) RT presented with a largely equal fold increase for both p62 (5.60 ± 0.67 , $W = 0$, $p = 0.666$) and T7RNAP (5.58 ± 0.67 , $W = 0$, $p = 0.666$). In conclusion, HTLV-1 pET-Duet-1 p62/p49 Gs, HTLV-1 pET15b and HTLV-1 pET42a (+) RT seem able to produce mRNA of the HTLV-1 pr-pol protein of HTLV-1 RT at equal or higher levels than HIV-1 pETDuet-1 p66/p51 does. In contrast, HTLV-1 pET-Duet-1 p62/p49 in-house is clearly unable to express HTLV-1 RT upon IPTG induction.

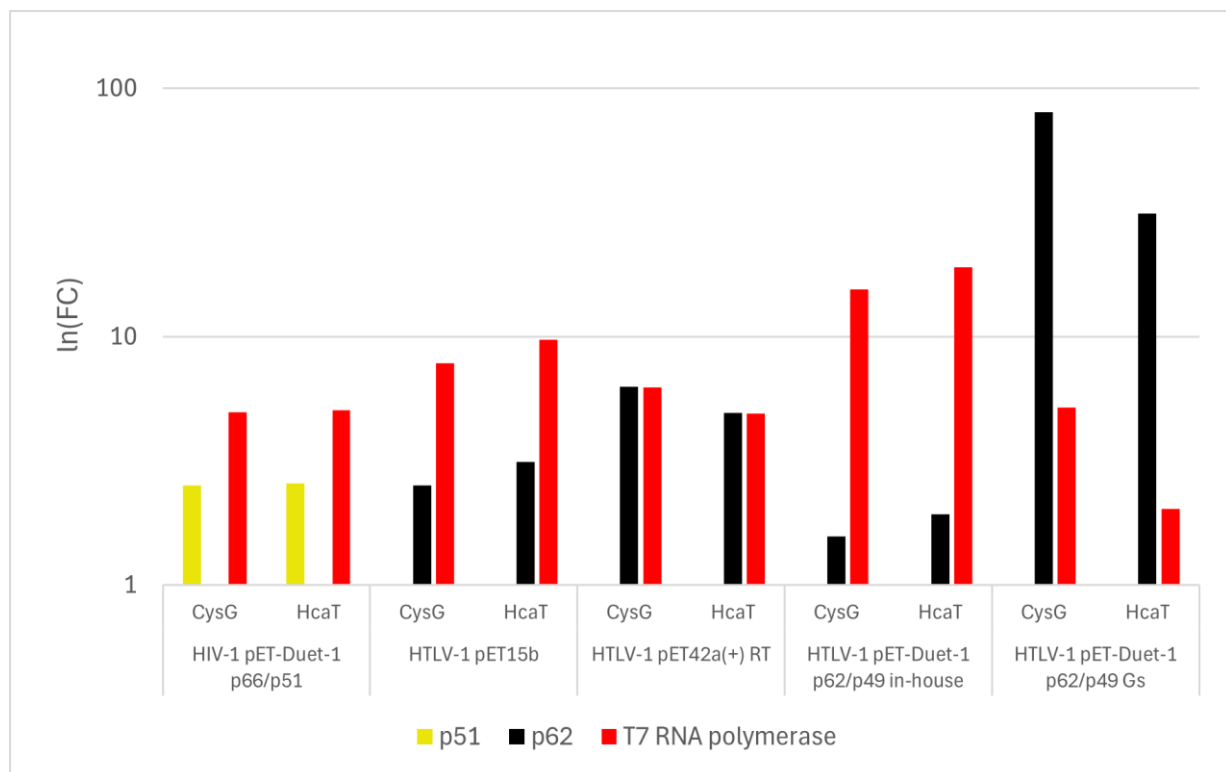


Figure 24. **pET-Duet-1 p62/p49 Gs shows very high level expression of p62.**

RNA was isolated from plasmid transformed One Shot BL21 (DE3) star cells before and after IPTG induction. Prior to qPCR, reverse transcription was performed to obtain cDNA from the bacterial cell's mRNA. Expression levels of HIV-1 p51, HTLV-1 p62 and T7 RNA polymerase were measured, using both CysG and HcaT as reference gene to which expression was normalised. Change in expression level is described as the natural logarithm (ln). Each sample had one biological replicate and each target was measured in three technical replicates.

3.7.2 HTLV-1 5' mRNA shows highly stable secondary structure

As RT-qPCR showed the presence of HTLV-1 p62 mRNA in equal or higher expression levels to HIV-1 p51 mRNA (Figure 24), we hypothesized that the mRNA might be inaccessible for the ribosomes to bind, reducing the translation. For this, the full-length mRNA of both HIV-1 RT p66 and HTLV-1 RT p62 was given as input to the *MFold RNA folding* webserver (Figure 25). Up to base 72, both sequences are identical and therefore should fold similarly. Interestingly, HTLV-1 p62 mRNA seems to fold into a highly stable conformation ($\Delta G = -717.12 \pm 7.52$ kcal/mol, $n = 42$), much more stable than HIV-1 p66 mRNA ($\Delta G = -494.05 \pm 7.76$ kcal/mol, $n = 50$) (Figure 25A). In each of the HTLV-1 p62 mRNA predictions, region 65-160 folds into a large 96 bp GC-rich hairpin loop, which is not seen in HIV-1 P66 mRNA (Figure 25B-C). This region was isolated and given as input to the *MFold RNA folding webserver* and resulted in a very similar structure with very high stability ($\Delta G = -34.02 \pm 0.77$ kcal/mol, $n = 5$). These data suggest even if sufficient HTLV-1 RT encoding mRNA is

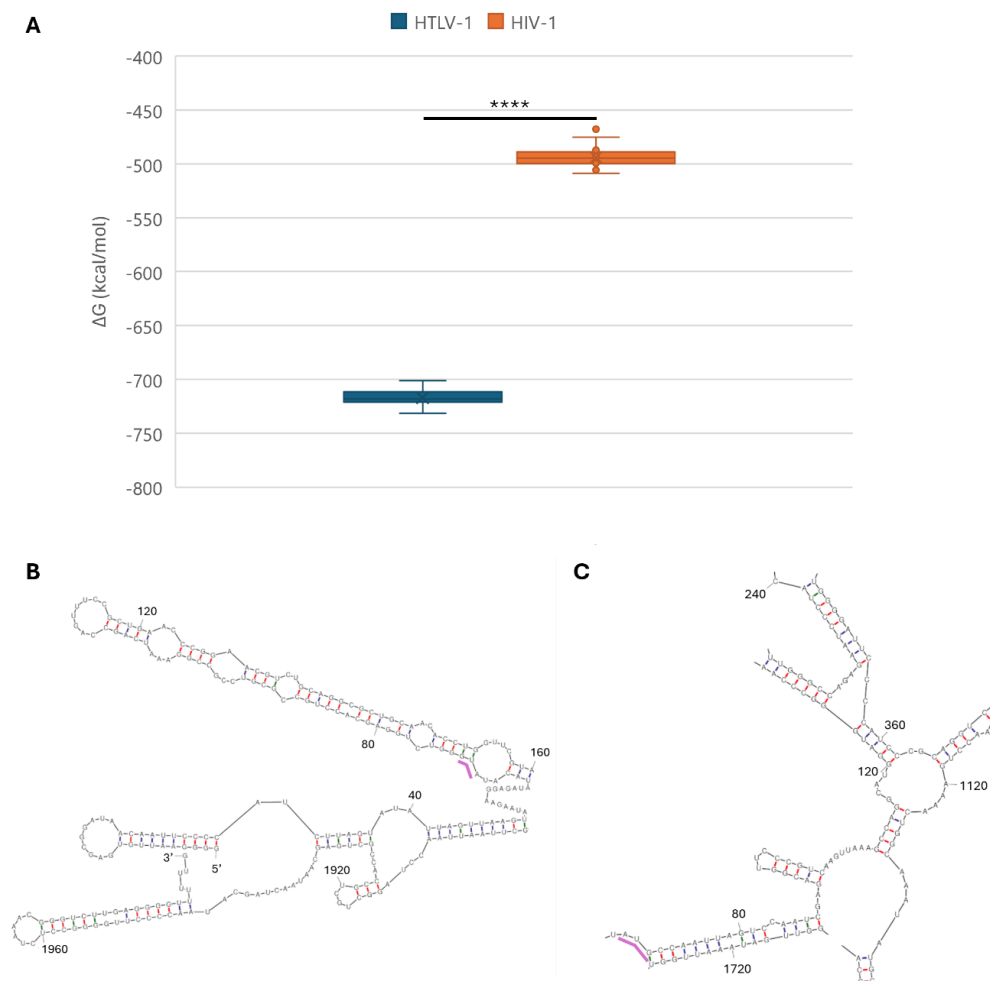


Figure 25. p62 mRNA folds into a highly stable hairpin structure, possibly interfering with protein expression.

Full mRNA sequence of HTLV-1 p62 and HIV-1 p66 was given as input into the *MFold RNA folding* server. (A) boxplots presenting the distribution of change in Gibbs free energy (ΔG) for HTLV-1 p62 mRNA ($n = 42$) and HIV-1 p66 mRNA ($n = 50$). Statistical significance was determined by parametric two-sample t-test. (B) 5' region of p62 mRNA. (C) 5' region of p66 mRNA. Start codon (AUG) is indicated in purple on both (B) and (C).

present after induction, the GC-rich hairpin loop containing the AUG start codon might interfere with the ribosomes ability to translate the mRNA into protein. Altogether, this might explain the low HTLV-1 RT expression seen previously.

