

PHOMOARCHERIN B AS A NOVEL HIV-1 REVERSE TRANSCRIPTASE RNASE H ACTIVITY INHIBITOR; CONCLUSIONS FROM COMPREHENSIVE COMPUTATIONAL ANALYSIS

Naeem Abdul Ghafoor¹, Ömür Baysal^{1*}, Barış Ethem Süzek², Ragıp Soner Silme³

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

²Department of Computer Engineering, Faculty of Engineering, Muğla Sıtkı Koçman University, 48121 Muğla, Turkey

³Center for Research and Practice in Biotechnology and Genetic Engineering, Istanbul University, 34119 Istanbul, Turkey

*Corresponding author: omurbaysal@mu.edu.tr

Abstract

The HIV-1 and its variants have claimed more than 32.7 million lives since its emergence in 1981, while many highly/ active antiretroviral therapies are available but most of these therapeutics have long-term side effects. In this study, genomic analysis was performed on 98 HIV-1 genomes to determine the most coherent target, which could be utilized for termination of the viral replication and the reverse transcriptase enzyme. Following the identification of the target protein, the RNase H activity of the reverse transcriptase was selected as the potential target based on its low mutation rate and high conservation determined using MAUVE analysis. Afterwards, a library of around 94.000 small molecule inhibitors was investigated and virtual screening was performed against the RNase domain of the reverse transcriptase to identify potential hits. Four compounds with the best scores were considered and their interaction within the active site was analysed. Subsequently, all-atom molecular dynamics simulations and MM-PBSA was performed to validate the stability and binding free energy of the hits within the RNase H active site. In computational analyses, ADMET assays were performed on the hit compounds to analyse their drug candidacy based on their physicochemical and pharmacological properties. Phomoarcherin B, a pentacyclic aromatic sesquiterpene naturally found in the endophytic fungus *Phomopsis archeri*, known for its

anticancer properties scored the best in all the experiments and was nominated as a potential inhibitor of the HIV-1 reverse transcriptase RNase H activity.

Keywords: HIV-1, Reverse Transcriptase, Computational biology, Drug discovery

Introduction

The Human immunodeficiency virus 1 (HIV-1) remains a global public health issue ever since its first identification in 1981, its infection results in progression to Acquired Immunodeficiency Syndrome (AIDS) that leaves the immune system of the host defenceless against secondary infection¹. The World Health Organization refers to HIV-1 as a “global epidemic”², and according to the Joint United Nations Programme on HIV/AIDS (UNAIDS) global statistics around 27.2–47.8 million individuals have died due to AIDS-related illnesses from its emergence upto 2020, a total of 30.2–45.1 million individuals still live with the virus³, most of them living in sub-Saharan Africa^{3,4}.

HIV is a complex retrovirus, like other retroviruses it stores its genome as a pair of ssRNA molecule of ~9kb. The genome contains the gag gene which encodes the structural proteins, mainly the protein capsid, the matrix protein, and the nucleocapsid, the genome also contains the pol gene which encodes the reverse transcriptase enzyme (RT), the protease enzyme, and the integrase enzyme and the env gene encode the membrane glycoprotein 120 and glycoprotein 41. HIV genome also encodes 6 regulatory proteins such as tat, rev, nef, vif, vpr, and vpu which are responsible for its pathogenicity and replication in the host⁵. Once the virus has been inside the host, the capsid disintegrates and the viral RNAs are reverse transcribed into DNA molecules via RT which starts with an RNA/DNA hybrid followed by further cleavage of the RNA, and synthesis of dsDNA takes place. The proviral dsDNA is further integrated into the host genome via the

integrase enzyme where it remains as 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 viral RNA genomes⁶⁻⁸. Nucleoside, nucleotide reverse transcriptase inhibitors (NRTIs), non-nucleoside reverse transcription inhibitors (NNRTIs), protease inhibitors (PIs), entry or fusion inhibitors, and integrase strand transfer inhibitors (INSTIs) are the major six classes of Antiretroviral therapy drugs targeting 5 different phases of the HIV life cycle⁹.

Many additional trials with longer time are necessary to develop novel drugs, but nowadays computational biology has shorten these processes period. In cure of AIDS, life-span medicines should be administered, however; they cause also immune-suppressive effect with possible other expected disorders. Therefore development of new drugs targeting the exact issue is of importance using computational biology techniques. Antiretroviral therapy against HIV infection has changed a uniformly fatal disease into a potentially chronic disease. There are now 17 drugs in common use for HIV treatment⁹. Patients who can access and adhere to combination therapy should be able to achieve durable, potentially lifelong suppression of HIV replication. Despite the unquestioned success of antiretroviral therapy, limitations persist. Treatment success needs strict lifelong drug adherence. Although the widely used drugs are generally well tolerated, most have some short-term toxic effects and all have the potential for both known and unknown long-term toxic effects. Drug and administration costs limit treatment in resource-poor regions, and are a growing concern even in resource rich settings. Finally, complete or near complete control of viral replication does not fully restore health. Long-term treated patients who are on an otherwise effective regimen often show persistent immune dysfunction and have higher than expected risk for various non-AIDS-related complications, including heart, bone, liver, kidney, and neurocognitive diseases⁹.

The computational drug discovery methods have gained huge momentum in recent years, especially with the availability of supercomputers with less time-intensiveness and lower the cost. Virtual screening to identify lead compounds that potentially inhibit several HIV-1 proteins and enzymes had been previously researched, including against the RNase domain, however, such researches were based solely on molecular docking, pharmacophore, and ADMET assays which considers mainly the best docking pose the ligands can have against the target protein in a vacuum-like condition and are trained on limited training datasets, which may or may not reflect their interaction in the physiological cell condition¹⁰⁻¹². Extensive computational analysis and molecular dynamics leads to finding of potential HIV-1 RT RNase domain has been previously performed by Zhang et al. (2016), however, the experiment was limited only to 77 α -hydroxytropolone derivatives, which limited the efforts of discovering novel small molecule inhibitors, hence, no large-scale extensive computational analysis with all-atom molecular dynamics simulations, and/or some form in silico binding free energy calculation validating the potential lead compounds have been performed against the HIV-1 RT RNase domain¹³.

In our previous studies, we have used MAUVE analysis to determine stable region of SARS-CoV-2 genome aimed for discovering potential drug candidate matching to proteins playing role in its virulence¹⁴⁻¹⁶. In this present study, a similar approach was applied for determining the rationale of targeting RT for the drug discovery and development efforts of anti-HIV-1 drugs, following the establishment of the RT 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 RT 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 Figure 1.

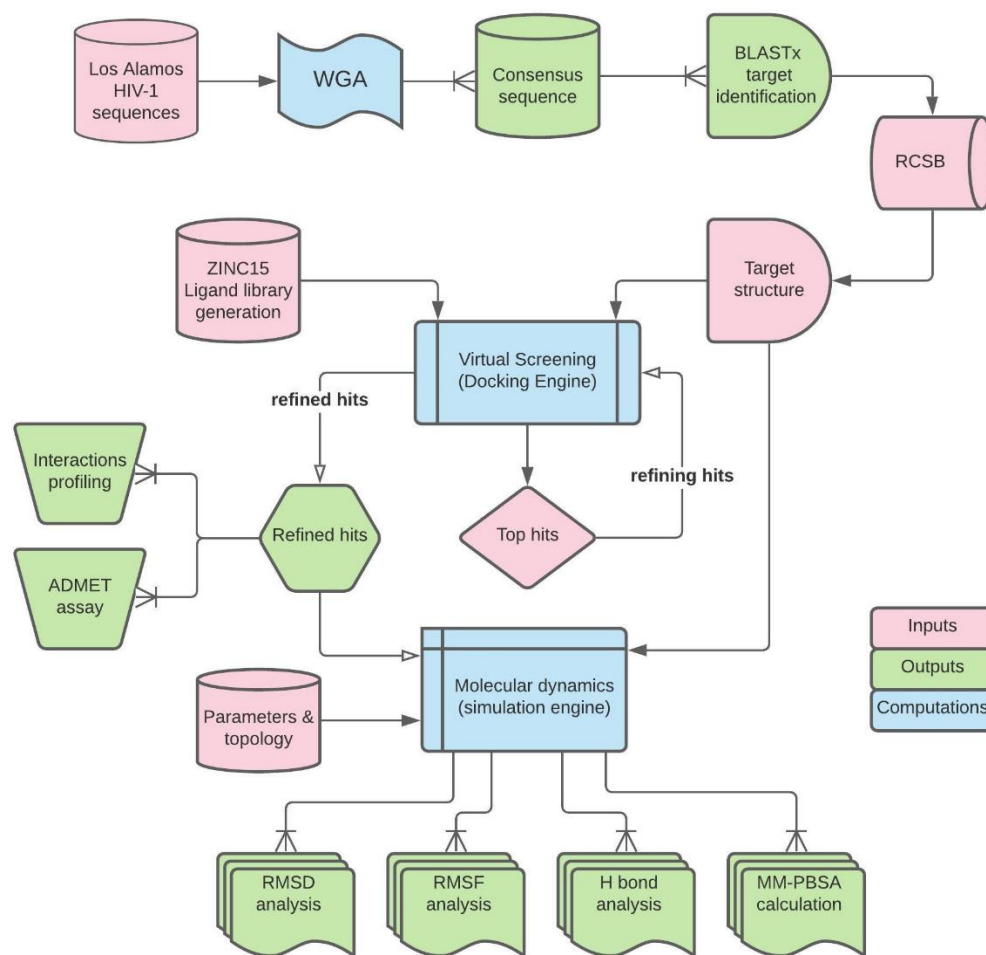


Figure 1. General workflow of the study. Pink shapes indicate external/intermediate inputs, green shapes indicate outputs that were analysed, and sky-blue shapes indicate the computational steps.

