Molecular Diversity (2024) 28:1869–1888
https://doi.org/10.1007/s11030-023-10707-6

ORIGINAL ARTICLE

Data mining and molecular dynamics analysis to detect HIV-1 reverse transcriptase RNase H activity inhibitor

Naeem Abdul Ghafoor[1] · Kevser Kübra Kırboğa[2,3] · Ömür Baysal[4] · Barış Ethem Süzek[5] · Ragıp Soner Silme[6]

Received: 30 May 2023 / Accepted: 26 July 2023 / Published online: 10 August 2023

Abstract

HIV-1 is a deadly virus that affects millions of people worldwide. In this study, we aimed to inhibit viral replication by targeting one of the HIV-1 proteins and identifying a new drug candidate. We used data mining and molecular dynamics methods on HIV-1 genomes. Based on MAUVE analysis, we selected the RNase H activity of the reverse transcriptase (R.T) enzyme as a potential target due to its low mutation rate and high conservation level. We screened about 94,000 small molecule inhibitors by virtual screening. We validated the hit compounds' stability and binding free energy through molecular dynamics simulations and MM/PBSA. Phomoarcherin B, known for its anticancer properties, emerged as the best candidate and showed potential as an HIV-1 reverse transcriptase RNase H activity inhibitor. This study presents a new target and drug candidate for HIV-1 treatment. However, in vitro and in vivo tests are required. Also, the effect of RNase H activity on viral replication and the interaction of Phomoarcherin B with other HIV-1 proteins should be investigated.

Keywords Computational biology · Drug discovery · HIV-1 · Molecular dynamics · Molecular docking · Pol gene · Reverse-transcriptase RNase H

Introduction

Human immunodeficiency virus 1 (HIV-1) has been a global public health issue since its first identification in 1981. HIV-1 infection leads to acquired immunodeficiency syndrome (AIDS), which makes the host's immune system vulnerable to secondary infections [1]. The World Health Organization refers to HIV-1 as a "global epidemic" [2], and according to the Joint United Nations Program on HIV/AIDS (UNAIDS) global statistics, around 33.6–48.6 million individuals have died of AIDS-related illnesses since its emergence until 2021, and a total of 33.9–43.8 million individuals are still living with the virus, most of them in sub-Saharan Africa [3, 4]. H.I.V. is a complex retrovirus; like other retroviruses, it stores its genome as a pair of ssRNA molecules of ~ 9 kb. The genome contains (1) the gag gene, which encodes the structural proteins, mainly the capsid protein, the matrix protein, and the nucleocapsid protein; (2) the pol gene, which encodes the reverse transcriptase enzyme (R.T.), the protease enzyme, and the integrase enzyme (IN); and (3) the env gene, which encodes the membrane glycoprotein 120 and glycoprotein 41. The H.I.V. genome also encodes six regulatory proteins, such as Tat, Rev, Nef, Vif, Vpr, and Vpu, responsible for its pathogenicity and

The study is dedicated to the memory of the late Prof. Dr. Randall J. Cohrs.

✉ Ömür Baysal
omurbaysal@mu.edu.tr

[1] Department of Molecular Biology and Genetics, Faculty of Science, Muğla Sıtkı Koçman University, Kötekli, 48121 Muğla, Turkey

[2] Bioengineering Department, Bilecik Seyh Edebali University, 11230 Bilecik, Turkey

[3] Informatics Institute, Istanbul Technical University, Maslak, 34469 Istanbul, Turkey

[4] Molecular Microbiology Unit, Department of Molecular Biology and Genetics, Faculty of Science, Muğla Sıtkı Koçman University, Kötekli, 48121 Muğla, Turkey

[5] Department of Computer Engineering, Faculty of Engineering, Muğla Sıtkı Koçman University, Kötekli, 48000 Muğla, Turkey

[6] Center for Research and Practice in Biotechnology and Genetic Engineering, Istanbul University, Vezneciler, Fatih, 34119 Istanbul, Turkey

replication in the host [5]. Once the virus enters the host, the capsid disintegrates, and the viral R.N.A.s are reverse transcribed into D.N.A. molecules by R.T., which starts with an RNA/DNA hybrid followed by further cleavage of the R.N.A. and synthesis of dsDNA. The proviral dsDNA is then integrated into the host genome by the integrase enzyme, which remains a reservoir for the virus until the cell is activated, upon which the cellular transcription and translation mechanisms are hijacked to produce the viral proteins and R.N.A. genomes [6–8]. Nucleoside R.T. inhibitors (NRTIs), nonnucleoside reverse transcriptase inhibitors (NNRTIs), protease inhibitors (P.I.s), entry or fusion inhibitors, and integrase strand transfer inhibitors (INSTIs) are the major classes of antiretroviral therapy drugs targeting five different phases of the H.I.V. life cycle. There are 17 registered drugs available for treatment against H.I.V. [9].

In the management of AIDS, lifelong drugs are administered. However, they cause immune-suppressive effects with possible other disorders. Therefore, developing new drugs targeting the exact issue is essential. Antiretroviral therapy against H.I.V. has various properties that change a fatal disease into a chronic form. Therefore, patients should be treated with continuous, potentially lifelong drugs to suppress H.I.V. replication. However, therapy is unsuccessful due to limitations that require strict lifelong drug adherence. The drugs are generally well tolerated, but they have some short-term toxic effects that could result in potential known and unknown long-term toxic effects. This situation causes persistent immune dysfunction and carries the risk of various non-AIDS-related complications, such as heart, bone, liver, kidney, and neurocognitive diseases [9].

RNase H is a family of ribonuclease enzymes that hydrolyse R.N.A. in RNA–DNA hybrids. These enzymes provide D.N.A. synthesis by cleaving the R.N.A. template [10]. In retroviruses such as HIV-1, RNase H is encoded as part of the R.T. enzyme and is required for viral replication. RNase H cuts the R.N.A. template during the conversion of viral R.N.A. to D.N.A., allowing the synthesis of the second D.N.A. strand. It also inhibits the translation of viral R.N.A. and the formation of new virions. Therefore, RNase H is an essential target for antiviral drug development. However, no approved drug specifically targeting RNase H. RNase H inhibition can prevent HIV-1 infection by stopping the viral replication cycle.

Computational drug discovery methods have gained huge momentum recently, especially with the availability and accessibility of vast computation resources for lower costs [11]. Virtual screening has been previously used to identify lead compounds potentially inhibiting several HIV-1 proteins and enzymes, such as the RNase domain. However, such studies were based solely on molecular docking, pharmacophore, and ADMET assays that consider mainly the best docking pose that the ligands can have against the target protein in a vacuum-like condition, which may not reflect their interaction in the physiological cell condition [12–17]. Zhang et al. performed an extensive computational analysis of molecular dynamics simulations and identified a potential HIV-1 R.T. RNase domain inhibitor. However, the experiment was limited to only 77 α-hydroxytropolone derivatives, which restricted the discovery of novel small molecule inhibitors. Therefore, no large-scale comprehensive computational analysis with all-atom molecular dynamics simulations and binding free energy calculations validating the potential lead compounds has been done against the HIV-1 R.T. RNase domain [18, 19]. The HIV-1 R.T. RNase domain plays a critical role in cleaving the viral RNA/DNA hybrid and synthesising the viral D.N.A. Inhibiting this protein can disrupt the H.I.V. life cycle and stop viral replication. However, the current antiretroviral drugs do not target this protein but rather affect other regions of the R.T. enzyme or different stages of H.I.V. This study presents a comprehensive computational drug discovery method to discover potential small molecule inhibitors against the HIV-1 R.T. RNase domain. This method combines virtual screening, molecular dynamics simulations, and binding free energy calculations [20]. This allows for obtaining a more realistic and reliable interaction model of the target protein under physiological conditions. This study fills an essential gap in developing new antiretroviral drugs targeting the HIV-1 R.T. RNase domain. The compounds evaluated in this study can offer unique and practical options for H.I.V. treatment and reduce AIDS-related complications.

Materials and methods

Phylogenetic tree

We retrieved 190 complete HIV-1 genomic sequences from the Los Alamos H.I.V. sequence database [21]. We manually selected the sequences to include only two sequences (where applicable) from each country and all the available geographic regions. We converted the whole genome sequences to FASTA format and aligned them using ClustalW [22] (Suppl. 14). We constructed a phylogenetic tree and visualised it using Ugene software v. 38. [23] (Suppl. 13 ve Suppl. 16). This study aimed to estimate the gene pool variation and identify common genomic regions representing different subtype diversity depending on the origin of the selected genomes.

Whole-genome alignment and BLASTx

This study was applied to determine the rationale of targeting the R.T. enzyme for the drug discovery and development efforts of anti-HIV-1 drugs. Following the establishment of

the R.T. enzyme as the best candidate for drug targeting, a dataset of around 94.000 small drug-like molecules was obtained from the ZINC15 database, structure-based virtual screening against the R.T. RNase domain was performed via molecular docking, the interaction, and dynamics of the top lead molecules were further validated using molecular dynamics. The general workflow of the study is illustrated in Fig. 1.

The dataset had some bias towards certain regions. For instance, the U.S. had higher coverage for its size and the number of high-quality sequences deposited. In contrast, despite its population size, other countries like India had a lower range due to the limited number of high-quality sequences deposited. Whole-genome alignment was performed using the progressive MAUVE algorithm with match seed weight set to automatic calculation, minimum Locally Collinear Blocks (L.C.B.) set to default (3 times the minimum match size), progressive MUSCLE (v3.6) was selected as a gap aligner for each L.C.B., and minimum island size, maximum backbone gap size, minimum backbone size were set to 50 [24, 25]. The list of all sequences used for the alignment is included in Supplementary Data 1, and the whole genome alignment result is included in Supplementary Data 2. The alignment result was visualised in Geneious Prime (v2020.1), and the most conserved continuous region with no gaps in the alignment was extracted from the alignment and visualised separately in-depth [26]. A consensus identity sequence from the conserved fragment was generated using Jalview and submitted to NCBI BLASTx with the default parameters (max target sequences 100, expected threshold 0.05, word size 6, max match in a query range 0, matrix BLOSUM62, gap costs for existence 11, an extension of 1, and compositional adjustments via conditional compositional

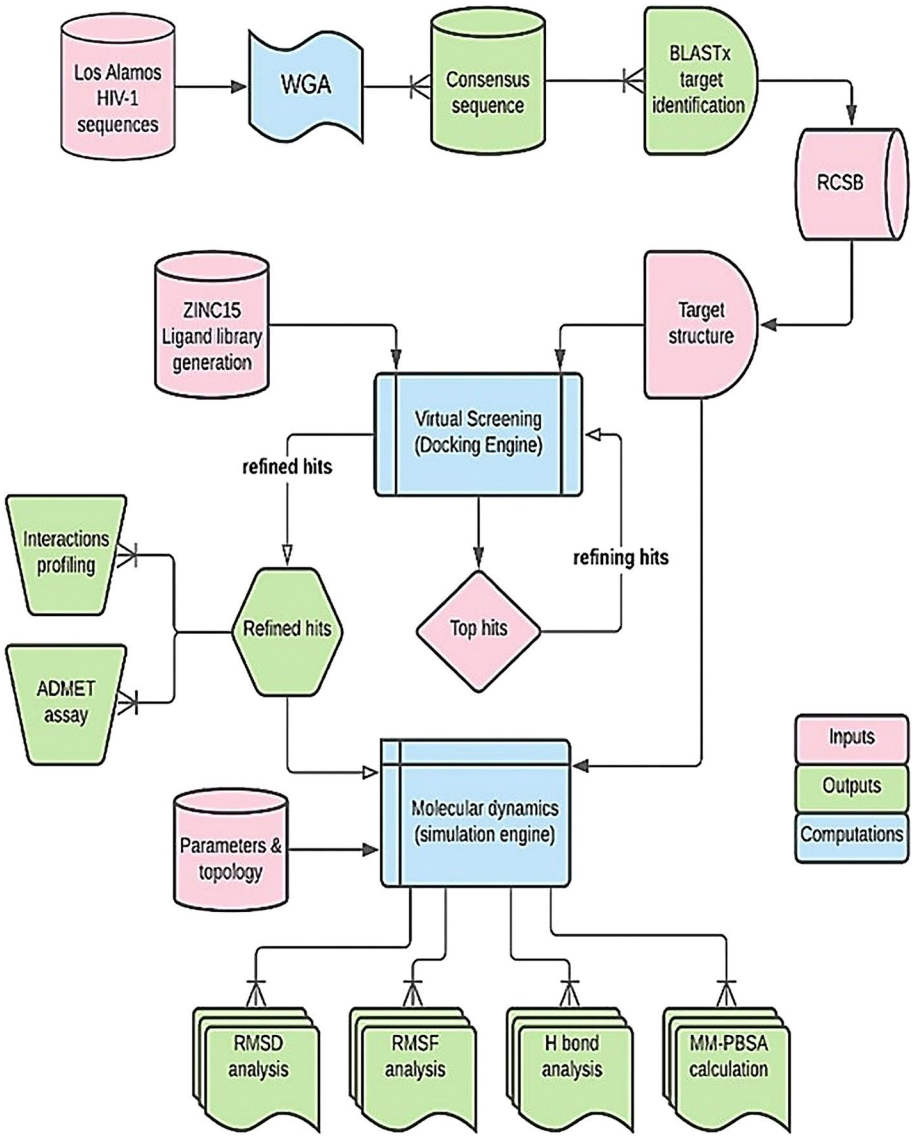


Fig. 1 The general workflow of the study. The rationale for targeting the R.T. enzyme for the drug discovery and development efforts of anti-HIV-1 drugs, following the establishment of the R.T. enzyme as the best candidate for drug targeting

score matrix adjustment), the alignment for the extracted fragments are provided in Supplementary Data 3, and the consensus sequence is provided in Supplementary Data 4 [27–29] that were analysed, and sky-blue shapes indicate the computational steps.