4. Discussion

4.1 Great homology with HIV-1 RT and HERV-K RT

A multiple sequence alignment was conducted on 34 retroviral reverse transcriptase (RT) sequences, revealing that HTLV-1 RT exhibits high homology with the RT of each of the STLV strains. This finding is expected, given that HTLV-1 and STLV-1 originate from a common ancestor and collectively form the primate T-cell lymphotropic viruses group¹¹². Additionally, a distant relationship to HIV-1 was noted due to the different taxonomic classifications of these viruses. The link between HIV-1 and HTLV-1 has already been established as it is commonly referred to as HIV-1's "distant cousin". Both share similar spreading mechanisms and very similar configuration of the viral particle and genome¹¹³. However, great differences still remain as HTLV-1 shows much lower infectivity and slower disease onset. Also, HIV-1 infection has become manageable, not curable, whereas HTLV-1 infection still remains largely untreatable^{8,10,114,115}. Domain identification and gene ontology analysis were performed on the HTLV-1 p62 sequence. This analysis revealed the presence of a polymerase domain, RNase H domain and three catalytic residues (D98, D173, D174). These metal coordinating residues are highly conserved among retroviral RTs, as confirmed by the MSA. Great homology was observed with HIV-1 RT and, surprisingly, HERV-K RT which was further analysed by pairwise sequence alignment. This showed great conservation of catalytically important residues, particularly in the polymerase active site. Also, HERV-K RT has been shown to hold great similarity to HIV-1 RT⁹⁵. Furthermore, NRTIs and NNRTIs were proven effective in decreasing HERV-K RT activity, thus blocking HERV-K infectivity¹¹⁶. The similarities, conservation and homology between these three RTs might indicate a possibility for the use of NRTIs and NNRTIs in prophylactic anti-HTLV-1 therapy.

4.2 In silico HTLV-1 RT model

Due to the lack of a high-resolution model of HTLV-1 RT, an *in silico* model was constructed. For this, six protein structure prediction algorithms were employed, three homology-based and three ab initio, and thoroughly scored (VERIFY3D, PROCHECK) to assess quality, reliability and nativeness. A comparison with an experimentally solved high-resolution structure of HIV-1 RT was done¹⁰⁶. The five highly conserved subdomains (finger, palm, thumb, connection and RNase H)

could be distinguished and great similarity to HIV-1 RT was observed, indicating structural conservation. The Mg^{2+} coordinating residues in HIV-1 RT spatially overlapped with those predicted by ScanProSite (D98, D173, D174), and were also found by MIB2. DNA docking was successful for four out of six models. The main reason for improper DNA docking in ESMFold and OmegaFold models is the oddly configured RNase H domain. The final model containing both catalytically required Mg^{2+} ions and a nucleic acid dimer was constructed from the AF2 model as it was the overall highest scoring structure.

Homology-based structure prediction considers evolutionary conservation by performing MSA to find a good template to construct a model for the query sequence¹⁰². This could mask HTLV-1 RT-specific structures that might hold physiological importance. Previously, HERV-K RT was found to have a strong evolutionary link to HTLV-1 RT by domain identification (InterProScan). This finding was substantiated as each homology-based structure prediction algorithm as all used HERV-K RT as template holding ~27% sequence identity. Ab initio structure prediction generates a model by solely using the fundamental rules of protein folding¹⁰². Even though it is computationally very heavy, it decreases the intrinsic bias in the model as no template is used. One major disadvantage is the loss of evolutionary important structural information, which is retained in homology-based structure prediction. This disadvantage was omitted as both homology-based and ab initio structure prediction was performed. Yet, the best overall structure was predicted by AF2, an algorithm that combines techniques from ab initio and homology-based model prediction. The strength of AF2 has been proven repeatedly, as it considers evolutionary information, but is not biased by it. Yet, it still holds some limitations as it does not consider other molecules that interact with the protein (e.g. nucleic acids, ions, co-factors) or post-translational modifications¹⁰¹. These are physiologically very relevant as they influence a protein's final structure or hold a functional relevance. In addition, its inability to construct multimeric complexes limited us to only generate an HTLV-1 RT p62 model, instead of the full HTLV-1 RT heterodimer.

All models were scored on equal grounds by employing both VERIFY3D and PROCHECK¹⁰³⁻¹⁰⁵. Surprisingly, the experimentally determined HIV-1 RT model used as reference failed the VERIFY3D test, whereas both the model by AF2 and ESMFold passed. A possible reason for this is that the HIV-1 RT model still holds some intrinsic strain and incorrectly configured residues due to inadequate modelling. Further optimisation of the model by extra energy minimisation (GROMACS, AMBER) or manually modelling might help solve this problem. Many stereochemical parameters were checked by PROCHECK¹⁰³. This showed that all models, HIV-1 RT included, held no problems in backbone residue configuration, but that the model by AF2 was best overall.

4.3 PoC: Antiviral compound docking

The final model of HTLV-1 RT, with Mg^{2+} and DNA, was used in proof-of-concept antiviral compound docking experiments with PMPA-DP, ZDV-TP, d4T-Tp and 3TC-TP and showed that all were able to bind into the polymerase active site at high affinity. To reduce bias, a large volume surrounding the polymerase active site was defined.

A similar study was conducted earlier, where it was shown that NRTI-TPs were effective in binding to HTLV-1 RT¹¹⁷. This, however, was performed on HTLV-1 RT predicted structures without the presence of Mg^{2+} or a nucleic acid dimer. Therefore, our model is much more physiologically relevant as NRTI-TPs bind into the polymerase active site upon reverse transcription, where a nucleic acid dimer is already present. In addition to the previously found high homology of HTLV-1 RT to HERV-K RT and HIV-1 RT, these results show that NRTI-TPs might be effective in blocking HTLV-1 RT upon infection.

The greatest limitation of this study is the lack of in vitro, in cellulo and in vivo validation, as this is rather a proof-of-concept. To prove efficacy in vitro, reverse transcription assays (e.g. primer extension assay) can be performed once protein of sufficient quality is obtained. It was also shown earlier that the active site of the predicted HTLV-1 RT model was slightly larger than HIV-1 RT. This indicates that the NRTI-TP would not bind as tightly as seen in HIV-1 RT, which might influence incorporation efficiency.

4.4 Plasmid design and protein expression

Five plasmids were used, each able to express HTLV-1 RT in a different manner (e.g. heterodimer, pol polyprotein, pr-pol polyprotein). For all constructs, restriction digest, PCR and Sanger sequencing proved the presence of the inserts and the sequence integrity. Only HTLV-1 pET15b failed the restriction digest validation test. Possible reasons for this failed test lay in the experimental procedure where the restriction enzymes might not have been in the most optimal conditions (e.g. inability to cut, intrinsically dysfunctional, absence of necessary cofactors). Luckily, the presence of the inserts and validity of the sequence was validated by PCR and Sanger sequencing.

After, each plasmid was transformed into multiple bacterial strains. Initially, no protein expression was seen in any of the tested conditions for any of the constructs. This was in stark contrast with the high expression in HIV-1 pET-Duet-1 p66/p51, and the presence of T7RNAP on all SDS-PAGE gels. Later, it was found that protein expression for both HTLV-1 pET15b and HTLV-1 pET-Duet-1 p62/p49 Gs could only be detected on western blot after expression in One Shot BL21 (DE3) star

cells. It is very likely that this also extends to the other constructs (e.g. HTLV-1 pET42a (+) RT, HTLV-1 pET42 (+) pol, HTLV-1 pET-Duet-1 p62/p49 in-house) and other bacterial strains (ArcticExpress (DE3) and BL21-CodonPlus-(DE3)-RIL), this however was not tested to date.

On the initial and the Ni-His purification western blot, 6xHis positive degradation products were observed for both HTLV-1 pET-Duet-1 p62/p49 Gs and HTLV-1 pET15b, which held a size of ~16 kDa and ~30 kDa respectively. The Pr domain that is encoded in the HTLV-1 pET15b constructs is ~14 kDa in size, exactly the difference in size between the two degradation products. This indicates that the degradation product seen in both are the same but the one in HTLV-1 pET15b holds the additional Pr domain. To prevent the visible degradation of the expressed protein the responsible bacterial protease could be identified (ExPASy's *PeptideCutter*) and specific protease inhibitor could be added to the *E. coli* culture. Examples of commonly used protease inhibitors are leupeptin, pepstatin A, E64 and phenylmethylsulfonyl fluoride each targeting various proteases (e.g. cysteine-, serine-, threonine- or aspartic acid-based). However, the possible effect of these inhibitors on *E. coli* physiology cannot be predicted.

Lastly, the effect of IPTG concentration on recombinant protein production was explored^{118,119}. Here, it was found that IPTG concentration had no observable impact on the level of HTLV-1 RT protein production. As this test was performed at both 20°C and 37°C, it was observed that induction temperature rather than IPTG concentration had an effect. For HTLV-1 RT production, the optimal induction conditions were found to be 37°C for six hours at 0.5 mM IPTG.

4.5 Successful protein purification

The induced HTLV-1 RT was purified using Ni-His affinity chromatography eluting the protein with 0-250 mM imidazole gradient. WB proved the presence of 6xHis-p49 at higher imidazole concentrations. This indicates that the 6xHis-tag on the p49 subunit is accessible and binds the Ni-NTA beads with high affinity. However, some contaminants remain present after elution, ArnA being the most abundant. To improve the purity of the sample, further purification will need to be done (e.g. size-exclusion or ion exchange chromatography).

4.6 GC-rich hairpin loop likely at the basis of low protein expression.

In silico modelling of the HTLV-1 RT mRNA sequence showed the presence of a highly stable GC-rich hairpin loop that contained the start codon for protein translation. By RT-qPCR, the mRNA expression ability of the constructs was validated, showing the greatest expression for the HTLV-1 pET-Duet-1 p62/p49 Gs. Yet, protein expression could only be detected on WB, indicating only a

small amount is present. This suggests that the One Shot BL21 (DE3) star cells are able to transcribe the plasmid, but unable to efficiently translated the mRNA. The GC-rich hairpin loop could significantly impact translation by obstruction of ribosomal binding to the start codon.

A possible way that HTLV-1 might get rid of this structure in cellulo is through chaperone proteins. Hypothetically, a chaperone protein could bind and destabilize the complex leading to the exposure of the start codon. Previously, it was found that HTLV-1 Ma exhibits strong chaperone protein activity in vitro to destabilise the PBS secondary structure, a similar mechanism that could happen here^{35,120-122}. This, however, would need further testing to confirm. This GC-rich hairpin loop is absent in HIV-1 RT p66 mRNA, and does not interfere with recombinant protein production as seen on multiple occasions. A possible approach in omitting this secondary structure would be to substitute this GC-rich hairpin loop encoding region by the 5' end sequence of the HIV-1 RT p66 mRNA. Due to the great homology present between HIV-1 RT and HTLV-1 RT, substitution of this region should not have significant impact on protein function. In addition, the loss of the GC-rich hairpin might cause increased protein expression.

