

QSAR-based drug designing studies on HIV-1 integrase inhibitors

Salam Pradeep Singh¹ · Chitta Ranjan Deb² · Lakshmi N. Kakati³ · Bolin Kumar Konwar^{4,5}

Received: 11 August 2016 / Accepted: 26 October 2016 / Published online: 5 November 2016
© Springer-Verlag Wien 2016

Abstract In this study, QSAR modeling was performed for predicting the IC₅₀ value for a set of HIV-1 integrase inhibitors using multiple regression and partial least square method obtaining an optimized model for each method. These models were used to predict a set of test compounds obtained by performing a chemical similarity search of the training set from the NCBI PubChem database subjected to the Lipinski rule of five filters. The predicted IC₅₀ value for the test set compounds was further analyzed for molecular docking simulation against HIV-1 integrase revealing that the test set compounds have a more binding affinity than the training set compounds and the market approved drug raltegravir. The stability of the docked compounds (protein–ligand complexes) was further validated by performing molecular dynamics simulations for 20 ns using Gromacs 5.0, and the RMSD backbone was analyzed. Last, the ADME–toxicity analysis was carried out for the top docking hit compounds and the market approved drug

raltegravir revealing that the docked compounds have enhanced pharmacological parameters than raltegravir.

Keywords QSAR · Molecular docking · Molecular dynamics · ADME–toxicity

1 Introduction

Human immunodeficiency virus type 1 (HIV-1) is a lentivirus that caused the acquired immunodeficiency syndrome (AIDS) in humans (Douek et al. 2009; Weiss 1993). The virus encodes three enzymes that are required for viral replication: reverse transcriptase, protease, and integrase (IN). Although there are reports on limited use of antiviral targeting reverse transcriptase and protease, the development of resistant strains has caused a great havoc and limited their effectiveness (Imamichi 2004). Therefore, there is a vital requirement for the discovery of new antiviral agents directed against alternative sites in the viral life cycle, such as the HIV-1 integrase, and hence, HIV-1 integrase has come to light as an attractive target for anti-HIV therapy, because this enzyme is necessary for stable infection and its known functional analogs are not found in the human host (Imamichi 2004). There are reports on a wide variety of integrase inhibitors; however, the Food and Drug Administration (FDA), USA, has approved only a very few (Anthony 2004; Johnson et al. 2004). And hence, the discovery and development of integrase inhibitors have led to the first FDA approved integrase inhibitor—raltegravir (Fikkert et al. 2003; Grobler et al. 2002; Murray et al. 2007). Even though raltegravir succeed in suppressing the virus in HIV patients (Murray et al. 2007), but HIV phenotypes resistance to raltegravir have evolved very rapidly (Malet et al. 2008).

Electronic supplementary material The online version of this article (doi:10.1007/s13721-016-0141-6) contains supplementary material, which is available to authorized users.

✉ Salam Pradeep Singh
salampradeep@gmail.com

¹ Bioinformatics Infrastructure Facility, Nagaland University, Lumami, 798627, Nagaland, India

² Department of Botany, Nagaland University, Lumami, 798627 Nagaland, India

³ Department of Zoology, Nagaland University, Lumami, 798627 Nagaland, India

⁴ Department of Molecular Biology and Biotechnology, Tezpur University, Tezpur, 784028 Assam, India

⁵ Nagaland University, Lumami, 798627 Nagaland, India

Therefore, there is a vital requirement for developing novel integrase inhibitors with alternative mechanisms of action.

In the present investigation, a set of 47 HIV-1 integrase inhibitors with their known experimental IC_{50} values was retrieved from the NCBI Bioassay database. These sets of diverse 47 compounds were used as the training set to construct the quantitative structure activity relationship (QSAR) model using multiple regression (MR) and partial least square (PLS). And the generated QSAR models were used to predict a set of compounds obtained from a chemical similarity search of the training set at the NCBI PubChem database subjected to the Lipinski rule of five parameters (Lipinski 2000). A molecular docking simulation was carried out against the training set and test set compounds and its subsequent protein–ligand interaction analysis of the top docking hits. In addition, a molecular dynamics simulation was performed for the best docked compounds to check the stability and dynamic behavior of the docked protein–ligand complexes.

2 Materials and method

2.1 Data set and QSAR modeling

A set of diverse HIV-1 integrase inhibitors (47 compounds) with known experimental IC_{50} values was retrieved from the NCBI Bioassay database (Wang et al. 2012). These 47 compounds were used to construct the QSAR model training set. The chemical structure, PubChem Compounds ID and the IC_{50} values of these 47 compounds are shown in supplementary material table ST1 and ST2. For the generation of the QSAR model, initially, the two-dimensional file format (SDF) of the training set compounds was optimized using MM2 force field method (Ulrich and Norman 1982) and converted to its corresponding three-dimensional file format (sybyl mol2) using (ChemOffice 2010). These 3D geometries of the compounds were then used to construct the QSAR model with two different methods, viz., multiple regression (MR) and partial least square (PLS) method using MDL QSAR 2.3 (MDL QSAR 2006). Multiple regression computes the least squares fit in an independent variable 'Xk' to a dependent variable 'Y'. It is commonly used in problems with a small number of independent variables that are not close to linearly dependent (Fisher 1922). In the multiple regression QSAR model, several models were obtained and the best model was chosen after barring 15 outlier molecules, and hence, only 25 molecules were considered in the regression model (shown in supplementary material Table ST1). The regression model quality was then accessed by multiple R-squared and cross validated with RSS. Furthermore, a randomization test was performed to check the validity of

the descriptor set. Finally, the cross-validation regression coefficient Q-Squared and the non-cross-validation coefficient R-Squared were used to estimate the model qualities.

On the other hand, a second model was constructed using partial least squares (PLS) regression. PLS is used where regression, principal components are used. In PLS, instead of forming regression equations, it produces a sequence of models with a different number of variables for predicting the dependent variable (Ildiko and Friedman 1993). Employing PLS QSAR model, the best model was chosen after barring 15 outliers, and hence, only 25 molecules were considered for the constructing the QSAR model (shown in supplementary material Table ST2).

2.2 Test set prediction

For predicting the IC_{50} of the test set compounds, a chemical similarity search was performed for each of the training set at the NCBI PubChem database for related structures and analogs (Bolton et al. 2008). The search parameters were set at 95% similarity subjected to the Lipinski rule of five filters (Lipinski 2000). In addition, only the compound whose IC_{50} value has not been reported in the PubChem Bioassay was retrieved (55 compounds). These compounds were used as the test set for predicting their IC_{50} values. The CID and 2D structures of these test set compounds are shown in Table ST1 and ST2 of supplementary material.

