Scientific Research Publishing

Computational Chemistry, 2025, 13(1), 1-31
https://www.scirp.org/journal/cc
ISSN Online: 2332-5984
ISSN Print: 2332-5968

QSAR Studies on HIV-1 Protease—A Battle against HIV Using Computational Chemistry

Pokala Sai Kovela Sanjana[1,2], Radhika Vangala[2], Sree Kanth Sivan[2*]

[1]Department of Chemistry, St. Francis College for Women, Hyderabad, India
[2]Department of Chemistry, University College for Women, Hyderabad, India
Email: *sanjanapokala@gmail.com

How to cite this paper: Sanjana, P.S.K., Vangala, R. and Sivan, S.K. (2025) QSAR Studies on HIV-1 Protease—A Battle against HIV Using Computational Chemistry. *Computational Chemistry*, **13**, 1-31.
https://doi.org/10.4236/cc.2025.131001

Received: December 5, 2024
Accepted: January 24, 2025
Published: January 27, 2025

Abstract

HIV/AIDS is a major public health problem. Despite the advances in HIV treatment, the cure for all HIV patients still possesses a major challenge, which needs to be surpassed in coming years. One attractive target is HIV-1 Protease. Herein we report a new series of HIV-1 Protease Inhibitors incorporating stereo chemically defined tetrahydrofuran-tertiary amine-acetamide P2 ligand. Structure activity relationship studies like 2D QSAR, ADME studies and Molecular Docking were performed to design a chemical entity. A total of 37 compounds were chosen for this study divided into Training and Test set. A total of 27 training set molecules were taken for the SAR analysis (2D QSAR). The model was seen to have an incredible Correlation coefficient (R = 0.942) and also exhibited good predictive power confirmed by the high value of cross validated correlation coefficient (Q2 = 0.701). The outcome of this study gives us an insight for designing novel and peculiar HIV-1 Protease Inhibitors and provides us a guideline for designing compounds with improved novel HIV virus inhibitory potential.

Keywords

HIV, HIV-1 Protease, ADME, MLR, 2D QSAR, Molecular Docking

1. Introduction

1.1. HIV (Human Immunodeficiency Virus)

HIV (Human Immunodeficiency Virus) is a virus that attacks the body's immune system, called CD4 cells which help the body's response to infection. Within the CD4 cells, the HIV virus replicates and thus damages and destroys the cell. If HIV is not treated, it can lead to AIDS (acquired immunodeficiency syndrome) [1].

HIV infection in humas has come from a Chimpanzee in Central Africa. SIV (Simian Immunodeficiency Virus) is a Chimpanzee version of virus transmitted to humans when they hunted for this Chimpanzee's and encountered this in their infected blood. Over the decades, HIV slowly spread across Africa and later into other parts of the world [2].

HIV is found in certain bodily fluids of people including blood, semen, vagina fluids, rectal fluids and breast milk. HIV can be transmitted by i) Unprotected sexual intercourse with a person living with HIV, ii) Blood transfusion of contaminated blood, iii) Sharing of needles, syringes, other infected equipment, surgical equipment or other sharp instruments, iv) From a mother living with HIV to her infant during pregnancy, childbirth and breast feeding [3].

1.2. Structure of HIV

The two HIV viruses HIV-1 and HIV-2 are members of family of Retroviruses in the genus of Lentivirus. HIV virus is roughly spherical in diameter of 120 nm. It is composed of 2 copies of positive sense single stranded RNA enclosed by capsid which is composed of viral protein P24. Inside the capsid three enzymes required for HIV replication—Reverse transcriptase, Integrase, Protease are present. A matrix composed of viral protein P17 surrounds the capsid. Matrix is surrounded by phospholipids. The glycoproteins layer has 2 spikes G41 and G120 [4].

1.3. Pathophysiology

HIV attaches and penetrates the host T cells through CD4 molecules releasing HIV, RNA and enzymes into host cell. HIV reverse transcriptase copies the viral RNA as proviral DNA. This enters the nucleus of host cell and enzyme integrase facilitates the integration of proviral DNA into host DNA. The host cell now produces HIV, RNA and HIV proteins. These then assemble into HIV virions. HIV protease cleaves the viral proteins converting the immature virion to a mature infectious virion [5]. When HIV is not treated the disease typically will progress through three stages (Table 1).

Table 1. Stages of HIV Infection [6].

Stage 1—Acute HIV Infection	Stage 2—Chronic HIV Infection	Stage 3—AIDS
• Large amount of HIV in blood. • Very contagious • Some have flu-like symptoms, body's natural response to infection • Some may not feel sick right away.	• This stage is called asymptomatic, HIV infection or clinical latency. • HIV is active at this stage but reproduces at low levels. • At end of this stage, the amount of blood is high while the CD4 cell count is low and the person moves to stage 3.	• Most severe stage of HIV infection. • People suffering with AIDS have such damaged immune system that they can be affected with opportunistic infection. • Without receiving any treatment people with AIDS can survive for 3 years typically

1.4. Symptoms

Approximately 80% of people infected with HIV experience flu like illness that

occurs 2 - 6 weeks after infection. Most common symptoms are Raised temperature (fever), Sore throat, Body rash. Other symptoms include Tiredness, Joint pain, Muscle pain, Swollen glands. Once immune system is severely damaged, symptoms can include weight loss, chronic diarrhea, Night sweat, Skin problems, recurrent infection, serious life-threatening illnesses. Post disappearance of initial symptoms, HIV may not cause any symptoms for many years, but the virus continues to be active and causes damage to the immune system. This process varies in different ways and may take upto 10 years during which an individual will feel and appear normal and well [7].

1.5. Diagnosis and Treatment

HIV is diagnosed through blood or saliva using Antigen-Antibody tests, Antibody tests, Nucleic acid tests (NATs), CD4 T cell count, Drug resistance. Diagnosis of HIV in stage 3 can be done by presence of opportunistic infection—Fungal diseases, Candidiasis, Histoplasmia, Recurrent pneumonia, and TB [8].

Anti-Retroviral therapy (ART) is the most effective treatment. It is a combination of several medicines that aim to control the number of viruses in the body. Goals for the medicine are—Control the growth of virus, improve working of immune system, Slow or stop the symptoms, Prevent transmission of HIV to others. The most common drugs used in ART are—Nucleotide reverse transcriptase inhibitors (NRTI's), Protease inhibitors (PI), Integrase inhibitors [9].

1.6. Epidemiology

HIV is a major global public health issue having claimed almost 33 million lives so far. Clinically AIDS was first recognized in US in 1981. In 1983, HIV was discovered to be the causative agent of AIDS. Since then, the number of HIV cases has increased in both US and other countries. The Center for Disease control and prevention estimated that around 1.2 million people aged 13 years and older are living with HIV infection. The discovery of antiviral drug therapy in 1996 has resulted in decreased number of deaths due to AIDS, among people receiving the therapy. The medication is expensive and requires strict dosing schedules. In developing countries many suffering with HIV do not have access to newer drug therapies [10].

1.7. HIV-1 Protease

HIV-1 protease is a dimeric enzyme from the family of aspartic proteases. The enzyme has been widely exploited as a drug target and exhibits broad substrate recognition. Each subunit is made up of nine β strands and a single α helix. It consists of subunits of 99 amino acids residues [11]. HIV-1 Protease is responsible for processing of the gag and gag-pol polyproteins during virion maturation. The activity of this enzyme is essential for virus infectivity, rendering the protein a therapeutic target for AIDS treatment. The HIV protease is unique that it cleaves between a phenylalanine and tyrosine or proline [12].

1.8. Computer Aided Drug Design (CADD)

Introducing a new drug in a market is a complex, risky and costly process in terms of time, money and manpower. Generally it is found that drug discovery and development process will take 10 - 14 years and more than 1 billion dollars. To reduce time, cost, risk-borne factors, computational approach in drug design is getting very rapid exploration, implementation and admiration [13]. CADD represents computation methods and resources used to facilitate the design and discovery of new therapeutic drugs. CADD methodologies and techniques are used to calculate molecular properties to aid in drug design process. These methods use 3D structure of a ligand, a target and then design new compound as well as construct large combinational libraries of compounds that can be screened computationally before attempting to synthesize and test them [14]. CADD makes used of structural knowledge of either the target (Structure-based) or known ligands with bioactivity (Ligand-based) to facilitate the determination of promising candidate drugs [15].

Lead Identification

During drug discovery for a particular disease, the molecular mechanism behind the disease is studied. These studies include the identification of cellular, genetic factors involved in disease which is followed by identification of potential targets. For ensuring the involvement of a biological target which is involved both *in vitro* (cell) and *in vivo* (animals) tests are performed. This is also termed as Target Validation. The result of this stage aids in lead compound identification [16].

Lead Compounds

Chemical compounds that show desired biological or pharmacological activity and may initiate the development of a new clinically relevant compound called as lead compound. These compounds are typically used as starting points in discovery of drugs to produce new entities. Resulting compounds from drug design undergo a series of pre-clinical studies and become clinical candidates if they do not exhibit toxic properties followed by their release in the marker as a new drug entity. The possible sources of lead compounds and novel drugs include natural products—Plants, Animals, Microbes, Chemical libraries, Medicinal chemistry [17].

Lead Optimization

The most common promising lead compounds advance into the lead optimization stage of drug discovery. It is an extremely important process that culminates in the identification of pre clinical candidates. Goal of this stage is to extensively optimize biological properties of lead series through *in vitro* and *in vivo* assays [18]. This is done by pharmacokinetics parameters, ADMET—Absorption, Distribution, Metabolism, Excretion, Toxicity [19]. Whereas study of biochemical, physiologic, molecular effects of drugs on the body is called as pharmacodynamics. Pharmacodynamics with pharmacokinetics helps in explanation of relationship between dose and response [20].

1.9. Techniques used in CADD

There are two main techniques used in CADD are

- **Ligand Based Drug Design**—This includes:
  - ➢2D QSAR
  - ➢3D QSAR
- **Structure Based Drug Design**—This includes:
  - ➢Docking
  - ➢Virtual Screening
  - ➢Similarity Search

Ligand Based Drug Design

Ligand based drug design is also called Indirect drug design. This method is dependent on the awareness of different new ligand molecules that can bind to a target protein molecule. A target protein molecule is produced based on the compound ligand. Ligand based drug design relies on the information of molecules that bind to biological target active site with interest. This model of drug design is used in the development of novel compounds that interact with biologically active molecules. Quantitative Structure Activity Relationship (QSAR) is defined as a correlation between calculated properties of a molecule and the experimentally determined biological activity. The use of QSAR studies is done to predict the activity of new molecules [21].

Quantitative Structure Activity Relationship (QSAR)

Quantitative Structure Activity Relationship is said to be mathematical relationship in form of an equation between biological activity and measurable physiochemical parameters. QSAR attempts to identify and quantify the physiochemical properties of a drug and to assess whether any of these properties influences the drugs biological activity. To modify chemical structure of lead compound to reinforce the desirable pharmacologic effect while minimizing unwanted physical and chemical properties this may result in a therapeutic agent [22]. 2D QSAR is an extremely useful technique for explaining the relationships between chemical structures and experimental observations. 2D QSAR techniques are very often used during the process of optimization of a chemical series towards a candidate for clinical trials. The main elements of this method include:

- Numerical descriptors used to translate chemical structure into mathematical variables.
- Quality of the data observed.
- Statistical methods used for deriving the relationship between observations and descriptors [23].