Molecular docking studies

We conducted molecular docking studies using AutoDock Vina (v1.1.2) [30] to identify novel RNase H inhibitors from an extensive library of small molecules and to evaluate the anti-RNase H activity of a selected compound (Phomoarcherin B) against different RNase H enzymes and a critical target structure (the transmembrane domain of HIV-1 gp41). The general docking procedure was as follows: we retrieved the experimental X-ray diffraction structure of HIV-1 R.T. with Protein Data Bank (PDB) I.D. 3IG1 from the Research Collaboratory for Structural Bioinformatics (RCSB) PDB website [31, 32]. This structure consists of two subunits, A and B, of HIV-1 R.T. with 555 and 428 amino acids, respectively. Each subunit contains sequence and structural domains, such as R.T., RNase H, and connection domains [33]. However, this structure does not represent the full-length HIV-1 R.T., which has 560 amino acids. In subunit A, 45 amino acids at the N-terminus and 5 amino acids at the C-terminus; in subunit B, 45 amino acids at the N-terminus and 132 amino acids at the C-terminus are missing regions. The missing residues from the structure were added using PyMol's builder plugin (open source v2.5.0), and the loop regions where the residues were added were refined using MODELLER (v10.1) [34–36]. We cleaned the structure from all heteroatoms except cofactor atoms, added polar hydrogens where necessary, and computed Kollman charges [37]. We minimised the energy on Phomoarcherin B using Avogadro software, considering the MMFF94 force field [38].

Molecular docking-based virtual screening

We performed molecular docking-based virtual screening to identify novel RNase H inhibitors from an extensive library of small molecules. The specific docking parameters and settings were as follows: we generated a library of 94.545 annotated anodyne small molecules (ligands) that are stable at physiological pH and have a 0, − 1, or − 2 charge from the ZINC15 database [39]. We calculated a grid box with a size of 25 Å X 32 Å X 32 Å along the X, Y, and Z-axis (a box around the RNase H active site). We conducted virtual screening with HIV-1 R.T. structure against the ligand dataset within the grid box calculated at an exhaustiveness of 64 using AutoDock Vina (v1.1.2) [30].

Additional docking studies

We performed additional docking studies to evaluate the anti-RNase H activity of Phomoarcherin B against different RNase H enzymes and a critical target structure. The specific docking parameters and settings were as follows: we used the R.T. enzyme of Feline Immunodeficiency Virus (FIV), which also exhibits a similar DEDD motif (PDB: 5OVN) [40], the RNase H of Bacteriophage T4 (PDB: 1TFR) [41], the monomeric R.T of Moloney murine leukaemia virus (MLV, PDB: 4MH8) [10] and the transmembrane domain of HIV-1 gp41 (PDB: 5JYN) [42] as protein structures; we predicted the affinity of Phomoarcherin B to the transmembrane domain of HIV-1 gp41 by performing blind docking using a grid box for the receptor coordinates that fit within the following volume: X = 78.0, − 17.2, − 8.8, Y = 128.0, 19.3, 26.2, Z = 103.0, 1.1, 8.7; ligand coordinates; X = 76.0, − 28.2, − 12.5, Y = 131.4, 21.5, 27.6, Z = 103.7, − 3.4, 7.5; and grid box X, Y, Z dimensions; 55.4 Å, 49.7 Å, 40.1 Å; and we conducted molecular docking studies according to these values.

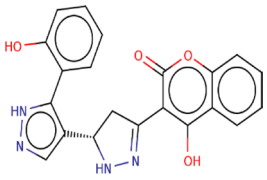
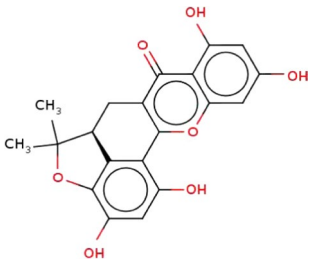
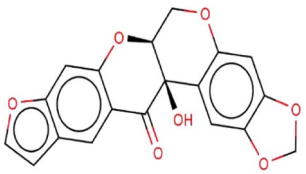
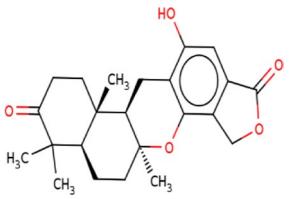
Protein–ligand interaction profiling

The best dock pose of the top hit ligands was loaded with HIV-1 R.T. to PyMol, and all the residues within 4 Å from the lead compounds were visualised (i.e., all potential hydrophobic interactions, hydrogen bonds, and ionic interactions) and evaluated; the manually predicted bonds were also cross-validated with TU Dresden's Protein–Ligand Interaction Profiler (PLIP) webserver and only overlapping interactions were considered [43].

Molecular dynamics simulation

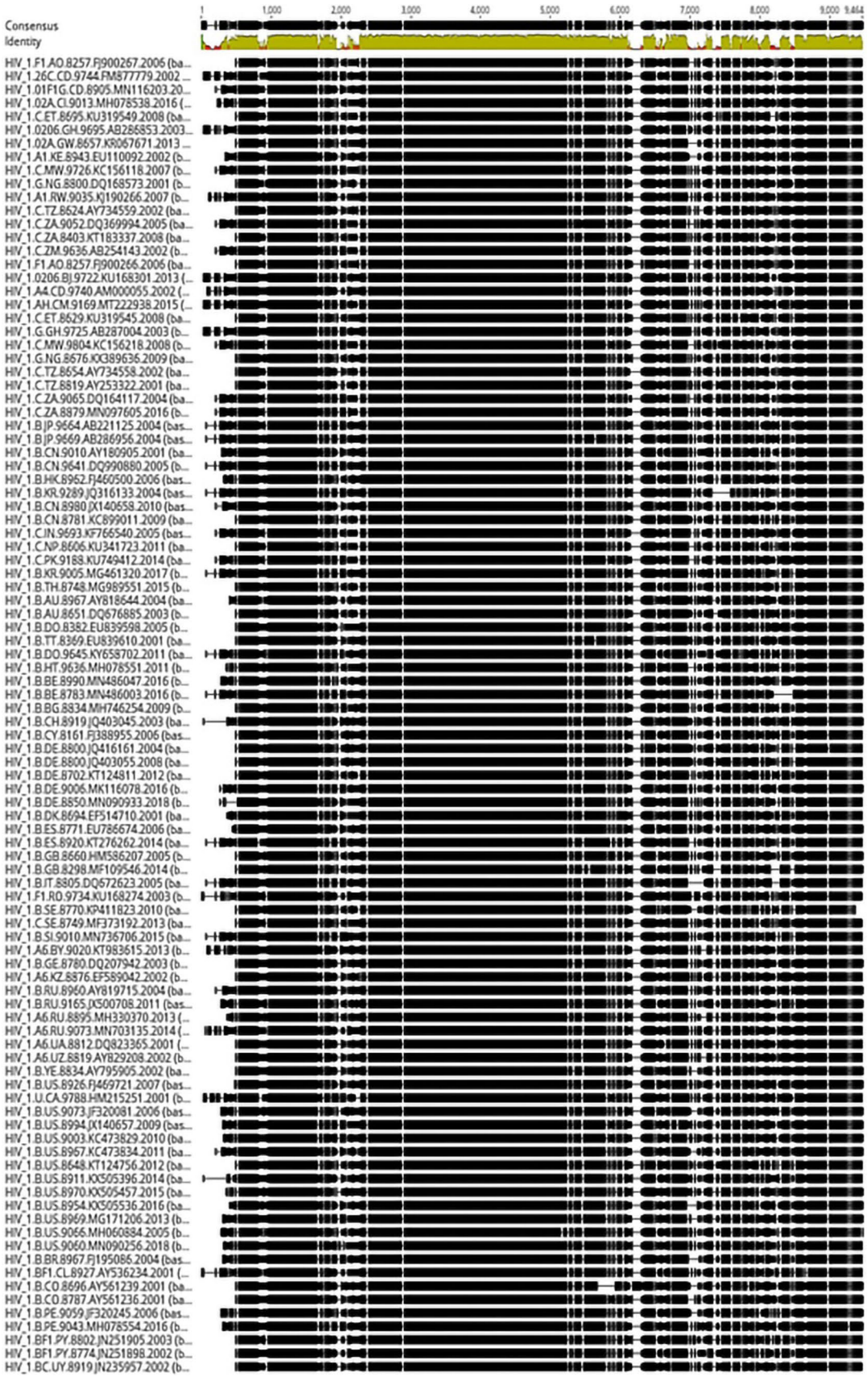
Studies on docking do not take into account the protein's flexibility. To ensure complete binding mode and check for stability, Molecular dynamics (M.D.) should be evaluated. M.D. simulations are computational techniques that allow the study of the dynamic behaviour of molecular systems over time [44, 45]. In this study, we used the Desmond module of the academic version of the Maestro program for M.D. simulations [46]. This way, we could examine the interactions, conformational changes, and system behaviour between protein and ligand. In this study, we used M.D. simulations to investigate the stability and behaviour of a protein–ligand complex. The protein of interest is PDB ID 3IG1, and we used four natural compounds (Table 1) as ligands for simulation. We aimed to determine the effect of ligands on the protein's stability and identify which protein

Table 1 A brief description of the top hit compounds is shown in Fig. 3A, B

Compound no.	ZINC ID	IUPAC name	Docking score (kcal/mol)	2D chemical structure
Compound **1**	ZINC000103288276	4-hydroxy-3-[5-[5-(2-hydroxyphenyl)-1H-pyrazol-4-yl]-4,5-dihydro-1H-pyrazol-3-yl]chromen-2-one	− 8.6	
Compound **2**	ZINC000013373252	1,3,8,10-tetrahydroxy-5,5-dimethyl-5a,6-dihydro-5H,7H-[1]benzofuro[3,4-bc]xanthen-7-one	− 8.5	
Compound **3**	ZINC000015147377	1-hydroxy-5,7,11,14,18-pentaoxahexacyclo[11.11.0.02,10.04,8.015,23.017,21]tetracosa-2,4(8),9,15(23),16,19,21-heptaen-24-one	− 8.5	
Compound **4**	ZINC000071318700	(6aR,6bS,10aR,12aS)-5-Hydroxy-6b,10,10,12a-tetramethyl-6a,7,8,10,10a,11,12,12a-octahydro-1H-benzo[a]furo[3,4-h]xanthene-3,9(6H,6bH)-dione	− 8.5	

regions fluctuated the most. We used an N.P.T. ensemble with a temperature of 300 K and a simulation time of 100,102 ns to perform the M.D. simulations. We used the OPLS-AA force field [47] for the protein–ligand complex, and the parameters for the ligands were generated using the LigPrep module of Maestro [46]. The N.P.T. ensemble is a type of molecular dynamics simulation that considers the system's particle number, pressure, and temperature. It allows the simulation of a system in a constant volume and temperature environment where pressure can fluctuate. This ensemble is helpful in examining the behaviour of systems in solution and provides a more realistic simulation environment.