2.3 Molecular docking computation

Molecular docking simulation was carried out using Molegro Virtual Docker. For molecular docking simulations, the three-dimensional crystal structure of HIV-1 integrase catalytic core domain (Kessl et al. 2012) (PDB ID: 4DMN) was retrieved from the Protein Data Bank (<http://www.rcsb.org/>). The protein was then imported in the Molegro Virtual Docker (MVD) (Molegro APS 2011) and the search space was set for the active site residues (Asp64, Asp116 and Glu152). Furthermore, the binding site was set inside a restricted sphere of radius 13 Å (X: -34.57, Y: -10.85, Z: 5.52) with the MolDock Score (GRID) scoring function having a grid resolution of 0.30 Å. Then, the test set and training set compounds along with experimental known inhibitor raltegravir (for validation purpose) were also imported in MVD. The bond flexibility of the compounds was set along with the side chain flexibility of the protein for search space (Asp64, Asp116 and Glu152) was also set with to a tolerance of 1.10 and a strength of 0.90 for docking simulations. RMSD threshold for multiple cluster poses was set at 2.00 Å. The docking algorithm was set at a maximum iteration of 1500 with a simplex evolution size of 50 and a minimum of 10

Table 1 IC₅₀ values (experimental) and IC₅₀ values (predicted) generated using multiple regression (MR) and partial least square (PLS)

SN	Compound	Multiple regression (MR)		Residual	Compound	Partial least square (PLS)		Residual
		IC ₅₀ (Exp)	IC ₅₀ (Predicted)			IC ₅₀ (Exp)	IC ₅₀ (Predicted)	
1	CID11166962	0.004	0.0042	-0.0002	CID11166962	0.004	0.0040	0.0000
2	CID11225444	0.004	0.0042	-0.0002	CID11225444	0.004	0.0035	0.0005
3	CID11453518	0.002	0.0019	0.0001	CID11453518	0.002	0.0020	0.0000
4	CID25256802	0.002	0.0018	0.0002	CID25256802	0.002	0.0019	0.0001
5	CID25263157	0.004	0.0043	-0.0003	CID25263157	0.004	0.0041	-0.0001
6	CID25263158	0.004	0.0040	0.0000	CID25263158	0.004	0.0039	0.0001
7	CID25263160	0.004	0.0038	0.0002	CID25263160	0.004	0.0039	0.0001
8	CID25263162	0.004	0.0039	0.0001	CID25263162	0.004	0.0036	0.0004
9	CID25263169	0.004	0.0035	0.0005	CID25263169	0.004	0.0044	-0.0004
10	CID54677501	0.007	0.0066	0.0004	CID54677501	0.007	0.0068	0.0002
11	CID54677537	0.006	0.0060	0.0000	CID54677537	0.006	0.0056	0.0004
12	CID54682238	0.006	0.0061	-0.0001	CID54682238	0.006	0.0059	0.0001
13	CID54682241	0.007	0.0066	0.0004	CID54682241	0.007	0.0069	0.0001
14	CID54682247	0.007	0.0071	-0.0001	CID54682247	0.007	0.0068	0.0002
15	CID54682250	0.007	0.0071	-0.0001	CID54682250	0.007	0.0070	0.0000
16	CID54688265	0.005	0.0051	-0.0001	CID54688265	0.005	0.0054	-0.0004
17	CID54682263	0.005	0.0049	0.0001	CID54682263	0.005	0.0045	0.0005
18	CID54712035	0.006	0.0055	0.0005	CID54712035	0.006	0.0059	0.0001
19	CID54709826	0.004	0.0045	-0.0005	CID54686312	0.003	0.0035	-0.0005
20	CID54709838	0.005	0.0055	-0.0005	CID54709833	0.006	0.0062	-0.0002
21	CID54709842	0.007	0.0067	0.0003	CID54709839	0.007	0.0066	0.0004
22	CID54712034	0.004	0.0041	-0.0001	CID54709841	0.006	0.0066	-0.0006
23	CID25023920	0.007	0.0069	0.0001	CID25023913	0.004	0.0042	-0.0002
24	CID25263163	0.003	0.0031	-0.0001	CID25263144	0.004	0.0047	-0.0007
25	CID54682245	0.005	0.0055	-0.0005	CID54682260	0.007	0.0070	0.0000

runs were performed for each compound. The pose generation energy threshold was set to 100, and the simplex evolution maximum step was set to 300 by a neighbor distance factor of 1. The best pose of each compound was selected for the subsequent ligand–protein interaction energy analysis.

2.4 Molecular dynamics simulation

Molecular dynamics simulation was carried out using Gromacs 5.0.1 (Van Der Spoel et al. 2005) operated in Ubuntu Linux 14.0 LTS. MD Simulation was carried out for the enzyme 4DMN and the top docking complexes 4DMN-CID69851424, 4DMN-CID66861330, and 4DMN-CID54714968 complexes, respectively. In MD simulation studies, initially, the protein and the protein–ligand complexes were solvated using the TIP3P water model and the solvated structures were energy minimized using the steepest descent method, terminating when the maximum force is found in smaller than 100 kJmol⁻¹nm⁻¹. All simulations were performed in the NPT ensemble at constant temperature (300 K) and pressure (1 bar) with a time

step of 2 fs. NVT was performed for 1 ns (nanoseconds), and the minimized structure were equilibrated with a timescale of 20 ns (nanoseconds). The molecular dynamics simulation was performed to both the protein and the protein–ligand binding complex for 20 ns to understand the dynamic behavior of the protein and their stability.