2. Materials and Methods

2.1. Whole-genome alignment and BLASTx

A total of 98 HIV-1 complete genomic sequences was retrieved from the Los Alamos HIV sequence database¹⁷, the sequences were manually selected such that only 2 sequences (where applicable) were chosen from each country and the dataset included all the geographic regions available (however, this trend was not strictly followed as some countries like the US received higher coverage for its size and the number of high-quality sequences deposited whereas other countries like India despite its size, due to lack of abundant high-quality sequences deposited, received lower coverage). Whole-genome alignment was performed via progressive MAUVE algorithm with match seed weight set to automatic calculation, minimum Locally Collinear Blocks (LCB) set to default (3 times the minimum match size), progressive Muscle (v3.6) was selected for as gap aligner for each LCB, and minimum island size, maximum backbone gap size, minimum backbone size were set to 50^{18,19}. The list of all the sequences used for the alignment is included in Supplementary Data 1 (SD1) and the whole genome alignment result is included in SD2. The alignment result was visualized in Geneious Prime (v2020.1) and the highest conserved continuous region with no gaps in the alignment was excised from the alignment and visualized separately in-depth²⁰. 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 score matrix adjustment), the alignment for the excised fragments are provided in SD3 and the consensus sequence is provided in SD4²¹⁻²³.

2.2. Molecular docking based virtual screening

The experimentally determined X-ray diffraction structure of HIV-1 RT with PDB ID 3IG1 was retrieved from the Research Collaboratory for Structural Bioinformatics (RCSB) website^{24,25}. The missing residues from the structure were added via the PyMol's builder plugin (open-source v2.5.0), the loop regions where the residues were added was refined using MODELLER (v10.1)²⁶⁻²⁸. The structure was then cleaned from all heteroatoms except for the cofactor atoms, polar hydrogens were added where necessary, and Kollman charges were computed²⁹. A library of 94,545 annotated anodyne small molecules (ligands) stable at physiological pH and having a charge of 0, -1, or -2 was generated from the ZINC15 database³⁰. A grid box with a size of 25 Å X 32 Å X 32 Å along the X, Y, Z-axis was calculated (a box around the RNase H active site). Virtual screening was performed with HIV-1 RT structure against the ligand dataset within the grid box calculated at exhaustiveness of 64 via AutoDock Vina (v1.1.2)³¹. The top 7 molecules with the highest affinity scores were screened again with the same configuration but with exhaustiveness of 256, compounds that successfully reproduced their scores in the same pose were retained for further analysis.

2.3. Protein-ligand interactions profiling

The best dock pose of the top hit ligands was loaded with the HIV-1 RT to PyMol and all the residues within 4 Å from the lead compounds were visualized (i.e. all potential, hydrophobic interactions, hydrogen bonds, and ionic interactions) and evaluated, the manually predicted bonds were also cross-validated with the TU Dresden's Protein-Ligand Interaction Profiler (PLIP) webserver and only overlapping interactions were considered³².

2.4. Molecular dynamics simulation

The molecular dynamics simulation was performed with the University of Illinois's Nanoscale Molecular Dynamics (NAMD v2.14 CUDA) tool³³. OPLS-AA/M force field from William L. Jorgensen research group was utilized to generate the topology and parameters for the RT enzyme, the same force field was used to parameterize the ligand molecules as well (via LigParGen server with 1.14 CM1A charge model)³⁴⁻³⁷. Each pair of protein-ligand complex was immersed in a square box with explicit TIP3P water with a distance of 5 Å was maintained between the edge of the box to the protein-ligand complex along each axis, the system was neutralized with Na⁺ and Cl⁻ ions and their final concentration was maintained at 0.15 mol/L (physiological salt concentration). The system was minimized for 2 ns to reach its lowest energy relaxed state from the X-ray diffraction state, the system was then equilibrated for 5 ns at 310K with periodic boundary conditions, Langevin dynamics, particle mesh Ewald (PME) for electrostatics, and Langevin piston (at 1 atm) with the protein-ligand complex constrained to allow the water and ions equilibrate around the complex, this step was followed by 10 ns equilibration with the constraints on the protein side chains released to allow the side chains to relax. The system was then subjected to 50 ns equilibration with constraints only on the cofactor Mn²⁺ cations to allow the system reach its equilibrium while maintaining the cofactor in the active site, the root mean square deviation (RMSD) of the RT's backbone (C α) and root mean square fluctuation (RMSF) of RT's C α from equilibration simulation was calculated using the 1st frame as the reference point to monitor the RT's behaviour under the simulation system. Finally, a 30 ns production simulation with no constraints was performed from which the trajectory was collected for analysis. The outputs were written to the trajectory every 1 ps, and RMSD of RT's C α and the lead ligand as a function of time elapsed was plotted along with RMSF of the C α for each residue of the RT throughout the

production simulation. The number of hydrogen bonds between the RT enzyme and the respective lead ligand throughout the production simulation was also plotted (with thresholds set to, donor-acceptor distance $< 3 \text{ \AA}$ and angle cut-off = 20°), all the statistical analysis and visualizations were performed using the *matplotlib* and *seaborn* libraries^{38,39}.

2.5. Binding free energy calculation via MM/PBSA

The binding free energy (ΔG_{bind} , Gibbs free energy) between the lead compounds and RT enzyme was calculated from the last 10 ns (stable RMSD interval) for each simulation system comprising of 1001 snapshots via MM/PBSA single trajectory protocol, the formula in equation (2) was followed to calculate the energy terms for each of the RT, lead compound and RT-lead complex using the CaFE plugin (v1.0) and finally equation (4) was used to calculate the ΔG_{bind} ^{40,41,42}. The equations derivation and approximation have been provided in SD5.

2.6. In silico physicochemical and ADMET profile analysis

The top lead compounds were submitted to the SwissADME webserver from the Swiss Institute of Bioinformatics to calculate their physicochemical properties and drug-likeness⁴³. The lead compounds drug-likeness were evaluated based on 5 filters, Lipinski rule of 5, Ghose filters, Veber filter, Egan filter, and Muegge filter⁴⁴⁻⁴⁸. As 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^{49,50}.

2.7. Additional docking studies

To evaluate the anti-RNase H activity of Phomoarcherin B, the reverse transcriptase enzyme of Feline immunodeficiency virus (FIV) which also exhibits a similar DEDD motif (PDB: 5OVN)⁵¹ was investigated using the same docking approach. Furthermore, docking of Phomoarcherin B with RNase H of Bacteriophage T4 (PDB: 1TFR)⁵² and monomeric reverse transcriptase of Moloney murine leukemia virus (MLV, PDB: 4MH8)⁵³ was also investigated.

RESULTS

Whole-genome alignment & BLASTx

The results from the whole-genome alignment are provided in Figure 2, the alignment is visualized such that each LCB is clustered together, the alignment is zoomed out so as each clustered LCB is shown as a black continuous bar, a region with a length of around 2.4 kb within the \approx 2.8-5.3 kb range from the consensus sequence is highly conserved with almost no gaps in most of the sequences. The alignment for this conserved region further shows a very high degree of conservation within this region throughout all of the HIV-1 genomes aligned with a gap only in one of the sequences (Figure 3).

The BLASTx results from the consensus sequence have indicated that the highly conserved region belongs to the HIV-1 pol gene (NCBI GenBank: QMX87928.1), hence nominating the functional proteins from the HIV-1 pol gene (RT, IN, and late-phase protease) as a promising target for the development of therapeutics, therefore, further therapeutic screening and analysis were performed on one of the main pol gene products, the reverse transcriptase enzyme⁵⁴.