Structure-Based Drug Design

Structure Based Drug Design is also called as Direct Method of drug design. In this method of drug design structure is initially identified by X-ray crystallography which improves the tendency to produce new drugs which can fight against new diseases. The lead molecule is synthesized after X-ray crystallography which is used to examine the structure of target protein bound to the unknown lead molecule.

The following steps are involved in Structure based drug design:

- Three-Dimensional structure of biological target.
- It is obtained through X-crystallography or NMR spectroscopy.
- In case the experimental structure is unavailable create a homology model of target based on the experimental structure of related protein.
- Various automated computational procedures may be used [24].

Docking

It is a method used to know the orientation of one molecule to other when they are bound to form a stable complex. The better orientation may then be used to identify the binding affinity between two molecules using scoring functions. Docking is one the most often used methods in structure-based drug designing because of its binding-confirmation of small molecule ligand to appropriate target binding site (**Figure 1**). Characterization of this binding behavior is particularly important in rational drug designing and to know fundamental biochemical processes [25].

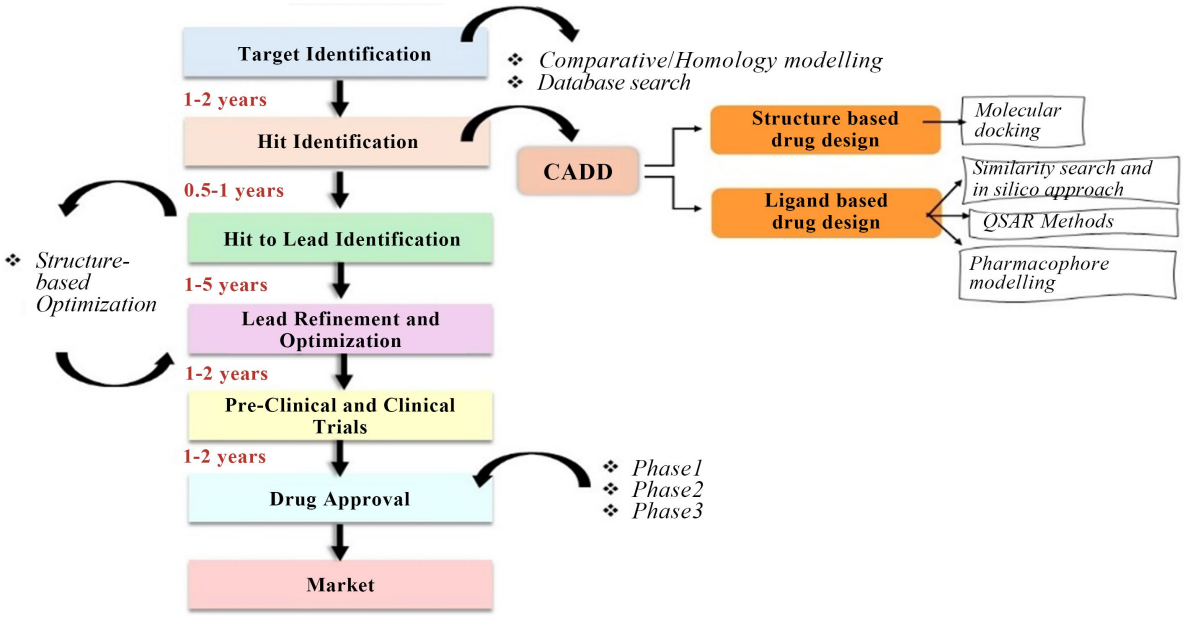


Figure 1. CADD workflow.

The chemical compounds investigated in this study were selected from the work of X. Bai *et al.* [26]. One effective design strategy of eliminating drug existence is to maximize interaction between inhibitor and the protease especially to promote extensive H bonds interaction with backbone atoms and residues in the protease active site. For example as mentioned in the work done by X. Bai *et al.*—For example, the stereo chemically defined P2-ligands bis-tetrahydrofuran (THF) of Darunavir (DRV), mono-THF of amprenavir (APV), tris-THF of GRL-0519 and Tp-THF of GRL-0476 introduce more hydrogen bonds through cyclic ether oxygens with main chain atoms of the protease residues in S2-subsite, such as amide NH of Asp29/Asp30. A number of potent inhibitors

have been reported with a substituted-phenyl-tertiary amine-acetamide structural template, splitting from the bis-THF of DRV by scaffold hopping method, as the P2-ligand to enhance interactions with the protein backbone in the S2-subsite.

One inhibitor containing N-(2,6-di-methylphenyl)-N-(2-morpholino-2-oxoethyl) acetamide group as P2-ligand along with a 4-methoxylphenylsulfonamide as P2'-ligand displayed a decent inhibitory activity with enzymatic IC50 of 15 nM. While the referenced study provided valuable insights into the chemical and biological properties of these compounds, computational analysis such as Ligand based drug design (2D QSAR) and Structure based drug design (Docking) were not explored. Recognizing the potential of these molecules we undertook CADD studies to further investigate their potential as HIV-1 protease inhibitors. This complementary approach aims to provide deeper insights into their molecular behavior and identify potential of these for treatment of HIV.

In this study, the HIV-1 Protease was combined with stereo chemically specified tetrahydrofuran tertiary amine-acetamide P2 ligand inhibitors, and QSAR and docking investigations were carried out on them.

2. Materials and Methods

2.1. Chem Sketch

Advanced Chemistry Department (ACD)/Chem Sketch freeware is a drawing package that allows us to draw chemical structures including organics, organometallics, polymers and Markush structures. It is a comprehensive structure editor with a variety of tools and functionality that ease the communication of scientific and chemical information. This software can also be used to generate names from molecular structure, calculate molecular properties from chemical structure and for 2D and 3D structure drawing and viewing. It can be used to convert 2-Dimensional drawings into 3-Dimensional wire frame structures using a modified molecular mechanics approach, and these pictures can be rotated and moved in three dimensions [27].

From button bars users may choose an array of bond types. Left hand buttons consist of chemical element symbols to select any element of periodic table. On the right-hand side of the drawing area are buttons to select sub structures such as rings, chains of various lengths, common groups such as COOH and amino acids [28]. The structures of potent inhibitors of HIV-1 protease were drawn using Chem Sketch.

2.2. Data Set Collection

A set comprising forty-two stereo chemically defined tetrahydrofuran-tertiary amine acetamide P2-ligand with HIV-1 protease inhibitory activity was utilized for QSAR analysis. Inhibitory concentrations (IC50) of compounds included in a data set varied and were changed over into corresponding pIC50 values utilizing the formula given underneath.

$$pIC50 = -\log 10\ [IC50]$$

2.3. SMILES

SMILES (Simplified Molecular Input Line Entry System) are a chemical notation that allows a user to represent a chemical notation that allows a user to represent a chemical structure in a way that can be used by the computer. The notation consists of a series of characters containing no spaces. Smiles system was designed to be truly computer interactive. SMILES is interpreted in a fast compact manner, thereby satisfying the machine objectives of time and space saving. Thus, resulting in a great improvement in the efficiency of information processing as compared to conventional methods. It is an easily learned and a flexible notation. Smiles have five basic syntax rules that must be followed. There are five generic Smiles encoding rules, corresponding to specification of atoms, bonds, branches, ring closures and disconnection [29] [30].

Forty-two stereo chemically defined tetrahydrofuran-tertiary amine acetamide P2-ligand molecules with HIV-1 protease inhibitory activity were converted to Smiles notation.

2.4. Descirptor Calculation Using pkCSM

pkCSM is a novel method for predicting and optimizing small-molecule pharmacokinetic and toxicity properties that relies on distance-based graph signatures. pkCSM is a user-friendly web server that enables researchers to freely predict ADMET properties for molecules of interest [31].

LogP, the Partition coefficient or Molecular lipophilicity is an important molecular characteristic in CADD. It estimates the lipophilicity of chemical compounds. It is obtained by measuring the partitioning of the molecule between an aqueous phase and a lipophilic phase [32].

Skin being a main barrier between internal body and external environment and thus its permeability is widely recognized as an essential parameter that must be considered for delivery of active substances like drugs [33].

CNS Permeability is an important parameter. BBB permeability is the ability of the drug to cross the blood-brain barrier surrounding the brain measured in-vivo condition is an important parameter for decreasing the side effects of drugs [34].

LogP, Water Solubility, Skin Permeability, CNS Permeability, BBB Permeability, Surface area and Surface tension were calculated. The ADME properties and descriptors of the ligands are determined using pkCSM.

2.5. 2D QSAR Model Development

The biological activity of dataset molecules *i.e.*, pIC50 of the datas*et al*ongside their calculated descriptors were imported into Build QSAR spread sheet and this 2D QSAR model was developed utilizing Build QSAR software. A correlation analysis is carried out between the calculated descriptors (an independent variable, X) and biological activity (dependent variable, Y) to comprehend the relation

between the two.

2.6. 2D QSAR Model Development Using MLR

The dataset of molecule was partitioned into training and test set molecule. This was done to ensure that models developed are reliable, generalizable and unbiased. The molecules were thus split randomly to ensure representative sample of data is included in both sets after which only the training set of molecules were used for MLR analysis. pIC50 is considered as dependent variable and different molecular descriptors were used as independent variable. A multi linear regression was carried out by selecting different descriptor combinations. The best model was chosen depending upon the statistical parameters such as correlation coefficient R and regression coefficient Q2.

2.7. Calculation of Predicted Activity

The data set molecules were partitioned into training set of 27 molecules and test set molecules of 10 molecules. The training set molecules were utilized for Multiple Linear Regression (MLR) analysis method choosing 10 out of 16 different molecular descriptors that include LogP, Surface Area, Water Solubility, $CaCO_2$ Permeability, VDss (human), Skin Permeability, BBB Permeability, CNS Permeability, Index of Refraction and Surface Tension. This equation was used to calculate the predicted activity of test set molecules.

2.8. Calculation of SD (Standard Deviation)

Standard deviation is a statistic that measures the dispersion of a dataset relative to its mean and is calculated as square root of variance [35]. SD of the test set was determined.

- STEP 1—The mean activity of training was calculated using the formula
  = sum (pIC50)/Total number of molecules in training set
- STEP 2—The SD of test set was calculated using the formula
  $= \sqrt{\sum (y_i - \hat{y}_i) 2/n}$

2.9. Calculation of PRESS

PRESS is the predictive residual sum of squares or the predicted extra sum of squares. It is the sum of overall compounds of the squared difference between the actual and predicted values for independent variables [36].

The PRESS value of test set was calculated using the formula

- $= \sum (\text{Ypred} - \text{Y obs}) 2$

2.10. Calculation of R_2

The extent to which the proposed relationship could explain the variance of biological activity is presented with R_2, the coefficient of determination [37]. R_2 value of the test set molecules was calculated using the formula

- $= \sum (\text{SD} - \text{PRESS})/\text{SD}$

2.11. Docking Methodlogy

2.11.1. Protein Preparation

The HIV-1 Protease (PDB ID-5yok) was prepared using the protein preparation wizard of *Schrödinger* suite. Hydrogen atoms were added to the protein. The missing side chains of residues were corrected using built interface incorporated in Maestro. For each structure optimization was carried out with Impact Refinement module, using the OPLS3 force field to alleviate steric clashes that may exist in the structures. Since this protein was associated with ligand, the ligand was selected to define the position and size of active site.

2.11.2. Ligand Preparation

The HIV-1 Protease incorporated with stereo chemically defined tetrahydrofuran tertiary amine-acetamide P2 ligand inhibitors were built in Schrodinger suite using Ligprep module using OPLS3 force field. The conceivable conformers of ligand were generated at physiological pH 7.0 ± 2.0. Indicated chiralities were retained to generate low energy conformers.