2.5 ADME toxicity studies

Molecular docking may result with a compound having a favorable binding affinity and strong interactions. However, an ideal oral drug should possess certain factors such as quick gastrointestinal tract absorption. The potential drug candidate should also specifically distributed to its target and metabolized in such a way that it does not instantly eliminate its activity, and removed in an appropriate manner without causing any damage (Ghasemi et al. 2015). Hence, ADME–toxicity (absorption, distribution, metabolism, excretion, and toxicity), which describes the nature of a pharmaceutical compound within an organism, is an important step in the drug discovery process. It is also evident that poor human pharmacokinetics (ADME–toxicity) is a main reason for the

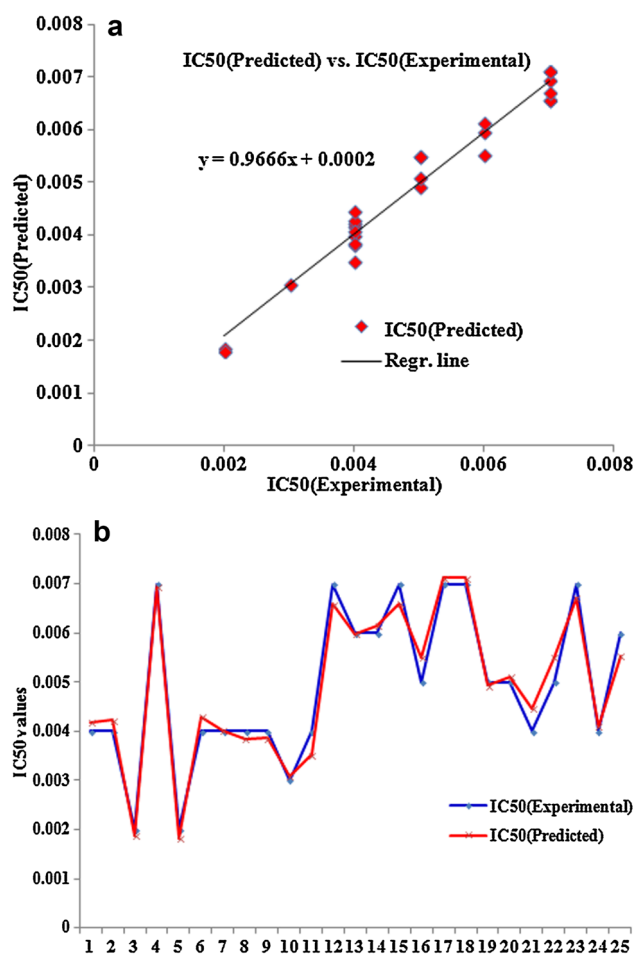


Fig. 1 a Regression line and b regression model graph of IC₅₀ (predicted) vs IC₅₀ (experimental) generated employing multiple regression (MR)

majority of the drug failure (Alavijeh et al. 2005). For these purposes, ADME–toxicity predictions were calculated for the top docking hits using ACD/I-Lab 2.0 (Advanced Chemistry Development, Inc). ACD/I-Lab 2.0 is a web-based service that provides instant access to spectral and chemical databases, and predicts properties, including physicochemical, ADME, and toxicity characteristics. In addition, comparative analyses were performed for the LD₅₀ mouse (intraperitoneal, oral, intravenous, subcutaneous) and probability of health effect of blood, cardiovascular system, gastrointestinal system, kidney, liver, and lung.

3 Results and discussion

3.1 QSAR modeling

QSAR modeling was carried out using MDL QSAR and two optimized models were generated using MLR and

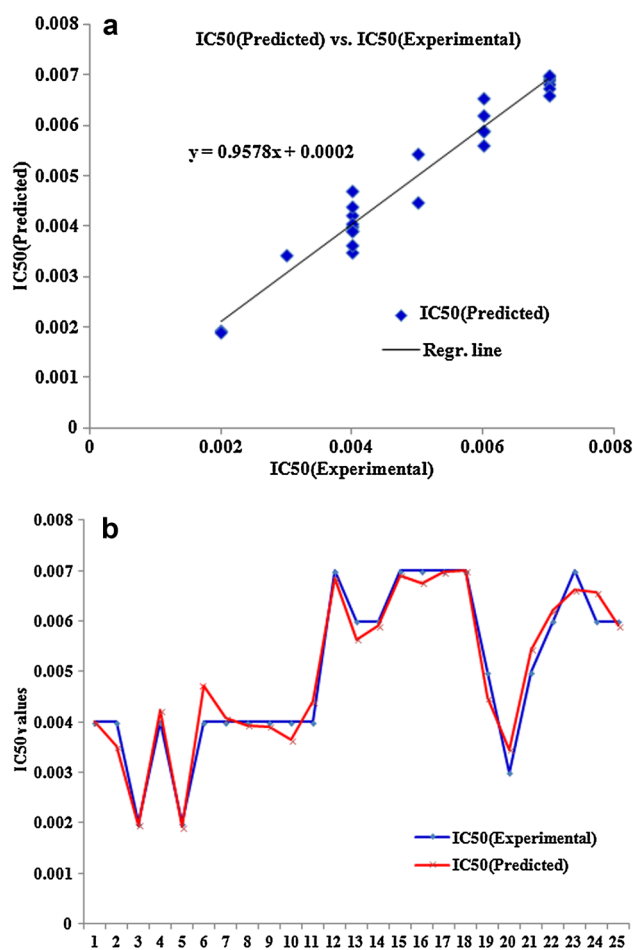


Fig. 2 a Regression line and b regression model graph of IC₅₀ (predicted) vs IC₅₀ (experimental) generated employing partial least square method (PLS)

PLS. First, the generated regression equation of the MLR is given below:

$$pIC_{50} = 0.0004949 * SHBint5_Acnt - 0.002895 * nelelem + 0.002142 * xvp6 + 0.0001224 * SdO - 0.009271 * SsF - 0.0008987 * SHsOH + 0.134736 \quad (1)$$

where SHBint5_Acnt is the count of internal hydrogen bonds with five skeletal bonds between donor and acceptor, nelelem is the number of chemical elements, xvp6 is the valence 6th order path Chi indices, SdO is the sum of all (=O) E-State values in molecule, SsF is the sum of all (-F) E-State values in molecule, and SHsOH:nSum of all [-OH] E-State values in molecule].

The regression quality of the optimized model generated using MR is described with a multiple R-Squared value of 0.9666 with a standard error of estimation of 0.0003275 (the smaller the better). The equation has an F-statistic value of 86.9 with a *P* value of 2.657E-012. The multiple

Table 2 Predicted IC₅₀ values of the test set compounds using multiple regression (MR) method and partial least square (PLS) method