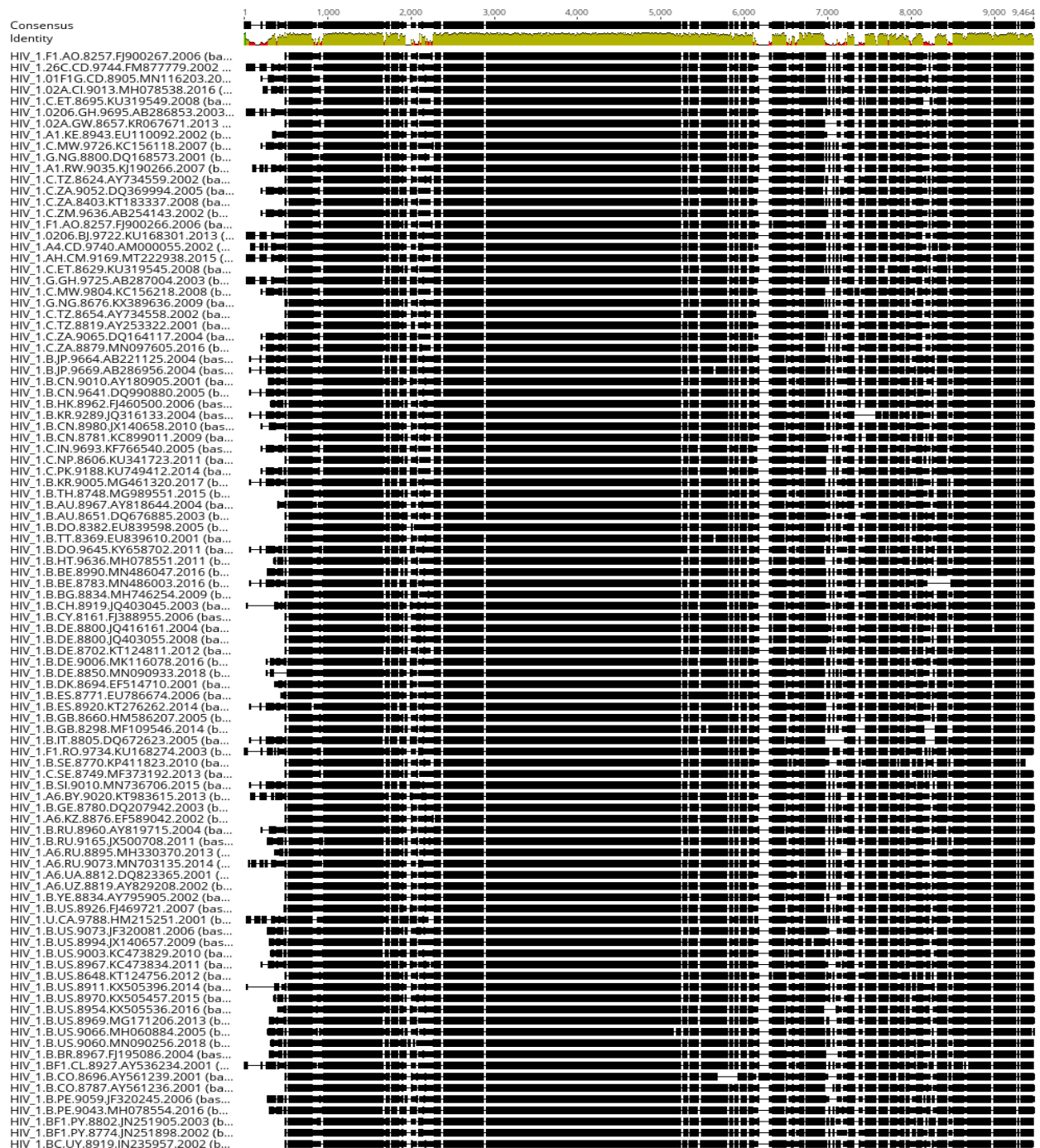


Figure 2. Whole-genome alignment of 98 HIV-1 genomes via progressive MAUVE algorithm, with headers for each sequence on the left (header format: organism, subtype, country code, sequence length, accession no. and year separated by periods), and identity percentage on top (red for low matches and green for high matches).

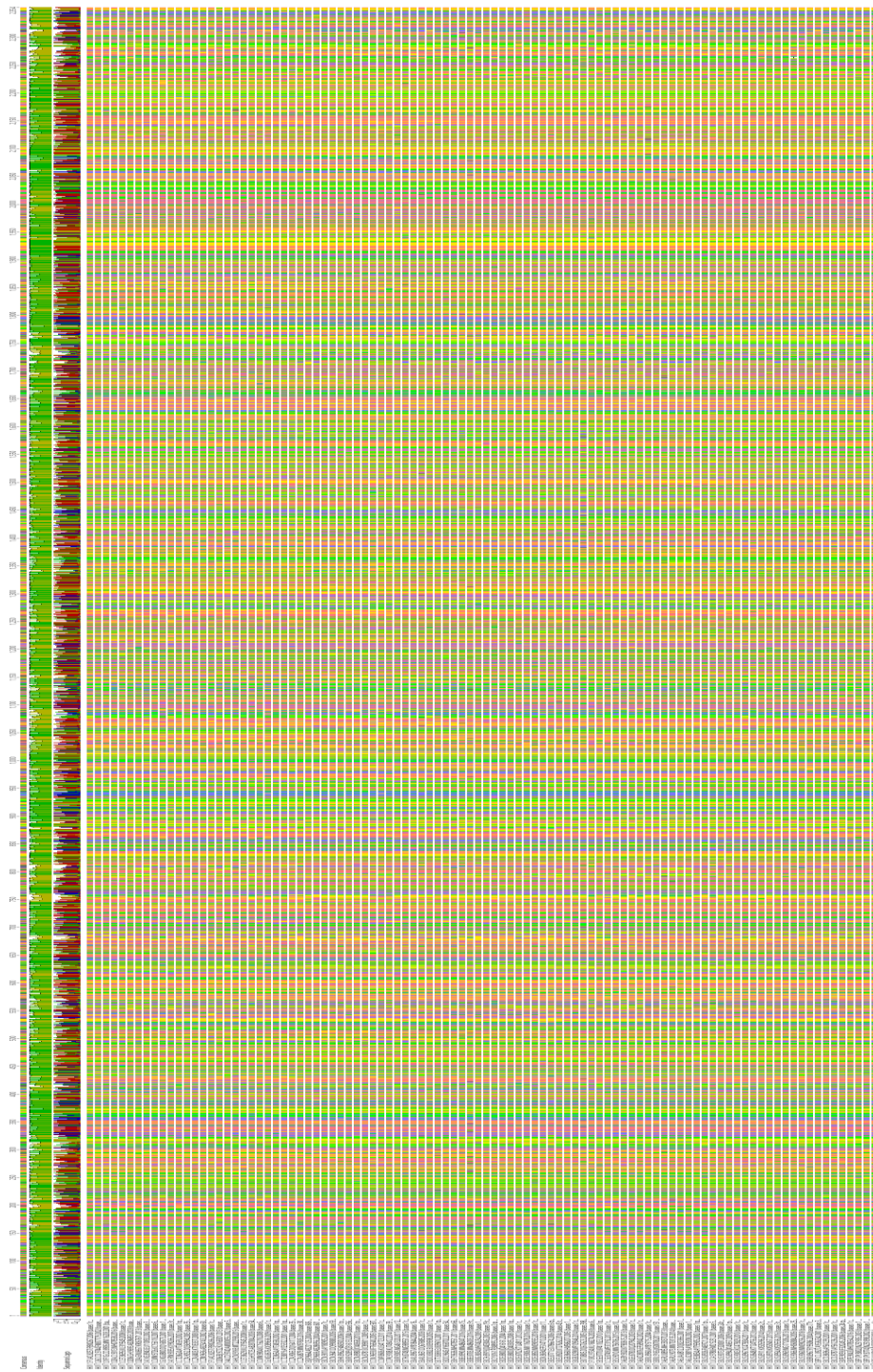


Figure 3. Alignment of the longest highly conserved region within the 98 HIV-1 genomes aligned via progressive MAUVE (Adenine in pink, guanine in yellow, thymine in green, and cytosine in blue).

Virtual Screening and molecular docking

The virtual screening of the ligand dataset against the HIV-1 RT enzyme nominated 7 compounds with significantly high affinities (≥ -8.5 kcal/mol), among them only 4 compounds successfully achieved the same affinity in 3 subsequent runs, hence only these 4 compounds were selected and further analysed, the top docking poses' for these 4 compounds are shown in Figure 4, a summary of each compound along with their chemical structures is given in Table 1.

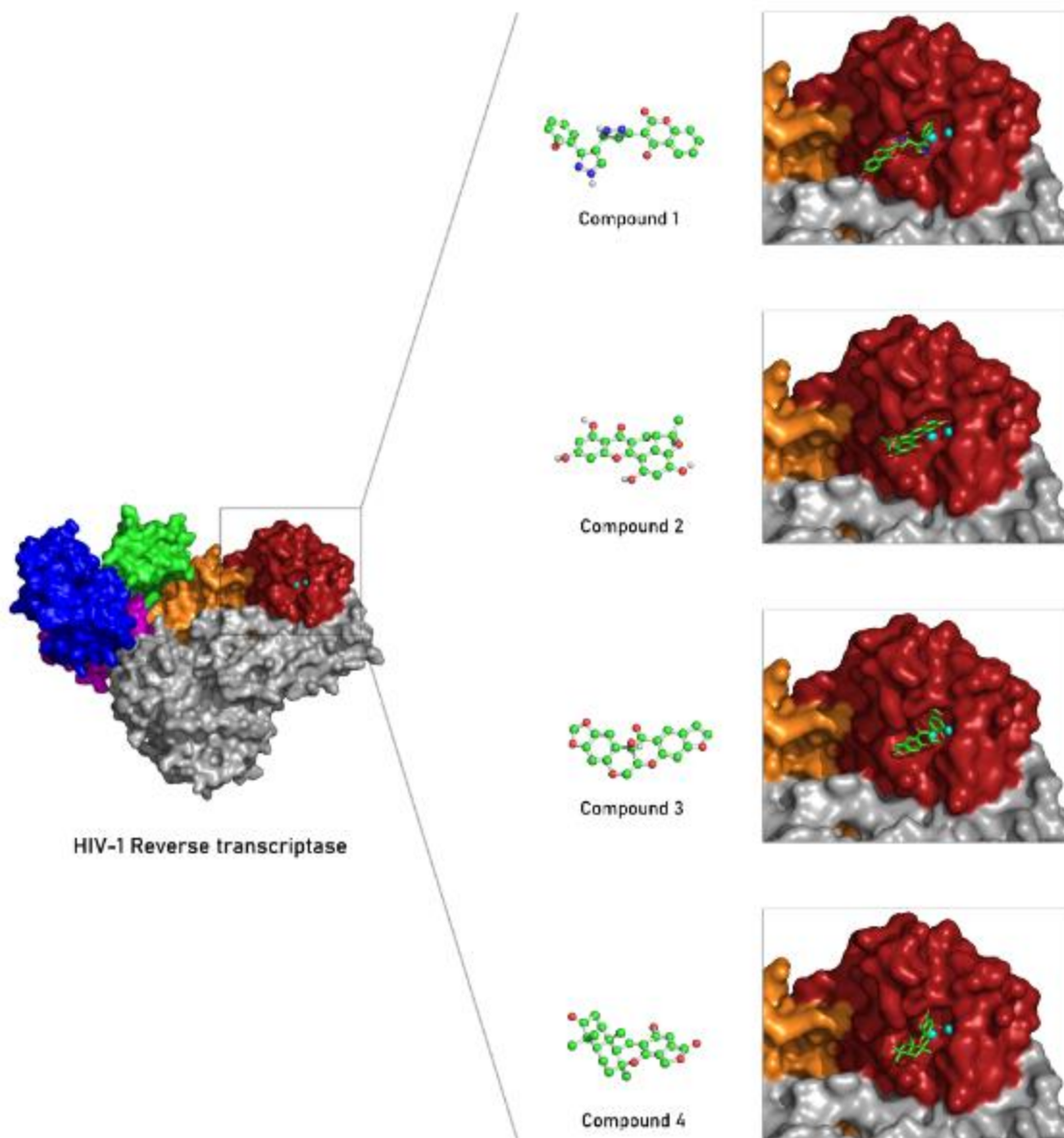
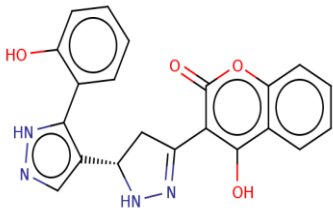
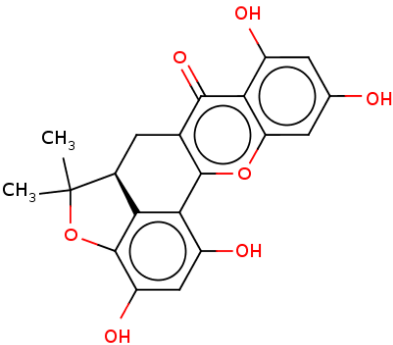
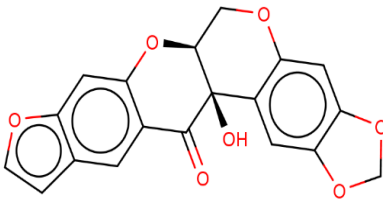
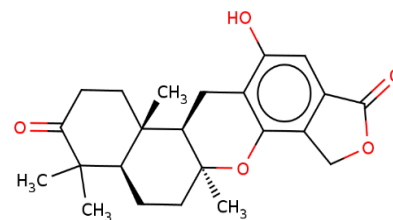