We monitored the RMSD of both protein and ligand throughout the simulation. We also calculated the RMSF of the protein to identify local changes along the protein chain. The RMSF values for each residue were monitored throughout the simulation, and the peaks allowed us to identify the protein areas that fluctuated the most during the simulation (Fig. 7). In addition, we analysed ligand contacts throughout the simulation to identify residues in the protein that interacted with the ligand. We marked the ligand-interacting residues with vertical bars in green. This allowed us to visualise which residues were involved in binding the ligand to the protein. We also monitored the secondary structure of the protein throughout the simulation. The secondary structure of a protein refers to local conformation patterns such as alpha helices and beta sheets. A protein's secondary structure can play an essential role in its function, and changes in secondary structure can affect the protein's activity. Finally, we reported the presence of counterions and salt concentration in the solvent medium. Counterions are ions added to a system to maintain charge neutrality, and their presence can affect the system's behaviour. In conclusion, our M.D. simulation study provided insight into the stability and behaviour of a protein–ligand complex under certain simulation conditions. We monitored the RMSD and RMSF of the protein and ligand, identified ligand-interacting residues, monitored the secondary structure of the protein, and reported


1
1,000
2,000
3,000
4,000
5,000
6,000
7,000
8,000
9,000
9,464
Consensus
Identity
HIV_1.F1.AO.8257.FJ900267.2006 (ba...
HIV_1.26C.CD.9744.FM877779.2002 ...
HIV_1.01F1G.CD.8905.MN116203.20...
HIV_1.02A.CI.9013.MH078538.2016 (...
HIV_1.C.ET.8695.KU319549.2008 (ba...
HIV_1.0206.GH.9695.AB286853.2003...
HIV_1.02A.GW.8657.KR067671.2013 ...
HIV_1.A1.KE.8943.EU110092.2002 (b...
HIV_1.C.MW.9726.KC156118.2007 (b...
HIV_1.G.NG.8800.DQ168573.2001 (b...
HIV_1.A1.RW.9035.KJ190266.2007 (b...
HIV_1.C.TZ.8624.AY734559.2002 (ba...
HIV_1.C.ZA.9052.DQ369994.2005 (ba...
HIV_1.C.ZA.8403.KT183337.2008 (ba...
HIV_1.C.ZM.9636.AB254143.2002 (b...
HIV_1.F1.AO.8257.FJ900266.2006 (ba...
HIV_1.0206.BJ.9722.KU168301.2013 (...
HIV_1.A4.CD.9740.AM000055.2002 (...
HIV_1.AH.CM.9169.MT222938.2015 (...
HIV_1.C.ET.8629.KU319545.2008 (ba...
HIV_1.G.GH.9725.AB287004.2003 (b...
HIV_1.C.MW.9804.KC156218.2008 (b...
HIV_1.G.NG.8676.KX389636.2009 (ba...
HIV_1.C.TZ.8654.AY734558.2002 (ba...
HIV_1.C.TZ.8819.AY253322.2001 (ba...
HIV_1.C.ZA.9065.DQ164117.2004 (ba...
HIV_1.C.ZA.8879.MN097605.2016 (b...
HIV_1.B.JP.9664.AB221125.2004 (bas...
HIV_1.B.JP.9669.AB286956.2004 (bas...
HIV_1.B.CN.9010.AY180905.2001 (ba...
HIV_1.B.CN.9641.DQ990880.2005 (b...
HIV_1.B.HK.8962.FJ460500.2006 (bas...
HIV_1.B.KR.9289.JQ316133.2004 (bas...
HIV_1.B.CN.8980.JX140558.2010 (bas...
HIV_1.B.CN.8781.KC899011.2009 (ba...
HIV_1.C.IN.9693.KF766540.2005 (bas...
HIV_1.C.NP.8606.KU341723.2011 (ba...
HIV_1.C.PK.9188.KU749412.2014 (ba...
HIV_1.B.KR.9005.MG461320.2017 (b...
HIV_1.B.TH.8748.MG989551.2015 (b...
HIV_1.B.AU.8967.AY818644.2004 (ba...
HIV_1.B.AU.8651.DQ676885.2003 (b...
HIV_1.B.DO.8382.EU839598.2005 (b...
HIV_1.B.TT.8369.EU839610.2001 (ba...
HIV_1.B.DO.9645.KY658702.2011 (ba...
HIV_1.B.HT.9636.MH078551.2011 (b...
HIV_1.B.BE.8990.MN486047.2016 (b...
HIV_1.B.BE.8783.MN486003.2016 (b...
HIV_1.B.BG.8834.MH746254.2009 (b...
HIV_1.B.CH.8919.JQ403045.2003 (ba...
HIV_1.B.CY.8161.FJ388955.2006 (bas...
HIV_1.B.DE.8800.JQ416161.2004 (ba...
HIV_1.B.DE.8800.JQ403055.2008 (ba...
HIV_1.B.DE.8702.KT124811.2012 (ba...
HIV_1.B.DE.9006.MK116078.2016 (b...
HIV_1.B.DE.8850.MN090933.2018 (b...
HIV_1.B.DK.8694.EF514710.2001 (ba...
HIV_1.B.ES.8771.EU786674.2006 (ba...
HIV_1.B.ES.8920.KT276262.2014 (ba...
HIV_1.B.GB.8660.HM586207.2005 (b...
HIV_1.B.GB.8298.MF109546.2014 (b...
HIV_1.B.IT.8805.DQ672623.2005 (ba...
HIV_1.F1.RO.9734.KU168274.2003 (b...
HIV_1.B.SE.8770.KP411823.2010 (ba...
HIV_1.C.SE.8749.MF373192.2013 (ba...
HIV_1.B.SI.9010.MN736706.2015 (ba...
HIV_1.A6.BY.9020.KT983615.2013 (b...
HIV_1.B.GE.8780.DQ207942.2003 (b...
HIV_1.A6.KZ.8876.EF589042.2002 (b...
HIV_1.B.RU.8960.AY819715.2004 (ba...
HIV_1.B.RU.9165.JX500708.2011 (bas...
HIV_1.A6.RU.8895.MH330370.2013 (...
HIV_1.A6.RU.9073.MN703135.2014 (...
HIV_1.A6.UA.8812.DQ823365.2001 (...
HIV_1.A6.UZ.8819.AY829208.2002 (b...
HIV_1.B.YE.8834.AY795905.2002 (ba...
HIV_1.B.US.8926.FJ469721.2007 (bas...
HIV_1.U.CA.9788.HM215251.2001 (b...
HIV_1.B.US.9073.JF320081.2006 (bas...
HIV_1.B.US.8994.JX140657.2009 (bas...
HIV_1.B.US.9003.KC473829.2010 (ba...
HIV_1.B.US.8967.KC473834.2011 (ba...
HIV_1.B.US.8648.KT124756.2012 (ba...
HIV_1.B.US.8911.KX505396.2014 (ba...
HIV_1.B.US.8970.KX505457.2015 (ba...
HIV_1.B.US.8954.KX505536.2016 (ba...
HIV_1.B.US.8969.MG171206.2013 (b...
HIV_1.B.US.9066.MH060884.2005 (b...
HIV_1.B.US.9060.MN090256.2018 (b...
HIV_1.B.BR.8967.FJ195086.2004 (bas...
HIV_1.BF1.CL.8927.AY536234.2001 (...
HIV_1.B.CO.8696.AY561239.2001 (ba...
HIV_1.B.CO.8787.AY561236.2001 (ba...
HIV_1.B.PE.9059.JF320245.2006 (bas...
HIV_1.B.PE.9043.MH078554.2016 (b...
HIV_1.BF1.PY.8802.JN251905.2003 (b...
HIV_1.BF1.PY.8774.JN251898.2002 (b...
HIV_1.BC.UY.8919.JN235957.2002 (b...

◂**Fig. 2** The figure represents the whole-genome alignment of HIV-1 genomes via the progressive MAUVE algorithm. Headers for each sequence on the left (header format: organism, subtype, country code, sequence length, accession no., and year separated by periods), and identity percentage on top (red for low matches and green for high matches)

the presence of counterions and salt concentration in the solvent medium.

Binding free energy calculation via MM/PBSA

The binding free energy (ΔGbind, Gibbs free energy) between the lead compounds and R.T. enzyme was calculated from the last 10 ns (stable RMSD interval) for each simulation system consisting of 1001 snapshots using the MM/PBSA single trajectory protocol. The formula in equation (2) was followed to calculate the energy terms for each of the R.T., lead compound, and R.T.-lead complex using the CaFE plugin (v1.0) of the Maestro program [46], and finally, equation (4) was used to calculate the ΔGbind [48–50]. The equation derivation and approximations have been provided in Supplementary Data 5.

Analysis of calculated physicochemical and ADMET properties

The top lead compounds were submitted to the SwissADME webserver of the Swiss Institute of Bioinformatics to calculate their physicochemical properties and drug-likeness [51]. The lead compounds' drug-likeness were evaluated based on 5 filters: Lipinski's rule of 5, Ghose filter, Veber filter, Egan filter, and Muegge filter [52–56]. For the ADMET analysis, admetSAR (which is also used by DrugBank to evaluate drugs) and ADMETlab (v2.0) webserver were collectively used to analyse each lead compound [57, 58].

Results and discussion

Phylogenetic tree

Supplementary Data 6 shows the phylogenetic tree of the HIV-1 sequences, which displays two main clusters and sub-clusters that vary by the origins of the subtypes. This data indicates that HIV-1 variants have geographic genetic diversity, which helps to identify the least mutated regions in the HIV-1 genome, the aim of our article.

Whole-genome alignment and BLASTx

Figure 2 shows the results from the whole-genome alignment using the progressive MAUVE algorithm. The alignment clusters each locally collinear block (L.C.B.) and shows each clustered L.C.B. as a continuous black bar. A region of about 2.4 kb within the 2.8–5.3 kb range from the consensus sequence is highly conserved, with almost no gaps in most sequences. The BLASTx results from the consensus sequence showed that the highly conserved region belongs to the R.T domain of the HIV-1 pol gene (NCBI GenBank: QMX87928.1), which suggests that the functional proteins from this domain are promising targets for developing therapeutics [59].

Molecular docking-based virtual screening

The virtual screening of the ligand dataset against the HIV-1 R.T. enzyme nominated 7 compounds with significantly high affinities (≥ -8.5 kcal/mol), which is an essential criterion for drug discovery as it indicates high binding strength and inhibition potential of the ligands against the target protein [60–62]. Among them, only 4 compounds successfully achieved the same affinity in 3 subsequent runs, hence only these 4 compounds [(1) 4-hydroxy-3-[5-[5-(2-hydroxyphenyl)-1H-pyrazol-4-yl]-4,5-dihydro-1H-pyrazol-3-yl]chromen-2-one ($C_{21}H_{16}N_4O_4$), (2) Artoindonesianin P ($C_{20}H_{16}O_7$), (3) 12a-Hydroxydolineone ($C_{19}H_{12}O_7$), (4) Phomoarcherin B ($C_{23}H_{28}O_5$)] were selected and further analysed, the top docking poses' for these 4 compounds are shown in Fig. 3A, a summary of each compound along with their chemical structures is given in Table 1. These compounds showed specific and tight interactions with the target protein and had stable conformations in the solution. They also blocked the catalytic centre of the target protein and stopped viral replication. The top 7 molecules with the highest affinity scores were screened again with the same configuration but with an exhaustiveness of 256, compounds that successfully reproduced their scores in the same pose were retained for further analysis.

According to the docking results, compound 1 forms hydrophobic interactions with the hydrophobic amino acids A445, Q500 and Y501 in the binding site of the target protein. In addition, compound 1 forms hydrogen bonds with S499, Q500 and Y501. These hydrogen bonds ensure the compatibility of the drug candidate with the target protein and increase the binding specificity. Compound 1 also forms ionic interactions with two Mn^{2+} cations. These ionic interactions allow the drug candidate to block the catalytic centre of the target protein and enhance its inhibition potential. Compound 2 forms a hydrophobic interaction with the hydrophobic amino acid Q475 in the binding site of the target protein. This interaction helps the drug candidate to settle into the binding site. In addition, compound 2 forms hydrogen bonds with R448, Q500 and Y501. These hydrogen bonds ensure the compatibility of