SN	Compound ID	pIC ₅₀ (MR)	pIC ₅₀ (PLS)	Residual IC ₅₀ (PLS-MLR)
1	CID11188356	0.0036	0.0035	-0.0001
2	CID49768913	0.0035	0.0035	0.0
3	CID54683051	0.0039	0.0035	-0.0004
4	CID54684210	0.0067	0.0067	0.0
5	CID54686299	0.0060	0.0031	-0.0029
6	CID54686300	0.0051	0.0034	-0.0017
7	CID54709855	0.0054	0.0057	0.0003
8	CID54712001	0.0059	0.0059	0
9	CID54712049	0.0059	0.0056	-0.0003
10	CID54712050	0.0062	0.0059	-0.0003
11	CID54714968	0.0053	0.0066	0.0013
12	CID54735067	0.0071	0.0060	-0.0011
13	CID54751387	0.0044	0.0033	-0.0011
14	CID58899512	0.0042	0.0027	-0.0015
15	CID59382804	0.0047	0.0053	0.0006
16	CID59513436	0.0042	0.0027	-0.0015
17	CID59513438	0.0032	0.0033	0.0001
18	CID59513441	0.0018	0.0013	-0.0005
19	CID59513448	0.0052	0.0038	-0.0014
20	CID59513450	0.0042	0.0034	-0.0008
21	CID59513457	0.0043	0.0032	-0.0011
22	CID59513489	0.0031	0.0063	0.0032
23	CID59513494	0.0037	0.0033	-0.0004
24	CID59513499	0.0019	0.0013	-0.0006
25	CID59513503	0.0043	0.0030	-0.0013
26	CID59513512	0.0031	0.0033	0.0002
27	CID59771033	0.0062	0.0063	0.0001
28	CID59771039	0.0055	0.0059	0.0004
29	CID59771043	0.0061	0.0063	0.0002
30	CID59771054	0.0064	0.0063	-0.0001
31	CID59771063	0.0069	0.0059	-0.001
32	CID59771080	0.0059	0.0062	0.0003
33	CID59771093	0.0050	0.0061	0.0011
34	CID66860413	0.0043	0.0030	-0.0013
35	CID66860724	0.0043	0.0035	-0.0008
36	CID66860834	0.0049	0.0041	-0.0008
37	CID66860876	0.0042	0.0034	-0.0008
38	CID66860937	0.0041	0.0033	-0.0008
39	CID66860979	0.0019	0.0013	-0.0006
40	CID66860988	0.0052	0.0038	-0.0014
41	CID66861008	0.0042	0.0027	-0.0015
42	CID66861015	0.0052	0.0038	-0.0014
43	CID66861120	0.0042	0.0029	-0.0013
44	CID66861173	0.0036	0.0035	-0.0001
45	CID66861176	0.0043	0.0030	-0.0013
46	CID66861221	0.0031	0.0033	0.0002
47	CID66861241	0.0043	0.0032	-0.0011
48	CID66861266	0.0032	0.0033	0.0001

Table 2 continued

SN	Compound ID	pIC ₅₀ (MR)	pIC ₅₀ (PLS)	Residual IC ₅₀ (PLS-MLR)
49	CID66861321	0.0042	0.0027	-0.0015
50	CID66861330	0.0038	0.0033	-0.0005
51	CID66861408	0.0034	0.0035	0.0001
52	CID66861663	0.0018	0.0013	-0.0005
53	CID66861863	0.0042	0.0032	-0.001
54	CID66861928	0.0041	0.0031	-0.001
55	CID66861937	0.0037	0.0033	-0.0004
56	CID66862537	0.0042	0.0034	-0.0008
57	CID66862568	0.0031	0.0033	0.0002
58	CID66862596	0.0042	0.0032	-0.001
59	CID67010999	0.0048	0.0061	0.0013
60	CID67011494	0.0064	0.0063	-0.0001
61	CID67120646	0.0052	0.0062	0.001
62	CID67478780	0.0041	0.0040	-0.0001
63	CID68223283	0.0060	0.0063	0.0003
64	CID68223466	0.0025	0.0039	0.0014
65	CID68311349	0.0031	0.0067	0.0036
66	CID69234998	0.0066	0.0062	-0.0004
67	CID69310201	0.0056	0.0061	0.0005
68	CID69319961	0.0036	0.0056	0.002
69	CID69322713	0.0036	0.0059	0.0023
70	CID69610495	0.0068	0.0035	-0.0033
71	CID69803743	0.0040	0.0059	0.0019
72	CID69850060	0.0022	0.0063	0.0041
73	CID69850430	0.0054	0.0058	0.0004
74	CID69850486	0.0040	0.0064	0.0024
75	CID69850630	0.0042	0.0063	0.0021
76	CID69851424	0.0063	0.0064	0.0001
77	CID69852302	0.0024	0.0061	0.0037
78	CID69852333	0.0029	0.0061	0.0032
79	CID69852653	0.0042	0.0060	0.0018
80	CID69853144	0.0028	0.0060	0.0032
81	CID69861341	0.0043	0.0061	0.0018

Q-Squared value is 0.932 and the RSS value for cross validation is 3.936E-006. The R-Square and Q-Square values closer to unity give a conclusion that the training set is very well described by the regression equation, which is statistically very significant and the cross-validation showed that the constructed model is very efficient for predicting the IC₅₀ values. On the other hand, the generated equation for the PLS is given below:

$$\begin{aligned} \text{pIC}_{50} = & 0.3857 * \text{tp1} + 1.141 * \text{tp2} + 0.2949 * \text{tp3} \\ & + 0.3149 * \text{tp4} + 0.4892 * \text{tp5} \end{aligned} \quad (2)$$

The regression quality of the optimized model generated using PLS is described with a multiple R-Squared of 0.9578 and an RSS value of 2.531E-006. The multiple

Q-Squared value is 0.8793, and the RSS value for cross validation is 7.242E-006 with a skewness of -0.4975, which describes that the training set is very well, and the cross validation showed that the constructed model is efficient for predicting the IC₅₀ values.

The IC₅₀ experimental and predicted using MR and PLS is shown in Table 1 and also the regression graph is shown in Fig. 1 for multiple regression and Fig. 2 for partial least square method.

3.2 IC₅₀ prediction

Applying both the optimized QSAR models, the IC₅₀ values for the 55 test set compounds were predicted, which is