Figure 4. Docking poses' for the top 4 hit compounds with highest scores against the HIV-1 RT enzyme (surface representation on the left, RNase H domain in firebrick red, fingers subdomain in blue, thumb subdomain in green, palm subdomain in magenta, connection subdomain in orange, cofactor Mn in cyan beads, chain B in gray, and lead compounds with ball and stick representation in light green).

Table 1. Brief description of the top hit compounds as shown in Figure 4.

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[11.11.0.0.2,10.0.4,8.0.15,23.0.17,21]tracosan-2,4(8),9,15(23),16,19,21-heptaen-24-one	8.5	

Compound 4	ZINC000071318700	(6aR,6bS,10aR,12a	8.5
		S)-5-Hydroxy-	
		6b,10,10,12a-	
		tetramethyl-	
		6a,7,8,10,10a,11,1	
		2,12a-octahydro-	
		1H-	
		benzo[a]furo[3,4-	
		h]xanthene-	
		3,9(6H,6bH)-dione	



Interactions profiling

The interaction between the RT's RNase H catalytic site with the divalent cation and the top 4 potential lead compounds are (as shown in Figure 4) was closely analysed and visualized, Figure 5 (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 RT backbone are further expanded (stick representations in dark wild willow) to visualize the interacting atoms. The right columns in Figure 5 (e-h) shows 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 visualize the proximity (sky blue sticks) of the interactions, a summary table of all the interactions between the lead compounds within the RNase H active site has been listed in Table 2 along with their distances.

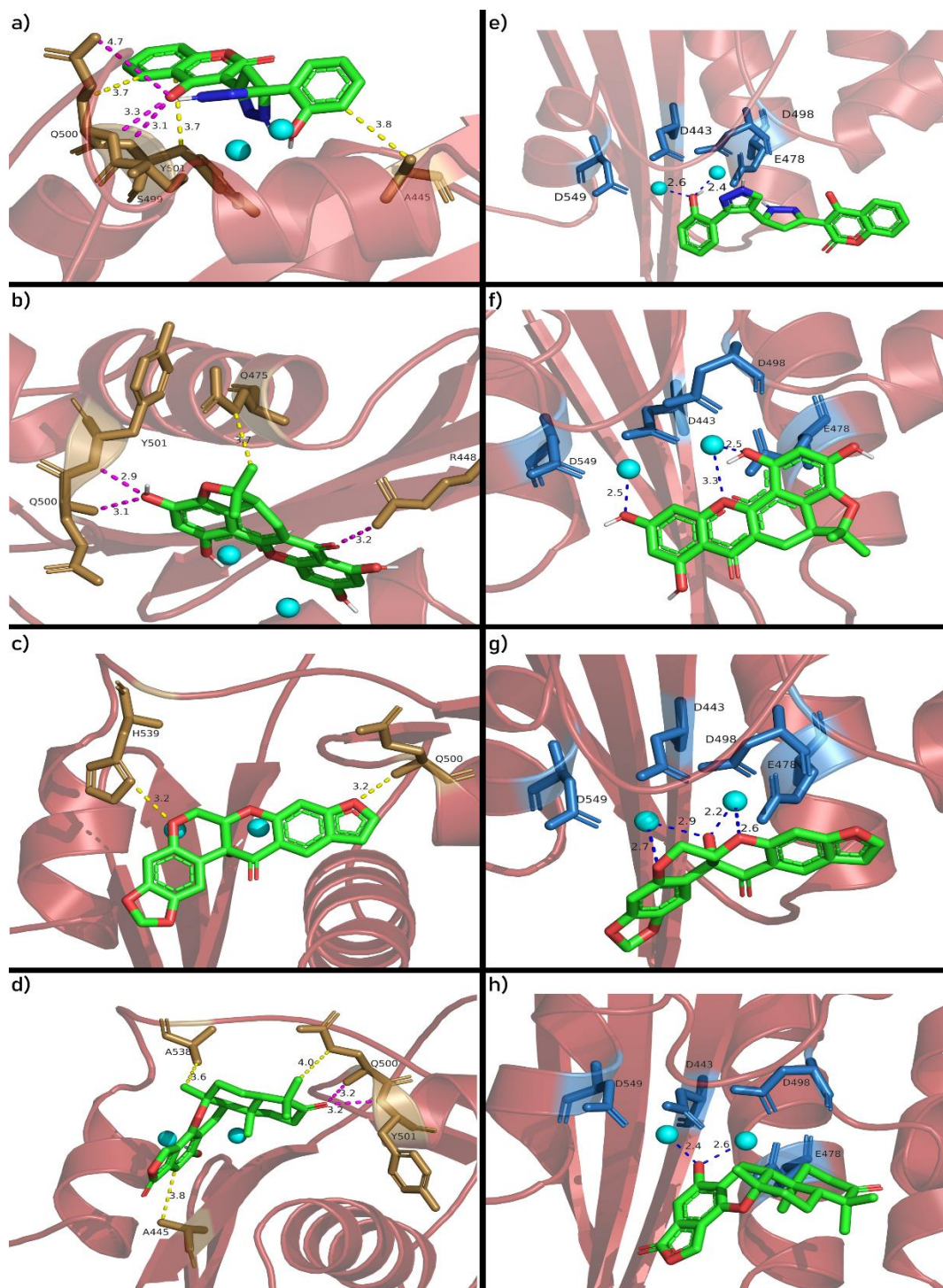


Figure 5. Predicted interactions of each lead compound within the RNase H active site of HIV-1 RT enzyme from their respective top binding pose. all the potential hydrophobic interactions (yellow dashes), hydrogen bonds (magenta dashes), and ionic interactions (dark blue dashes)

within RNase H active site for (a & e) compound 1, (b & f) compound 2, (c & g) compound 3, and (d & h) compound 4 are visualized along with their distances, all lead compounds have shown as green sticks with oxygen atoms colored 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 Mn²⁺ is shown as cyan beads, and the catalytic site residues holding the Mn²⁺ cations (the DEDD motif) shown as sky blue sticks emerging from the protein backbone (e-h), all measurements are in Å unit.

Table 2. Summary of all the interactions between the RNase H active site and the respective lead compounds as visualized in Figure 5.

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	2.40	Ionic
	Mn	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	2.50, 3.30*	Ionic

	Mn	2.50	Ionic
Compound 3	Q500	3.23	Hydrophobic
	H539	3.22	Hydrophobic
	Mn	2.21, 2.64*	Ionic
	Mn	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	2.41	Ionic
	Mn	2.60	Ionic

* Lead compounds making more than 1 interaction with the same Mn²⁺ cation have their distances mentioned within the same cell separated by a comma.

Molecular dynamics analysis

The RMSD for the C α of the RT from each frame throughout the 50 ns equilibration simulation was extracted and calculated using the 1st frame in each trajectory as the reference point, Figure 6-a shows the RT equilibration state throughout the equilibration simulation (plateau in the 40-50 ns interval). Figure 6-b shows the C α RMSF calculated using the same 1st frame as the reference point for each equilibration simulation. An average RMSF plot for the 4 simulations was also calculated by averaging the RMSF of each residue over all of the simulations (Figure 6-c) to analyse the overall RMSF of the RT through the course of equilibration under the simulation conditions.