2.11.3. Receptor Grid Generation

A grid is generated for the protein which includes the specific binding/active site. The active site of a protein is the binding pocket of the protein where the ligands bind to show a specific activity and substrate specificity. Substrates bind to the active site of the enzyme or specificity pocket through hydrogen bonds, hydrophobic interactions, or a combination of all of these to form the enzyme-substrate complex.

Residues of the active site will act as donors or acceptors of protons or other groups for the substrate to facilitate the reaction.

A grid is generated using the receptor grid generation module of Glide software around crystal structure ligand. The scaling factor was set to 0.9. The scoring Grid was generated using the dimensions of 32 Å so that the ligands bind in that groove specifically.

2.11.4. Ligand Docking

Ligand docking is an important step which predicts the most stable low energy binding state between the ligand and the protein. Glide is used for docking to generate the possible conformers for each ligand and gives rise to possible orientations and positions over the active site for docking. The ligands are docked flexibly in SP (standard precision) mode. 37 ligand molecules were docked flexibly generating several poses. The ligand with the highest Glide score or lowest Dock Score is considered as the possible lead molecule.

3. Results and Discussion

2D QSAR: Two dimensional QSAR studies were carried out to determine the relation between molecular properties of molecules with HIV-1 Protease inhibitory activity. Table 2 shows the correlation matrix between selected descriptors. Table 3 the experimental and calculated pIC50 values for QSAR model is provided. A

total of 10 out of 15 descriptors were used to obtain the best fit. The best QSAR model built using multiple linear regression (MLR) method is represented by the following equation:

pIC50 = −0.5565 (±1.4204) LOGP − 0.0201 (±0.0172) Surface Area + 0.9165 (± 1.3641) Water Solubility − 1.3443 (± 0.9165) Caco2 Permeability − 66.2009 (± 30.6041) Skin Permeability − 3.0414 (± 1.5477) VDss (human) + 6.5167 (± 1.8499) BBB Permeability − 3.6362 (± 2.7842) CNS Permeability − 74.1118 (± 41.0647) Index of refraction + 0.1233 (± 0.0786) Surface Tension − 56.3965 (± 130.1698) (n = 27; R = 0.942; s = 0.300; F = 12.490; p < 0.0001; Q2 = 0.701; SPress = 0.486; SDEP = 0.381)

Where "n" is the number of observations, R' is the correlation coefficient, R_2 is the squared correlation coefficient, p is the statistical significance >99.9% with Fisher's statistic F and Q2 is Regression coefficient, SPRESS is the standard deviation of sum of squared error of prediction and SDEP is the standard deviation of error of prediction. A graph representing correlation of observed versus calculated activities using MLR is shown in Figure 2.

Table 2. Correlation matrix between selected descriptors.

COLUMN 1	pIC50	LogP	Surface Area	Water Solubility	$CaCO_2$ Permeability	Skin Permeability	VDss (human)	BBB Permeability	CNS Permeability	Index of Refraction	Surface Tension
pIC50	1	0.282	0.193	0.3	0.256	0.25	0.443	0.323	0.225	0.067	0.124
LogP	0.282	1	0.488	0.775	0.011	0.604	0.102	0.299	0.881	0.731	0.747
Surface Area	0.193	0.488	1	0.631	0.217	0.601	0.084	0.481	0.715	0.235	0.543
Water Solubility	0.3	0.775	0.631	1	0.347	0.348	0.469	0.378	0.893	0.234	0.564
$CaCO_2$ Permeability	0.256	0.011	0.217	0.347	1	0.257	0.295	0.197	0.266	0.195	0.217
Skin Permeability	0.25	0.604	0.601	0.348	0.257	1	0.283	0.237	0.574	0.652	0.646
VDss (human)	0.443	0.102	0.084	0.469	0.295	0.283	1	0.274	0.128	0.261	0.195
BBB Permeability	0.323	0.299	0.481	0.378	0.197	0.237	0.274	1	0.496	0.128	0.208
CNS Permeability	0.225	0.881	0.715	0.893	0.266	0.574	0.128	0.496	1	0.433	0.637
Index of Refraction	0.067	0.731	0.235	0.234	0.195	0.652	0.261	0.128	0.433	1	0.631
Surface Tension	0.124	0.747	0.543	0.564	0.217	0.646	0.195	0.208	0.637	0.631	1

Table 3. Experimental and calculated pIC50 values for QSAR model.

Molecule	Experimental Activity	Predicted Activity	Residual Activity	SD
20a	8.652	8.791	−0.139	−0.464
20d	7.357	7.068	0.289	0.965
21a	7.335	7.318	0.017	0.056

Continued

21d	7.008	7.077	−0.069	−0.23
22a	8.386	8.191	0.195	0.651
22d	6.796	6.974	−0.178	−0.595
23a	7.991	7.856	0.135	0.451
23d	6.959	6.974	−0.016	−0.052
24a	6.538	6.484	0.054	0.18
25a	6.509	6.61	−0.101	−0.337
26d	7.566	7.479	0.087	0.289
20b	7.69	7.7	−0.009	−0.031
21b	7.474	6.984	0.49	1.635
22b	7.244	7.356	−0.112	−0.374
23b	7.715	7.442	0.273	0.911
24b	6.276	6.404	−0.128	−0.429
26e	9.114	8.645	0.468	1.563
20c	7.544	7.7	−0.156	−0.521
20f	7.9	7.726	0.173	0.579
21c	6.569	6.984	−0.415	−1.386
21f	6.536	6.645	−0.109	−0.363
22f	7.223	7.101	0.123	0.41
26c	7.245	7.195	0.05	0.166
26f	7.64	7.86	−0.219	−0.732
27h	8.164	8.045	0.12	0.399
28h	7.642	8.04	−0.398	−1.328
29h	8.222	8.645	−0.423	−1.414

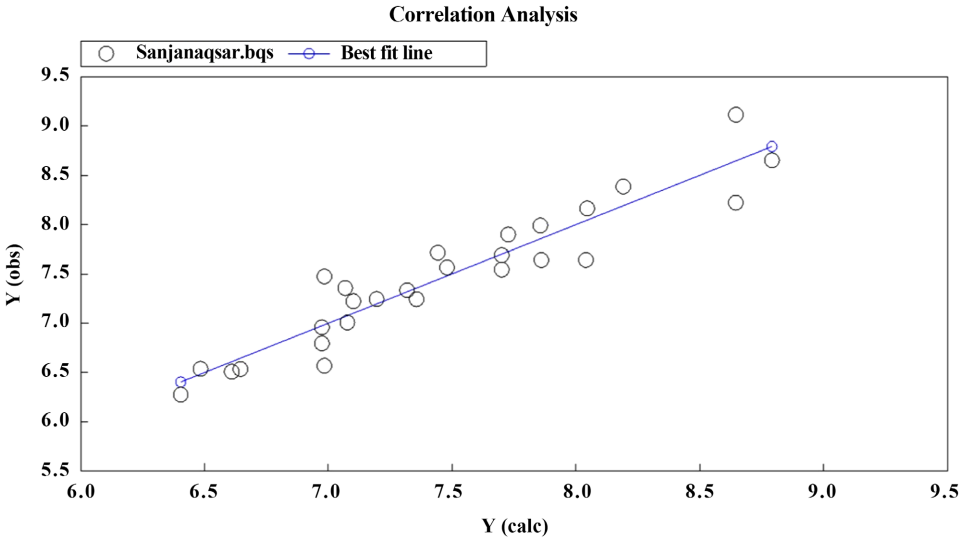
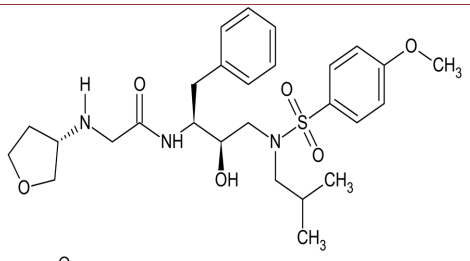
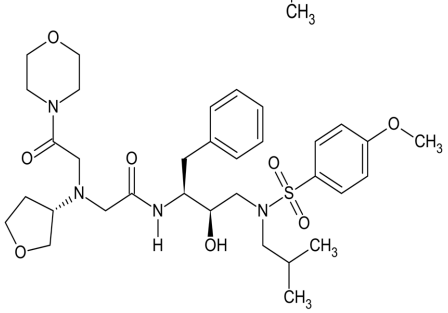
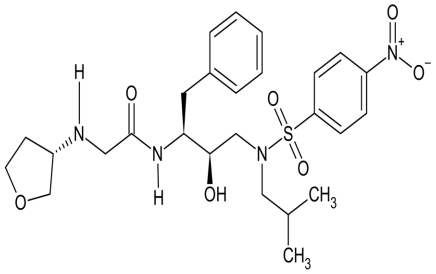
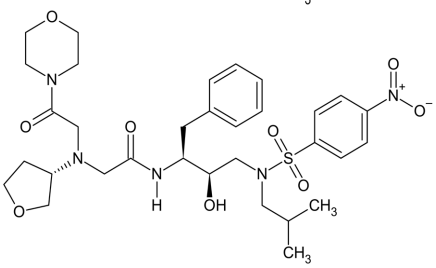
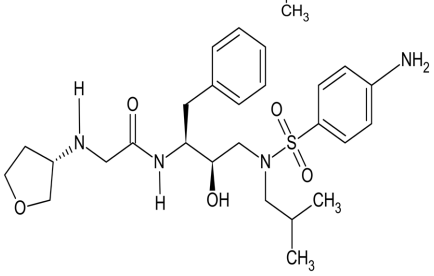
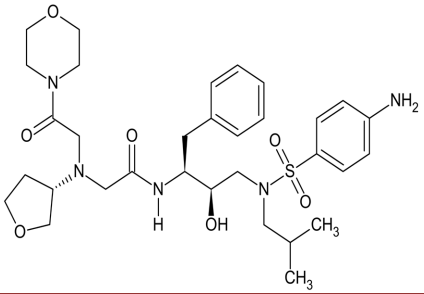
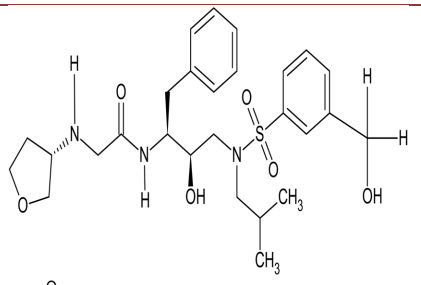
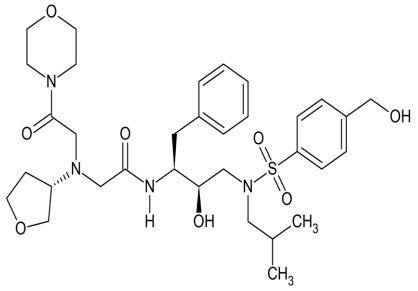
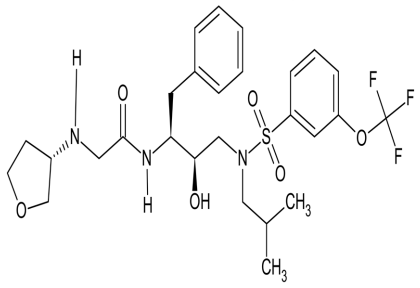
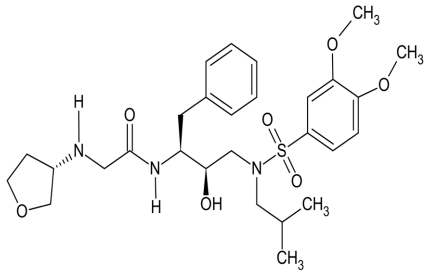
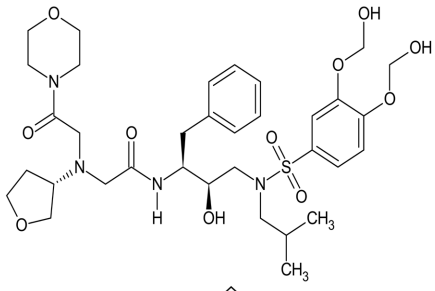
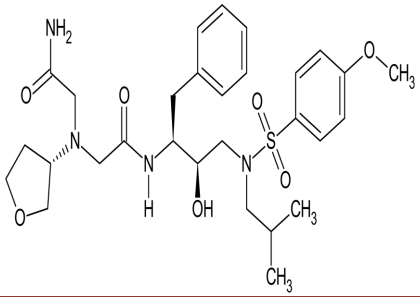


Figure 2. Correlation of observed vs calculated activities in 2D QAR model by MLR.