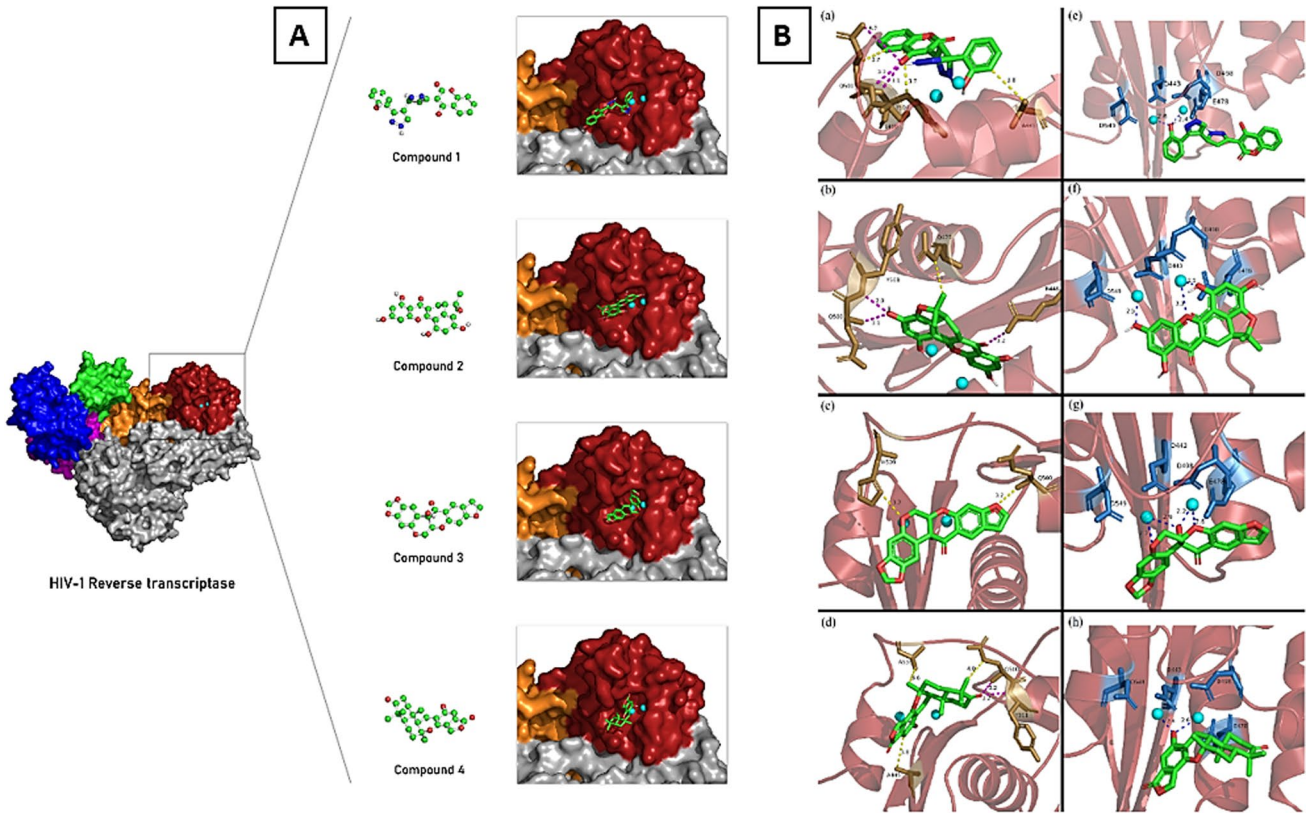


Fig. 3 **A** Docking poses for the top 4 hit compounds with the highest scores against the HIV-1 R.T. enzyme. Surface representation on the left, RNase H domain in firebrick red, fingers subdomain in blue, thumb subdomain in green, palm subdomain in magenta, connection subdomain in orange, cofactor Mn^{2+} in cyan beads, chain B in grey, and lead compounds with ball and stick representation in light green. **B** Predicted interactions of each lead compound within the RNase H active site of the HIV-1 R.T. enzyme from their respective top binding pose. All the potential hydrophobic interactions (yellow dashes), hydrogen bonds (magenta dashes), and ionic interactions (dark blue dashes) within the RNase H active site for (a & e) compound **1**, (b and f) compound **2**, (c and g) compound **3**, and (d and h) compound **4** are visualised along with their distances. All lead compounds are shown as green sticks with oxygen atoms coloured pink and nitrogen atoms blue. RNase H domain with cartoon representation (firebrick red, semi-transparent), all residues interacting with their respective lead compound are represented with sticks emerging from the protein backbone in dark wild willow colour (a-d), the cofactor Mn^{2+} is shown as cyan beads, and the catalytic site residues holding the Mn^{2+} cations (the DEDD motif) shown as sky blue sticks emerging from the protein backbone (e–h), all measurements are in Å unit

the drug candidate with the target protein and increase the binding specificity. Compound 2 also forms ionic interactions with two Mn^{2+} cations. These ionic interactions allow the drug candidate to block the catalytic centre of the target protein and enhance its inhibition potential.

Compound 3 forms hydrophobic interactions with the hydrophobic amino acids Q500 and H539 in the binding site of the target protein. In addition, compound 3 forms two ionic interactions with the two Mn^{2+} cations. These ionic interactions allow the drug candidate to block the catalytic centre of the target protein and enhance its inhibition potential. Compound 4 forms hydrophobic interactions with the hydrophobic amino acids A445, Q500 and A538 in the binding site of the target protein. Compound 4 forms hydrogen bonds with Q500 and Y501. These hydrogen bonds ensure the compatibility of the drug candidate with the target protein and increase the binding specificity. Compound 4 also forms ionic interactions with two Mn^{2+} cations. These ionic interactions allow the drug candidate to block the catalytic centre of the target protein and enhance its inhibition potential (Table 2).

Protein–ligand interaction profiling

The interaction between the R.T.'s RNase H catalytic site with the divalent cation and the top 4 potential lead compounds (as shown in Fig. 3) was closely analysed and visualised, Fig. 3B a–d shows all the potential hydrophobic interactions (with yellow dashes) and hydrogen bonds (with magenta dashes) between each residue and lead compound within the RNase H active site, the interacting residues from the R.T. backbone are further expanded (stick

Table 2 A summary of all interactions between the RNase H active site and the respective lead compounds is visualised in Fig. 3B

Interacting compound	Interacting residue	Distance (Å)	Interaction type
Compound **1**	A445	3.80	Hydrophobic
	Q500	3.69	Hydrophobic
	Y501	3.72	Hydrophobic
	S499	3.59	Hydrogen bond
	Q500	2.58	Hydrogen bond
	Y501	2.28	Hydrogen bond
	Mn^{2+}	2.40	Ionic
	Mn^{2+}	2.60	Ionic
Compound **2**	Q475	3.62	Hydrophobic
	R448	2.64	Hydrogen bond
	Q500	2.47	Hydrogen bond
	Y501	2.03	Hydrogen bond
	Mn^{2+}	2.50, 3.30*	Ionic
	Mn^{2+}	2.50	Ionic
Compound **3**	Q500	3.23	Hydrophobic
	H539	3.22	Hydrophobic
	Mn^{2+}	2.21, 2.64*	Ionic
	Mn^{2+}	2.73, 2.91*	Ionic
Compound **4**	A445	3.85	Hydrophobic
	Q500	3.96	Hydrophobic
	A538	3.66	Hydrophobic
	Q500	2.50	Hydrogen bond
	Y501	2.25	Hydrogen bond
	Mn^{2+}	2.41	Ionic
	Mn^{2+}	2.60	Ionic

*Lead compounds making more than one interaction with the same Mn^{2+} cation have their distances mentioned within the same cell separated by a comma

representations in dark wild willow) to visualise the interacting atoms. The right columns in Fig. 3B e–h show all the potential interactions between the lead compounds and the cofactor Mn^{2+} cations (cyan beads, dark blue dashes), the residues D443, E478, D498, and D549 (DEDD motif) interacting with the cofactor cations are also expanded to visualise the proximity (sky blue sticks) of the interactions, a summary table of all interactions between the lead compounds within the RNase H active site has been listed in Table 2 along with their distances.

Molecular dynamics simulation

The stability of the interaction between the four compounds and HIV-1 R.T. with the Inhibitor beta-Thujaplicinol Bound at the RNase H Active Site was systematically investigated by simulation of 100ns molecular dynamics in Fig. 4. The RMSD values allowed us to determine if the simulation had stabilised and if there were any significant conformational changes. We found that the RMSD values of the protein and ligand stabilised after about 20 ns of simulation, indicating that the system has reached equilibrium. We found that some residues had higher RMSF values than others, suggesting that certain protein regions are more resilient. These highly flexible regions could play a role in ligand binding to the protein. When we traced the ligand contacts, we found that the ligands interacted with different residues in the protein, and the specific interactions varied depending on the ligand. When we examined the secondary structure of the protein throughout the simulation, we found that its secondary structure remained constant throughout the simulation without any significant change in the distribution of the secondary structural elements. Finally, in the results of the behaviour of counterions on the system, we found that counterions were present in the solvent medium, and their concentrations remained relatively constant throughout the simulation. Our findings show that ligands interact with different residues in the protein, and the specific interactions vary depending on the ligand. We also found that certain protein regions are more flexible than others. As shown in Fig. 5, during the molecular dynamics simulation of the four complexes, the Structure of HIV-1 R.T. with the Inhibitor beta-Thujaplicinol Bound at the RNase H Active Site showed a stable trend after 20 ns, and for compound 4-protein, the fluctuation of the RMSD value was significantly higher than for the other three compound groups. As seen in Fig. 6, higher RMSF values in compound 4 mean that the ligand moves more in that region. In this way, it was revealed that the interactions of compound 4 on the protein were higher (Fig. 7).

Binding free energy calculation via MM/PBSA

Using the single trajectory approach for MM/PBSA calculation, the 8 energy terms (ΔE_elec, ΔE_vdw, ΔG_PB, ΔG_SA, ΔG_gas, ΔG_sol, ΔG_pol, ΔG_npol) were calculated for the R.T. enzyme, lead compound, and R.T.-lead complex separately from each production simulation, and their total sum was used in equation (4) to calculate the binding free energy ΔG_bind/mmpbsa. The sums of each energy term are provided in Table 3**,** along with their standard deviations. The values of ΔG_mmpbsa indicate the spontaneity of the interaction between R.T. and lead compounds (i.e. more negative = more spontaneous). Detailed values for each energy term for the protein, ligand, and complex are separately provided in Supplementary Data 7, Supplementary Data 15. Table 3 shows the binding free energies and their components of the drug candidates with the target protein. The binding free energy consists of the sum of the electrostatic and van der Waals energies in the gas phase and the polar and apolar solvation energies in the solution phase. According to the table, compound 1 is the drug candidate with the lowest binding free energy (− 41.84 ± 3.90 kcal/

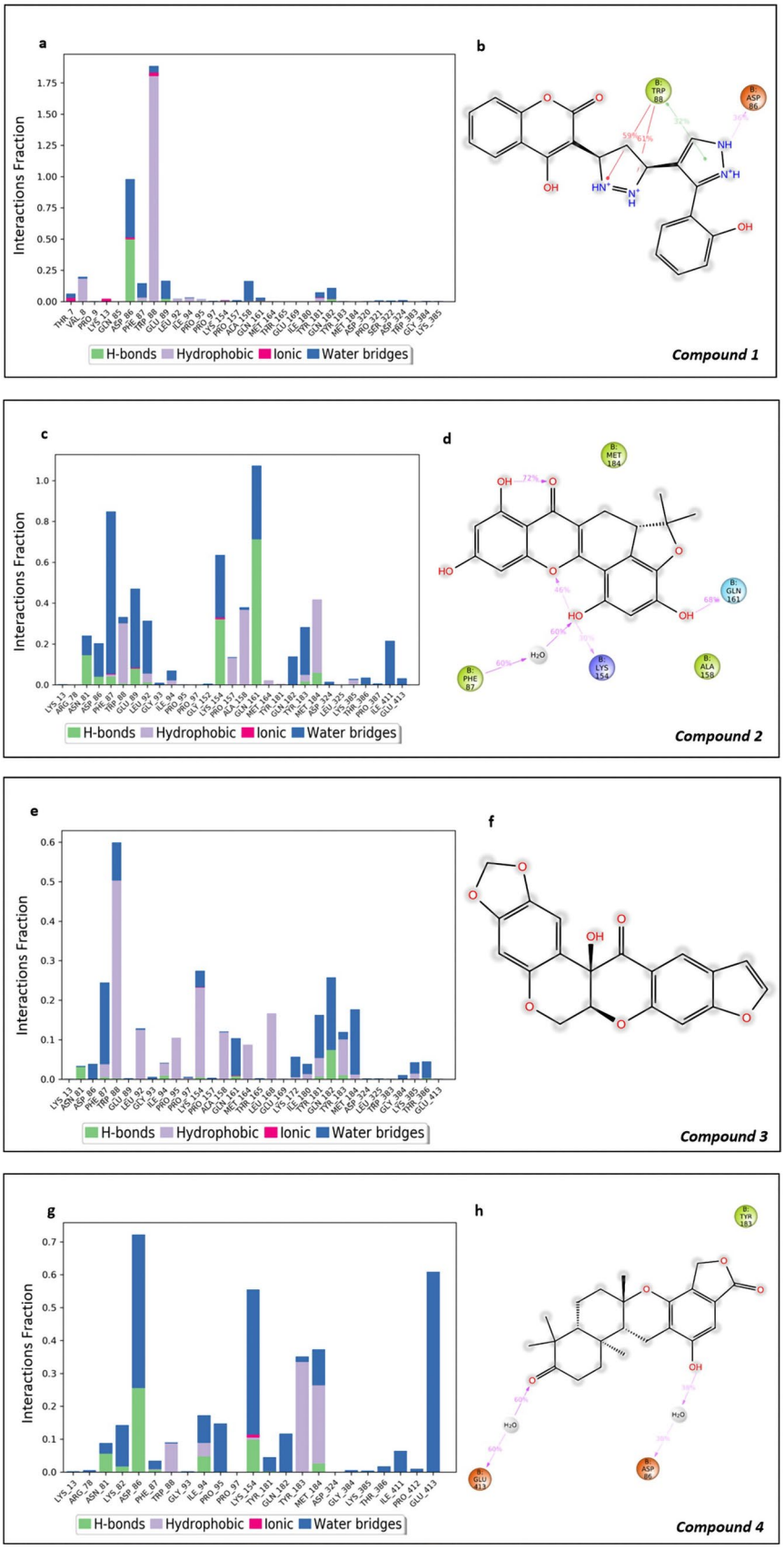


Fig. 4 **a**, **c**, **e**, and **g** Histogram of protein–ligand contact along the orbit. **b**, **d**, **f**, and **h** The structural protein–ligand complex system is produced by molecular dynamics simulation of the 2D interaction mode. The green, pink, and dark green lines represent hydrogen bonds, salt bridges, and π-cation interaction, respectively. **a** and **b** represent compound 1, **c** and **d** represent compound 2, **e** and **f** represent compound 3, and **g** and **h** represent compound 4