Table 3 Molecular docking score of the test set and training set compounds

Ligand	MolDock score ^a	Rerank score ^b	Interaction ^c	Internal ^d	HBond ^e	LE1 ^f	LE3 ^g
CID69851424 (test)	-98.97	-74.99	-110.52	11.55	-10.05	-3.81	-2.88
CID66861330 (test)	-96.26	-68.85	-124.05	27.79	-1.69	-2.83	-2.03
CID54714968 (test)	-96.19	-66.93	-99.87	3.68	-7.49	-3.01	-2.09
CID25263165 (training)	-83.19	-64.36	-115.10	31.91	-2.38	-2.45	-1.89
CID66861221 (test)	-90.47	-61.40	-120.37	29.90	0.00	-2.58	-1.75
CID54682245 (training)	-81.65	-60.91	-108.87	27.21	-6.25	-2.63	-1.96
CID54698277 (training)	-67.83	-60.18	-95.26	27.44	-3.94	-2.61	-2.31
CID68223283 (test)	-81.26	-59.14	-93.18	11.91	-4.28	-2.71	-1.97
CID69322713 (test)	-77.40	-55.48	-95.70	18.30	-4.43	-2.87	-2.05
CID54709842 (training)	-79.65	-55.43	-95.07	15.42	-5.00	-2.41	-1.68
CID54712049 (test)	-70.55	-54.22	-95.50	24.96	-5.21	-2.61	-2.01
CID54684210 (test)	-89.00	-53.81	-107.41	18.41	-9.73	-2.70	-1.63
CID54712049 (training)	-71.10	-52.77	-83.95	12.85	-8.00	-2.63	-1.95
CID54712035 (training)	-87.18	-51.80	-95.16	7.98	-3.99	-2.81	-1.67
CID54709833 (training)	-72.74	-46.66	-87.04	14.30	-1.46	-2.27	-1.46
CID54709830 (training)	-69.86	-45.25	-79.40	9.54	-8.40	-2.12	-1.37
CID11166962 (training)	-77.81	-44.14	-109.65	31.84	-4.00	-2.29	-1.30
CID67478780 (test)	-65.94	-43.59	-73.64	7.70	-11.05	-2.36	-1.56
Raltegravir	-74.35	-34.18	-65.34	-9.01	-0.07	-2.32	-1.07

^a Moldock score is derived from the PLP scoring functions with a new hydrogen bonding term and new charge schemes (Thomsen and Christensen 2006)

^b The Rerank score is a linear combination of E-inter (steric, Van der Waals, hydrogen bonding, electrostatic) between the ligand and the protein, and E-intra (torsion, sp²-sp², hydrogen bonding, Van der Waals, electrostatic) of the ligand weighted by pre-defined coefficients Weiss (1993)

^c The total interaction energy between the pose and the protein (kJ mol⁻¹)

^d The internal energy of the pose

^e Hydrogen bonding energy (kJ mol⁻¹)

^f Ligand efficiency 1: MolDock score divided by heavy atoms count

^g Ligand efficiency 3: Rerank score divided by heavy atoms count

Table 4 Molecular interaction analysis of the top docking hits

SN	Compound ID	Ligand-protein interaction	Interaction energy (kJ mol ⁻¹)	Interaction distance (Å)
1	CID69851424	O(1)—OD1(Asp64)	-0.81	2.4
		N(6)—N(Cys65)	-2.1	3.19
		O(1)—OG1(Thr115)	-0.78	3.44
		O(2)—OG1(Thr115)	-1.5	3.31
		O(1)—N(Asp116)	-0.75	3.02
		O(2)—ND2(Asp120)	-2.5	2.68
2	CID66861330	N(7)—OD1(Asp64)	-2.82	2.82
		O(2)—OD2(Asp64)	-2.5	2.66
3	CID54714968	N(8)—OD1(Asp116)	-1.5	3.30
		O(2)—ND2(Asn155)	-2.5	3.10

shown in Table 2. From Table 2, it is revealed that there is not much variation in the IC₅₀ values predicted applying MLR and PLS methods. Hence, confirmed the reliability of both the model for the IC₅₀ value prediction.

3.3 Molecular docking analysis

Molecular docking simulation was carried out using MVD for the test set and training set compounds using MVD.

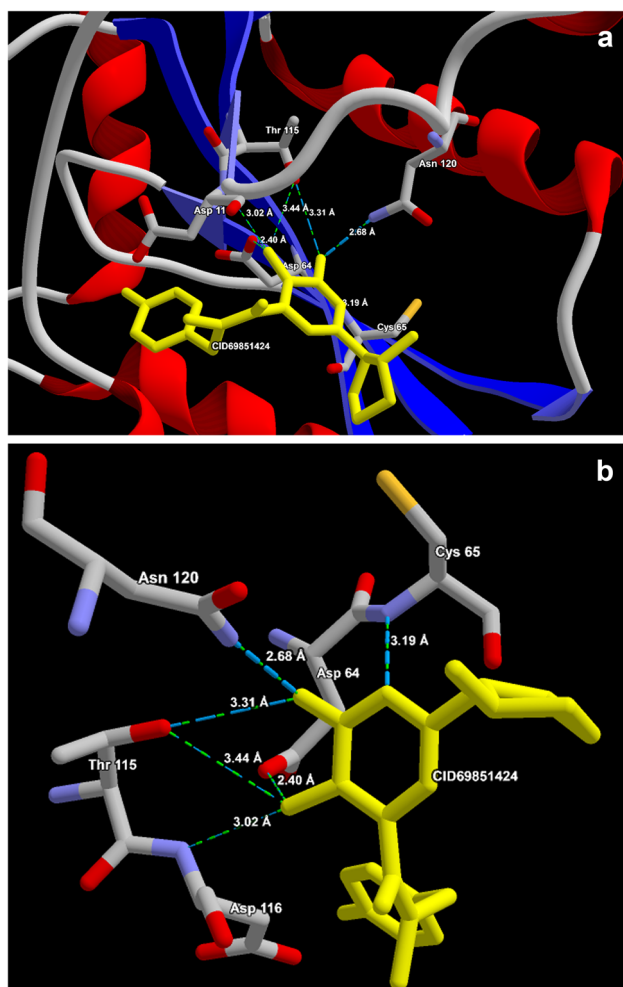


Fig. 3 **a** Binding mode of CID69851424 (yellow) to 4DMN active site region. **b** Predicted bonded interactions (dashed lines) between CID69851424 and Asp64, Cys65, Thr115, Asp116, and Asp120 residues

MVD is a molecular visualization and molecular docking software based on a differential evolution algorithm; the solution of the algorithm takes into account the sum of the intermolecular interaction energy between the ligand and the protein and the intramolecular interaction energy of the ligand. The docking energy scoring function is based on the modified piecewise linear potential (PLP) with new hydrogen bonding and electrostatic terms included. Full description of the algorithm and its reliability as compared to the other common docking algorithm can be found in the literature (Thomsen and Christensen 2006). The molecular docking score indicates that the test set compounds have favorable Rerank score than the training set as well as from the experimentally known integrase inhibitors raltegravir (Table 3). The test sets, compound, viz., CID69851424, CID66861330, and CID54714968 have Rerank score of -74.99 , -68.85 , and 66.93 , respectively. While CID25263165 (training set) has a Rerank score of -66.93