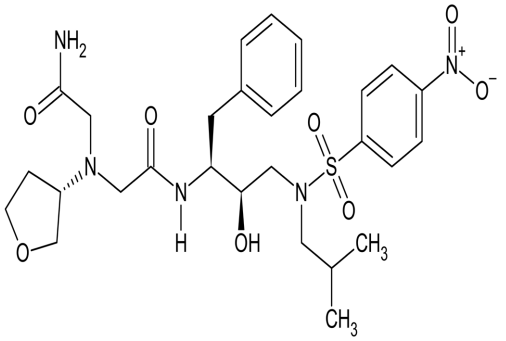
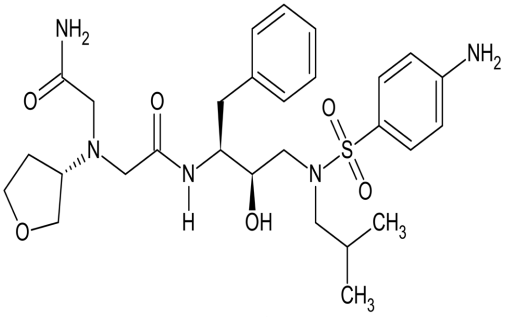
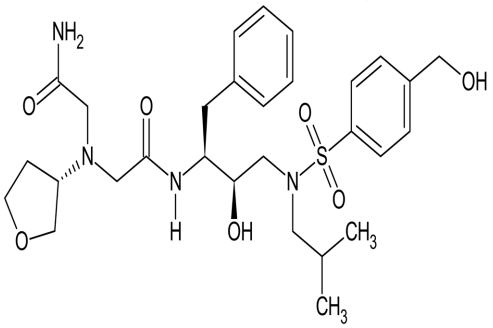
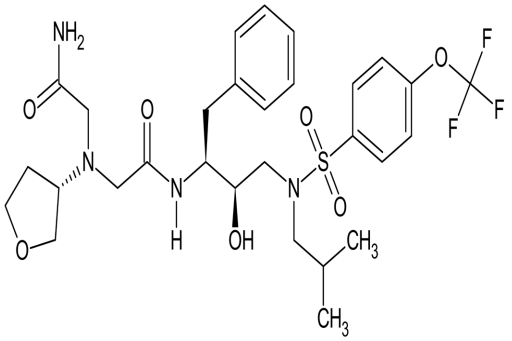
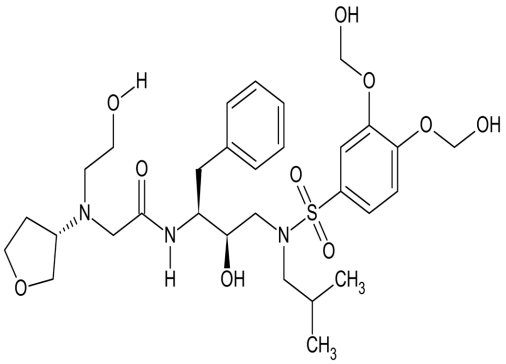
Table 4. HIV-1 protease inhibitors.

Molecule	Structure	Ic50 (nm)	Experimental Activity	Predicted Activity
20a		2.23	8.652	8.791
20d		44.0	7.357	7.068
21a		46.3	7.335	7.318
21d		98.1	7.008	7.077
22a		4.11	8.386	8.191
22d		160	6.796	6.974

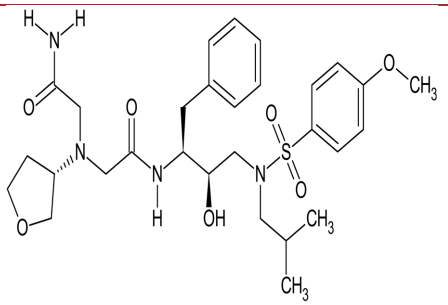
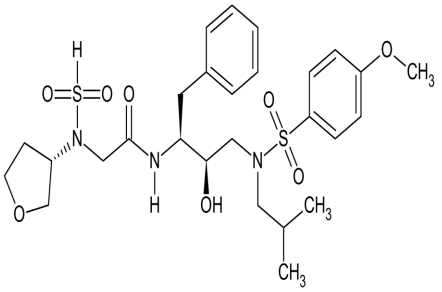
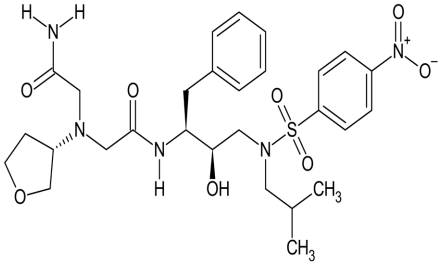
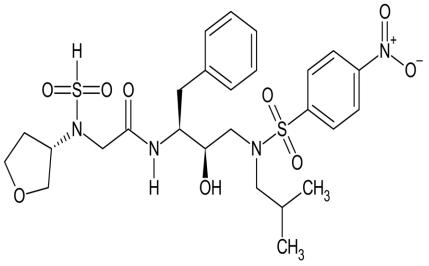
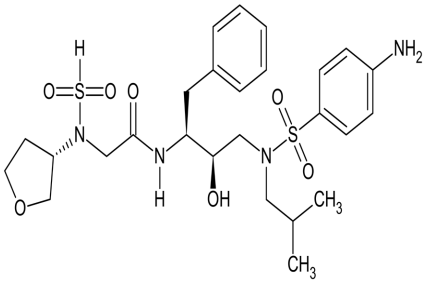
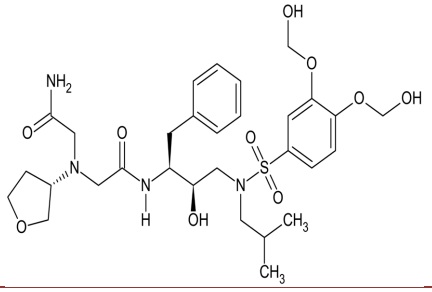
Continued

23a		10.2	7.991	7.856
23d		110	6.959	6.974
24a		290	6.538	6.484
25a		310	6.509	6.61
26d		27.2	7.566	7.479
20b		20.4	7.69	7.7

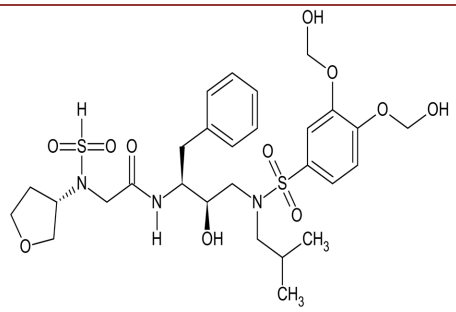
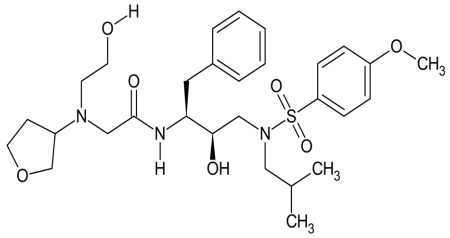
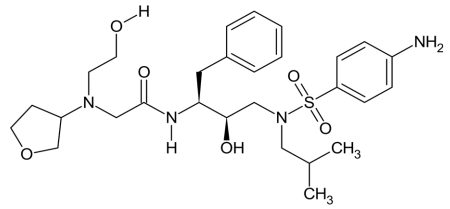
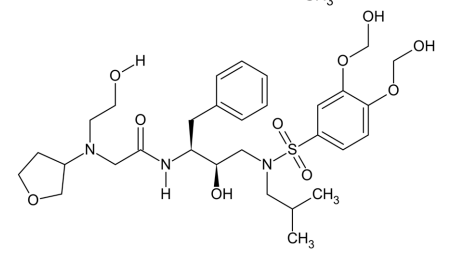
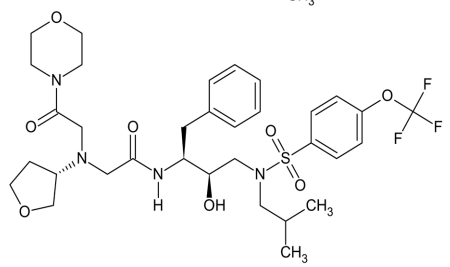
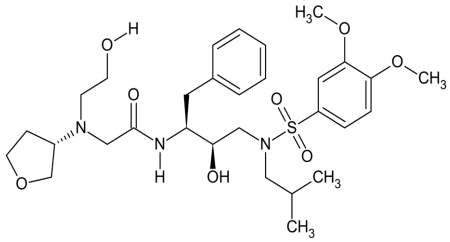
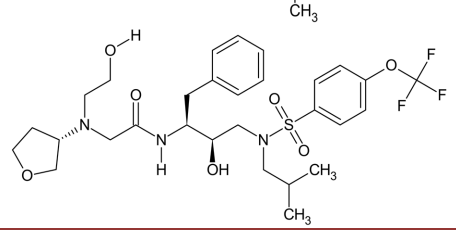
Continued

21b		33.6	7.474	6.984
22b		57.0	7.244	7.356
23b		19.3	7.715	7.442
24b		530	6.276	6.404
26e		0.77	9.114	8.645

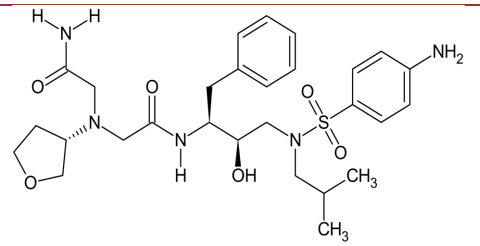
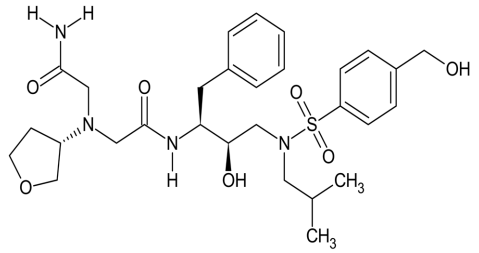
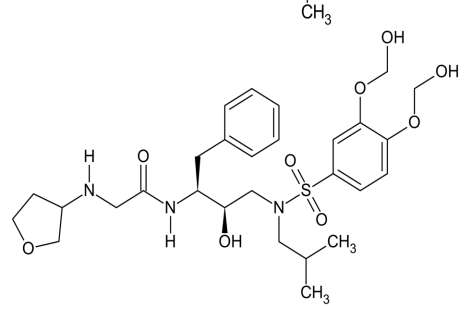
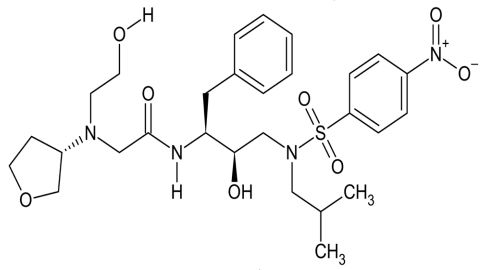
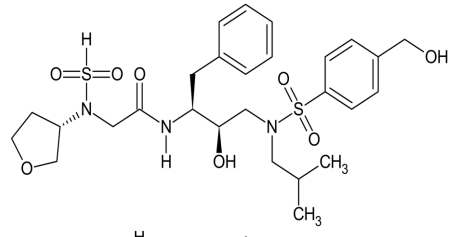
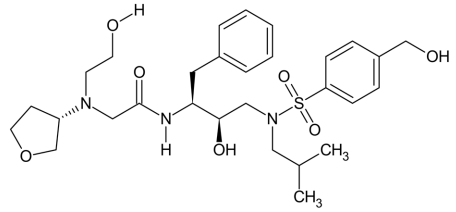
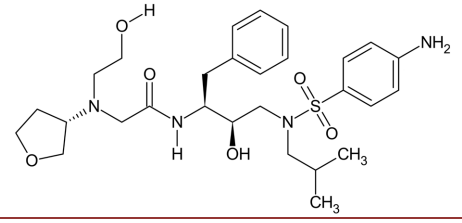
Continued