5. Conclusion

There is a growing need for new antiviral drugs to target HTLV-1 for which HTLV-1 RT provides a very interesting target. By solving its structure, new ways of screening possible drug candidates along with SBDD can be explored. In this thesis, the link between HIV-1 RT and HTLV-1 RT was validated. Furthermore, an in silico model of HTLV-1 RT in complex with DNA was constructed and tested for antiviral compound docking, showing possible inhibitory effect of existing NRTI-TPs. Lastly, various plasmid constructs were designed to express HTLV-1 RT, and HTLV-1 RT was obtained at low amounts. The desired protein was successfully isolated using Ni-His affinity chromatography and further experiments to obtain high-quality sample for structural studies were designed. Low level expression might be explained by the presence of a GC-rich hairpin loop at the 5' end of the p62 mRNA. Future steps lay in optimising protein expression by adapting the constructs to destabilize the highly stable hairpin loop. Providing sufficient quality of protein will lead to preliminary structural studies using single particle Cryo-EM to obtain a true model of HTLV-1 RT. This model can then be utilised in high-throughput drug screening, structure-based drug design, basic mechanistic studies on reverse transcription and biochemical studies providing fundamental understanding of HTLV-1 RT.

Bibliography

1. Martin, J. L., Maldonado, J. O., Mueller, J. D., Zhang, W. & Mansky, L. M. Molecular studies of HTLV-1 replication: An update. *Viruses* **8**, 31 (2016).
2. Eusebio-Ponce, E., Anguita, E., Paulino-Ramirez, R. & Javier Candel, F. HTLV-1 infection: An emerging risk. Pathogenesis, epidemiology, diagnosis and associated diseases. *Rev Esp Quimioter* **32**, 485–496 (2019).
3. Manzari, V. & Gallo, R. C. Human T Cell Leukemia-Lymphoma Virus: Characterization, Biology and Significance. *Surv. immunol. Res* 122–125 (1982).
4. Iwanaga, M. Epidemiology of HTLV-1 Infection and ATL in Japan: An Update. *Front Microbiol* **11**, (2020).
5. Rosadas, C. *et al.* Adult T-cell leukaemia/lymphoma in Brazil: A rare disease or rarely diagnosed? *Br J Haematol* **188**, 46–49 (2020).
6. Kato, T., Imaizumi, Y. & Miyazaki, Y. Nationwide Hospital-Based Survey of Adult T-Cell Leukemia/Lymphoma in Japan. *Viruses* **14**, 791 (2022).
7. El Hajj, H. *et al.* Novel Treatments of Adult T Cell Leukemia Lymphoma. *Front Microbiol* **11**, 1062 (2020).
8. Nakamura, T. HAM/TSP Pathogenesis: The Transmigration Activity of HTLV-1-Infected T Cells into Tissues. *Pathogens* **12**, 492 (2023).
9. Osame, M. *et al.* HTLV-I associated myelopathy, a new clinical entity. *Lancet* **1**, 1031–1032 (1986).
10. Schwalb, A. *et al.* Early-onset HTLV-1-associated myelopathy/tropical spastic paraparesis. *Pathogens* **9**, 1–11 (2020).
11. Araújo, A., Fontenelle, L., Pádua, P., Filho, H. S. M. & de Araújo, A. Juvenile Human T Lymphotropic Virus Type 1-Associated Myelopathy. *Clinical Infectious Diseases* **35**, 201–204 (2002).
12. Ahmadi Ghezeldasht, S. *et al.* HTLV-1 oncovirus-host interactions: From entry to the manifestation of associated diseases. *Rev Med Virol* **31**, 2235 (2021).
13. Sánchez-Montalvá, A., Salvador, F., Caballero, E. & Molina, I. Cyclosporine for the treatment of HTLV-1-induced HAM/TSP: An experience from a case report. *Medicine* **94**, e382 (2015).

14. Sato, T. *et al.* Multicenter, randomized, double-blind, placebo-controlled phase 3 study of mogamulizumab with open-label extension study in a minimum number of patients with human T-cell leukemia virus type-1-associated myelopathy. *J Neurol* (2024) doi:10.1007/s00415-024-12239-x.
15. Sun, M., Grigsby, I. F., Gorelick, R. J., Mansky, L. M. & Musier-Forsyth, K. Retrovirus-Specific Differences in Matrix and Nucleocapsid Protein-Nucleic Acid Interactions: Implications for Genomic RNA Packaging. *J Virol* **88**, 1271–1280 (2014).
16. Maldonado, J. O., Cao, S., Zhang, W. & Mansky, L. M. Distinct morphology of human T-cell leukemia virus type 1-like particles. *Viruses* **8**, (2016).
17. Verdonck, K. *et al.* Human T-lymphotropic virus 1: recent knowledge about an ancient infection. *Lancet Infect Dis* **7**, 266–281 (2007).
18. Mesnard, J. M., Barbeau, B., Césaire, R. & Pélouponèse, J. M. Roles of HTLV-1 basic Zip factor (HBZ) in viral chronicity and leukemic transformation. potential new therapeutic approaches to prevent and treat HTLV-1-related diseases. *Viruses* **7**, 6490–6505 (2015).
19. Bai, X. T. & Nicot, C. Overview on HTLV-1 p12, p8, p30, p13: Accomplices in persistent infection and viral pathogenesis. *Front Microbiol* **3**, (2012).
20. Matsuoka, M. & Jeang, K.-Teh. Human T-cell leukemia virus type 1 (HTLV-1) and leukemic transformation: Viral infectivity, Tax, HBZ and therapy. *Oncogene* **30**, 1379–1389 (2011).
21. Thulson, E. *et al.* An RNA pseudoknot stimulates HTLV-1 pro-pol programmed –1 ribosomal frameshifting. *RNA* **26**, 512–528 (2020).
22. Jones, K. S. *et al.* Molecular aspects of HTLV-1 entry: Functional domains of the HTLV-1 surface subunit (SU) and their relationships to the entry receptors. *Viruses* **3**, 794–810 (2011).
23. Jones, K. S., Petrow-Sadowski, C., Bertolette, D. C., Huang, Y. & Ruscetti, F. W. Heparan Sulfate Proteoglycans Mediate Attachment and Entry of Human T-Cell Leukemia Virus Type 1 Virions into CD4 + T Cells. *J Virol* **79**, 12692–12702 (2005).
24. Lesbats, P., Engelmann, A. N. & Cherepanov, P. Retroviral DNA Integration. *Chem Rev* **116**, 12730–12757 (2016).
25. Passos, D. O., Li, M., Craigie, R. & Lyumkis, D. Retroviral integrase: Structure, mechanism, and inhibition. in *Enzymes* vol. 50 249–300 (Academic Press, 2021).

26. Meekings, K. N., Leipzig, J., Bushman, F. D., Taylor, G. P. & Bangham, C. R. M. HTLV-1 integration into transcriptionally active genomic regions is associated with proviral expression and with HAM/TSP. *PLoS Pathog* **4**, e1000027 (2008).
27. Bangham, C. R. M., Cook, L. B. & Melamed, A. HTLV-1 clonality in adult T-cell leukaemia and non-malignant HTLV-1 infection. *Semin Cancer Biol* **26**, 89–98 (2014).
28. Trainor, B. M. & Shcherbik, N. Short and Sweet: Viral 5' -UTR as a Canonical and Non-Canonical Translation Initiation Switch. *J Cell Immunol* **3**, 296–304 (2021).
29. Holcik, M. & Sonenberg, N. Translational control in stress and apoptosis. *Nat Rev Mol Cell Biol* **6**, 318–327 (2005).
30. Menéndez-Arias, L., Sebastián-Martín, A. & Álvarez, M. Viral reverse transcriptases. *Virus Res* **234**, 153–176 (2017).
31. Marquet, R., Lsel, C., Ehresmann, C. & Ehresmann, B. tRNAs as primer of reverse transcriptases. *Biochimie* **77**, 113–124 (1995).
32. Figiel, M. *et al.* Mechanism of polypurine tract primer generation by HIV-1 reverse transcriptase. *Journal of Biological Chemistry* **293**, 191–202 (2018).
33. Vallejos, M. *et al.* Activity of the human immunodeficiency virus type 1 cell cycle-dependent internal ribosomal entry site is modulated by IRES trans-acting factors. *Nucleic Acids Res* **39**, 6186–6200 (2011).
34. Liu, S. *et al.* Rous sarcoma virus genomic RNA dimerization capability in vitro is not a prerequisite for viral infectivity. *Viruses* **12**, 568 (2020).
35. Wu, W. *et al.* Human T-cell leukemia virus type 1 Gag domains have distinct RNA-binding specificities with implications for RNA packaging and dimerization. *Journal of Biological Chemistry* **293**, 16261–16276 (2018).
36. Shippen-Lentz, D. & Blackburn, E. H. Functional Evidence for an RNA Template in Telomerase. *Science (1979)* **247**, 546–552 (1990).
37. Rubio Gomez, M. A. & Ibba, M. Aminoacyl-tRNA synthetases. *RNA* **26**, 910–936 (2020).
38. Ruggero, K. *et al.* Small Noncoding RNAs in Cells Transformed by Human T-Cell Leukemia Virus Type 1: a Role for a tRNA Fragment as a Primer for Reverse Transcriptase. *J Virol* **88**, 3612–3622 (2014).
39. Schorn, A. J., Gutbrod, M. J., LeBlanc, C. & Martienssen, R. LTR-Retrotransposon Control by tRNA-Derived Small RNAs. *Cell* **170**, 61–71 (2017).