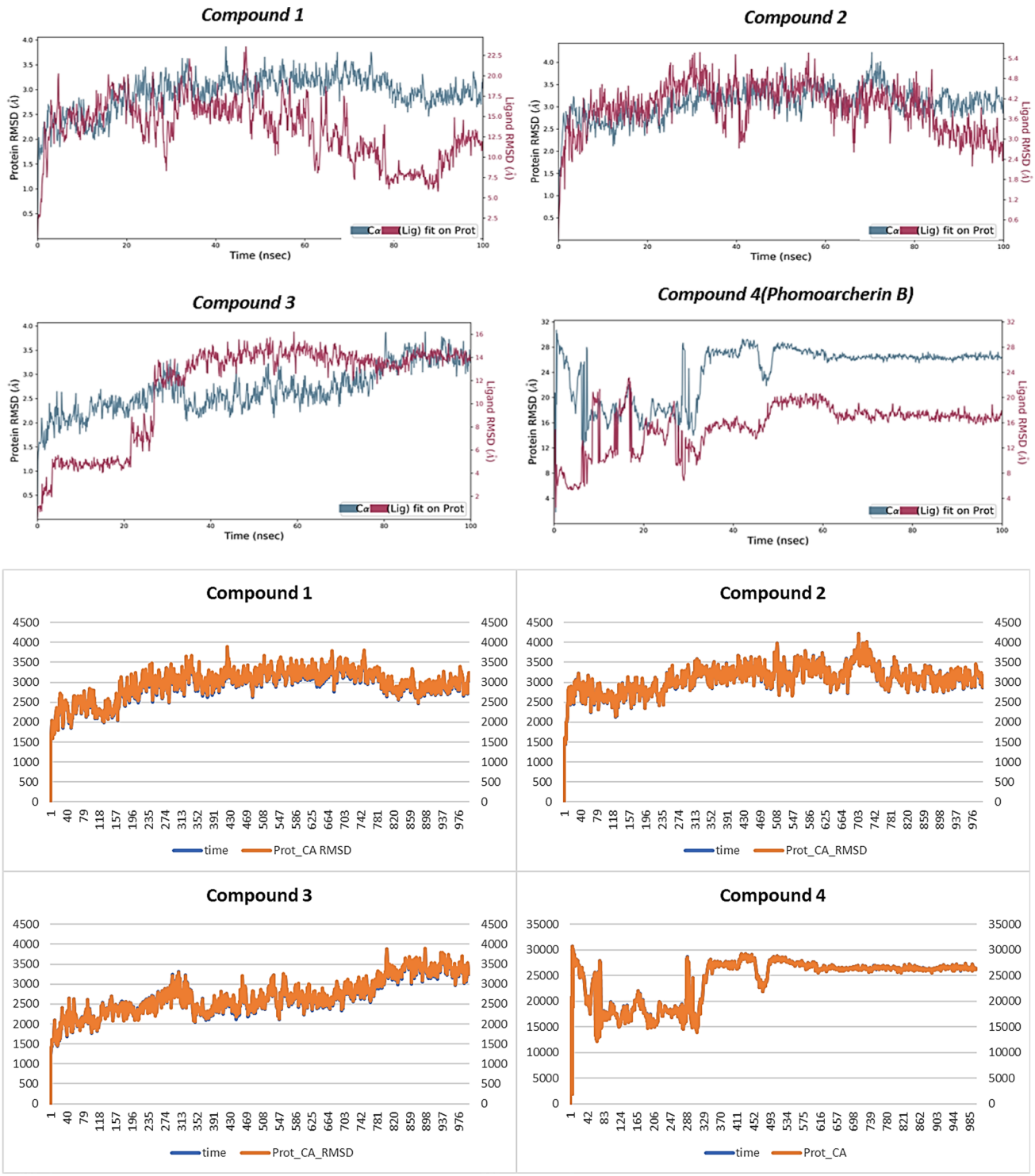
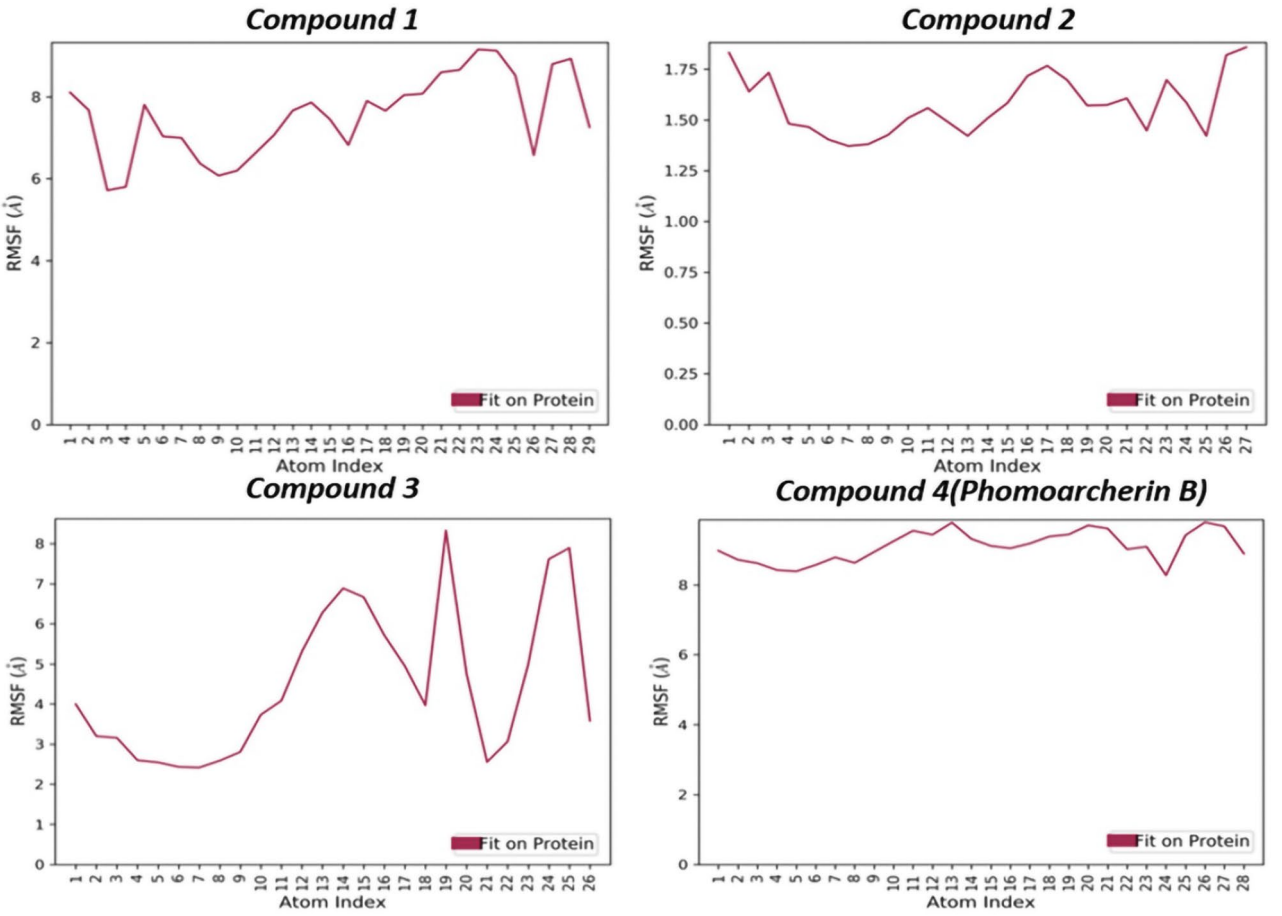
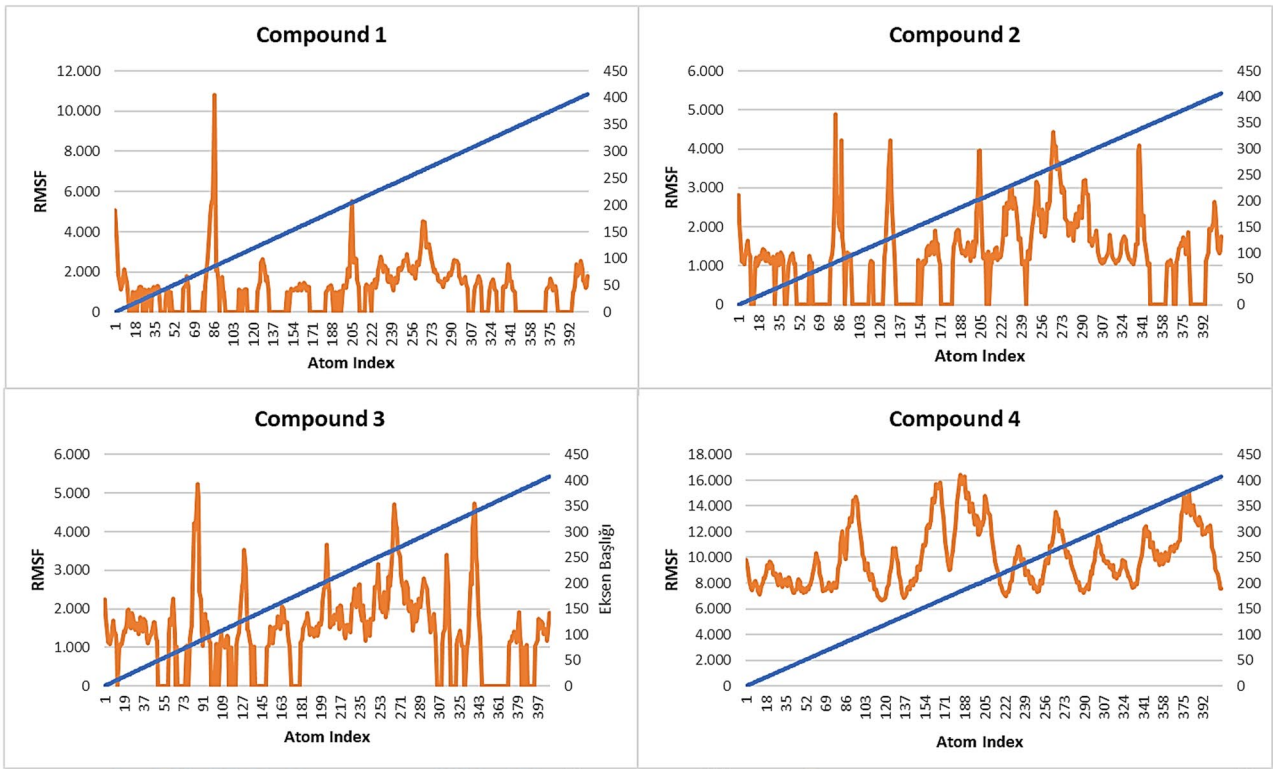


Fig. 5 R.T. backbone (bb) RMSD plot in reference to the first frame throughout the 30 ns production simulation for compound 1, compound 2, compound 3, and compound 4. The cyan line represents the motion of the lead compound in each system (similarly, less variation along the Y-axis implies less deviation from its initial docked position). R.T. backbone RMS (red line) shows the movement of the R.T. backbone during the simulation


Compound 1
RMSF (Å)
Atom Index
Fit on Protein
Compound 2
RMSF (Å)
Atom Index
Fit on Protein
Compound 3
RMSF (Å)
Atom Index
Fit on Protein
Compound 4(Phomoarcherin B)
RMSF (Å)
Atom Index
Fit on Protein


Compound 1
RMSF
Atom Index
Compound 2
RMSF
Atom Index
Compound 3
RMSF
Eksen Başlığı
Atom Index
Compound 4
RMSF
Atom Index

◂**Fig. 6** RMSF plot for each residue of the R.T. backbone in reference to the first frame throughout the 100 ns production simulation for compound 1, compound 2, compound 3, and compound 4

mol). The low binding free energy of compound 1 results from the high electrostatic and van der Waals energies in the gas phase and the low polar and apolar solvation energies in the solution phase. These energy components indicate that compound 1 interacts specifically and tightly with the target protein and has a stable solution conformation. Compounds 2, 3 and 4 have higher binding free energies (-29.20 ± 4.80, -42.88 ± 3.21 and -35.58 ± 4.84 kcal/mol). This indicates that these drug candidates bind more weakly and have lower inhibition potential with the target protein. These drug candidates' high binding free energies result from the low electrostatic and van der Waals energies in the gas phase and the high polar and apolar solvation energies in the solution phase. These energy components indicate that these drug candidates do not interact specifically and tightly with the target protein and have unstable conformations in the solution.

Analysis of calculated physicochemical and ADMET properties

The physicochemical properties of the 4 lead compounds are listed in Table 4, along with their drug-likeness results. All 4 lead compounds passed the Lipinski rule of 5, Ghose filters, Veber filter, Egan filter, and Muegge filter without any violations. The ADMET profiles of each lead compound are also summarised in Tables 4 and 5 based on the results from admetSAR and ADMETlab.

Additional docking studies

Additional docking studies are performed to validate the results obtained from the molecular dynamics simulations and to determine the most favourable conformations of the drug candidates against the target protein. These studies help us better understand the drug candidates' binding sites, binding modes, and binding strengths. Moreover, these studies show how the drug candidates affect the parameters such as displacement, orientation and angle of the atoms in the critical regions with the target protein. These parameters help us to determine how the drug candidates contribute to the binding free energy and inhibition potential. FIV also exhibited a similar DEDD motif (PDB: 5OVN) and was investigated using the same docking approach. Similar binding poses with an affinity of − 9.0 kcal/mol were observed. Detailed results are included in Supplementary Data 9.