20c		28.6	7.544	7.7
20f		12.6	7.9	7.726
21c		270	6.569	6.984
21f		291	6.536	6.645
22f		59.8	7.223	7.101
26c		56.9	7.245	7.195

Continued

26f		22.9	7.64	7.86
27h		6.85	8.164	8.045
28h		22.8	7.642	8.04
29h		6.00	8.222	8.645
24d		910	6.041	5.93379
25e		190	6.7213	7.339545
24e		68.0	7.1675	6.8691

Continued

22c		38.0	7.4203	7.3628
23c		65.0	7.1871	7.448538
29g		53.4	7.2725	7.56238
21e		17.7	7.7521	7.896739
23f		22.7	7.644	7.516305
23e		4.26	8.3706	8.168279
22e		0.88	9.0555	8.046661

3.1. Model Validation

The predictability and validity of the model (test set) based on active compounds were using cross validation coefficient (Q2 = 0.701). Regression coefficient of the training set was 0.722, which was relevant to the model. Stability of the generated model ranges from 0.6612 to 0.9475 on a maximum scale of 1. F value was found to be 12.490. Pearson-R value of 0.942 indicated greater degree of confidence on model. Scatter plots for experimental and predicted activities of the ligands elicited significant linear correlation and moderate difference between experimental and the predicted values shown in Figure 3 for test set.

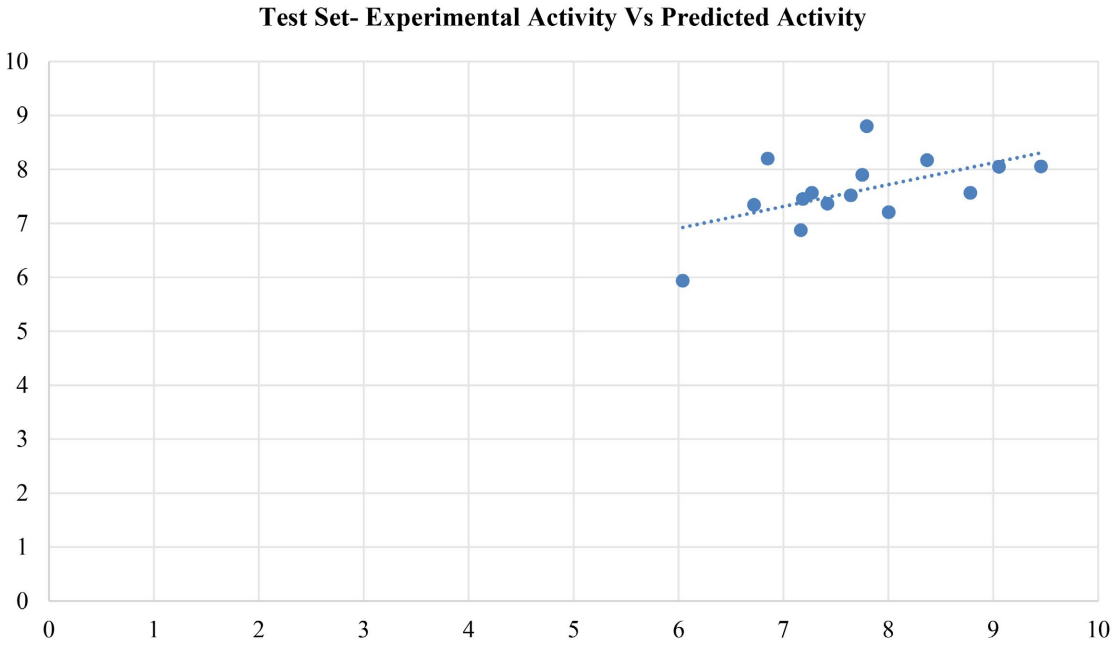
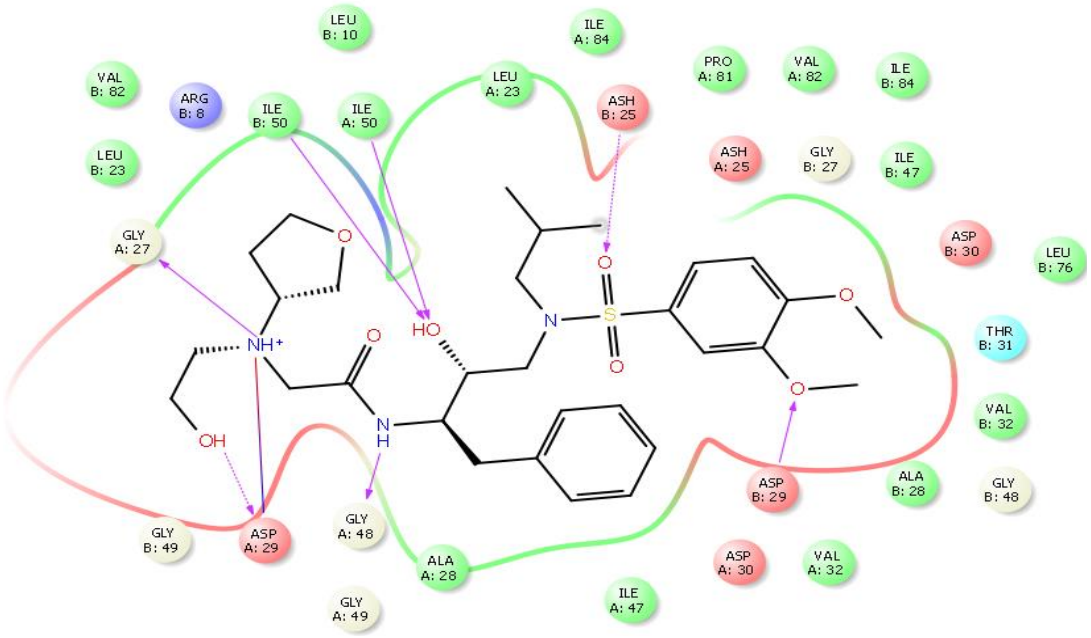


Figure 3. Scatter plot of observed vs predicted activity for test set.

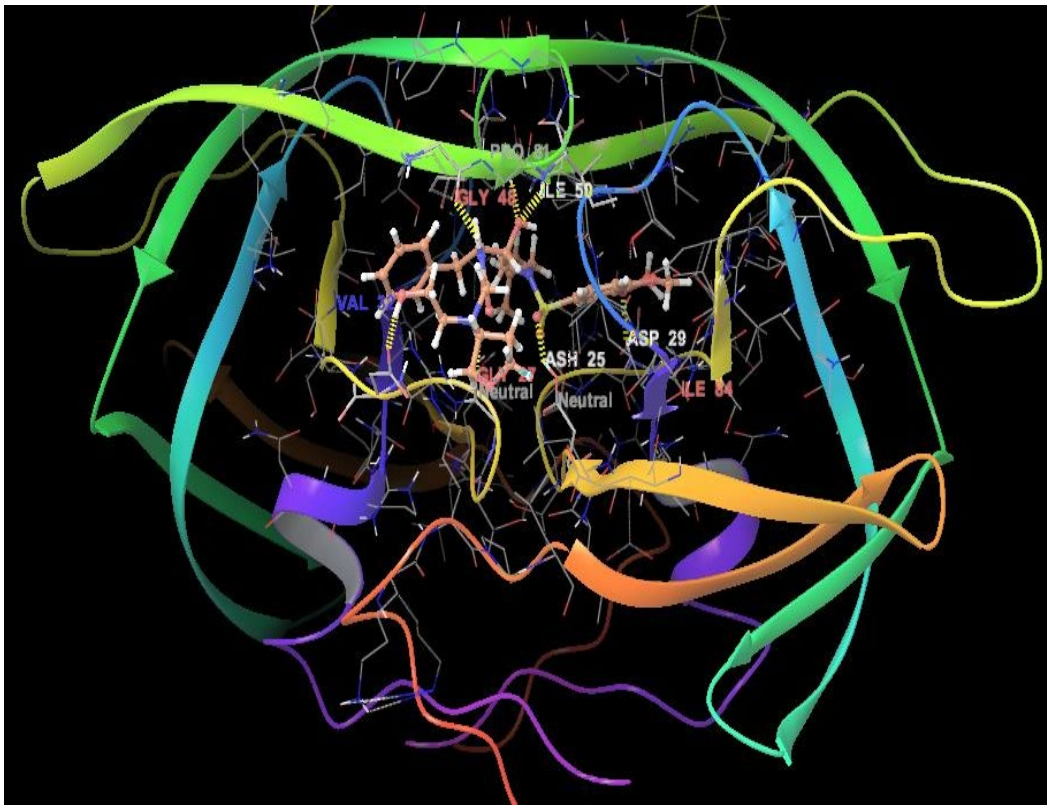
3.2. Docking

In Molecule number **25e** (N-((2S, 3R)-3-Hydroxy-4-(N-isobutyl-3,4-dimethoxyphenylsulfonamido)-1-phenylbutan-2-yl)-2-((2-methoxyethyl)((S)-tetrahydrofuran-3-yl)amino)acetamide) shown in Figure 4, Molecular number 20e (N-((2S, 3R)-3-Hydroxy-4-(N-isobutyl-4-methoxyphenylsulfonamido)-1-phenylbutan-2-yl)-2-((2-methoxyethyl)((S)-tetrahydrofuran-3-yl)amino)acetamide) shown in Figure 5; the NH+ group shows interaction with amino acid ASP A:29 and GLY A:27. In Molecule number **20b** shown in Figure 6, Molecule number **23a** shown in Figure 7, Molecule number 26a shown in Figure 8, Molecule number 26d shown in Figure 9 Molecule number 26f shown in Figure 10; the OH group shows interaction with amino acid ILE A:50. Similarly in Molecule number 25e, 20e, 26f; the Hydroxyl group shows hydrogen bonding with amino acids ASP A:29. In Molecule number 24e shown in Figure 11 and Molecule number 20b show interaction with amino acid ASH A:25. In Molecule number 22b shown in Figure 12; the amine group shows interaction with amino acid ASP A:29. In Molecule number

20d shown in Figure 13; the OH group shows interactions with amino acid GLY B:27.



(a)



(b)

Figure 4. (a) Ligand interaction diagram for molecule 25e with dock score −11.262; (b) Dock Pose for molecule 25e.