40. Larsen, K. P. *et al.* Architecture of an HIV-1 reverse transcriptase initiation complex. *Nature* **557**, 118–122 (2018).
41. Das, K., Martinez, S. E., DeStefano, J. J. & Arnold, E. Structure of HIV-1 RT/dsRNA initiation complex prior to nucleotide incorporation. *Proceedings of the National Academy of Sciences* **116**, 7308–7313 (2019).
42. Lipferta, J. *et al.* Double-stranded RNA under force and torque: Similarities to and striking differences from double-stranded DNA. *Proc Natl Acad Sci U S A* **111**, 15408–15413 (2014).
43. Harteis, S. & Schneider, S. Making the bend: DNA tertiary structure and protein-DNA interactions. *Int J Mol Sci* **15**, 12335–12363 (2014).
44. Isel, C., Ehresmann, C. & Marquet, R. Initiation of HIV reverse transcription. *Viruses* **2**, 213–243 (2010).
45. Lanchy, J. M., Ehresmann, C., Le Grice, S. F. J., Ehresmann, B. & Marquet, R. Binding and kinetic properties of HIV-1 reverse transcriptase markedly differ during initiation and elongation of reverse transcription. *EMBO J* **15**, 7178–7187 (1996).
46. Isel, C. *et al.* Specific initiation and switch to elongation of human immunodeficiency virus type 1 reverse transcription require the, post-transcriptional modifications of primer tRNA³LyS. *EMBO J* **15**, 917–924 (1996).
47. Liu, S., Harada, B. T., Miller, J. T., Le Grice, S. F. J. & Zhuang, X. Initiation complex dynamics direct the transitions between distinct phases of early HIV reverse transcription. *Nat Struct Mol Biol* **17**, 1453–1460 (2010).
48. Larsen, K. P. *et al.* Distinct Conformational States Underlie Pausing during Initiation of HIV-1 Reverse Transcription. *J Mol Biol* **432**, 4499–4522 (2020).
49. London, R. E. Structural maturation of HIV-1 reverse transcriptase—a metamorphic solution to genomic instability. *Viruses* **8**, 260 (2016).
50. Chung, S. *et al.* Examining the role of the HIV-1 reverse transcriptase p51 subunit in positioning and hydrolysis of RNA/DNA hybrids. *Journal of Biological Chemistry* **288**, 16177–16184 (2013).
51. Dash, C., Scarth, B. J., Badorrek, C., Götte, M. & Le Grice, S. F. J. Examining the ribonuclease H primer grip of HIV-1 reverse transcriptase by charge neutralization of RNA/DNA hybrids. *Nucleic Acids Res* **36**, 6363–6371 (2008).
52. Hermann, T., Meier, T., Gotte, M. & Heumann, H. The ‘helix clamp’ in HIV-1 reverse transcriptase: a new nucleic acid binding motif common in nucleic acid polymerases. *Nucleic Acids Res* **22**, 4625–4633 (1994).

53. Lin, S., Long, S., Ramirez, S. M., Cotter, R. J. & Woods, A. S. Characterization of the 'helix clamp' motif of HIV-1 reverse transcriptase using MALDI-TOF MS and surface plasmon resonance. *Anal Chem* **72**, 2635–2640 (2000).
54. Bebenek, K. *et al.* A minor groove binding track in reverse transcriptase. *Nat Struct Biol* **4**, 194–197 (1997).
55. Powell, M. D. *et al.* Residues in the H and I Helices of the HIV-1 Reverse Transcriptase Thumb Subdomain Required for the Specificity of RNase H-catalyzed Removal of the Polypurine Tract Primer. *The Journal of Biological Chemistry* **274**, 19885–19893 (1999).
56. Wang, Y.-X., Zhang, H.-J., Xu, J., Zheng, B.-J. & Wen, Y.-M. Mutational analysis of the 'turn' of helix clamp motif of HIV-1 reverse transcriptase. *Biochem Biophys Res Commun* **377**, 915–920 (2008).
57. Sarafianos, S. G. *et al.* Structure and Function of HIV-1 Reverse Transcriptase: Molecular Mechanisms of Polymerization and Inhibition. *J Mol Biol* **385**, 693–713 (2009).
58. Gutiérrez-Rivas, M. & Menéndez-Arias, L. A mutation in the primer grip region of HIV-1 reverse transcriptase that confers reduced fidelity of DNA synthesis. *Nucleic Acids Res* **29**, 4963–4972 (2001).
59. Yu, Q., Ottmann, M., Pechoux, C., Le Grice, S. & Darlix, J.-L. Mutations in the Primer Grip of Human Immunodeficiency Virus Type 1 Reverse Transcriptase Impair Proviral DNA Synthesis and Virion Maturation. *J Virol* **72**, 7676–7680 (1998).
60. Palaniappan, C. *et al.* Mutations within the primer grip region of HIV-1 reverse transcriptase result in loss of RNase H function. *Journal of Biological Chemistry* **272**, 11157–11164 (1997).
61. Singh, K., Marchand, B., Kirby, K. A., Michailidis, E. & Sarafianos, S. G. Structural aspects of drug resistance and inhibition of HIV-1 reverse transcriptase. *Viruses* **2**, 606–638 (2010).
62. Klarmann, G. J., Smith, R. A., Schinazi, R. F., North, T. W. & Preston, B. D. Site-specific Incorporation of Nucleoside Analogs by HIV-1 Reverse Transcriptase and the Template Grip Mutant P157S. *Journal of Biological Chemistry* **275**, 359–366 (2000).
63. Smith, R. A. *et al.* A New Point Mutation (P157S) in the Reverse Transcriptase of Human Immunodeficiency Virus Type 1 Confers Low-Level Resistance to (-)-2,3-Dideoxy-3-Thiacytidine. *Antimicrob Agents Chemother* **43**, 2077–2080 (1999).

64. Deval, J. *et al.* The molecular mechanism of multidrug resistance by the Q151M human immunodeficiency virus type 1 reverse transcriptase and its suppression using α -boranophosphate nucleotide analogues. *Journal of Biological Chemistry* **277**, 42097–42104 (2002).
65. Quan, Y. *et al.* Dominance of the E89G Substitution in HIV-1 Reverse Transcriptase in Regard to Increased Polymerase Processivity and Patterns of Pausing. *Journal of Biological Chemistry* **273**, 21918–21925 (1998).
66. Rajotte, D. *et al.* Identification and characterization of a novel HIV-1 nucleotide-competing reverse transcriptase inhibitor series. *Antimicrob Agents Chemother* **57**, 2712–2718 (2013).
67. Delviks-Frankenberry, K. A., Nikolenko, G. N., Barr, R. & Pathak, V. K. Mutations in Human Immunodeficiency Virus Type 1 RNase H Primer Grip Enhance 3'-Azido-3'-Deoxythymidine Resistance. *J Virol* **81**, 6837–6845 (2007).
68. Rausch, J. W. *et al.* Altering the RNase H primer grip of human immunodeficiency virus reverse transcriptase modifies cleavage specificity. *Biochemistry* **41**, 4856–4865 (2002).
69. Boyer, P. L., Sarafianos, S. G., Arnold, E. & Hughes, S. H. Analysis of mutations at positions 115 and 116 in the dNTP binding site of HIV-1 reverse transcriptase. *PNAS* **97**, 3056–3061 (1999).
70. Brown, J. A. & Suo, Z. Unlocking the sugar 'steric gate' of DNA polymerases. *Biochemistry* **50**, 1135–1142 (2011).
71. Martin-Hernandez, A. M., Domingo, E. & Menendez-Arias, L. Human immunodeficiency virus type 1 reverse transcriptase: role of Tyr115 in deoxynucleotide binding and misinsertion fidelity of DNA synthesis. *EMBO J* **15**, 4434–4442 (1996).
72. Parkash, V. *et al.* A sensor complements the steric gate when DNA polymerase ϵ discriminates ribonucleotides. *Nucleic Acids Res* **51**, 11225–11238 (2023).
73. Garforth, S. J., Kim, T. W., Parniak, M. A., Kool, E. T. & Prasad, V. R. Site-directed Mutagenesis in the Fingers Subdomain of HIV-1 Reverse Transcriptase Reveals a Specific Role for the β 3- β 4 Hairpin Loop in dNTP Selection. *J Mol Biol* **365**, 38–49 (2007).
74. Goldschmidt, V. *et al.* Mg²⁺ dependency of HIV-1 reverse transcription, inhibition by nucleoside analogues and resistance. *Nucleic Acids Res* **34**, 42–52 (2006).

75. Gong, S. *et al.* Kinetic and thermodynamic analysis defines roles for two metal ions in DNA polymerase specificity and catalysis. *Journal of Biological Chemistry* **296**, (2021).
76. Berg, J. M., Tymoczko, J. L., Gatto Jr. Gregory J. & Stryer, L. DNA Replication, Repair and Recombination. in *Biochemistry* vol. 9 949–982 (2019).
77. Rungrotmongkol, T., Nunthaboot, N., Aruksakunwong, O. & Hannongbua, S. *Encyclopedia of Biophysics. Encyclopedia of Biophysics* (Springer Berlin Heidelberg, Berlin, Heidelberg, 2013). doi:10.1007/978-3-642-16712-6.
78. Castro, C. *et al.* Nucleic acid polymerases use a general acid for nucleotidyl transfer. *Nat Struct Mol Biol* **16**, 212–218 (2009).
79. Batra, V. K. *et al.* Amino acid substitution in the active site of DNA polymerase β explains the energy barrier of the nucleotidyl transfer reaction. *J Am Chem Soc* **135**, 8078–8088 (2013).
80. Achuthan, V. & Destefano, J. J. Alternative divalent cations (Zn^{2+} , Co^{2+} , and Mn^{2+}) are not mutagenic at conditions optimal for HIV-1 reverse transcriptase activity. *BMC Biochem* **16**, (2015).
81. Lockbaum, G. J. *et al.* Inhibiting HTLV-1 Protease: A Viable Antiviral Target. *ACS Chem Biol* **16**, 529–538 (2021).
82. Sohraby, F. & Aryapour, H. Reconstruction of the binding pathway of an anti-HIV drug, Indinavir, in complex with the HTLV-1 protease using unaggregated unbiased molecular dynamics simulation. *Comput Biol Chem* **96**, 107616 (2022).
83. Keikha, M. & Karbalaei, M. The Design of Novel Kni-10729 Based Compound for Inhibition of HTLV-1 Protease as Novel Insight for Development of Anti-HTLV Drugs. *International Journal of Health and Life Sciences* **5**, e89840 (2019).
84. Sohraby, F. & Aryapour, H. Comparative analysis of the unbinding pathways of antiviral drug Indinavir from HIV and HTLV1 proteases by supervised molecular dynamics simulation. *PLoS One* **16**, e0257916 (2021).
85. Kuhnert, M., Steuber, H. & Diederich, W. E. Structural basis for HTLV-1 protease inhibition by the HIV-1 protease inhibitor indinavir. *J Med Chem* **57**, 6266–6272 (2014).
86. Li, M. *et al.* Mechanisms of HIV-1 integrase resistance to dolutegravir and potent inhibition of drug-resistant variants. *Sci Adv* **9**, (2023).
87. Sinha, P. & Yadav, A. K. Repurposing integrase inhibitors against human T-lymphotropic virus type-1: a computational approach. *J Biomol Struct Dyn* 1–12 (2024) doi:10.1080/07391102.2024.2304681.