Furthermore, docking of Phomoarcherin B with RNase H of Bacteriophage T4 (PDB: 1TFR) and monomeric R.T. of MLV (PDB: 4MH8) produced affinities of − 8.1 kcal/mol and − 8.3 kcal/mol, respectively, further indicating the potency of Phomoarcherin B as a potential antiviral RNase H candidate. Detailed log files of the docking experiments are in Supplementary Datas 10 and 11, respectively. The highest affinity was − 5.08 kcal/mol with the docking analyses using local computational analysis of the transmembrane domain of HIV-1 gp41 (PDB: 5JYN) and Phomoarcherin B. These results were also confirmed using the SwissDock server; the highest value was − 6.36 kcal/mol. The cluster with the most elements contributing to the affinity was considered. Detailed log files of the docking experiments are in Supplementary Data 12.

H.I.V. infections have been threatening lives ever since their emergence. Although modern antiviral and HAART therapies provide some relief and support for patients, they also have disadvantages and limitations. Therefore, this study aimed to perform an extensive computational analysis to discover and evaluate potent novel inhibitors of HIV-1 replication within the host. This study aimed to discover and assess potent novel inhibitors of HIV-1 replication within the host. These inhibitors target the R.T. domain of the HIV-1 pol gene, which is the least mutated and most consistent target. This way, we aimed to provide therapeutic options for the patients while accelerating the drug development processes by providing potential candidate compounds. HIV-1's mutation rate has been reported to be between 10^{-4} and 10^{-2} mutants/clones, producing 109 virions/day in an infected individual. This shows that the virus mutates efficiently to develop resistance and evade the immune system. However, not all these mutants are expected to survive and replicate, as mutations in some genes can be lethal. This also shows us the importance of identifying regions with low mutation rates in the HIV-1 genome. The comparative genomics approach aims to answer the reasons behind the characteristic phenotypes of closely related species or even different variants of the same species by comparing their genotype (genome). This method has also been widely used to discover resistance genes in several bacterial genomes. We also used this method to analyse the genetic diversity among HIV-1 variants and identify the least mutated and most conserved regions in the HIV-1 pol gene [63, 64].

In this study, we performed whole-genome alignment with MAUVE to identify the regions within the HIV-1 genome that show the highest level of consensus among all the selected sequences from the Los Alamos database. An example of the whole-genome alignment is given in Fig. 2. We found that a genomic fragment of around 2.4 kb within the 2.8–5.3 kb range from the consensus sequence was the most conserved region. This region belongs to the R.T. domain of the HIV-1 pol gene, which encodes three essential functional proteins of HIV-1: the viral R.T., integrase, and late-phase protease. These proteins are major targets for drug discovery. The F.D.A. has already approved NNRTI

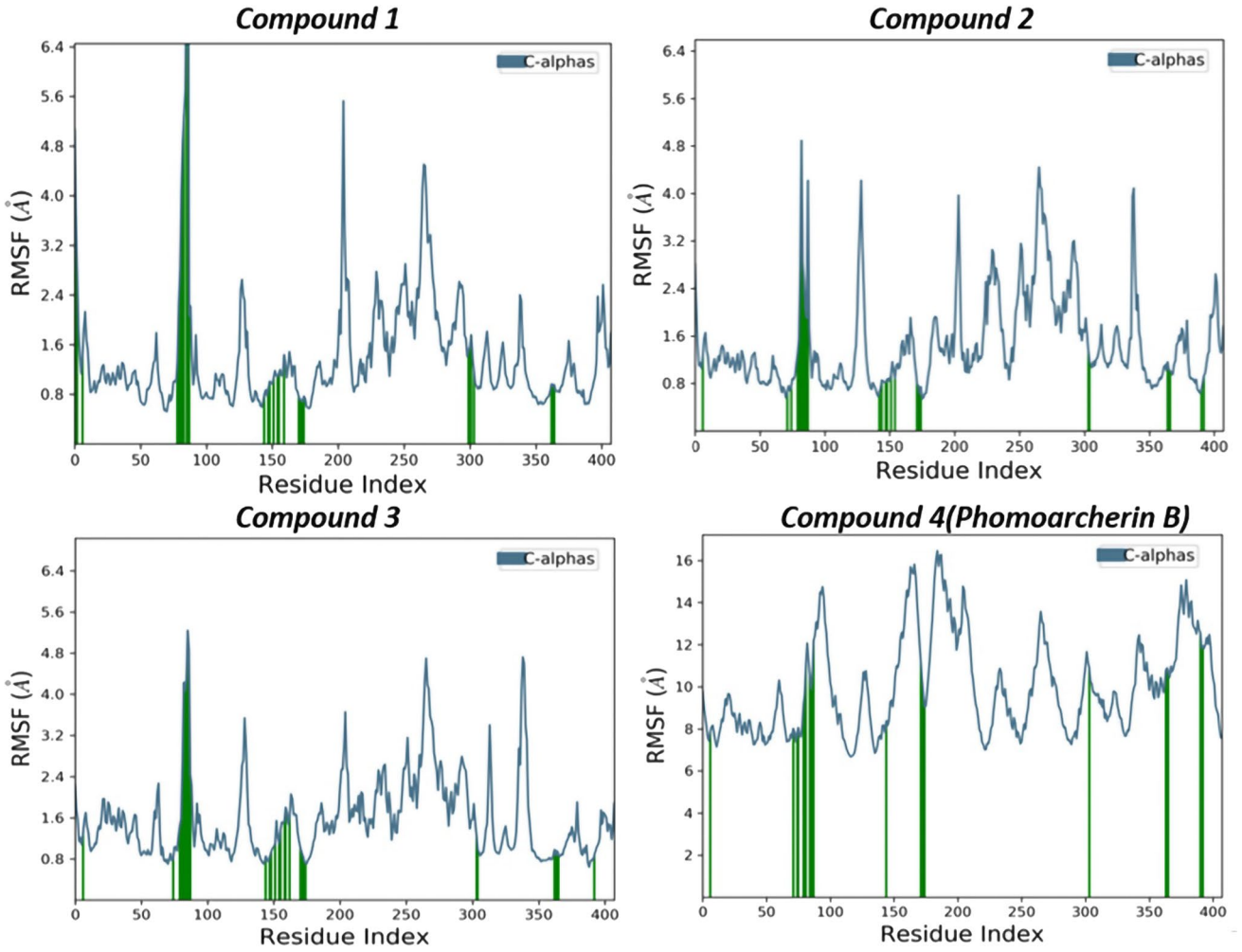


Fig. 7 Protein RMSF plot showing local changes useful for characterising local changes in the protein chain. Protein residues that interact with the ligand are marked with vertical bars in green

Table 3 MM/PBSA energy terms for 〖ΔG〗_complex were calculated for each R.T.-lead pair of HIV-1, and energies were calculated from the last 10 ns of the production trajectory using the single trajectory approach

Compound	ΔE_{elec}	ΔE_{vdw}	$\Delta G_{P.B.}$	$\Delta G_{S.A.}$	ΔG_{Gas}	ΔG_{sol}	ΔG_{pol}	ΔG_{npol}	ΔG_{mmpbsa}
1	− 28.73 ± 3.76	− 49.67 ± 3.40	41.65 ± 2.85	− 5.10 ± 0.12	− 78.40 ± 5.07	36.56 ± 2.81	12.92 ± 4.48	− 54.77 ± 3.44	− 41.84 ± 3.90
2	− 3.68 ± 3.15	− 50.89 ± 2.67	30.76 ± 3.84	− 5.38 ± 0.10	− 54.58 ± 3.97	25.38 ± 3.83	27.08 ± 4.11	− 56.28 ± 2.64	− 29.20 ± 4.80
3	− 2.27 ± 1.23	− 48.92 ± 2.66	13.21 ± 1.20	− 4.90 ± 0.11	− 51.20 ± 3.00	8.32 ± 1.18	10.94 ± 1.52	− 53.818 ± 2.70	− 42.88 ± 3.21
4	− 2.83 ± 1.95	− 54.86 ± 2.56	27.58 ± 3.64	− 5.48 ± 0.09	− 57.69 ± 3.02	22.11 ± 3.62	24.76 ± 4.22	− 60.33 ± 2.54	− 35.58 ± 4.84

[ΔE_elec] electrostatic energy difference, *[ΔE_vdw]* Van der Waals energy difference, *[ΔG_P.B.]* Poisson-Boltzmann solvation-free energy difference. *[ΔG_SA]* surface area free energy difference, *[ΔG_Gas]* free energy difference in gas phase, *[ΔG_sol]* free energy difference in solution. *[ΔG_pol]* polar solvation-free energy difference, *[ΔG_npol]* nonpolar solvation-free energy difference, *[ΔG_mmpbsa]* binding free energy difference calculated by MMPBSA method

and NRTI with inhibitory functions on the D.N.A. polymerase activity of R.T. Therefore, and our study focused on discovering and analysing potential R.T. RNase H inhibitors, which are critical for viral replication due to their polymerase activity [65–68]. We performed a virtual screening of the ligand dataset against the HIV-1 R.T. enzyme and identified

Table 4 Physicochemical and drug-likeness properties of the lead compounds based on the SwissADME results

Physicochemical properties	Compound **1**	Compound **2**	Compound **3**	Compound **4**
Molecular weight (g/mol)	388.38	368.34	352.29	384.47
No. heavy atoms	29	27	26	28
No. aromatic heavy atoms	21	16	15	6
No. rotatable bonds	3	0	0	0
No. H-bond acceptors[c]	6	7	7	5
No. H-bond donors[c]	4	4	1	1
Log S (ESOL)	− 4.15	− 4.31	− 3.92	− 4.68
Solubility (mg/mL)	2.72e−02	1.82e−02	4.24e−02	8.12e−03
Solubility class[a]	Moderately soluble	Moderately soluble	Soluble	Moderately soluble
Lipophilicity (Log $P_{o/w}$)[b]	2.30	2.42	2.15	3.67
Lipinksi rule of 5[d]	Pass (0)	Pass (0)	Pass (0)	Pass (0)
Ghose filters[d]	Pass (0)	Pass (0)	Pass (0)	Pass (0)
Veber filters[d]	Pass (0)	Pass (0)	Pass (0)	Pass (0)
Egan filters[d]	Pass (0)	Pass (0)	Pass (0)	Pass (0)
Muegge filters	Pass	Pass	Pass	Pass

[a]Based on the Log S (ESOL) scale, insoluble < − 10 < poor < − 6 < moderate < − 4 < soluble < − 2 < very < 0 < highly soluble

[b]The values are average of iLOGP, XLOGP3, WLOGP, MLOGP, and SILICOS-IT

[c]Based on the BOILED-Egg model [59]

[d]Numbers within parentheses indicate the number of violations of the respective filter/rule

four compounds with high affinities (≥ − 8.5 kcal/mol). An example of the four compounds with the highest affinity is given in Fig. 3A. These compounds interacted specifically and tightly with the target protein and had stable conformations in the solution. They also blocked the catalytic centre of the target protein by forming ionic interactions with two Mn^{2+} cations and stopped viral replication. The details of the interactions for each compound are given in Table 2. We performed molecular dynamics simulation and binding free energy calculation to evaluate further the stability and spontaneity of the interactions between the four compounds and the target protein. We found that compound 4 (Phomoarcherin B) was the best-performing lead compound, with a docking score of − 8.5 kcal/mol, less than 1 Å deviation from its initial docked pose throughout the 30 ns molecular dynamic simulation, binding free energy of ≈ − 35.58 ± 4.84 kcal/mol and near-perfect scores on each ADMET profile. The backbone, RMSD of the R.T. enzyme, also reached a plateau after the 20 ns time-lapse of the production simulation [69, 70].

Phomoarcherin B is a natural compound found in the endophytic fungus Phomopsis archeri, and it was first isolated and characterised as a pentacyclic aromatic sesquiterpene by Hemtasin et al. [71]. They tested it for antimalarial activity against Plasmodium falciparum and anticancer activities against cholangiocarcinoma cell lines. Bedi et al. [72] also reported its anticancer activity, whereas no in vitro or in vivo assay was performed regarding its antiviral or RNase H inhibitory activity. As shown in Table 5, Phomoarcherin B can penetrate the blood–brain barrier, increasing its drug-candidate potential. However, further in vitro assays and clinical trials are needed to confirm its pharmaceutical potential as an HIV-1 RNase H inhibitor. We also performed additional docking studies of Phomoarcherin B with FIV and monomeric R.T. of MLV belonging to the Retroviridae family and RNase H of Bacteriophage T4 belonging to the Myoviridae family. We found that Phomoarcherin B also showed promising results with these targets, indicating its potency as a potential antiviral RNase H candidate. The details of these docking studies are provided in Supplementary Datas 9, 10, and 11.

Moreover, we performed a docking analysis of Phomoarcherin B with the transmembrane domain of HIV-1 gp41, another critical target for drug discovery. We found that Phomoarcherin B had a high affinity for this target protein as well, confirmed by docking server analysis (Supplementary Data 12) (Fig. 8a). However, we should stress that computational studies are not sufficient alone to show the inhibitory effect of our suggested candidate compound. Even though Phomoarcherin B has a higher affinity than the other three compounds over the active site, we also need some experimental studies and future works for validation (Fig. 9).