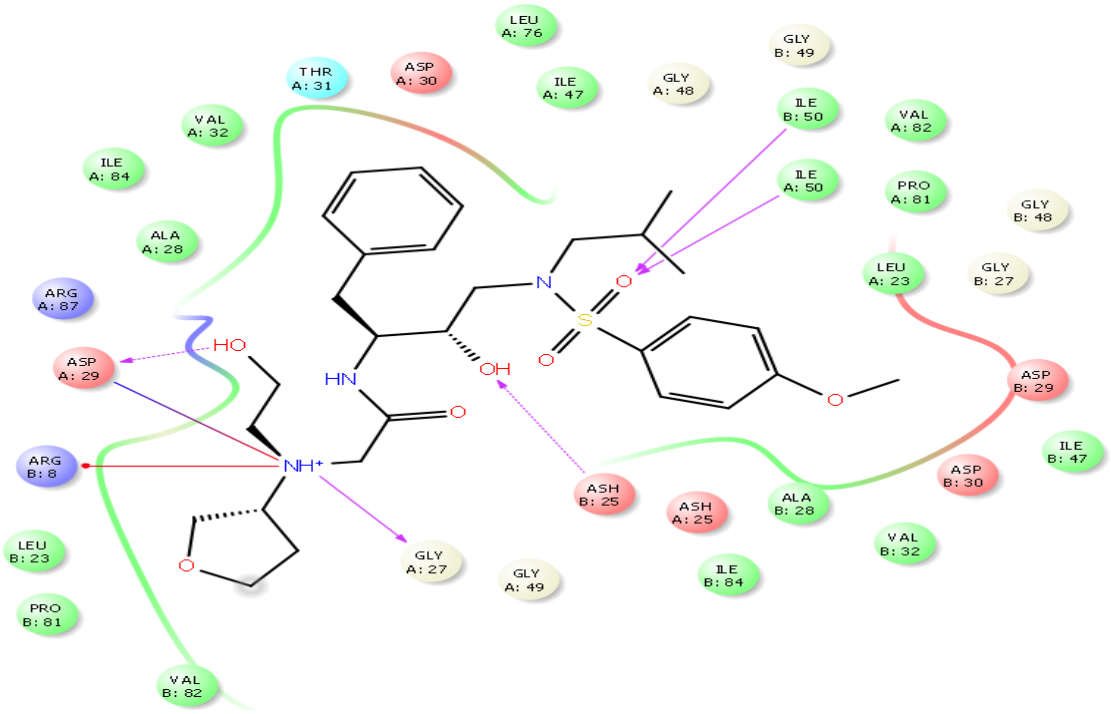


Figure 5. Ligand interaction diagram for molecule 20e with dock score −10.959.

In Molecule number 20b shown in **Figure 6**, Molecule number 23a shown in **Figure 7**, Molecule number 26a shown in **Figure 8**, Molecule number 26d shown in **Figure 9**, Molecule number 26f shown in **Figure 10**; the OH group shows interaction with amino acid ILE A:50. Similarly in Molecule number 25e, 20e, 26f; the Hydroxyl group shows hydrogen bonding with amino acid. ASP A:29.

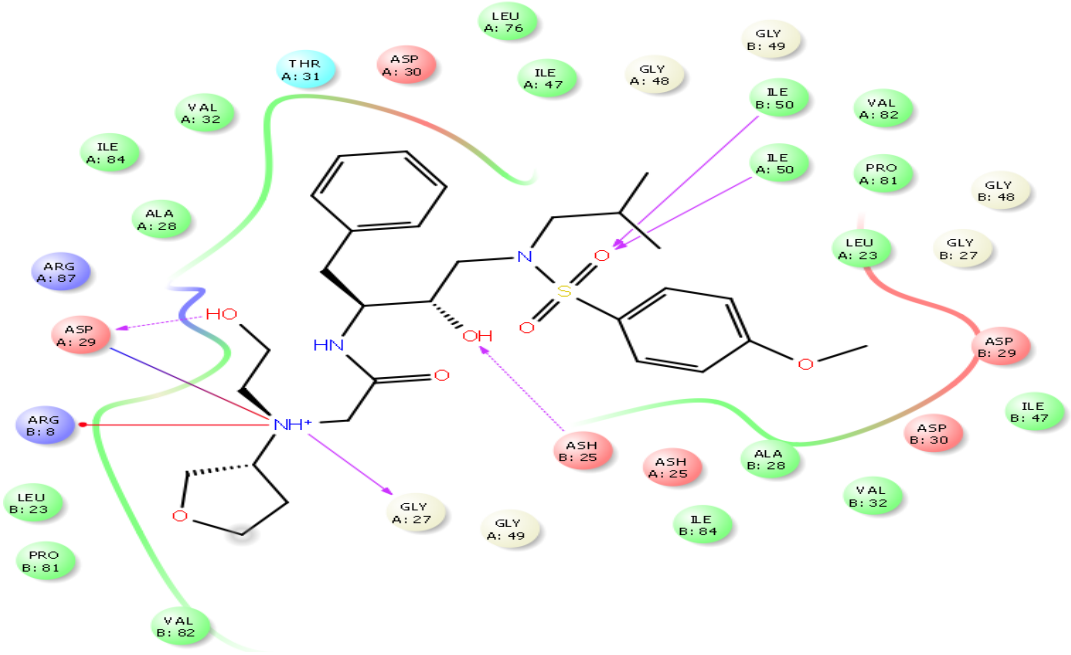


Figure 6. Ligand interaction diagram for molecule 20b with dock score −10.442.

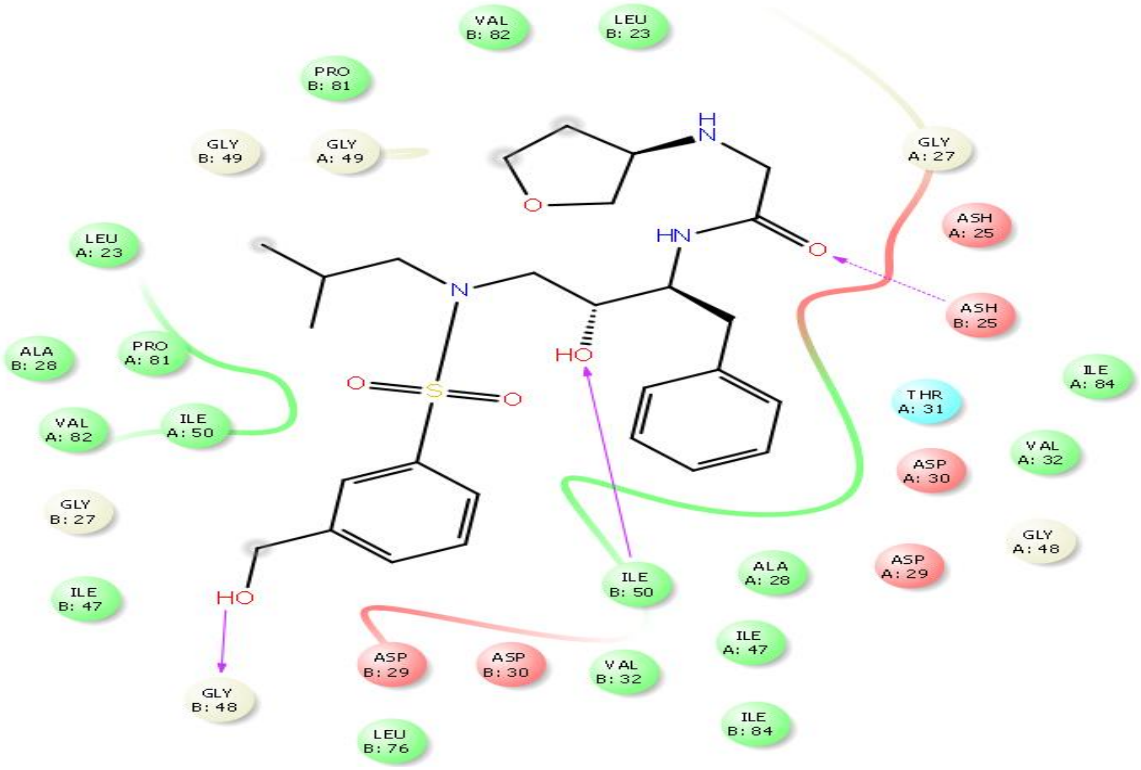


Figure 7. Ligand interaction diagram for molecule 23a with dock score −9.685.

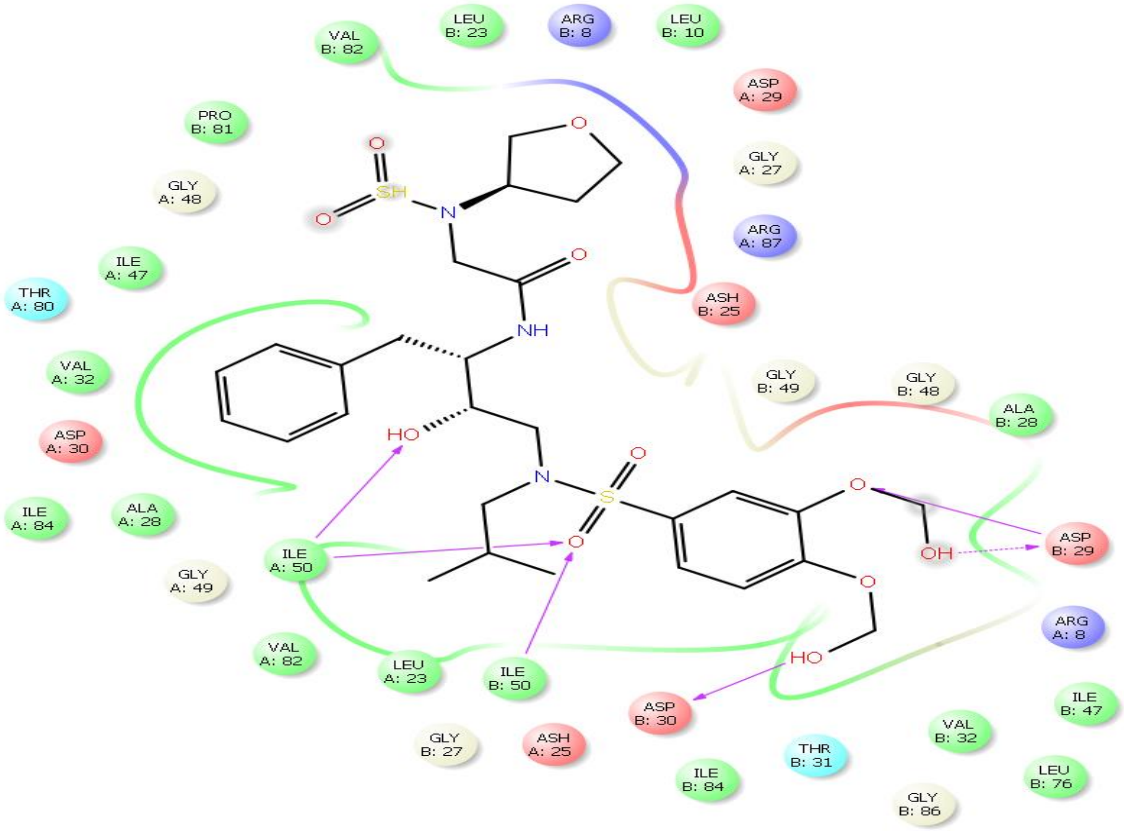


Figure 8. Ligand interaction diagram for molecule 26a with dock score −10.387.