88. Barski, M. S. *et al.* Structural basis for the inhibition of HTLV-1 integration inferred from cryo-EM deltaretroviral intasome structures. *Nat Commun* **12**, (2021).
89. Barski, M. *et al.* Cryo-EM structure of the deltaretroviral intasome in complex with the PP2A regulatory subunit B56γ. *Nat Commun* **11**, 5043 (2020).
90. Barski, M. S., Minnell, J. J. & Maertens, G. N. Inhibition of HTLV-1 infection by HIV-1 first-and second-generation integrase strand transfer inhibitors. *Front Microbiol* **10**, (2019).
91. Madeira, F. *et al.* Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res* **50**, W276–W279 (2022).
92. Wilkins, M. R. *et al.* Protein Identification and Analysis Tools in the ExPASy Server. in *2-D Proteome Analysis Protocols* (ed. Walker, J. M.) 531–552 (Humana Press, New Jersey, 2005). doi:10.1385/1-59259-584-7:531.
93. de Castro, E. *et al.* ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res* **34**, W362–W365 (2006).
94. Junier, T. & Pagni, M. Dotlet: diagonal plots in a Web browser. *Bioinformatics* **16**, 178–179 (2000).
95. Baldwin, E. T. *et al.* Human endogenous retrovirus-K (HERV-K) reverse transcriptase (RT) structure and biochemistry reveals remarkable similarities to HIV-1 RT and opportunities for HERV-K-specific inhibition. *Proceedings of the National Academy of Sciences* **119**, e2200260119 (2022).
96. Waterhouse, A. *et al.* SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res* **46**, W296–W303 (2018).
97. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* **10**, 845–858 (2015).
98. Webb, B. & Sali, A. Comparative protein structure modeling using MODELLER. *Curr Protoc Bioinformatics* **54**, 5.6.1-5.6.37 (2016).
99. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* (1979) **379**, 1123–1130 (2023).
100. Wu, R. *et al.* High-resolution de novo structure prediction from primary sequence. *bioRxiv* (2022) doi:https://doi.org/10.1101/2022.07.21.500999.
101. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

102. Agnihotry, S., Pathak, R. K., Singh, D. B., Tiwari, A. & Hussain, I. Protein structure prediction. in *Bioinformatics* 177–188 (Elsevier, 2022). doi:10.1016/B978-0-323-89775-4.00023-7.
103. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Crystallogr* **26**, 283–291 (1993).
104. Lüthy, R., Bowie, J. U. & Eisenberg, D. Assessment of protein models with three-dimensional profiles. *Nature* **356**, 83–85 (1992).
105. Bowie, J. U., Lüthy, R. & Eisenberg, D. A Method to Identify Protein Sequences That Fold into a Known Three-Dimensional Structure. *Science* (1979) **253**, 164–170 (1991).
106. Hung, M. *et al.* Elucidating molecular interactions of L-nucleotides with HIV-1 reverse transcriptase and mechanism of M184V-caused drug resistance. *Commun Biol* **2**, 469 (2019).
107. Lu, C.-H. *et al.* MIB2: metal ion-binding site prediction and modeling server. *Bioinformatics* **38**, 4428–4429 (2022).
108. Rodríguez-Lumbreras, L. A., Jiménez-García, B., Giménez-Santamarina, S. & Fernández-Recio, J. pyDockDNA: A new web server for energy-based protein-DNA docking and scoring. *Front Mol Biosci* **9**, (2022).
109. Trott, O. & Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* **31**, 455–461 (2010).
110. Eberhardt, J., Santos-Martins, D., Tillack, A. F. & Forli, S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J Chem Inf Model* **61**, 3891–3898 (2021).
111. Bauman, J. D. *et al.* Crystal engineering of HIV-1 reverse transcriptase for structure-based drug design. *Nucleic Acids Res* **36**, 5083–5092 (2008).
112. Vandamme, A.-M., Salemi, M. & Desmyter, J. The simian origins of the pathogenic human T-cell lymphotropic virus type I. *Trends Microbiol* **6**, 477–483 (1998).
113. Kalinichenko, S., Komkov, D. & Mazurov, D. HIV-1 and HTLV-1 Transmission Modes: Mechanisms and Importance for Virus Spread. *Viruses* **14**, 152 (2022).
114. Mohanty, S. & Harhaj, E. W. Mechanisms of Oncogenesis by HTLV-1 Tax. *Pathogens* **9**, 543 (2020).

115. de la Mora, L., Mallolas, J. & Ambrosioni, J. Epidemiología, tratamiento y pronóstico de la infección VIH en 2024: revisión práctica. *Med Clin (Barc)* (2024) doi:10.1016/j.medcli.2023.12.007.
116. Tyagi, R., Li, W., Parades, D., Bianchet, M. A. & Nath, A. Inhibition of human endogenous retrovirus-K by antiretroviral drugs. *Retrovirology* **14**, 21 (2017).
117. Tardiota, N., Jaberolansar, N., Lackenby, J. A., Chappell, K. J. & O'Donnell, J. S. HTLV-1 reverse transcriptase homology model provides structural basis for sensitivity to existing nucleoside/nucleotide reverse transcriptase inhibitors. *Virol J* **21**, 14 (2024).
118. D'Angelo, C. *et al.* Cold-Azurin, a New Antibiofilm Protein Produced by the Antarctic Marine Bacterium *Pseudomonas* sp. TAE6080. *Mar Drugs* **22**, 61 (2024).
119. Mahmoud, H. E., El-Far, S. W. & Embaby, A. M. Cloning, expression, and in silico structural modeling of cholesterol oxidase of *Acinetobacter* sp. strain RAMD in *E. coli*. *FEBS Open Bio* **11**, 2560–2575 (2021).
120. Stewart-Maynard, K. M. *et al.* Retroviral Nucleocapsid Proteins Display Nonequivalent Levels of Nucleic Acid Chaperone Activity. *J Virol* **82**, 10129–10142 (2008).
121. Pachulska-Wieczorek, K., Błaszczuk, L., Biesiada, M., Adamiak, R. W. & Purzycka, K. J. The matrix domain contributes to the nucleic acid chaperone activity of HIV-2 Gag. *Retrovirology* **13**, 18 (2016).
122. Cen, S. *et al.* Retrovirus-Specific Packaging of Aminoacyl-tRNA Synthetases with Cognate Primer tRNAs. *J Virol* **76**, 13111–13115 (2002).
123. Huang, H., Chopra, R., Verdine, G. L. & Harrison, S. C. Structure of a Covalently Trapped Catalytic Complex of HIV-1 Reverse Transcriptase: Implications for Drug Resistance. *Science* (1979) **282**, 1669–1675 (1998).

Supplementary material

SI. Overview of used informatics tools

Table SI. Overview of all used informatics tools mentioning the tool, its parameter setting, input, source and usage.

Tool	Parameters	Input	Source	Usage
ProtParam	default	P03362 [583-1167] & P03362 [583-1021]	Gasteiger E. et al. (2005)	Sequence characteristics
PepStats	default		Madeira, F. et al. (2022)	
PepInfo	default	P03362 [583-1167]	Madeira, F. et al. (2022)	
Clustal Omega	Sequence Type = "Protein"	HIV-1 RT p66 (P03366.3, 600-1159), HIV-2 RT p66 (P04584.3, 613-1171), RSV RT α subunit (P03354.2, 709-1280), MMLV RT (P03355.5, 660-1330), MPMV RT (P07572.2, 912-1496), XMLV RT p80 (Q2F7J0.1, 658-1328), ALV RT (Q7SQ98.2, 709-1567), HTLV-1 RT p62 (P03362.3, 583-1167), JSR RT (P31623.2, 866-1446), SMR RT (P03364.3, 999-1880), SR-1 RT (P04025.2, 913-1497), SR-2 RT (P51517.2, 909-1493), EIAV RT (P32542.1, 196-914), WMSV RT (P03359.2, 642-1310), BEV RT (P10272.2, 659-1335), REV RT (P03360.2, 79-751), JDV RT (Q82851.1, 523-1151), HTLV-3 RT (Q4U0X6.4, 562-1145), MMTV RT (P03365.3, 861-1436), BLV RT (P03361.2, 546-1120), VMV EV1 RT (P35956.2, 541-1091), HTLV-2 RT (P03363.4, 580-1166), MIA RT (P11368.1, 1-592), CAEV RT (P33459.1, 153-865), FIV RT (P31822.1, 154-709), FEV RT (P31792.1, 1-647), OL (P16901.1, 121-671), GHA (P04026.1, 1-546), SIV RT (P05895.2, 617-1176), BIV RT (P19560.2, 563-1193), STLV-1 RT (BBM28410.1, 2-205), STLV-2 RT (NP_056907.1, 595-809 & 1042-1163), STLV-3 RT (AAO62102.1, 25-236 & 496-577), FLV RT (P10273.2, 628-1297),	Madeira, F. et al. (2022)	Multiple sequence alignment and phylogeny

ScanProsite	default	HTLV-1 RT p62 (P03362, 583-1167), HIV-1 RT p66 (P05961, 594-1153), HERV-K RT (PDB: 7SR6)	De Castro, E. et al. (2006)	Domain identification, functional protein analysis and homology search
InterProScan	default		EMBL-EBI	
PfamScan	default		EMBL-EBI	
Dotlet JS	Window size = 50	P03362 [583-1167]	Junier, T. & Pagni, M. (2000)	Graphical pairwise sequence alignment
ProtScale	Window size: α -helix (11), β -sheet (6), polarity & hydropathy (5)	P03362 [583-1167] PDB files of all protein model made by ESMFold, AF2, OmegaFold, Modeller, SWISS-MODEL and Phyre ² .	Gasteiger E. et al. (2005)	Secondary structure prediction
ESMFold	default		Lin, Z. et al. (2023)	Ab initio structure prediction
AF2	default		Jumper, J. et al. (2021)	
OmegaFold	default		Wu, R. et al. (2022)	
Modeller via ModWeb	default		Webb, B. & Sali, A. (2017)	Homology-based structure prediction
SWISS-MODEL	default		Waterhouse, A. et al. (2018)	
Phyre²	default		Kelley, L. et al. (2015)	
ERRAT	default	PDB files of all protein model made by ESMFold, AF2, OmegaFold, Modeller, SWISS-MODEL and Phyre ² . PDB files of all protein model made by ESMFold, AF2, OmegaFold, Modeller, SWISS-MODEL and Phyre ² , and DNA PDB file containing DNA from HIV-1 RT (PDB: 6UIT) model. Restrictions were set using residues showed in Table 9.	Colovos, C. & Yeates, T. O. (1993)	Scoring of the obtained protein models
VERIFY 3D	default		Bowie et al., 1991.	
PROCHECK	default		Luethy et al., 1992	
MIB2	Metal Ion = Mg ²⁺		references	Mg ²⁺ docking
			Lin, Y. et al. (2016)	
PyDockDNA	default	SDF-files of antiviral compounds downloaded from Protein DataBank (PDB).	Rodríguez-Lumbreras, LA. et al. (2022)	DNA docking
SMILES	default	PDB file of final model of HTLV-1 RT:dsDNA:Mg ²⁺ , compound PDB files	Nation health institute (NIH). National Cancer institute (NCI). Chemical Biology Laboratory	An online tool to convert SDF-file format to 3D PDB file format.
AutoDockTools	default	Final model of HTLV-1 RT:dsDNA:Mg ²⁺ and antiviral compound, with configuration file containing coordinates of grid and size made with AutoDockTools.	Scripps institute	Dock preparation, grid configuration (size and position) needed for Autodock Vina.
AutoDock Vina	default	PDB files of predicted protein structures, output of AutoDock Vina, MIB2, PyDockDNA and SMILES.	Scripps institute	Compound docking into the model.
Chimera	x		Pettersen, E.F. et al. (2004)	Protein model visualisation, metal ion coordination analysis.
ChimeraX	x		Goddard, T. D. et al. (2018) Pettersen, E. F. et al. (2021)	Protein model visualisation and basic protein structure analysis