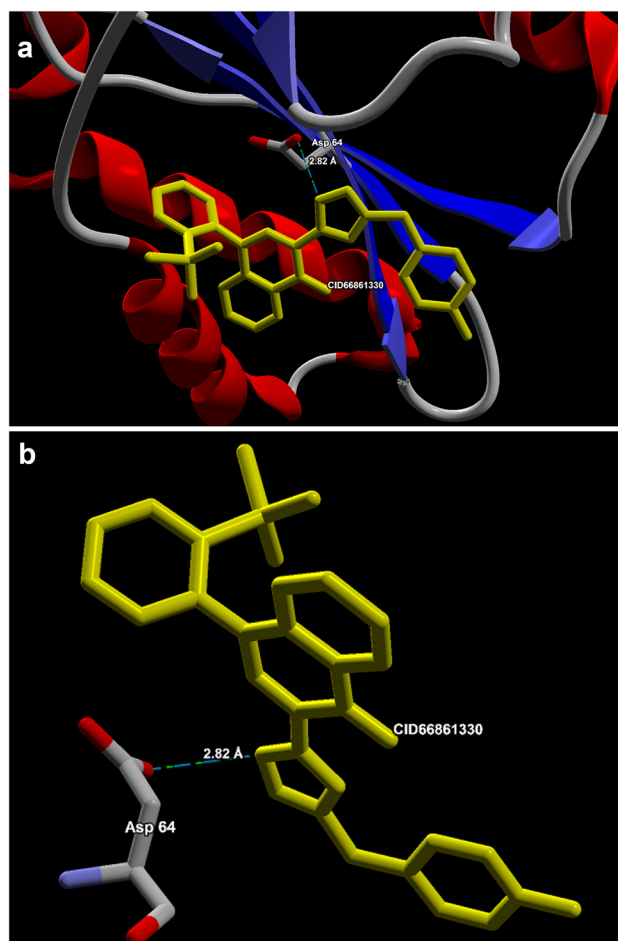


Fig. 4 **a** Binding mode of CID_66861330 (yellow) to 4DMN active site region. **b** Predicted bonded interactions (dashed lines) between CID_66861330 and Asp64 residues

compared to -34.18 of raltegravir. The Rerank score is a weighted combination of the terms used by the MolDock score mixed with a few addition terms. The Rerank Score includes the Steric term which is the Lennard–Jones approximations to the steric energy, while the MolDock score uses a piecewise linear potential to approximate the steric energy. The coefficients for the weighted Rerank Score are given in the Rerank Weight column, and the weighted terms and their summations are given in the Rerank Score column. The Reranking score function is computationally more expensive than the scoring function used during the docking simulation, but it is generally better than the docking score function at determining the best pose among several poses originating from the same ligand (Thomsen and Christensen 2006).

In addition, CID69851424 and CID54714968 possessed a hydrogen bonding energy of -10.05 kJ mol^{-1} and -7.49 kJ mol^{-1} , respectively, which accounts for the attractive intermolecular binding affinity and binding interaction. To understand the in-depth molecular

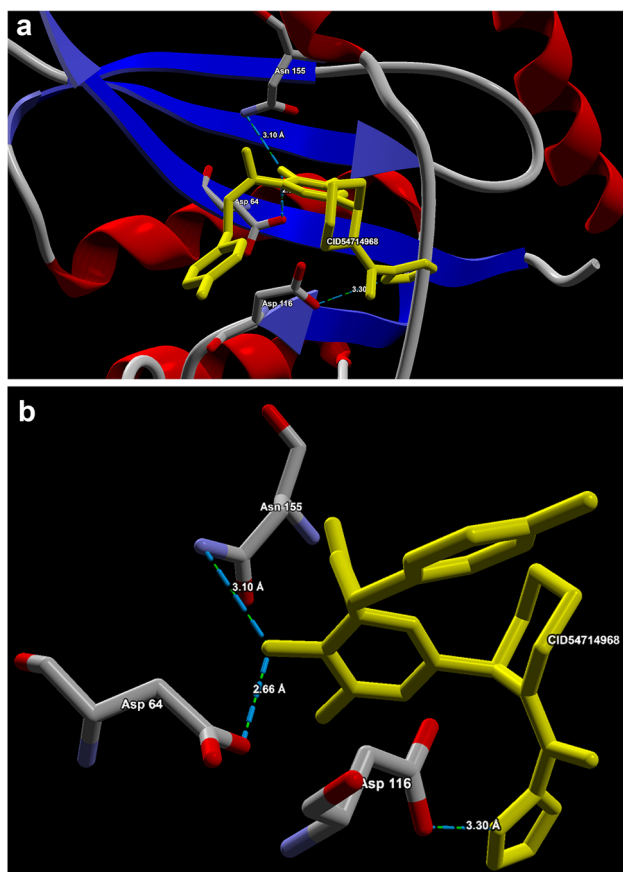


Fig. 5 **a** Binding mode of CID_54714968 (yellow) to 4DMN active site region. **b** Predicted bonded interactions (dashed lines) between CID_54714968 and Asp64, Asp116, and Asn155 residues

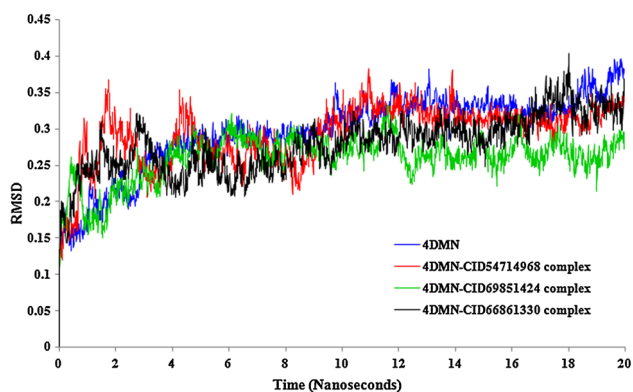


Fig. 6 MD Simulation analysis depicting the RMSD graph for 4DMN and 4DMN-ligand complexes

interaction, ligand–protein interaction analysis was carried out for the CID69851424, CID66861330, and CID54714968CID. The interaction analysis revealed the compounds have interaction with Asp64 (shown in Table 4) which is an active site residue. The snaps of the ligand–protein interaction analysis are shown in Figs. 3, 4, and 5 for CID69851424, CID66861330, and CID54714968CID, respectively.