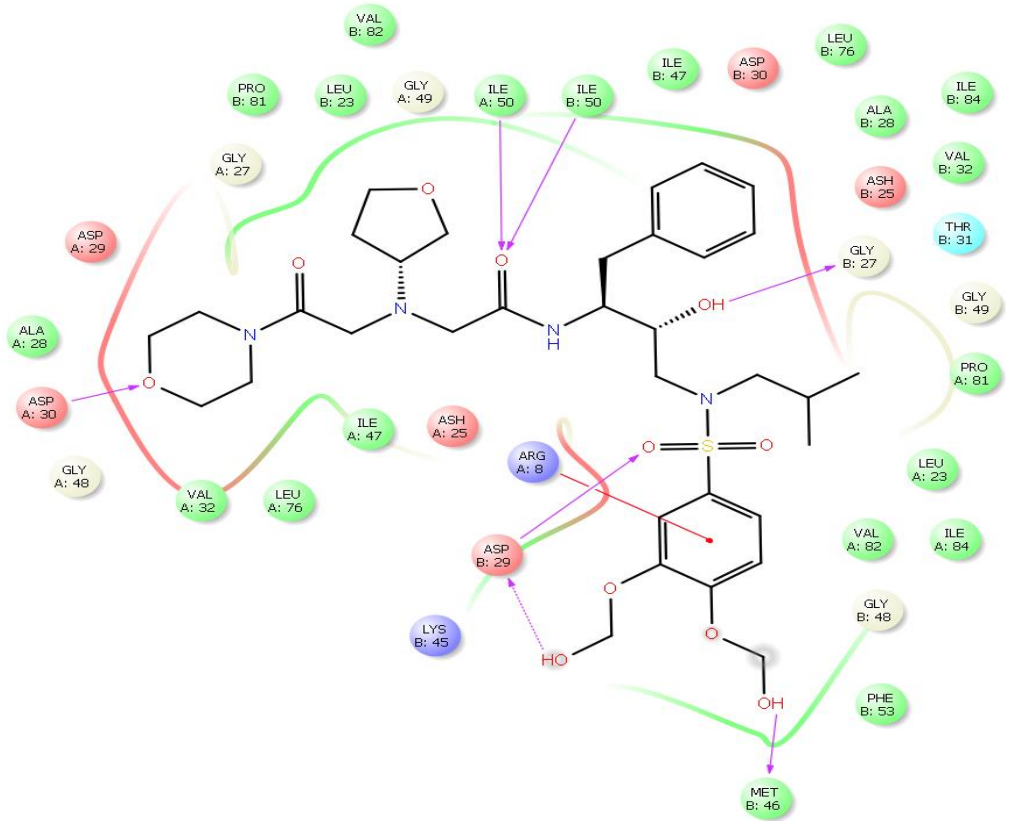


Figure 9. Ligand interaction diagram for molecule 26d with dock score −10.045.

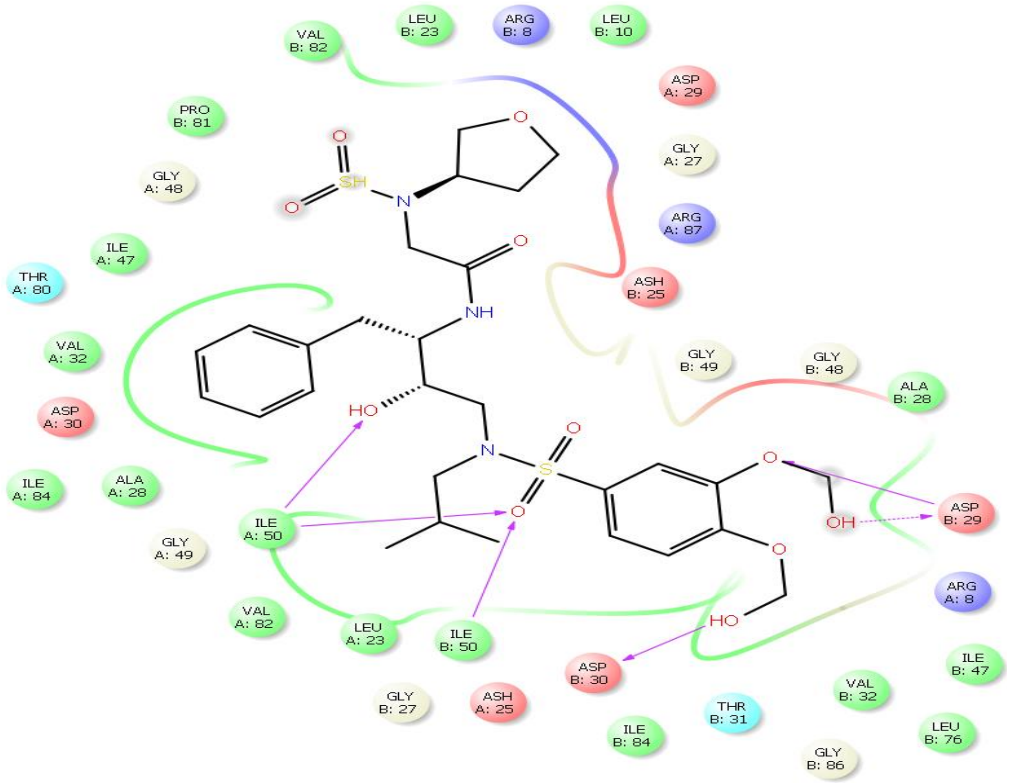


Figure 10. Ligand interaction diagram for molecule 26f with dock score −10.046.

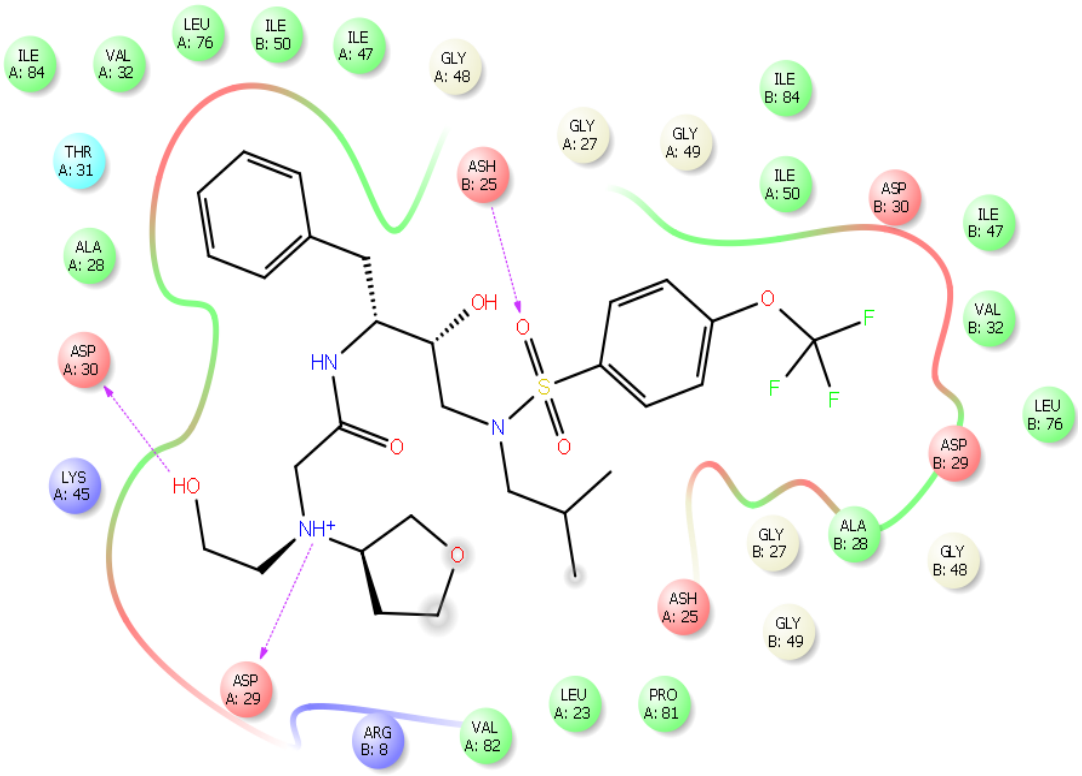


Figure 11. Ligand interaction diagram for molecule 24e with dock score −10.158.

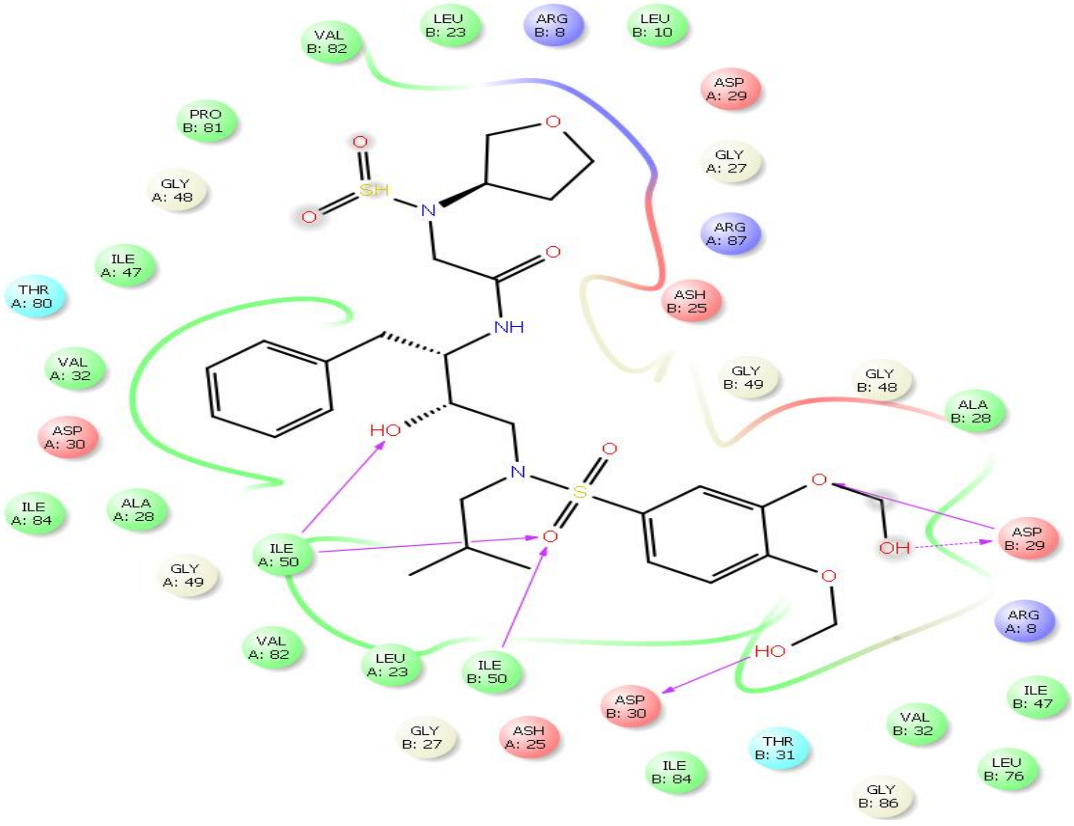


Figure 12. Ligand interaction diagram for molecule 22b with dock score −9.789.