Table 5 ADMET profiles of the lead compounds as per results from admetSAR and ADMETlab

ADMET properties	Compound **1**	Compound **2**	Compound **3**	Compound **4**
		Absorption		
Gastrointestinal absorption[d]	High	High	High	High
Blood–brain barrier permeation[d]	None	None	None	Yes
		Distribution		
Plasma binding protein[a]	96.58%	98.23%	95.53%	94.06%
Fraction unbound in plasma	2.52%	3.30%	5.60%	7.13%
		Metabolism		
CYP450 2C9 substrate	Non-substrate	Non-substrate	Non-substrate	Non-substrate
CYP450 2D6 substrate	Non-substrate	Non-substrate	Non-substrate	Non-substrate
CYP450 3A4 substrate	Non-substrate	Substrate	Non-substrate	Substrate
CYP450 1A2 inhibitor	Inhibitor	Inhibitor	Non-inhibitor	Non-inhibitor
CYP450 2C9 inhibitor	Inhibitor	Inhibitor	Non-inhibitor	Non-inhibitor
CYP450 2D6 inhibitor	Non-inhibitor	Non-inhibitor	Non-inhibitor	Non-inhibitor
CYP450 2C19 inhibitor	Inhibitor	Non-inhibitor	Non-inhibitor	Non-inhibitor
CYP450 3A4 inhibitor	Inhibitor	Non-inhibitor	Non-inhibitor	Non-inhibitor
C.Y.P. inhibitory promiscuity	High	High	Low	Low
		Excretion[a]		
$T_{1/2}$ (hours)[b]	< 3 (0.349)	> 3 (0.52)	< 3 (0.113)	> 3 (0.70)
		Toxicity		
Rat acute (LD_{50}, mol/kg)	2.43	2.38	2.41	2.71
TP[c] ($pIGC_{50}$, μg/L)	0.46	1.04	0.57	1.38
Acute oral (LD_{50} mg/kg)[d]	III	III	III	III
Carcinogenicity	None	None	None	None

The SMILES notation of the lead compounds was used as the input to calculate each property (provided in Supplementary Data 8)

[a]Based on the predictions of the ADMETLab 2.0 [58]

[b]Probability of half-life being greater than 3 h is given within parentheses, below 0.5 was considered to have $T_{1/2} < 3$

[c]Tetrahymena pyriformis toxicity

[d]Class I ≤ 50 mg/kg, class II > 50 mg/kg, class III > 500 mg/kg, and class IV > 5000 mg/kg

Conclusion

In this study, we aimed to discover and evaluate potent novel inhibitors of HIV-1 replication within the host by targeting the R.T. domain of the HIV-1 pol gene, the least mutated and most consistent region in the HIV-1 genome. We performed a whole-genome alignment, virtual screening, molecular docking, molecular dynamics simulation, binding free energy calculation and ADMET prediction to identify and analyse the potential inhibitors. We found that Phomoarcherin B, a natural compound from the endophytic fungus Phomopsis archeri, was the best-performing lead compound, with high affinity, stability and spontaneity for the target protein and its cofactor Mn^{2+} cations. We also found that Phomoarcherin B can inhibit the transmembrane domain of HIV-1 gp41, which is involved in viral fusion and entry. This study provides a comprehensive computational analysis of Phomoarcherin B as a potential HIV-1 reverse transcriptase RNase H inhibitor. It is a promising target for drug design due to its essential role in viral replication and low mutation rate. However, this study has some limitations, assumptions and sources that must be considered. For example, we used only one sequence for each subtype and group of HIV-1, which may not represent the whole diversity of the virus. We also assumed that the docking poses and binding energies reflect the true interactions between the compounds and the target protein, which may not be accurate in reality.

Moreover, we used only computational methods to analyse Phomoarcherin B, which may not account for all the factors that affect its biological activity and pharmacokinetics. Therefore, we suggest that further in vitro and in vivo studies and clinical trials are needed to confirm the pharmaceutical potential of Phomoarcherin B as an HIV-1 RNase H inhibitor. This study also provides valuable insights into the molecular mechanisms and interactions of Phomoarcherin B with the target proteins and its ADMET properties. These results can be a reference for future studies and applications of Phomoarcherin B as an antiviral agent.

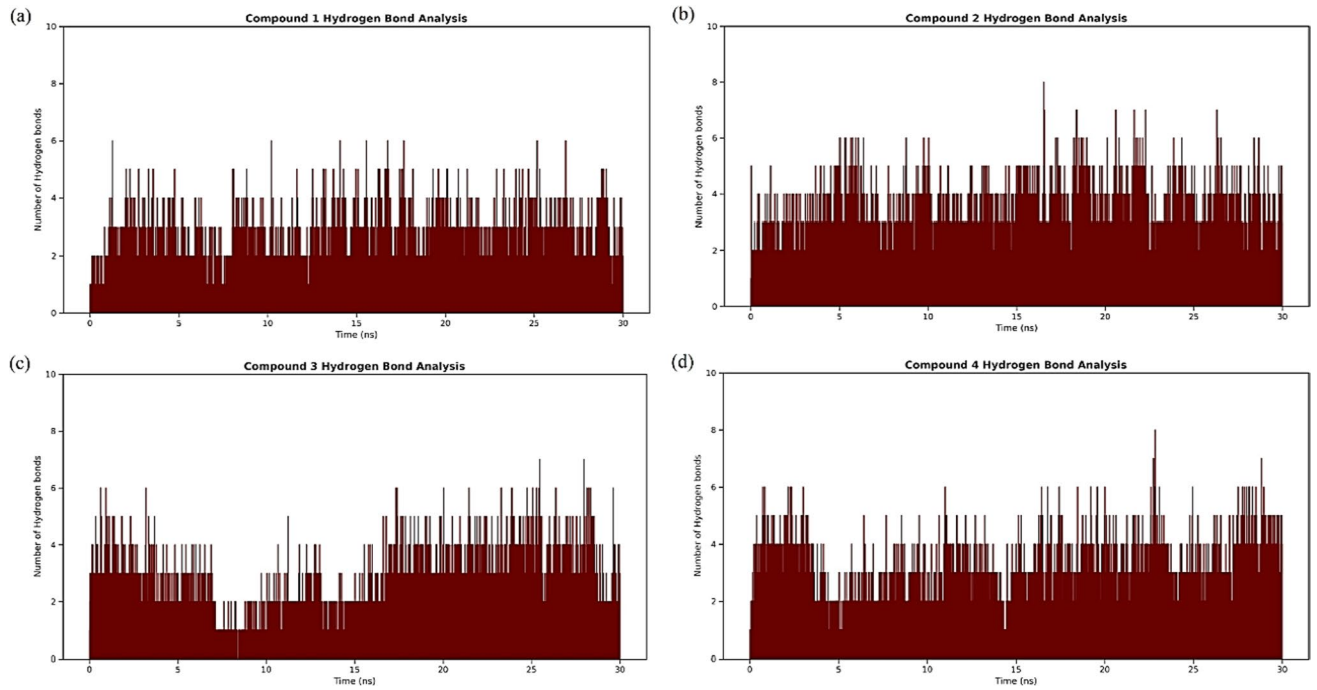


Fig. 8 Number of hydrogen bonds formed by each R.T.-lead compound pair throughout the production simulation for **a** compound **1**, **b** compound **2**, **c** compound **3**, and **d** compound **4**. The parameters used to calculate the H bonds were a donor–acceptor distance of less than 3 Å and an angle cut-off of 20° (the most guaranteed threshold for H bonds, hence low bias)

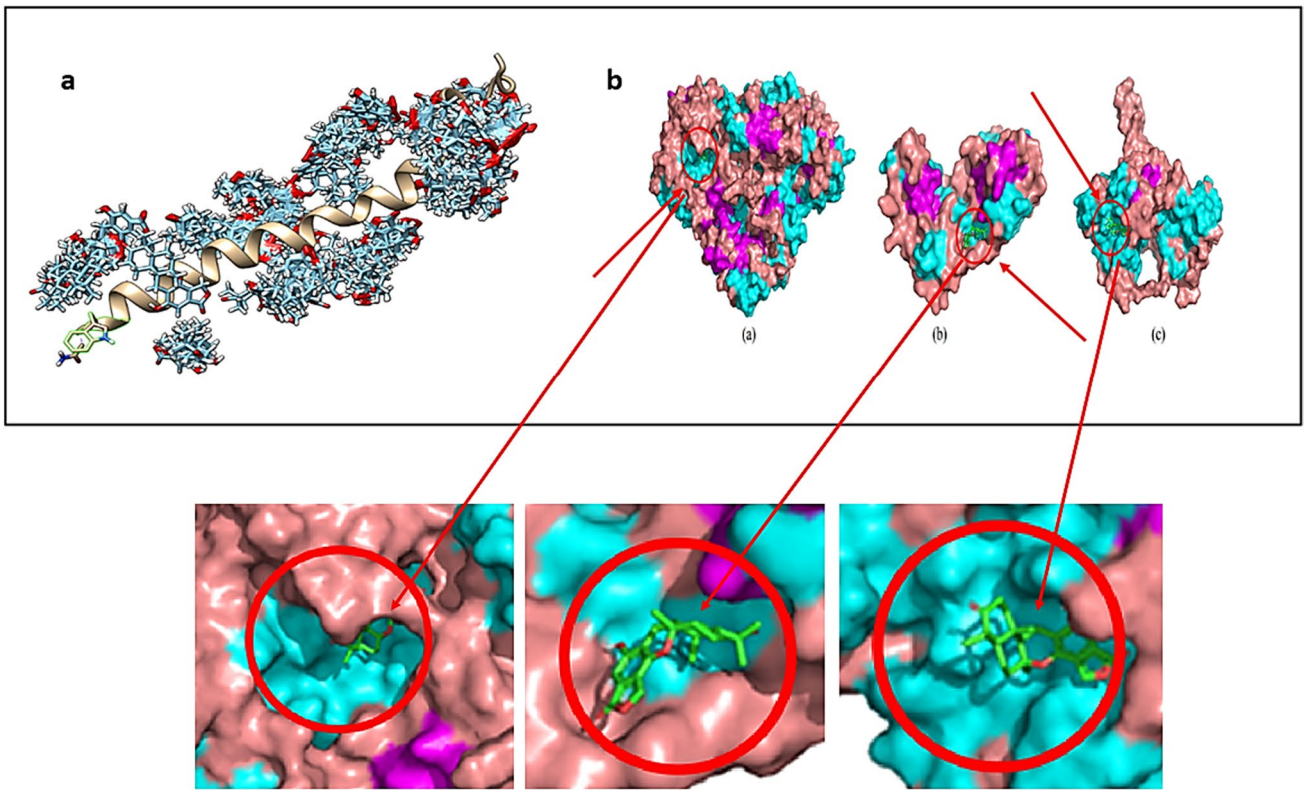


Fig. 9 **a** The whole docking pose of Phomoarcherin B (blue and red) with the transmembrane domain of HIV-1 gp41 (helix structure) considering n number of Phomoarcherin B compound bind on the target protein. **b** Top docking pose of Phomoarcherin B (red circle) with the (**a**) R.T. of FIV, (**b**) R.T. of MLV, and (**c**) RNase H of the Bacteriophage T4. Phomoarcherin B is represented as green sticks and bonds, proteins in the surface representation coloured based on their (helices in cyan, sheets in light pink, and loops in dry violet)

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11030-023-10707-6.

Author contributions NAG: formal analysis; writing—original draft preparation. KKK: conceptualisation; investigation; formal analysis; writing—review & editing—original draft preparation. ÖB: conceptualisation; investigation; formal analysis; writing—original draft preparation—writing—review & editing. BES: investigation; formal analysis. RSS: conceptualisation; formal analysis; writing—original draft preparation. All authors commented on previous versions and read and approved the final manuscript.

Data availability This work has been published as a preprint version (https://www.biorxiv.org/content/10.1101/2021.09.09.459559v1.full) and additional updates are available to the preprint version.

Declarations

Competing interests The authors declare no competing interests.