HIV-1

HIV-1

HIV-1

HIV-2

RSV

MMV

MPMV

MXLV

AIIV

HTLV-1

JSR

SMR

SR2

SR1

EIAV

WMSV

BEV

RV

JDV

HTLV-3

MMTV

BLV

VMV

HTLV-2

MIAP

CAEV

FIV

FEV

OL

CHAP

SIV

BIV

STLV-1

STLV-2

STLV-3

FLV

HIV-1

HIV-2

RSV

MMV

MPMV

MXLV

AIIV

HTLV-1

JSR

SMR

SR2

SR1

EIAV

WMSV

BEV

RV

HTLV-3

MMTV

BLV

VMV

HTLV-2

MIAP

CAEV

FIV

FEV

OL

CHAP

SIV

BIV

STLV-1

STLV-2

STLV-3

FLV

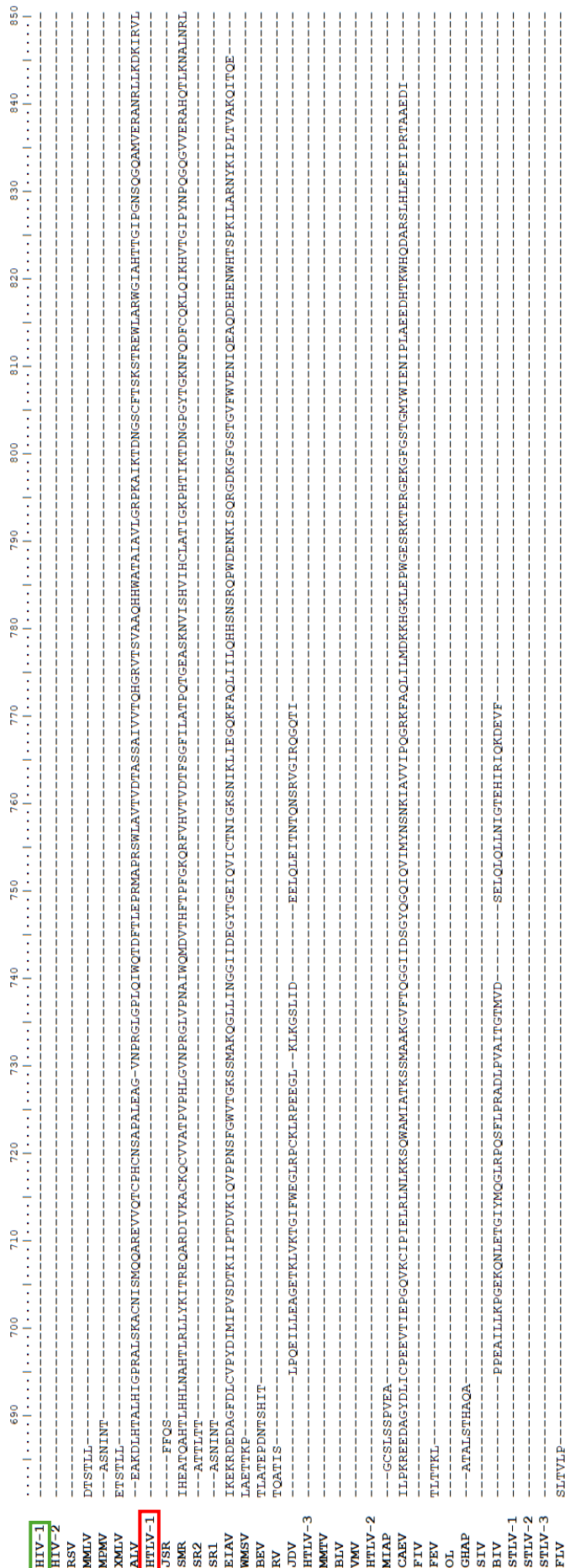
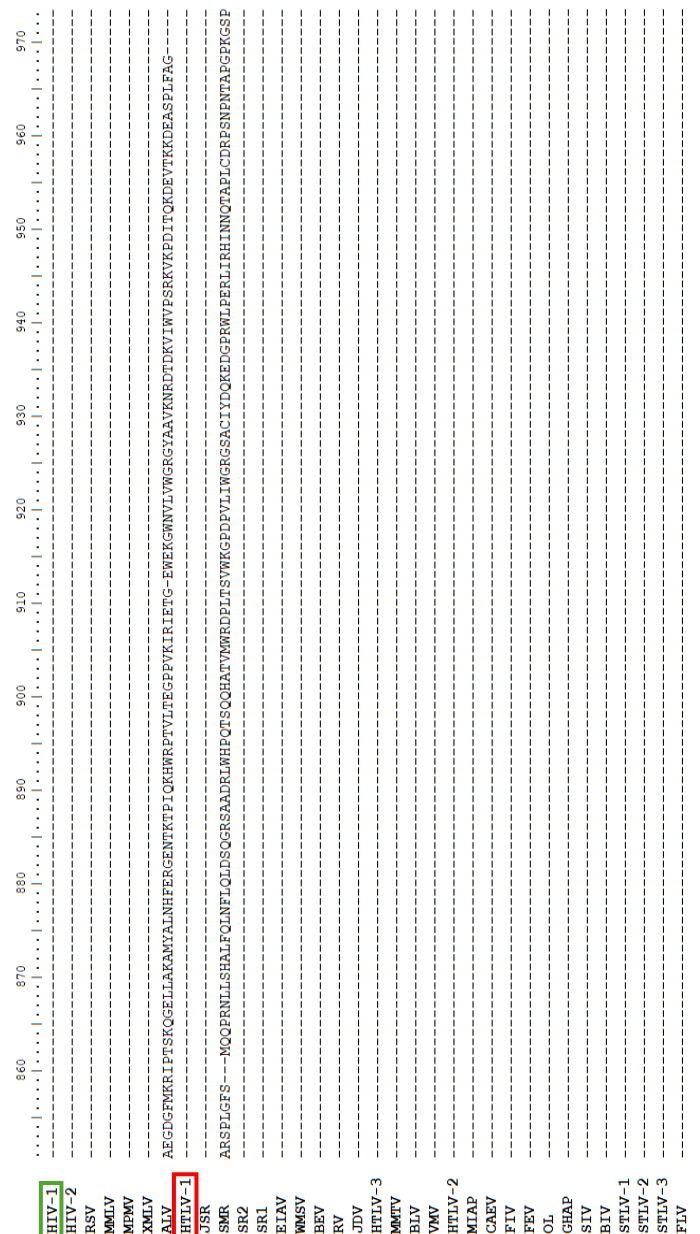


Figure S11. Multiple sequence alignment of 34 RT sequences by ClustalW.

Multiple sequence alignment was done according to ClustalW MSA protocol using 34 retroviral reverse transcriptase (RT) amino acid sequences. Colouring: similar (grey) and identical (black) residues were shaded from 85% conservation onwards. **Red box** highlights HTLV-1 RT, **green box** highlights HIV-1 RTRV: Reticuloendotheliosis virus, WMSV: Woolly monkey sarcoma virus, BEV: Baboon endogenous virus, FEV: Feline endogenous virus, MMLV: Moloney murine leukemia virus, XMLV: Xenotropic MuLV-related virus, FLV: Feline leukemia virus, JDV: Jembrana disease virus, BIV: Bovine immunodeficiency virus, CAEV: Caprine arthritis encephalitis virus, VMV: Visna/maedi virus, OL: Ovine lentivirus, EIAV: Equine infectious anemia virus, FIV: Feline immunodeficiency virus, SIV: Simian immunodeficiency virus, HIV-1: Human immunodeficiency virus type 1, HIV-2: Human immunodeficiency virus type 2, MIAP: Mouse intracisternal A-particle, GHAP: Golden hamster intracisternal A-particle, MMTV: Mouse mammary tumor virus, JSR: Jaagsiekte sheep retrovirus, SMR: Squirrel monkey retrovirus, SR-2: Simian retrovirus 2, MPMV: Mason-Pfizer monkey virus, SR-1: Simian retrovirus 1, RSV: Rous sarcoma virus, ALV: Avian leukosis virus, HTLV: Human T-cell lymphotropic virus (type 1, 2, 3), STLV: Simian T-cell lymphotropic virus (type 1, 2, 3).



SIII. Homology-based structure prediction per model scores

Table SIII. Per model scoring according to the algorithms' intrinsic scoring functions.

Algorithm	Type	Region	Score	Criteria	Number of templates /aligned sequences	Sequence identity of used template	template (PDB)
Phyre2	Secondary structure confidence score	1-435, 449-585	6-9	>5	99	27%	HERV-K RT (7SR6)
		436-448	3-5	>6			
SWISS-MODEL	Global Model Quality Estimate (GMQE)		0.6	> 0.7	1247	27.82%	HERV-K RT (7SR6)
	QMEANDisCo global		0.57	> 0.6			
	QMEAN Z-score		-3.9	-1 to 1			
Modeller	E-value		0	as low as possible	4	28%	HERV-K RT (7SR6)
	GA341		1	> 0.7			
	normalised Discrete Optimized Protein Energy score (z-DOPE)		0.18	< -1			
	TSVMod NO35		0.697	>0.4			
	ModPipe Quality score (MPQS)		1.2819	>1.1			

SIV. Comparative alignment: Alignment scores

Table SIV. Alignment score after pairwise structure alignment via ChimeraX Matchmaker.