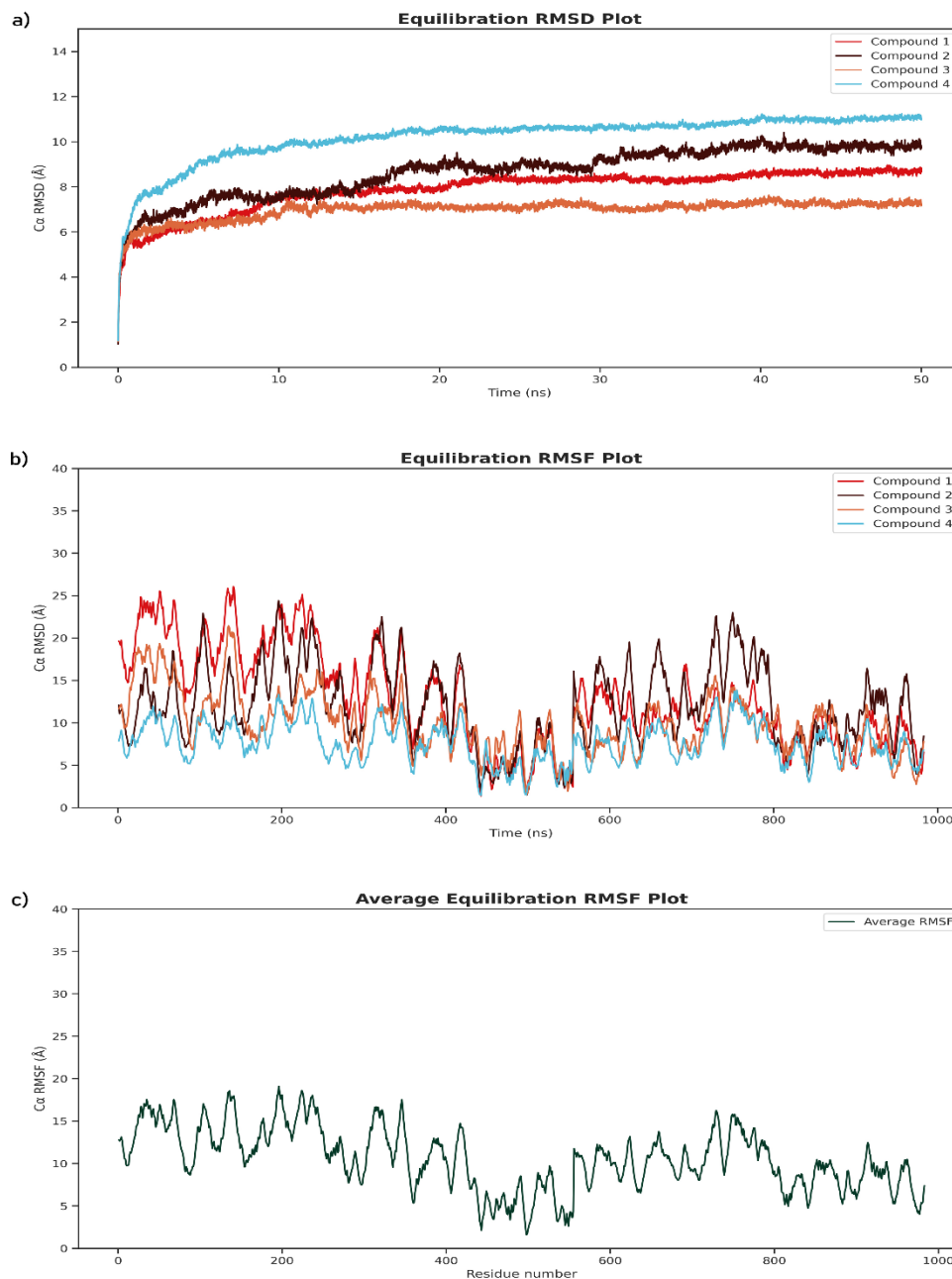


Figure 6. The behaviour of HIV-1 RT enzyme's backbone through the equilibration simulation. (a) plot of RT C α RMSD against time throughout the 50 ns equilibration simulation for each of the RT-lead compound system, (b) plot of RT C α RMSF against time throughout the 50 ns equilibration simulation for each of the RT-lead compound system, (c) plot of the average RT C α RMSF from the 4 RT-lead compound equilibration simulation.

The plots for C α RMSD against time for each RT-lead compound simulation (Figure 7), C α RMSF for each residue for each RT-lead compound (Figure 8), and the hydrogens bonds formed as a function of time in each RT-lead compound system (Figure 9) throughout the 30 ns production simulation were plotted to subsequently analyse the motion of the lead compound within the simulation system as well as the stability of the RT backbone in presence of the lead compounds.

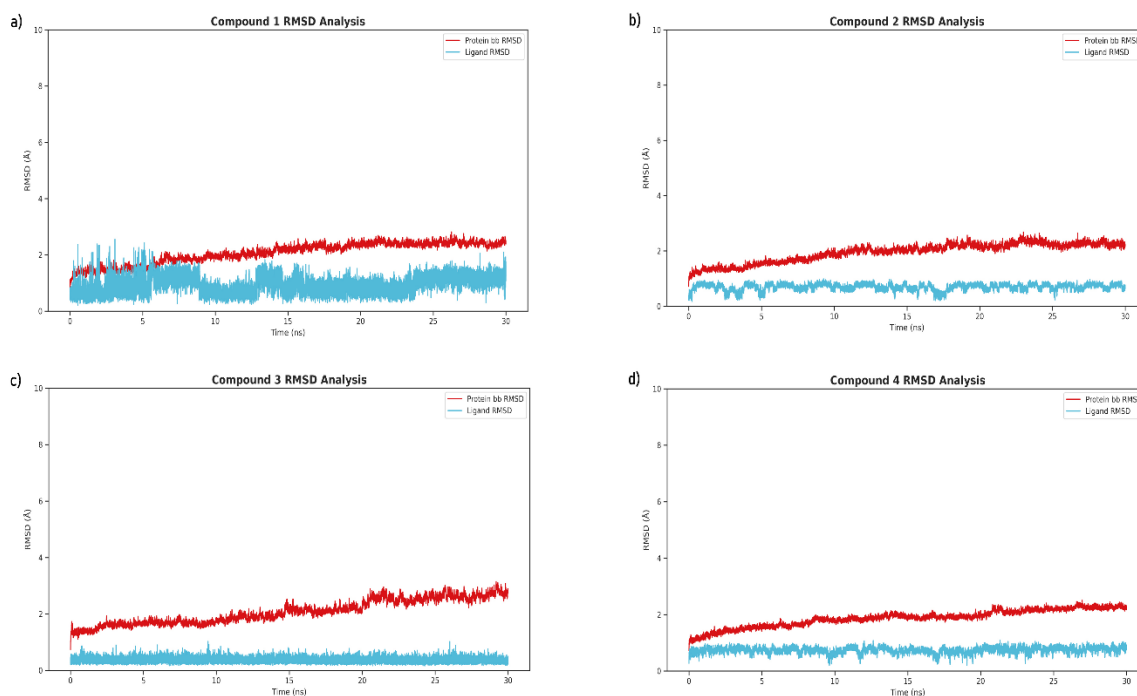


Figure 7. RT backbone (bb) RMSD plot in reference to the 1st frame throughout the 30 ns production simulation for (a) compound 1, (b) compound 2, (c) compound 3, and (d) compound 4. The cyan line represents the motion of the lead compound in each system (less variation along the Y-axis implies less deviation from its initial docked position), similarly. RT backbone RMSD (red line) shows the movement of the RT backbone during the simulation (since it plateaued in Figure 8, most of the motion is random loop movements and/or vibrations).

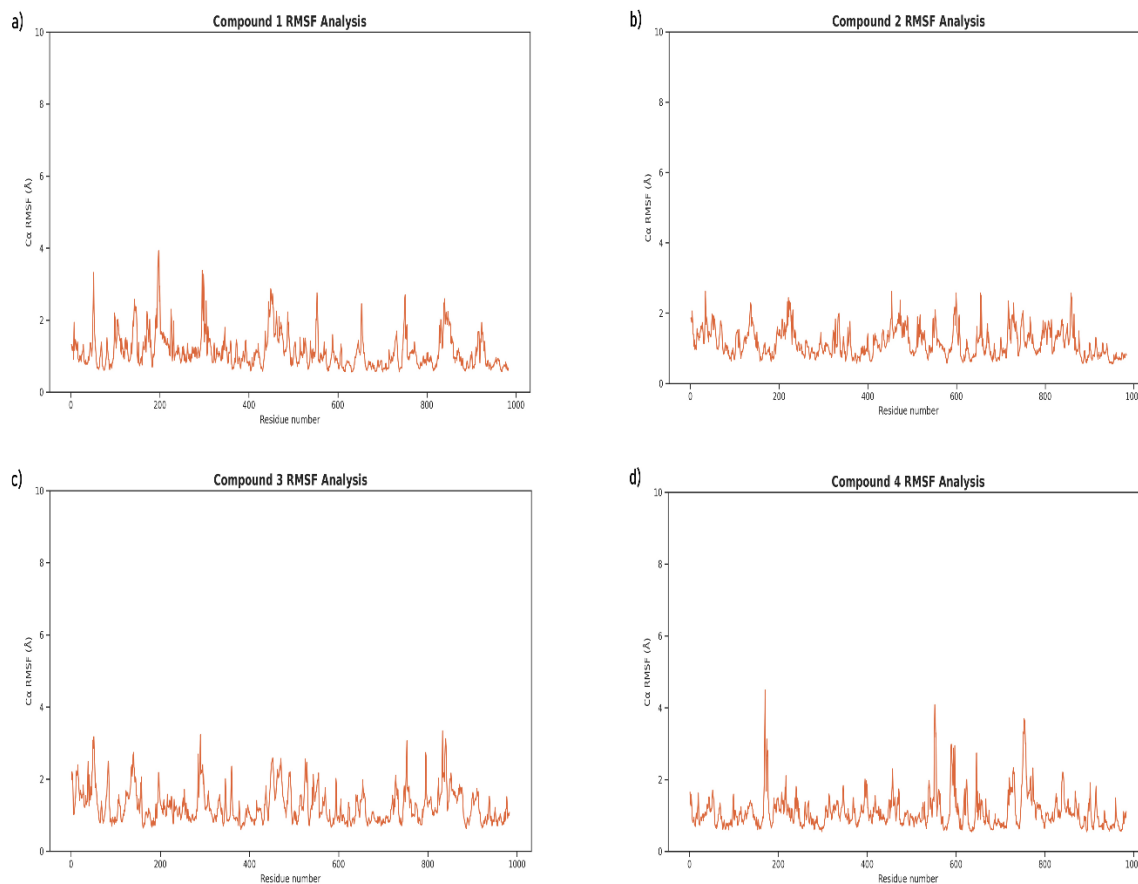


Figure 8. RMSF plot for each residue of RT backbone in reference to the 1st frame throughout the 30 ns production simulation for (a) compound 1, (b) compound 2, (c) compound 3, and (d) compound 4, the peaks (orange line) represent how much each residue within the RT moved from its initial state throughout the simulation (lower fluctuation implies more stable structure).