3.4 Molecular dynamics simulation

Molecular dynamics simulation was performed for 20 ns using Gromacs 5.0.1 to understand the conformational changes of the protein–ligand binding complex and the protein to understand the amendment occurring in the dynamic environment. The RMSD plot for 4DMN-CID69851424 complex, 4DMN-CID66861330 complex, 4DMN-CID54714968 complex, and 4DMN is shown in Fig. 6. The RMSD plot clearly explains the variations of the HIV-1 integrase enzyme and protein–ligand binding complexes. The average RMSD showed ~ 0.27 Å for 4DMN-CID66861330 complex, ~ 0.26 Å for 4DMN-CID69851424 complex, and ~ 0.29 Å for 4DMN-CID54714968 complex, respectively, while the HIV-1 integrase enzyme (4DMN) showed an average RMSD deviation of ~ 0.30 Å, which is an indication revealing that the protein–ligand complex is more stable in the dynamic equilibrium condition. Furthermore, the dynamics simulation revealed the 4DMN-CID69851424 complex is more stable than the 4DMN-CID66861330 complex and 4DMN-CID54714968 complex in the dynamic environment.

3.5 ADME–toxicity studies

From the ADME–toxicity analysis calculated using ACD/I-Lab 2.0 (Advanced Chemistry Development, Inc 1994), the three top docking hits from the test set compounds are readily soluble in water compared to the market available drug raltegravir (shown in Table 5). For absorption, CID66861330 and CID69851424 showed 100 and 53% passive absorption, respectively, compared to raltegravir (7%). Moreover, the apparent volume of distribution of these three compounds is higher than that of raltegravir (Table 5). The comparative graph plot on LD₅₀ mouse and probability of health effect is shown in Figs. 7 and 8, respectively. The LD₅₀ mouse oral indicates these compounds can be administered orally with more or less similar values with raltegravir. Last, the probability of health effects depicted CID69851424 is likely to have an overall low impact on health effects. Overall, the ADME toxicity analysis comes to a conclusion that these compounds in general have enhanced pharmacological parameters than the market approved drug raltegravir.

4 Conclusion

IC₅₀ values were predicted for a set of HIV-1 integrase inhibitors employing QSAR models generated using multiple regression and partial least square method. The prediction indicated that there is not much variation between these two methods, and cross validation confirms that these

Table 5 Predicted ADME–toxicity parameters for CID69851424, CID66861330, CID54714968, and market approved raltegravir

ADME–toxicity parameters	CID69851424	CID66861330	CID54714968	Raltegravir
Solubility H ₂ O (mg mL ⁻¹) ^a	−0.81	−5.14	−1.88	−2.03
BBB Log PS ^b	−3.8	−1.8	−5.4	−4.8
BBB Log PB ^b	−0.36	−0.75	−2	−0.26
BBB Log(PS*fu, brain) ^b	−3.9	−2.6	−5.4	−4.8
% Oral bioavailability ^c	>30	30–70	>30%	>30
Absorption (% passive absorption) ^d	53	100	3	7
Absorption (permeability, cm s ⁻¹) ^d	0.21 × 10 ⁻⁴	7.12 × 10 ⁻⁴	0.01 × 10 ⁻⁴	0.02 × 10 ⁻⁴
Distribution (L kg ⁻¹) ^e	0.46	1.6	0.37	0.7
LD 50 mouse (mg kg ⁻¹ , intraperitoneal) ^f	630	670	730	630
LD 50 mouse (mg kg ⁻¹ , oral) ^f	560	890	980	990
LD 50 mouse (mg kg ⁻¹ , intravenous) ^f	130	71	130	530
LD 50 mouse (mg kg ⁻¹ , subcutaneous) ^f	580	130	1200	210
Prob. of blood effect ^g	0.61	0.86	0.73	0.53
Prob. of cardiovascular system effect ^g	0.19	0.94	0.35	0.29
Prob. of gastrointestinal system effect ^g	0.46	0.99	0.96	0.83
Prob. of kidney effect ^g	0.41	0.68	0.72	0.28
Prob. of liver effect ^g	0.59	0.34	0.87	0.31
Prob. of lung effect ^g	0.39	0.38	0.51	0.89

^a Calculates compound's solubility in a buffer at a specified pH value

^b Calculates the blood brain barrier (BBB) transport (LogPS, Rate of brain penetration; LogPB, Extent of brain penetration; Log (PS*fu, brain) Brain/plasma equilibration rate)

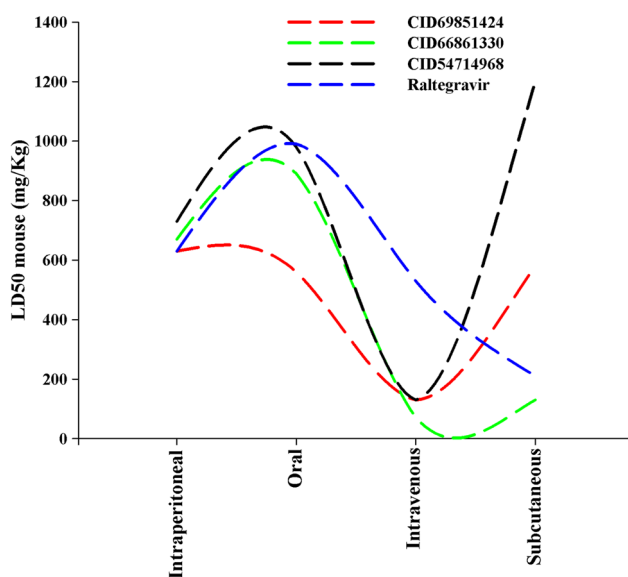
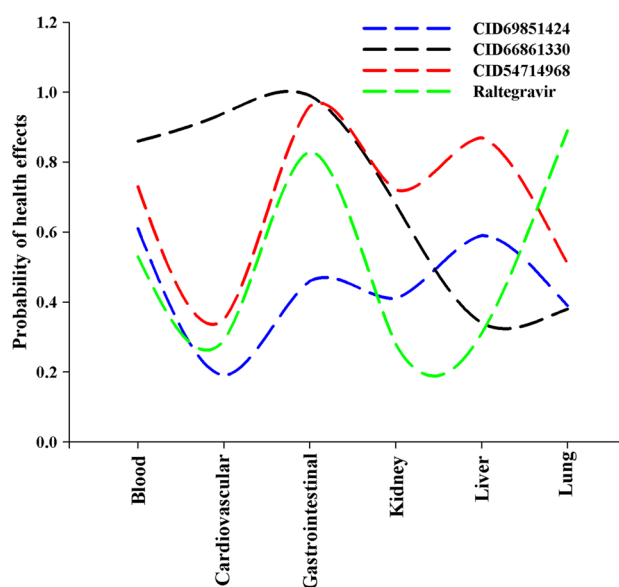
^c Estimates the probability of a compounds bioavailability being above 30 and 70%