Alignment score					
	Modeller	SWISS-MODEL	Phyre ²	AlphaFold2	HIV-1 RT
Modeller		2740.1	2784.5	2679.5	769.2
SWISS-MODEL			2806.8	2695.2	782.5
Phyre ²				2789.7	837.8
AlphaFold2					840.1

SV. Comparative alignment: Pruned pairs

Table SV. Number of pruned pairs after pairwise structure alignment via ChimeraX Matchmaker.

Pruned pairs					
	Modeller	SWISS-MODEL	Phyre ²	AlphaFold2	HIV-1 RT
Modeller		289	299	240	209
SWISS-MODEL			468	256	159
Phyre ²				250	156
AlphaFold2					221

SVI. Comparative alignment: RMSD for pruned pairs

Table SVI. RMSD after pairwise structure alignment via ChimeraX Matchmaker for pruned pairs only.

RMSD pruned pairs (Å)					
	Modeller	SWISS-MODEL	Phyre ²	AlphaFold2	HIV-1 RT
Modeller		0.885	0.896	1.109	1.195
SWISS-MODEL			0.316	1.169	0.953
Phyre ²				1.159	0.931
AlphaFold2					1.157

SVII. Comparative alignment: All pairs

Table SVII. Number of pairs for after pairwise structure alignment via ChimeraX Matchmaker.

All pairs					
	Modeller	SWISS-MODEL	Phyre ²	AlphaFold2	HIV-1 RT
Modeller		574	574	574	511
SWISS-MODEL			575	575	510
Phyre ²				577	522
AlphaFold2					517

SVIII. Comparative alignment: RMSD for all pairs

Table SVIII. RMSD after pairwise structure alignment via ChimeraX Matchmaker for all pairs.

RMSD all pairs (Å)					
	Modeller	SWISS-MODEL	Phyre ²	AlphaFold2	HIV-1 RT
Modeller		5.935	4.488	5.049	5.243
SWISS-MODEL			4.988	6.435	5.641
Phyre ²				5.549	6.732
AlphaFold2					6.069

SIX. Ca-Ca in HIV-1 RT active site for important residues

Table SIX. C-C distance between catalytically important residues in the HIV-1 RT active site

HIV-1 RT (PDB: 6UIT)							
distance(Ca-Ca, Å)	R72	D110	V111	Y115	Q151	D185	D186
K65	6.171	13.839	14.341	14.787	13.121	15.099	15.117
R72		15.317	14.24	11.245	8.629	15.098	16.597
D110			3.8	12.117	16.805	5.814	4.576
V111				9.231	14.947	5.393	6.659
Y115					7.281	8.681	12.234
Q151					13.476	16.275	
D185						3.826	

SX. Ca-Ca in HTLV-1 RT (AF2) active site for important residues

Table SX. Ca-Ca distance between catalytically important residues in the HTLV-1 RT active site.

HTLV-1 RT (AlphaFold2)							
Distance (Ca-Ca, Å)	R59	D98	L99	F103	Q139	D173	D174
K53	6.88	17.809	18.161	17.979	14.977	18.847	19.371
R59		17.674	16.547	13.609	10.2	17.473	19.298
D98			3.825	12.531	16.588	7.121	5.969
L99				9.346	14.29	6.139	7.155
F103					6.578	8.503	12.201
Q139						12.458	15.663
D173							3.856

SXI. HTLV-1 RT expression in various bacterial strains

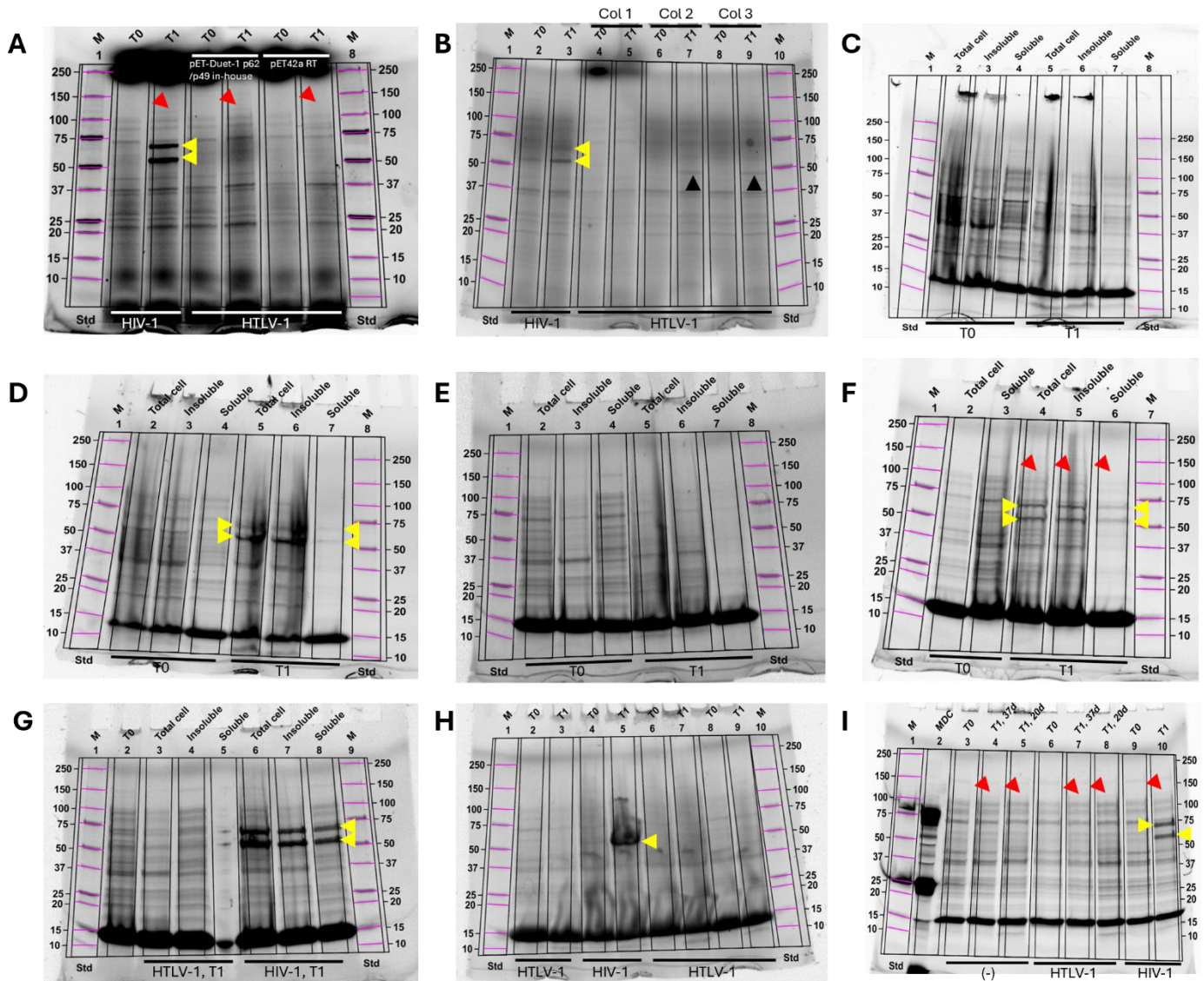


Figure SXI. HTLV-1 RT remains undetectable on SDS-PAGE after induction in different cell strains at various induction conditions.

(A) Induction of pET-Duet-1 p62/p49 in-house [lane 4 - 5] and pET42a (+) RT [lane 6 - 7] in BL21 (DE3) RIL cells at 37 °C for 2 hours with 1 mM IPTG. (B, C, D) Induction of HTLV-1 pET-Duet-1 p62/p49 in-house (B,C) and HIV-1 pET-Duet-1 p66/p51 (D) in BL21 (DE3) Arctic express cell at 13 °C for 18 hours at 1 mM IPTG. (E,F) Induction of HTLV-1 pET-Duet-1 p62/p49 in-house (E), HIV-1 pET-Duet-1 p66/p51 (F) at 20 °C overnight with 0.5 mM IPTG in BL21 (DE3) star cells. (G) Induction of HTLV-1 pET-Duet-1 p62/p49 in-house and HIV-1 pET-Duet-1 p66/p51 at 37 °C for 3 hours with 1 mM IPTG in BL21 (DE3) star cells. (H) Induction of HTLV-1 pET15b [lane 2 - 3], HTLV-1 pET42a (+) RT [lane 6 - 7], HTLV-1 pET42a (+) pol [lane 8 - 9] and HIV-1 pET-Duet-1 p66/p51 [lane 4 - 5] overnight at 20 °C with 0.5 mM IPTG in BL21 (DE3) star cells. (I) Induction of HTLV-1 pET-Duet-1 p62/p49 Gs and HIV-1 pET-Duet-1 p66/p51 at 37 °C for 6 hours and 20 °C overnight at 0.5 mM IPTG in BL21 (DE3) star cells. Negative control (-) is HTLV-1 pET-Duet-1 p62/p49 Gs plasmid with 6xHis out-of-frame with p49. T0: start point, T1: end point, M: Precision Plus protein unstained standards (Biorad), MDC: Precision plus protein dual color standards (BioRad). (A, B, D, G-I) yellow arrows: HIV-1 RT. (A, F, I) red arrows: T7RNAP. (B) black arrow: possibly 6xHis-p49.