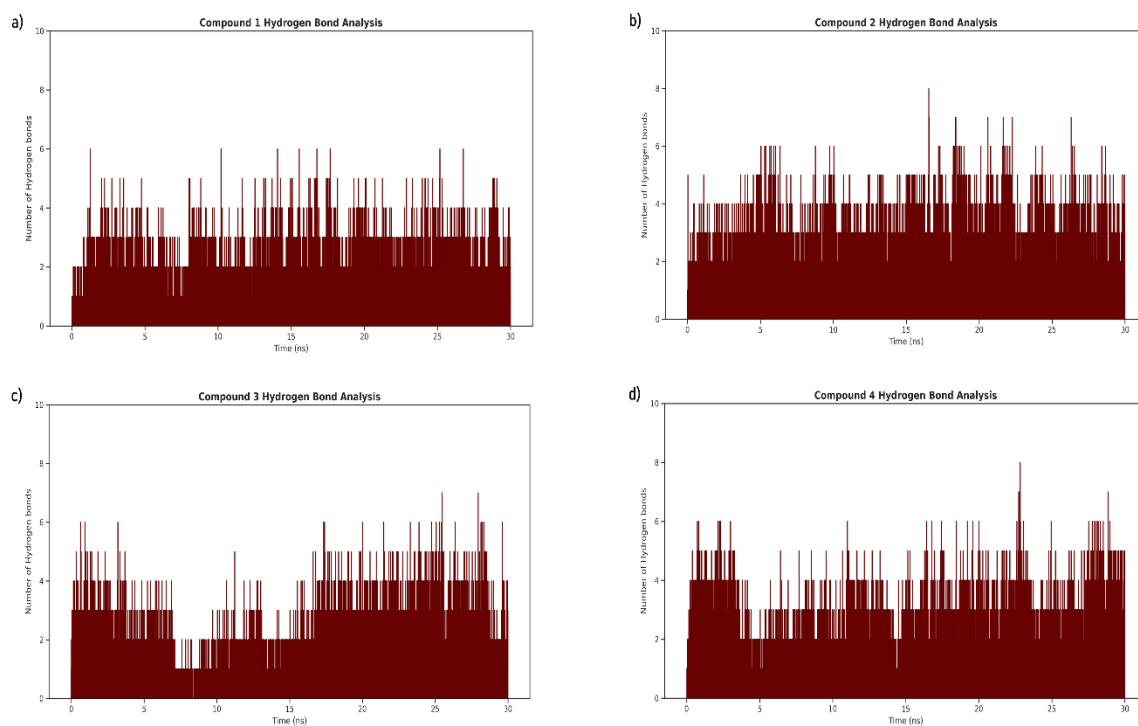


Figure 9. Number of hydrogen bonds formed by each RT-lead compound pair throughout the production simulation for (a) compound 1, (b) compound 2, (c) compound 3, and (d) compound 4. The parameters used to calculate the H bonds were a donor-acceptor distance of less than 3 Å and an angle cut-off of 20° (the most guaranteed threshold for H bonds, hence low bias).

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 RT enzyme, lead compound, and RT-lead complex separately from each production simulation, and their total sum was used in equation (4) to calculate the binding free energy $\Delta G_{bind/mmpbsa}$, the sums of each energy term is provided in Table 3 along with their standard deviations, the values of ΔG_{mmpbsa}

indicates the spontaneity of the interaction between the RT and lead compounds (i.e. more negative = more spontaneous). Detailed value for each energy term for the protein, ligand, and complex is separately provided in SD6.

Table 3. MM/PBSA energy terms for $\Delta G_{complex}$ calculated for each RT-lead pair, energies were calculated from the last 10 ns of the production trajectory using the single trajectory approach.

ΔE_{elec}	ΔE_{vdw}	ΔG_{PB}	ΔG_{SA}	ΔG_{Gas}	ΔG_{sol}	ΔG_{pol}	ΔG_{npol}	ΔG_{mmpbsa}
Compound 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
Compound 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
Compound 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
Compound 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

± indicates the standard deviations.

All values given energy values are for difference between the complex and sum of protein and ligand, $\Delta X_y = \Delta X_{y(complex)} - (\Delta X_{y(protein)} + \Delta X_{y(ligand)})$.

All values are in kcal/mol unit.

In silico ADMET assay

The physiochemical 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 summarized in Table 5 based on the results from admetSAR and ADMETlab.

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

Physiochemical 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	6	7	7	5
No. H-bond donors	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*	Moderately soluble	Moderately soluble	Soluble	Moderately soluble
Lipophilicity (Log P _{o/w}) [×]	2.30	2.42	2.15	3.67
Lipinski rule of 5 [#]	Pass (0)	Pass (0)	Pass (0)	Pass (0)
Ghose filters [#]	Pass (0)	Pass (0)	Pass (0)	Pass (0)
Veber filters [#]	Pass (0)	Pass (0)	Pass (0)	Pass (0)
Egan filters [#]	Pass (0)	Pass (0)	Pass (0)	Pass (0)
Muegge filters	Pass	Pass	Pass	Pass

*Based on the Log S (ESOL) scale, insoluble < -10 < poor < -6 < moderate < -4 < soluble < -2 < very < 0 < highly soluble.

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

[†] Based on the BOILED-Egg model⁵⁵.

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

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 [†]	High	High	High	High
Blood-brain barrier permeation [†]	None	None	None	Yes
Distribution				
Plasma binding protein [×]	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
CYP Inhibitory Promiscuity	High	High	Low	Low
Excretion [×]				
T _{1/2} (hours) [¶]	< 3 (0.349)	>3 (0.52)	< 3 (0.113)	>3 (0.70)

		Toxicity		
Rat acute (LD50, mol/kg)	2.43	2.38	2.41	2.71
TP* (pIGC50, ug/L)	0.46	1.04	0.57	1.38
Acute oral (LD50 mg/kg) [‡]	III	III	III	III
Carcinogenicity	None	None	None	None

[×] Based on predictions of ADMETLab 2.0⁵⁰.

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

* *Tetrahymena pyriformis* toxicity.

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

The SMILES notation of the lead compounds was used as the input to calculate each of the properties (provided in SD11).

Additional docking studies

Feline immunodeficiency virus (FIV) which also exhibits a similar DEDD motif (PDB: 5OVN) was investigated using the same docking approach a similar binding pose with an affinity of -9.0 kcal/mol was observed detailed results included in SD7. Furthermore, docking of Phomoarcherin B with RNase H of Bacteriophage T4 (PDB: 1TFR) and monomeric reverse transcriptase of Moloney murine leukemia virus (MLV, PDB: 4MH8) produced affinities of -8.1 kcal/mol and -8.3 kcal/mol respectively, further indicating the potency of Phomoarhcerin B as potential anti-viral RNase H candidate (detailed log files of the docking experiment in SD8 and SD9, respectively).

Discussion

As its well known, HIV infections have kept on claiming lives ever since its emergence, while modern anti-viral and HAART therapies provide some relief and support for the patients, it also comes with its disadvantages and limitations. This study aimed to perform extensive computational analysis to discover and evaluate potent novel inhibitors of HIV-1 replication within the host by targeting the most coherent target, providing therapeutic options for the patients while at the same time accelerating the drug development processes by providing potential leads.

The mutation rate for HIV-1 has been reported to be 10^{-4} to 10^{-2} mutants/clones and with the estimated production of 10^9 virions/day within an infected individual, the virus mutates quite efficiently to develop resistance and/or evade the immune system⁵⁶. However, not all of these mutants are expected to survive and replicate as mutations occurring on some genes could be lethal. In addition, the most coherent target for drug discovery and development efforts would be the phenotypes that mutate less frequently as their chance of developing resistance or evasion, which are lower than their highly mutating counterparts. All these information is useful to determine such regions with low mutation rate within the HIV-1 genome. The comparative genomics approach considers the correlations and differences between the genotype (genome) of closely related species or even different variants of the same specie to answer the reasons behind their characteristic phenotypes. This method has also been widely used to determine resistance genes for several bacteria in the past⁵⁷⁻⁵⁸.

In this study, we utilized 98 high quality HIV-1 sequence from Los Alamos database and performed whole-genome alignment with MAUVE to determine the regions within the HIV-1. The genome shows the highest level of consensus among all of the 98 sequences, as visualized in Figure 2, a genomic fragment of around 2.4 kb was the longest genomic fragment (the green bars

on top of the sequences in Figure 2) with the least variation among the selected 98 sequences. BLASTx was used to determine what this genomic region encoded, the pol gene which serves as the precursor for 3 important functional proteins of the HIV-1; the viral reverse transcriptase, integrase, and late-phase protease, that all of which are coherent targets for drug targeting. Given several NNRTI and NRTI has already been approved by the FDA for use and since all of them functions by inhibiting the DNA polymerase activity of the RT, our study focused specifically on discovering and analysing potent RT RNase H inhibitors, which is equally critical for the viral replication as its polymerase activity⁵⁹⁻⁶².

Considering all the computational analysis reported in this paper and the lead selection criteria applied, compound 4 is the best performing lead compound, with a docking score of -8.5 kcal/mol, several hydrophobic, hydrogen bond, and ionic interactions with active site residues of the HIV-1 RNase H and the cofactor Mn²⁺ cations, less than 1 Å deviation from its initial docked pose throughout the 30 ns molecular dynamic simulation, a binding free energy of $\approx -35.58 \pm 4.84$ kcal/mol and near-perfect scores on each of the ADMET profiles. To counter the potential bias that might rise from using the OPLS-AA force field and to ensure the reproducibility of the results, the molecular dynamic simulation of compound 4 with the RT enzyme was repeated with the CHARMM36m force field. The RT-Compound 4 system was generated with the same parameters via the charmm-gui, the system which was minimized and equilibrated for 10 ns followed by a production run of 30 ns from which the trajectory was collected. The outputs were written to the trajectory every 10th of a nanosecond as shown in Figure 10, compound 4 successfully reproduced its results and maintained less than 1 Å RMSD deviation (Figure 10, cyan line) from its docked pose. Then, further confirming its potential RNase H inhibitory activity, the RT enzyme's backbone RMSD also reached a plateau following the 20 ns time-lapse of the production

simulation^{63,64}. A video of the last 10 ns of the production simulation was also generated and is included in SD10 (HIV-1 RT RNase H domain as red cartoon, cofactor MN cations as cyan beads, and compound 4 as line-green sticks). The video also confirms the contribution of the loop regions and the terminal residues to the high protein bb RMSD as they're moving consecutively throughout the simulation. The enzyme-lead complex is also seen to rotate around its axis, further contributing to the backbone RMSD rise.