References

1. Hemelaar J (2012) The origin and diversity of the HIV-1 pandemic. Trends Mol Med 18(3):182–192. https://doi.org/10.1016/j.molmed.2011.12.001
2. WHO. H.I.V. data and statistics. https://www.who.int/teams/global-hiv-hepatitis-and-stis-programmes/hiv/strategic-information/hiv-data-and-statistics. Accessed 2023
3. UNAIDS. Global H.I.V. & AIDS statistics. https://www.unaids.org/en/resources/fact-sheet. Accessed 2023
4. Cohen MS, Hellmann N, Levy JA, DeCock K, Lange J (2008) "The spread, treatment, and prevention of HIV-1: evolution of a global pandemic. J Clin Invest 118(4):1244–1254. https://doi.org/10.1172/jci34706
5. Kirchhoff F (2013) H.I.V. Life cycle: overview, pp 1–9
6. Swanson CM, Malim MH (2008) SnapShot: HIV-1 proteins. Cell 133(4):742. https://doi.org/10.1016/j.cell.2008.05.005
7. Fanales-Belasio E, Raimondo M, Suligoi B, Buttò S (2010) H.I.V. virology and pathogenetic mechanisms of infection: a brief overview. Ann Ist Super Sanita 46(1):5–14. https://doi.org/10.4415/ann_10_01_02
8. Ruelas DS, Greene WC (2013) An integrated overview of HIV-1 latency. Cell 155(3):519–29. https://doi.org/10.1016/j.cell.2013.09.044
9. Volberding PA, Deeks SG (2010) Antiretroviral therapy and management of H.I.V. infection. Lancet 376(9734):49–62. https://doi.org/10.1016/s0140-6736(10)60676-9
10. Das D, Georgiadis MM (2004) The crystal structure of the monomeric reverse transcriptase from Moloney Murine Leukemia Virus. Structure 12(5):819–829. https://doi.org/10.1016/j.str.2004.02.032
11. Ahamad S, Islam A, Ahmad F, Dwivedi N, Hassan MI (2019) 2/3D-QSAR, molecular docking and M.D. simulation studies of FtsZ protein targeting benzimidazoles derivatives. Comput Biol Chem 78:398–413. https://doi.org/10.1016/j.compbiolchem.2018.12.017
12. Poongavanam V, Kongsted J (2013) Virtual screening models for prediction of HIV-1 RT associated RNase H inhibition. PLoS ONE 8(9):e73478. https://doi.org/10.1371/journal.pone.0073478
13. Shin Y et al (2021) Identification of aristolactam derivatives that act as inhibitors of human immunodeficiency virus type 1 infection and replication by targeting tat-mediated viral transcription. Virol Sin 36(2):254–263. https://doi.org/10.1007/s12250-020-00274-7
14. Poli G, Granchi C, Rizzolio F, Tuccinardi T (2020) Application of MM-PBSA methods in virtual screening. Molecules. https://doi.org/10.3390/molecules25081971
15. Gao Y et al (2019) In silico study of 3-hydroxypyrimidine-2,4-diones as inhibitors of H.I.V. RT-associated RNase H using molecular docking, molecular dynamics, 3D-QSAR, and pharmacophore models. New J Chem 43(43):17004–17017. https://doi.org/10.1039/C9NJ03353J
16. Panwar U, Singh SK (2018) An overview on Zika virus and the importance of computational drug discovery. J Explor Res Pharmacol 3(2):43–51. https://doi.org/10.14218/jerp.2017.00025
17. Frey KM et al (2022) Structural studies and structure activity relationships for novel computationally designed non-nucleoside inhibitors and their interactions with HIV-1 reverse transcriptase. Front Mol Biosci 9:805187. https://doi.org/10.3389/fmolb.2022.805187
18. Zhang B, D'Erasmo MP, Murelli RP, Gallicchio E (2016) Free energy-based virtual screening and optimisation of RNase H inhibitors of HIV-1 reverse transcriptase. ACS Omega 1(3):435–447. https://doi.org/10.1021/acsomega.6b00123
19. Seniya C, Yadav A, Khan GJ, Sah NK (2015) In-silico studies show potent inhibition of HIV-1 reverse transcriptase activity by a herbal drug. IEEE/ACM Trans Comput Biol Bioinform 12(6):1355–1364. https://doi.org/10.1109/tcbb.2015.2415771
20. Panwar U, Singh SK (2021) In silico virtual screening of potent inhibitor to hamper the interaction between HIV-1 integrase and LEDGF/p75 interaction using E-pharmacophore modeling, molecular docking, and dynamics simulations. Comput Biol Chem 93:107509. https://doi.org/10.1016/j.compbiolchem.2021.107509
21. Kuiken C, Korber B, Shafer RW (2003) HIV sequence databases. AIDS Rev 5(1):52–61
22. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22(22):4673–4680. https://doi.org/10.1093/nar/22.22.4673
23. Okonechnikov K, Golosova O, Fursov M (2012) Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics 28(8):1166–1167. https://doi.org/10.1093/bioinformatics/bts091
24. Darling AC, Mau B, Blattner FR, Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 14(7):1394–1403. https://doi.org/10.1101/gr.2289704
25. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinform 5(1):113. https://doi.org/10.1186/1471-2105-5-113
26. Geneious Prime 2023.1 (2023). https://www.geneious.com
27. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ (2009) Jalview Version 2–a multiple sequence alignment editor and analysis workbench. Bioinformatics 25(9):1189–1191. https://doi.org/10.1093/bioinformatics/btp033
28. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410. https://doi.org/10.1016/s0022-2836(05)80360-2
29. States DJ, Gish W (1994) Combined use of sequence similarity and codon bias for coding region identification. J Comput Biol 1(1):39–50. https://doi.org/10.1089/cmb.1994.1.39
30. Trott O, Olson AJ (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimisation, and multithreading. J Comput Chem 31(2):455–461. https://doi.org/10.1002/jcc.21334
31. Himmel DM et al (2009) Structure of HIV-1 reverse transcriptase with the inhibitor beta-Thujaplicinol bound at the RNase H active

site. Structure 17(12):1625–1635. https://doi.org/10.1016/j.str.2009.09.016
32. Berman H, Henrick K, Nakamura H (2003) Announcing the worldwide Protein Data Bank. Nat Struct Biol 10(12):980. https://doi.org/10.1038/nsb1203-980
33. Esposito F, Corona A, Tramontano E (2012) HIV-1 reverse transcriptase still remains a new drug target: structure, function, classical inhibitors, and new inhibitors with innovative mechanisms of actions. Mol Biol Int 2012:586401. https://doi.org/10.1155/2012/586401
34. Webb B, Sali A (2016) Comparative protein structure modeling using MODELLER. Curr Protoc Bioinform 54:561–5637. https://doi.org/10.1002/cpbi.3
35. Fiser A, Do RK, Sali A (2000) Modeling of loops in protein structures. Protein Sci 9(9):1753–1773. https://doi.org/10.1110/ps.9.9.1753
36. The PyMOL molecular graphics system. (015). https://pymol.org/2/
37. Morris GM et al (2009) AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. J Comput Chem 30(16):2785–2791. https://doi.org/10.1002/jcc.21256
38. Avogadro: an open-source molecular builder and visualisation tool. (2022). https://avogadro.cc/
39. Sterling T, Irwin JJ (2015) ZINC 15—ligand discovery for everyone. J Chem Inf Model 55(11):2324–2337. https://doi.org/10.1021/acs.jcim.5b00559
40. Galilee M, Alian A (2018) The structure of FIV reverse transcriptase and its implications for nonnucleoside inhibitor resistance. PLoS Pathogens 14(1):e1006849. https://doi.org/10.1371/journal.ppat.1006849
41. Bhagwat M, Meara D, Nossal NG (1997) Identification of residues of T4 RNase H required for catalysis and D.N.A. binding*. J Biol Chem 272(45):28531–28538. https://doi.org/10.1074/jbc.272.45.28531
42. Dev J et al (2016) "Structural basis for membrane anchoring of HIV-1 envelope spike. Science 353(6295):172–175. https://doi.org/10.1126/science.aaf7066
43. Salentin S, Schreiber S, Haupt VJ, Adasme MF, Schroeder M (2015) PLIP: fully automated protein-ligand interaction profiler. Nucleic Acids Res 43(W1):W443–W447. https://doi.org/10.1093/nar/gkv315
44. Ahamad S, Hassan MI, Dwivedi N (2018) Designing of phenol-based β−carbonic anhydrase1 inhibitors through QSAR, molecular docking, and M.D. simulation approach. 3 Biotech 8(5):256. https://doi.org/10.1007/s13205-018-1278-z
45. Ahamad S, Kanipakam H, Birla S, Ali MS, Gupta D (2021) Screening Malaria-box compounds to identify potential inhibitors against SARS-CoV-2 Mpro, using molecular docking and dynamics simulation studies. Eur J Pharmacol 890:173664. https://doi.org/10.1016/j.ejphar.2020.173664
46. Schrödinger Release 2022-3: Maestro (2021). New York. https://www.schrodinger.com/products/maestro
47. Pelemis S, Tomić B, Armaković S, Armaković S (2019) Reactive properties of ephedrine molecule—a frequently used stimulant
48. Liu H, Hou T (2016) CaFE: a tool for binding affinity prediction using end-point free energy methods. Bioinformatics 32(14):2216–2218. https://doi.org/10.1093/bioinformatics/btw215
49. Hou T, Wang J, Li Y, Wang W (2011) Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. J Chem Inf Model 51(1):69–82. https://doi.org/10.1021/ci100275a
50. Singh N, Warshel A (2010) Absolute binding free energy calculations: on the accuracy of computational scoring of protein-ligand interactions. Proteins 78(7):1705–1723. https://doi.org/10.1002/prot.22687
51. Daina A, Michielin O, Zoete V (2017) SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. Sci Rep 7(1):42717. https://doi.org/10.1038/srep42717
52. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 46(1–3):3–26. https://doi.org/10.1016/s0169-409x(00)00129-0
53. Ghose AK, Viswanadhan VN, Wendoloski JJ (1999) A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterisation of known drug databases. J Comb Chem 1(1):55–68. https://doi.org/10.1021/cc9800071
54. Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, Kopple KD (2002) Molecular properties that influence the oral bioavailability of drug candidates. J Med Chem 45(12):2615–2623. https://doi.org/10.1021/jm020017n
55. Egan WJ, Merz KM, Baldwin JJ (2000) Prediction of drug absorption using multivariate statistics. J Med Chem 43(21):3867–3877. https://doi.org/10.1021/jm000292e
56. Muegge I, Heald SL, Brittelli D (2001) Simple selection criteria for drug-like chemical matter. J Med Chem 44(12):1841–1846. https://doi.org/10.1021/jm015507e
57. Cheng F et al (2012) admetSAR: a comprehensive source and free tool for assessment of chemical ADMET properties. J Chem Inf Model 52(11):3099–3105. https://doi.org/10.1021/ci300367a
58. Xiong G et al (2021) ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties. Nucleic Acids Res 49(W1):W5-w14. https://doi.org/10.1093/nar/gkab255
59. Waterson AP (1983) Acquired immune deficiency syndrome. Br Med J 286(6367):743–746. https://doi.org/10.1136/bmj.286.6367.743
60. Wingert BM, Oerlemans R, Camacho CJ (2018) Optimal affinity ranking for automated virtual screening validated in prospective D3R grand challenges. J Comput Aided Mol Des 32(1):287–297. https://doi.org/10.1007/s10822-017-0065-y
61. Sarma H, Mattaparthi VSK (2019) Structure-based virtual screening of high-affinity ATP-competitive inhibitors against human Lemur Tyrosine Kinase-3 (LMTK3) Domain: a novel therapeutic target for breast cancer. Interdiscip Sci 11(3):527–541. https://doi.org/10.1007/s12539-018-0302-7
62. Méndez-Álvarez D et al (2022) Ligand-based virtual screening, molecular docking, and molecular dynamics of eugenol analogs as potential acetylcholinesterase inhibitors with biological activity against *Spodoptera frugiperda*. Mol Divers 26(4):2025–2037. https://doi.org/10.1007/s11030-021-10312-5
63. Fournier PE et al (2006) Comparative genomics of multidrug resistance in *Acinetobacter baumannii*. PLoS Genet 2(1):e7. https://doi.org/10.1371/journal.pgen.0020007
64. Hardison RC (2003) Comparative genomics. PLoS Biol 1(2):E58. https://doi.org/10.1371/journal.pbio.0000058
65. De Clercq E (2004) Nonnucleoside reverse transcriptase inhibitors (NNRTIs): past, present, and future. Chem Biodivers 1(1):44–64. https://doi.org/10.1002/cbdv.200490012
66. King RW, Klabe RM, Reid CD, Erickson-Viitanen SK (2002) Potency of nonnucleoside reverse transcriptase inhibitors (NNRTIs) used in combination with other human immunodeficiency virus NNRTIs, NRTIs, or protease inhibitors. Antimicrob Agents Chemother 46(6):1640–1646. https://doi.org/10.1128/aac.46.6.1640-1646.2002
67. De Clercq E (1999) Perspectives of nonnucleoside reverse transcriptase inhibitors (NNRTIs) in the therapy of HIV-1 infection.

Farmaco 54(1–2):26–45. https://doi.org/10.1016/s0014-827x(98)00103-7
68. Melikian GL et al (2014) Nonnucleoside reverse transcriptase inhibitor (NNRTI) cross-resistance: implications for preclinical evaluation of novel NNRTIs and clinical genotypic resistance testing. J Antimicrob Chemother 69(1):12–20. https://doi.org/10.1093/jac/dkt316
69. Huang J et al (2017) CHARMM36m: an improved force field for folded and intrinsically disordered proteins. Nat Methods 14(1):71–73. https://doi.org/10.1038/nmeth.4067
70. Jo S, Kim T, Iyer VG, Im W (2008) CHARMM-GUI: a web-based graphical user interface for CHARMM. J Comput Chem 29(11):1859–1865. https://doi.org/10.1002/jcc.20945
71. Hemtasin C et al (2011) Cytotoxic pentacyclic and tetracyclic aromatic sesquiterpenes from *Phomopsis archeri*. J Nat Prod 74(4):609–613. https://doi.org/10.1021/np100632g
72. Bedi A, Adholeya A, Deshmukh KS (2018) Novel anticancer compounds from Endophytic fungi. Curr Biotechnol 7(3):168–184. https://doi.org/10.2174/2211550105666160622080354