SXII. Amplification plot: HTLV-1 pET15b

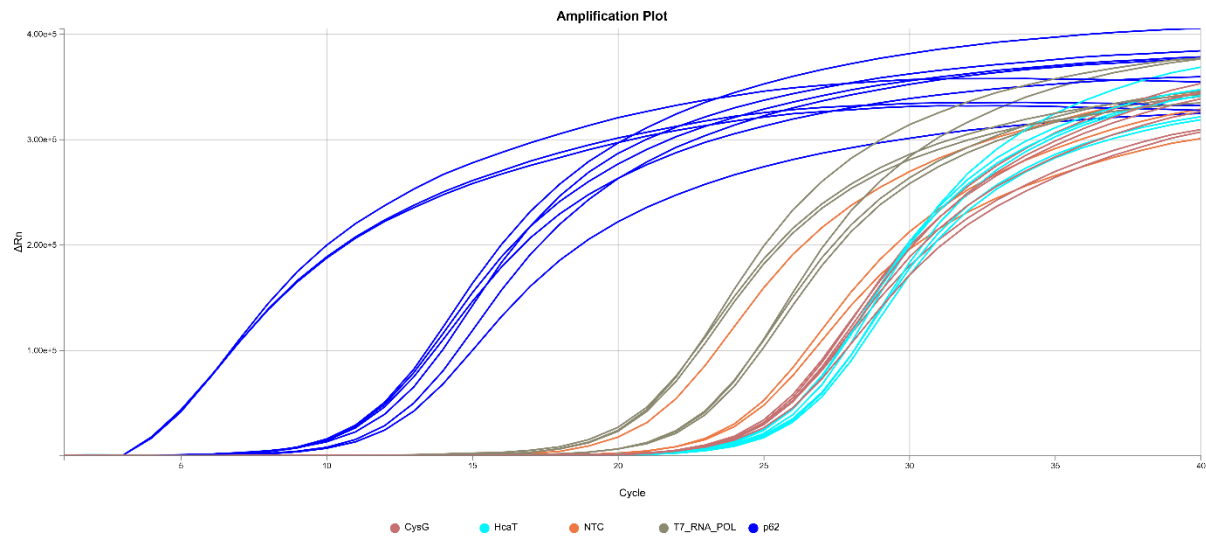


Figure SXII. RT-qPCR amplification plot of HTLV-1 pET15b.

SXIII. Amplification plot: HTLV-1 pET42a (+) RT

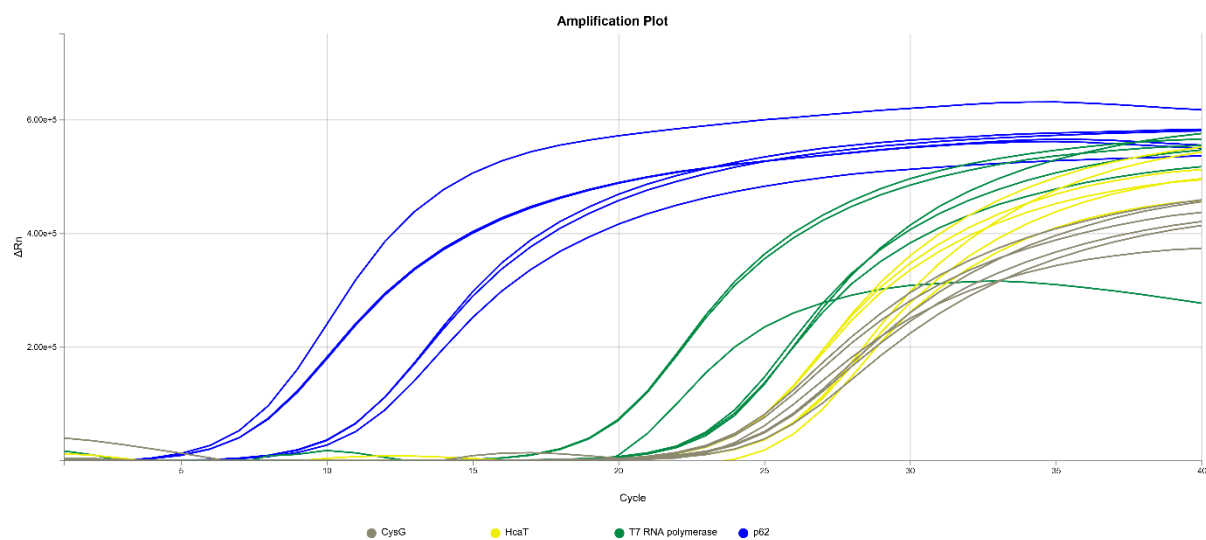


Figure SXIII. RT-qPCR amplification plot of HTLV-1 pET42a(+) RT.

SXIV. Amplification plot: HIV-1 pET-Duet-1 p66/p51

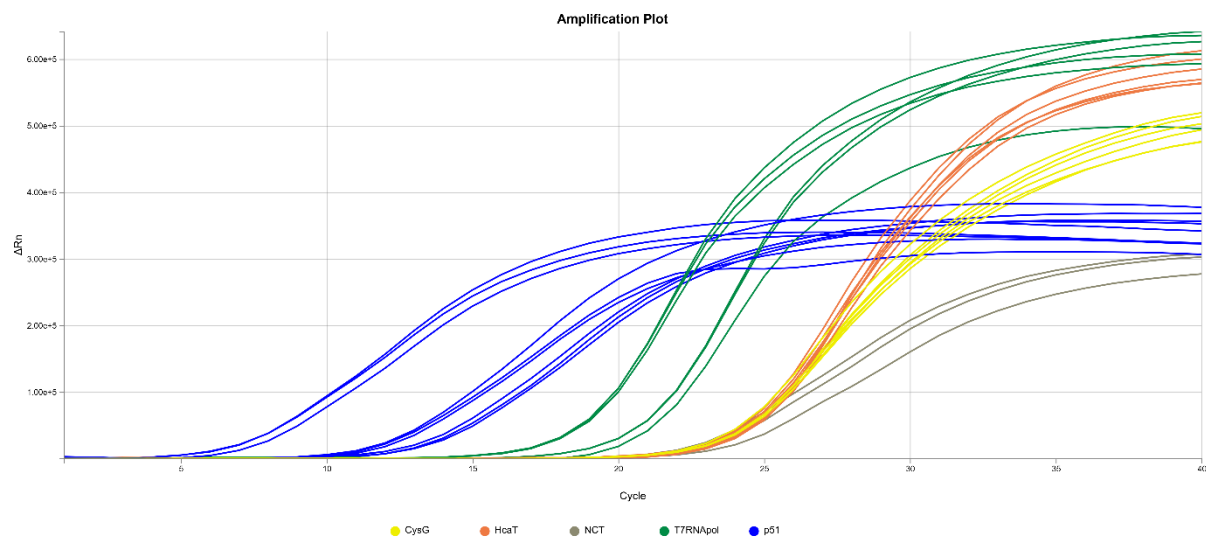


Figure SXIV. RT-qPCR amplification plot of HIV-1 pET-Duet-1 p66/p51.

SXV. Amplification plot: HTLV-1 pET-Duet-1 p62/p49 in-house

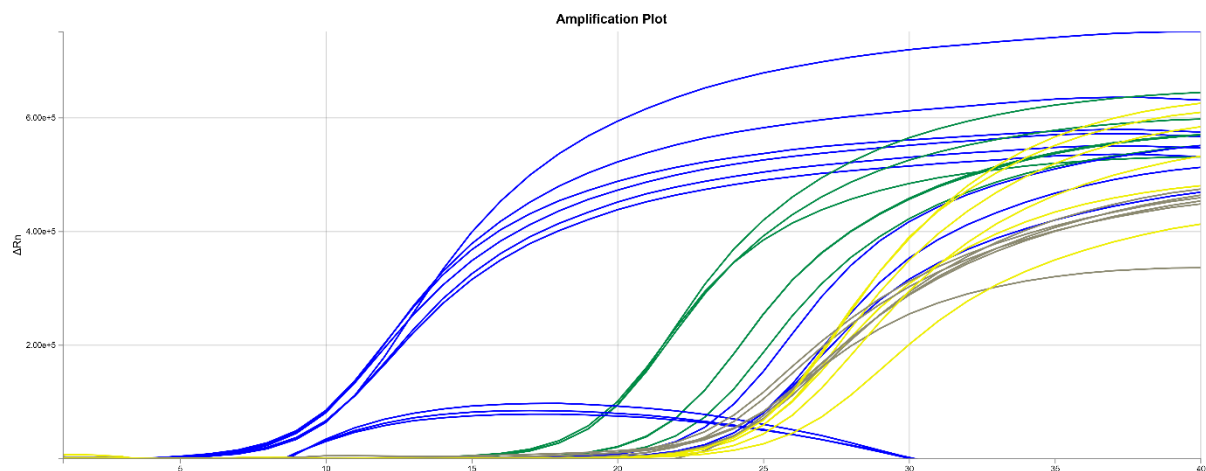


Figure SXV. RT-qPCR amplification plot of HTLV-1 pET-Duet-1 p62/p49 in-house.

SXVI. Amplification plot: HTLV-1 pET-Duet-1 p62/p49 Gs

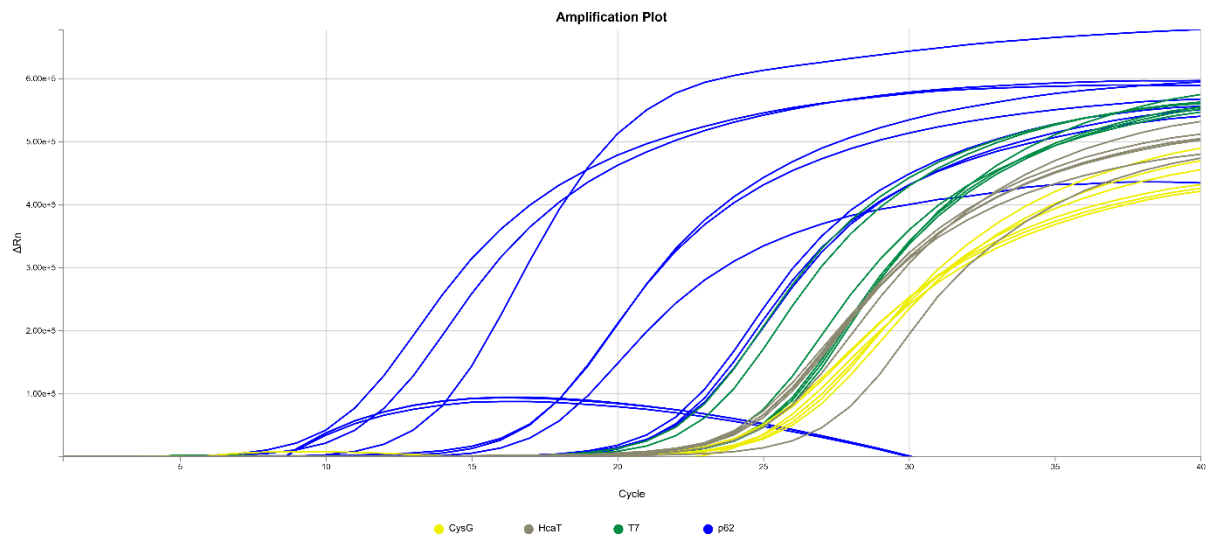


Figure SXVI. RT-qPCR amplification plot of HTLV-1 pET-Duet-1 p62/p49 Gs.