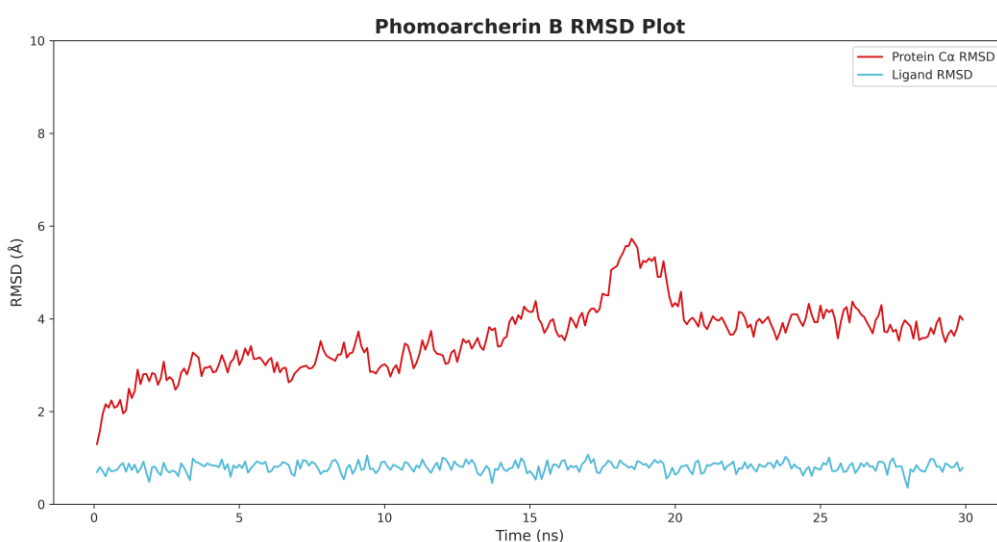


Figure 10. RT backbone (bb) RMSD plot in reference to the 1st frame throughout the 30 ns production simulation for compound 4 with the CHARMM 36m force field. The cyan line represents the motion of compound 4 (Phomoarcherin B). RT backbone RMSD (red line) shows the motion of the RT backbone during the simulation.

Compound 4 which is Phomoarcherin B (PubChem CID 52952104) that is a natural compound found in the endophytic fungus *Phomopsis archeri*, it was first isolated and characterized as a pentacyclic aromatic sesquiterpene via spectroscopic analysis by Hemtasin et al. (2011) and tested

for antimalaria activity against *Plasmodium falciparum* and anticancer activities on cholangiocarcinoma cell lines⁶⁵. Anticancer activity was also stated by Bedi et al. (2018) whereas there is no *in vitro* or *in vivo* assay performed regarding its antiviral or RNase H inhibitory activity, making it a potential novel drug candidate that could inhibit HIV-1 replication by inhibiting its RNase H activity⁶⁶. Hence further *in vitro* assays and clinical trials are necessary to confirm its potential as a HIV-1 RNase H inhibitor.

We believe that the additional docking studies of Phomoarcherin B with Feline immunodeficiency virus (FIV), monomeric reverse transcriptase of Moloney murine leukemia virus (MLV) and RNase H of Bacteriophage T4 has also provided promising results regarding its anti-RNase H, opening the doors for further studies to confirm its potential *in vitro* and/or *in vivo*, details provided in SD7, SD8 and SD9 (Figure 11).

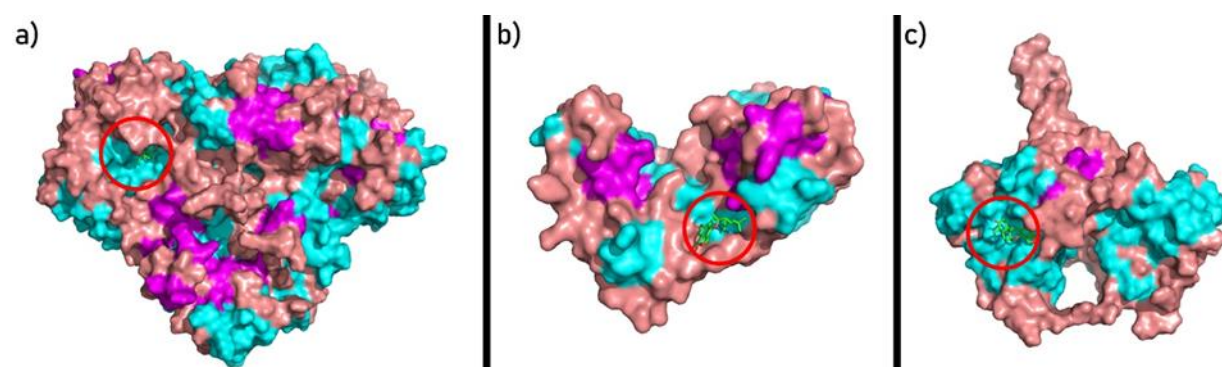


Figure 11. Top docking pose of Phomoarcherin B (red circle) with the (a) RT of FIV, (b) RT of MLV, and (c) RNase H of the Bacteriophage T4. Phomoarcherin B represented as green sticks and bonds, proteins in surface representation colored based on their (helices in cyan, sheets in light pink, and loops dry violet).

Data availability

Source data for all figures are available from the first author and/or the corresponding author.

References

1. Hemelaar, J. The origin and diversity of the HIV-1 pandemic. *Trends Mol. Med.* **18**, 182–192 (2012).
2. WHO. WHO HIV data and statistics. <https://www.who.int/teams/global-hiv-hepatitis-and-stis-programmes/data-use/hiv-data-and-statistics> (2019).
3. UNAIDS. Global HIV & AIDS statistics - 2020 fact sheet. <https://www.unaids.org/en/resources/fact-sheet> (2020).
4. Cohen, M. S., Hellmann, N., Levy, J. A., De Cock, K. & Lange, J. The spread, treatment, and prevention of HIV-1: evolution of a global pandemic. *J. Clin. Invest.* **118**, 1244-1254. (2008).
5. Kirchhoff, F. HIV Life Cycle: Overview in *Encyclopedia of AIDS*. (eds. Hope, T., Stevenson, M., Richman, D.) (Springer, 2016) https://doi.org/10.1007/978-1-4614-9610-6_60-1
6. Swanson, C. M. & Malim, M. H. SnapShot: HIV-1 Proteins. *Cell.* **133**, 9–10 (2008).
7. Fanales-Belasio, E., Raimondo, M., Suligoj, B. & Buttò, S. HIV virology and pathogenetic mechanisms of infection: a brief overview. *Ann. Ist. Super. Sanita.* **46**, 5-14 (2010).
8. Ruelas, D. S. & Greene, W. C. An integrated overview of HIV-1 latency. *Cell.* **155**, 519-529. (2013).
9. Volberding, P. A. & Deeks, S. G. Antiretroviral therapy and management of HIV infection. *Lancet* **376**, 49-62 (2010).
10. Poongavanam, V. & Kongsted, J. Virtual screening models for prediction of HIV-1 RT associated RNase H inhibition. *PLoS One* **8**, e73478; 10.1371/journal.pone.0073478 (2013).
11. Esposito, F. *et al.* Kuwanon-L as a new allosteric HIV-1 integrase inhibitor: molecular modeling and biological evaluation. *Chembiochem* **16**, 2507–2512 (2015).
12. Pinto, V. O. & de Azevedo, W. F. Optimized virtual screening workflow: towards target-based polynomial scoring functions for HIV-1 protease. *Comb. Chem. High Throughput Screen.* **20**, 820–827 (2017).
13. Zhang, B., D’Erasmus, M. P., Murelli, R. P. & Gallicchio, E. Free energy-based virtual screening and optimization of RNase H inhibitors of HIV-1 reverse transcriptase. *ACS Omega* **1**, 435–447 (2016).

14. Baysal, Ö., Abdul Ghafoor, N., Silme, R. S., Ignatov, A. N. & Kniazeva, V. Molecular dynamics analysis of N-acetyl-D-glucosamine against specific SARS-CoV-2's pathogenicity factors. *PLoS ONE* **16**, e0252571; 10.1371/journal.pone.0252571 (2021).
15. Baysal, Ö., Silme, R. S., Karaaslan, C. & Ignatov, A. Genetic uniformity of a specific region in SARS-CoV-2 genome and repurposing of N-Acetyl-D-Glucosamine. *Fresenius Environ. Bull.* **30**, 2848-2857 (2021).
16. Baysal, Ö. & Silme, R. S. Utilization from computational methods and omics data for antiviral drug discovery to control of SARS-CoV-2 [Online First]. IntechOpen <http://doi.org/10.5772/intechopen.98319> <https://www.intechopen.com/online-first/76991> (2021).
17. Kuiken, C., Korber, B. & Shafer, R. W. HIV sequence databases. *AIDS Reviews* **5**, 52–61 (2003).
18. Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14**, 1394–1403 (2004).
19. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).
20. Geneious Prime 2020.2.4. <https://www.geneious.com> (2020).
21. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).
22. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
23. States, D. J. & Gish, W. Combined use of sequence similarity and codon bias for coding region identification. *J. Comput. Biol.* **1**, 39–50 (1994).
24. Himmel, D. M. *et al.* Structure of HIV-1 reverse transcriptase with the inhibitor & B-Thujaplicinol bound at the RNase H active site. *Structure* **17**, 1625–1635 (2009).
25. Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide protein data bank. *Nat. Struct. Mol. Biol.* **10**, 980 (2003).
26. Webb, B. & Sali, A. Comparative protein structure modeling using MODELLER. *Curr. Protoc. Bioinforma.* **54**, 5.6.1-5.6.37 (2016).
27. Fiser, A., Do, R. K. & Sali, A. modeling of loops in protein structures. *Protein Sci.* **9**, 1753–1773 (2000).
28. Schrödinger, LLC. The PyMOL Molecular Graphics System, Version~1.8. (2015).
29. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **30**, 2785–2791 (2009).