^d Estimates maximum passive absorption and human jejunum permeability

^e Calculates the apparent volume of distribution for a compound in L kg⁻¹

^f Estimates LD₅₀ value in mg kg⁻¹ after intraperitoneal, oral, intravenous and subcutaneous administration to mice

^g Estimates probability of blood, gastrointestinal system, kidney, liver and lung effect at therapeutic dose range

**Fig. 7** Comparative analysis on LD₅₀ mouse (intraperitoneal, oral, intravenous, subcutaneous) for CID69851424, CID66861330, CID54714968, and raltegravir**Fig. 8** Comparative analysis on probability of health effect on blood, cardiovascular system, gastrointestinal system, kidney, liver, and lung for CID69851424, CID66861330, CID54714968, and raltegravir

models can be used to predict the IC₅₀ values. Furthermore, a molecular docking analysis against HIV-1 integrase showed the compounds have favorable molecular interaction than the market approved raltegravir. Furthermore, the molecular docking simulation was validated by molecular dynamics simulation for 20 ns, where the docked ligand–protein complex showed more stability than the HIV-integrase enzyme. Last, ADME–toxicity prediction confirms that the docked compounds have enhanced pharmacological parameters than raltegravir. Hence, the authors proposed for the clinical trial of these reported compounds, viz., CID69851424, CID66861330, and CID54714968.

Acknowledgements The authors acknowledged the Department of Biotechnology, Ministry of Science and Technology, Government of India, New Delhi for necessary support.

Compliance with ethical standards

Conflict of interest The authors declare that no conflicts of interest exist.

References

- Advanced Chemistry Development, Inc: ACD/I-Lab, Version 2.0 (1994). Toronto, ON, Canada
- Alavijeh MS, Chishty M, Qaiser MZ, Palmer AM (2005) Drug metabolism and pharmacokinetics, the blood–brain barrier, and central nervous system drug discovery. *NeuroRx*. 2:554–571
- Anthony NJ (2004) HIV-1 integrase: a target for new AIDS chemotherapeutics. *Curr Top Med Chem* 4:979–990
- Bolton E, Wang Y, Thiessen PA, Bryant SH (2008) PubChem: integrated platform of small molecules and biological activities. Chapter 12. In: *Annual Reports in Computational Chemistry*, vol 4. American Chemical Society, Washington
- ChemOffice (2010) CambridgeSoft Corporation, 875 Massachusetts Ave, Cambridge, MA 02139, USA
- Douek DC, Roederer M, Koup RA (2009) Emerging concepts in the immunopathogenesis of AIDS. *Ann Rev Med* 60:471–484
- Fikkert V, Van Maele B, Vercammen J, Hantson A, Van Remoortel B, Michiels M, Gurnari C, Pannecouque C, De Maeyer M, Engelborghs Y, De Clercq E, Debyser Z, Witvrouw M (2003) Development of resistance against diketo derivatives of human immunodeficiency virus type 1 by progressive accumulation of integrase mutations. *Science* 287:646–650
- Fisher RA (1922) The goodness of fit of regression formulae and the distribution of regression coefficients. *J R Stat Soc*. 85:597–612
- Ghasemi JB, Shiri F, Pirhadi S, Heidari Z (2015) Discovery of new potential antimalarial compounds using virtual screening of ZINC database. *Comb Chem High Throughput Screen*. 18:227–234
- Grobler JA, Stillmock K, Hu B, Witmer M, Felock P, Espeseth AS, Wolfe A, Egbertson M, Bourgeois M, Melamed J, Wai JS, Young S, Vacca J, Hazuda DJ (2002) Diketo acid inhibitor mechanism and HIV-1 integrase: implications for metal binding in the active site of phosphotransferase enzymes. *Proc Natl Acad Sci USA* 99:6661–6666
- Iidiko FE, Friedman JH (1993) A statistical view of some chemometrics regression tools. *Technometrics* 35:109–148
- Imamichi T (2004) Action of anti-HIV drugs and resistance: reverse transcriptase inhibitors and protease inhibitors. *Curr Pharm Des* 10:4039–4053
- Johnson AA, Marchand C, Pommier Y (2004) HIV-1 integrase inhibitors: a decade of research and two drugs in clinical trial. *Curr Top Med Chem* 4:1059–1077
- Kessl JJ, Jena N, Koh Y, Taskent-Sezgin H, Slaughter A, Feng L, de Silva S, Wu L, Le Grice SF, Engelman A, Fuchs JR, Kvaratskhelia M (2012) Multimode, cooperative mechanism of action of allosteric HIV-1 integrase inhibitors. *J Biol Chem* 287:16801–16811
- Lipinski CA (2000) Drug-like properties and the causes of poor solubility and poor permeability. *J Pharmacol Toxicol Methods* 44:235–249
- Malet I, Delelis O, Valantin MA, Montes B, Soulie C, Wirden M, Tchertanov L, Peytavin G, Reynes J, Mouscadet JF, Katlama C, Calvez V, Marcelin AG (2008) Mutations associated with failure of raltegravir treatment affect integrase sensitivity to the inhibitor in vitro. *Antimicrob Agents Chemother* 52:1351–1358
- MDL QSAR (2006) MDL Information Systems Inc, USA
- Molegro APS (2011) MVD 5.0 Molegro Virtual Docker. DK-8000 Aarhus C, Denmark
- Murray JM, Emery S, Kelleher AD, Law M, Chen J, Hazuda DJ, Nguyen BY, Tepler H, Cooper DA (2007) Antiretroviral therapy with the integrase inhibitor raltegravir alters decay kinetics of HIV, significantly reducing the second phase. *Aids* 21:2315–2321
- Thomsen R, Christensen MH (2006) MolDock: a new technique for high-accuracy molecular docking. *J Med Chem* 49:3315–3321
- Ulrich B, Norman LA (1982) ACS Monograph 177. American Chemical Society, Washington
- Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ (2005) GROMACS: fast, flexible, and free. *J Comput Chem* 26:1701–1718
- Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Zhou Z, Han L, Karapetyan K, Dracheva S, Shoemaker BA, Bolton E, Gindulyte A, Bryant SH (2012) PubChem's bioassay database. *Nucleic Acids Res* 40:D400–D4012
- Weiss RA (1993) How does HIV cause AIDS? *Science* 260:1273–1279