30. Sterling, T. & Irwin, J. J. ZINC 15 – Ligand discovery for everyone. *J. Chem. Inf. Model.* **55**, 2324–2337 (2015).
31. 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).
32. Salentin, S., Schreiber, S., Haupt, V. J., Adasme, M. F. & Schroeder, M. PLIP: fully automated protein–ligand interaction profiler. *Nucleic Acids Res.* **43**, W443–W447 (2015).
33. Phillips, J. C. *et al.* Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J. Chem. Phys.* **153**, 044130 (2020).
34. Robertson, M. J., Tirado-Rives, J. & Jorgensen, W. L. Improved peptide and protein torsional energetics with the OPLS-AA force field. *J. Chem. Theory Comput.* **11**, 3499–3509 (2015).
35. Dodda, L. S., Cabeza de Vaca, I., Tirado-Rives, J. & Jorgensen, W. L. LigParGen web server: an automatic OPLS-AA parameter generator for organic ligands. *Nucleic Acids Res.* **45**, W331–W336 (2017).
36. Dodda, L. S., Vilseck, J. Z., Tirado-Rives, J. & Jorgensen, W. L. 1.14*CM1A-LBCC: Localized bond-charge corrected CM1A charges for condensed-phase simulations. *J. Phys. Chem. B* **121**, 3864–3870 (2017).
37. Jorgensen, W. L. & Tirado-Rives, J. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proc. Natl. Acad. Sci.* **102**, 6665–6670 (2005).
38. Waskom, M. L. Seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).
39. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. & Eng.* **9**, 90–95 (2007).
40. Liu, H. & Hou, T. CaFE: a tool for binding affinity prediction using end-point free energy methods. *Bioinformatics* **32**, 2216–2218 (2016).
41. Hou, T., Wang, J., Li, Y. & Wang, W. 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**, 69–82 (2011).
42. Singh, N. & Warshel, A. Absolute binding free energy calculations: On the accuracy of computational scoring of protein–ligand interactions. *Proteins Struct. Funct. Bioinforma.* **78**, 1705–1723 (2010).
43. Daina, A., Michielin, O. & Zoete, V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci. Rep.* **7**, 42717 (2017).
44. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **46**, 3–26 (2001).

45. Ghose, A. K., Viswanadhan, V. N. & Wendoloski, J. J. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. *J. Comb. Chem.* **1**, 55–68 (1999).
46. Veber, D. F. *et al.* Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* **45**, 2615–2623 (2002).
47. Egan, W. J., Merz Kenneth M. & Baldwin, J. J. Prediction of drug absorption using multivariate statistics. *J. Med. Chem.* **43**, 3867–3877 (2000).
48. Muegge, I., Heald, S. L. & Brittelli, D. Simple selection criteria for drug-like chemical matter. *J. Med. Chem.* **44**, 1841–1846 (2001).
49. Cheng, F. *et al.* admetSAR: A comprehensive source and free tool for assessment of chemical ADMET properties. *J. Chem. Inf. Model.* **52**, 3099–3105 (2012).
50. Xiong, G. *et al.* ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties. *Nucleic Acids Res.* **49**, W5–W14; 10.1093/nar/gkab255 (2021).
51. Galilee, M. & Alian, A. The structure of FIV reverse transcriptase and its implications for non-nucleoside inhibitor resistance. *PLOS Pathog.* **14**, e1006849; 10.1371/journal.ppat.1006849 (2018).
52. Mueser, T. C., Nossal, N. G. & Hyde, C. C. Structure of bacteriophage T4 RNase H, a 5' to 3' RNA–DNA and DNA–DNA exonuclease with sequence similarity to the RAD2 family of eukaryotic proteins. *Cell* **85**, 1101–1112 (1996).
53. Das, D. & Georgiadis, M. M. The crystal structure of the monomeric reverse transcriptase from Moloney Murine Leukemia virus. *Structure* **12**, 819–829 (2004).
54. Chapter 26 - Acquired Immune Deficiency Syndrome in *Immunology for Pharmacy* (ed. Flaherty, D. K.) 214–223 (Mosby, 2012). <https://doi.org/10.1016/B978-0-323-06947-2.10026-4>
55. Daina, A. & Zoete, V. A BOILED-egg to predict gastrointestinal absorption and brain penetration of small molecules. *ChemMedChem* **11**, 1117–1121 (2016).
56. Yeo, J. Y., Goh, G.-R., Su, C. T.-T. & Gan, S. K.-E. The determination of HIV-1 RT mutation rate, its possible allosteric effects, and its implications on drug resistance. *Viruses* **12**, 297 (2020).
57. Fournier, P.-E. *et al.* Comparative genomics of multidrug resistance in *Acinetobacter baumannii*. *PLOS Genet.* **2**, e7; 10.1371/journal.pgen.0020007 (2006).
58. Hardison, R. C. Comparative Genomics. *PLOS Biol.* **1**, E58; 10.1371/journal.pbio.0000058 (2003).
59. De Clercq, E. Non-nucleoside reverse transcriptase inhibitors (NNRTIs): past, present, and future. *Chem. Biodivers.* **1**, 44–64 (2004).
60. King, R. W., Klabe, R. M., Reid, C. D. & Erickson-Viitanen, S. K. Potency of nonnucleoside reverse transcriptase inhibitors (NNRTIs) used in combination with other Human

Immunodeficiency Virus NNRTIs, NRTIs, or protease inhibitors. *Antimicrob. Agents Chemother.* **46**, 1640 LP – 1646 (2002).

61. De Clercq, E. Perspectives of non-nucleoside reverse transcriptase inhibitors (NNRTIs) in the therapy of HIV-1 infection. *Farm.* **54**, 26–45 (1999).

62. Melikian, G. L. *et al.* Non-nucleoside reverse transcriptase inhibitor (NNRTI) cross-resistance: implications for preclinical evaluation of novel NNRTIs and clinical genotypic resistance testing. *J. Antimicrob. Chemother.* **69**, 12–20 (2014).

63. Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **14**, 71–73 (2017).

64. Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: A web-based graphical user interface for CHARMM. *J. Comput. Chem.* **29**, 1859–1865 (2008).

65. Hemtasin, C. *et al.* Cytotoxic Pentacyclic and Tetracyclic Aromatic Sesquiterpenes from *Phomopsis archeri*. *J. Nat. Prod.* **74**, 609–613 (2011).

66. Bedi, A., Adholeya, A. & Deshmukh, S. K. Novel anticancer compounds from endophytic fungi. *Curr. Biotechnol.* **7**, 168–184 (2018).

Acknowledgment

The computation resources for performing the virtual screening were provided by the Suzek lab at Mugla Sitki Kocman University, the molecular dynamic simulation and MM-PBSA calculations reported in this paper were performed on TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources).

Authors information

Affiliations

Department of Molecular Biology and Genetics, Faculty of Science, Muğla Sıtkı Koçman University, 48121 Muğla, Turkey

Naeem Abdul Ghafoor

Molecular Microbiology Unit in Department of Molecular Biology and Genetics, Faculty of Science, Muğla Sıtkı Koçman University, 48121 Muğla, Turkey

Ömür Baysal

Department of Computer Engineering, Faculty of Engineering, Muğla Sıtkı Koçman University, 48121 Muğla, Turkey

Barış Ethem Süzek

Center for Research and Practice in Biotechnology and Genetic Engineering, Istanbul University, 34119 Istanbul, Turkey

Ragıp Soner Silme

Contributions

N.A.G. collected the data, devised the study method, performed the computational calculations and analyses', and created the figures and graphs. O.B. has given the philosophy of the study and scientific approach, then commented on the results and the whole data analysis, and the text along with R.S.S. and B.E.S. R.S.S. compiled the references to construct of discussion related to

findings. B.E.S. also provided technical assistance and computational resources. All authors contributed towards writing of the manuscript.

Corresponding author

Correspondence to Ömür Baysal

omurbaysal@mu.edu.tr

Ethics declarations

Competing interests

The authors declare no competing interests.

Supplementary Data

SD1: HIV-1 sequences used for the whole genome alignments (in fasta format) in the study.

SD2: MAUVE alignment result (in fasta format).

SD3: Excised conserved region from whole genome alignment (in fasta format).

SD4: Consensus sequences for the conserved region (in fasta format).

SD5: Equations used for calculation of binding free energy via MM/PBSA.

SD6: Each MM-PBSA energy terms calculated for each reverse transcriptase, lead compound and reverse transcriptase-lead compound (plain text file).

SD7: Log file of Phomoarcherin B docking with Feline immunodeficiency virus' reverse transcriptase enzyme.

SD8: Log file of Phomoarcherin B docking with Moloney murine leukemia virus' reverse transcriptase enzyme.

SD9: Log file of Phomoarcherin B docking with Bacteriophage T4 RNase H.

SD10: Simulation video of Phomoarcherin B (lime-green stick) with the HIV-1 reverse transcriptase (MP4 format) in cartoon representation. Chain B (silver) also called the P51 subunit contains no catalytic sites whereas chain A, also known as the P66 subunit, contains both the polymerase and ribonuclease domains, the DNA polymerase domain consists of 3 subdomains, the fingers subdomain (blue), the palm subdomain (pink), the thumb subdomain (green), a connecting region (orange) bridges the DNA polymerase domain to the RNase H domain (red) which also contains 2 Mn atoms as cofactor (cyan beads).

SD11: SMILES notations for the 4 lead compounds.