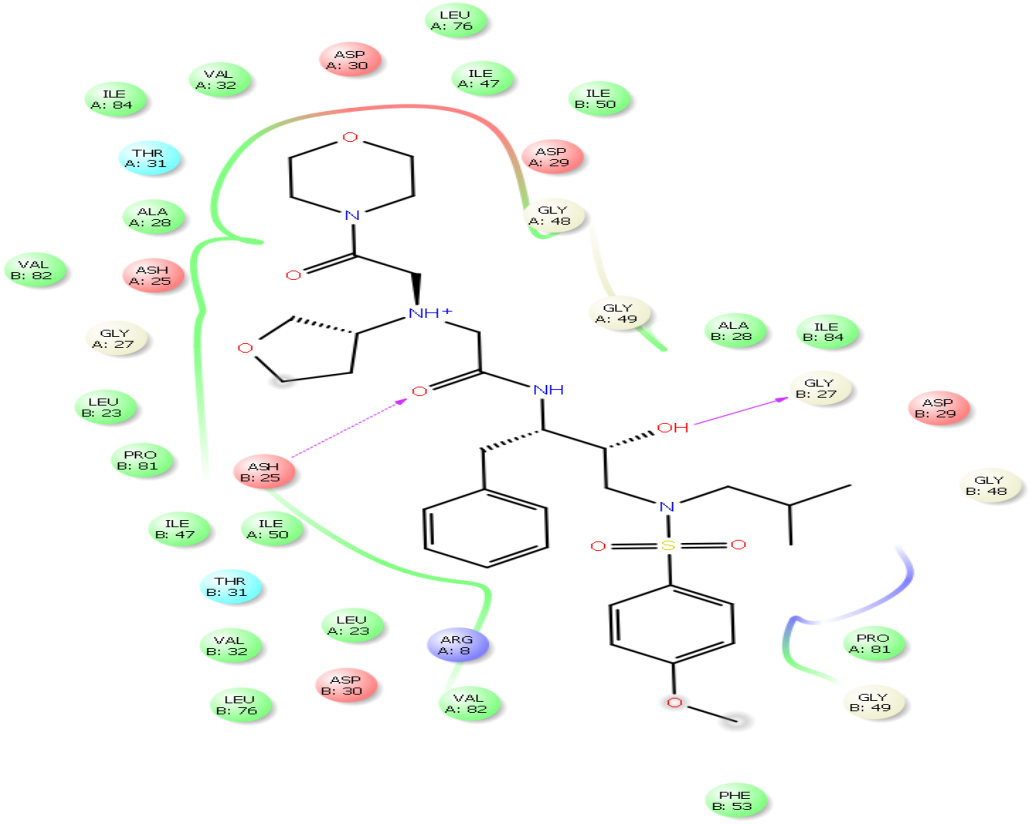


Figure 13. Ligand interaction diagram for molecule 20d with dock score −9.641.

Table 5. Docking and glide score of HIV-1 protease inhibitors.

Molecule	Docking score	Glide score
25e	−11.262	−11.312
20e	−10.959	−11.009
20b	−10.442	−10.606
26a	−10.387	−10.948
24e	−10.158	−10.208
26f	−10.046	−10.046
26d	−10.045	−10.807
22b	−9.789	−9.953
23a	−9.685	−10.245
20d	−9.641	−9.833
22e	−9.614	−9.665
21e	−9.607	−9.771
28h	−9.595	−9.645
20f	−9.445	−9.445
22d	−9.388	−9.530
27g	−9.222	−9.513

Continued

26c	−9.214	−9.378
23b	−9.203	−9.376
23c	−9.185	−9.348
21f	−9.079	−9.079
23f	−8.882	−8.882
21a	−8.847	−9.136
26e	−8.745	−8.796
21e	−8.708	−8.759
23a	−8.640	−8.690
22c	−8.601	−8.765
23d	−8.562	−8.753
22f	−8.527	−8.527
26b	−8.405	−9.248
22a	−8.397	−8.957
20a	−8.389	−8.950
21d	−8.383	−8.575
27h	−8.195	−8.245
24b	−8.063	−8.227
20c	−7.938	−8.101
21b	−7.905	−8.069
24a	−7.882	−8.443
28g	−7.880	−8.441
29g	−7.698	−8.259
29h	−7.631	−7.681
25a	−7.570	−8.131
24d	−7.484	−7.675

The findings of the present study offer significant insights for future inhibitor design against HIV-1 protease (Table 4). The developed 2D QSAR model characterized by a high correlation coefficient (R = 0.942) and a satisfactory Q2 value (0.701), highlights the influence of molecular descriptors such as Log P, Surface area, Water solubility and various permeability indices on inhibitory activity. These descriptors provide a foundational guidelines for rational optimization of lead molecules in early drug discovery.

The study's outcomes suggest that enhancing specific pharmacokinetic properties such as lipophilicity, blood brain barrier permeability, and aqueous solubility could improve the biological efficacy of potential inhibitors. Consequently the model offers a valuable tool for pre-synthetic screening, enabling the selection of molecules with favorable activity profiles before proceeding to synthesis and experi-

mental testing. This approach can help streamline the drug development pipeline by reducing time, cost, resource expenditure associated with traditional hit and trial methods.

Moreover, the molecular docking results support the QSAR predictions by revealing key binding interactions notably hydrogen bonding with residues such as Asp29 and Ile50 in the HIV-1 protease active site. These interactions provide a structural rationale for observed biological activities and can serve as pharmacophoric targets in future design efforts.

Therefore, the integrated QSAR and docking framework proposed in this study may complement existing drug development protocols by offering a data-driven, structure based rationale for inhibitor design. Further research incorporating more diverse chemical libraries and experimental validation would be instrumental in advancing these findings towards chemical translation.

3.3. Limitations and Future Directions

While the present demonstrates promising results in the application of 2D QSAR and molecular docking to identify potential HIV-1 protease inhibitors (**Table 5**), certain limitations should be noted.

i) The dataset comprised a relatively small number of compounds (n = 37), which may affect the generalization ability of QSAR model.

ii) Additionally, the use of multiple linear regression, while effective for initial modeling, may not fully capture more complex, non-linear relationships between descriptors and biological activity.

iii) Moreover, the findings are based entirely on computational methods and further experimental validation both *in vitro* and *in vivo* is essential to confirm the predicted biological efficacy and pharmacokinetic properties of the proposed compounds.

iv) Docking was performed using a single crystal structure of the HIV-1 protease, which may not account for conformational flexibility of target protein.

v) Expanding this approach to incorporate multiple protein conformation or molecular dynamics stimulation could offer deeper insights into binding interaction.

vi) Finally while ADME parameters were considered a more comprehensive evaluation including toxicity profiling would strengthen the drug-likeliness assessment.

These limitations do not detract from the value of current work but rather highlight the need for further more extensive studies. Further research involving a larger and more diverse compound dataset, additional modelling techniques and experimental validation will be critical to advance these findings towards therapeutic application.

4. Conclusions

The combining Ligand Based Drug Design and Structure Based Drug Design

computational approach was applied to obtain an insight into the structural basis for discovering novel lead HIV-1 Protease Inhibitors. The study reveals the pharmacophoric features responsible for biological activity. 2D QSAR analysis was performed to provide a structural framework for understanding the Structure activity Relationship of these compounds.

The 2D QSAR model revealed an excellent Correlation coefficient R = 0.942 and Cross Validation coefficient R_2 Pred = 0.722524. The model also exhibited a good Regression coefficient Q2 = 0.701.

Molecular Docking studies were also carried out to understand the intermolecular interaction between receptors. Therefore, the identified novel compounds could be further employed in not only designing novel and potent HIV-1 Protease Inhibitors but also is useful in optimization and discovery of same with new scaffold.

Acknowledgements

The authors acknowledge the Department of chemistry, St. Francis College for Women and Department of Chemistry, University College for Women, OU, Koti, Hyderabad for providing the lab facilities. The authors also want to acknowledge Schrodinger for providing the software required for computational work.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Seitz, R. (2016) Human Immunodeficiency Virus (HIV). *Transfusion Medicine and Hemotherapy*, **43**, 203-222. https://doi.org/10.1159/000445852

[2] Van Heuverswyn, F., Li, Y., Neel, C., Bailes, E., Keele, B.F., Liu, W., *et al.* (2006) Human Immunodeficiency Viruses: SIV Infection in Wild Gorillas. *Nature*, **444**, 164-164. https://doi.org/10.1038/444164a

[3] Deeks, S.G., Overbaugh, J., Phillips, A. and Buchbinder, S. (2015) HIV Infection. *Nature Reviews Disease Primers*, **1**, Article No. 15035. https://doi.org/10.1038/nrdp.2015.35

[4] Freed, E.O. (2015) HIV-1 Assembly, Release and Maturation. *Nature Reviews Microbiology*, **13**, 484-496. https://doi.org/10.1038/nrmicro3490

[5] Campbell-Yesufu, O.T. and Gandhi, R.T. (2011) Update on Human Immunodeficiency Virus (HIV)-2 Infection. *Clinical Infectious Diseases*, **52**, 780-787. https://doi.org/10.1093/cid/ciq248

[6] Hernandez-Vargas, E.A. and Middleton, R.H. (2013) Modeling the Three Stages in HIV Infection. *Journal of Theoretical Biology*, **320**, 33-40. https://doi.org/10.1016/j.jtbi.2012.11.028

[7] Mindel, A. (2001) ABC of AIDS: Natural History and Management of Early HIV Infection. *BMJ*, **322**, 1290-1293. https://doi.org/10.1136/bmj.322.7297.1290

[8] Fearon, M. (2005) The Laboratory Diagnosis of HIV Infections. *Canadian Journal of Infectious Diseases and Medical Microbiology*, **16**, 26-30. https://doi.org/10.1155/2005/515063

[9] Cihlar, T. and Fordyce, M. (2016) Current Status and Prospects of HIV Treatment. *Current Opinion in Virology*, **18**, 50-56. https://doi.org/10.1016/j.coviro.2016.03.004

[10] Fettig, J., Swaminathan, M., Murrill, C.S. and Kaplan, J.E. (2014) Global Epidemiology of HIV. *Infectious Disease Clinics of North America*, **28**, 323-337. https://doi.org/10.1016/j.idc.2014.05.001

[11] Yang, H., Nkeze, J. and Zhao, R.Y. (2012) Effects of HIV-1 Protease on Cellular Functions and Their Potential Applications in Antiretroviral Therapy. *Cell & Bioscience*, **2**, Article No. 32. https://doi.org/10.1186/2045-3701-2-32

[12] Gulnik, S., Erickson, J.W. and Xie, D. (2000) HIV Protease: Enzyme Function and Drug Resistance. *Vitamins & Hormones*, **58**, 213-256. https://doi.org/10.1016/s0083-6729(00)58026-1

[13] Surabhi, S. and Singh, B. (2018) Computer Aided Drug Design: An Overview. *Journal of Drug Delivery and Therapeutics*, **8**, 504-509. https://doi.org/10.22270/jddt.v8i5.1894

[14] Jain, A. (2017) Computer Aided Drug Design. *Journal of Physics*: *Conference Series*, **884**, Article ID: 012072. https://doi.org/10.1088/1742-6596/884/1/012072

[15] Macalino, S.J.Y., Gosu, V., Hong, S. and Choi, S. (2015) Role of Computer-Aided Drug Design in Modern Drug Discovery. *Archives of Pharmacal Research*, **38**, 1686-1701. https://doi.org/10.1007/s12272-015-0640-5

[16] Sinha, S. and Vohora, D. (2018) Drug Discovery and Development. In: Vohora, D. and Singh, G., Eds., *Pharmaceutical Medicine and Translational Clinical Research*, Elsevier, 19-32. https://doi.org/10.1016/b978-0-12-802103-3.00002-x

[17] Khazir, J., Mir, B.A., Mir, S.A. and Cowan, D. (2013) Natural Products as Lead Compounds in Drug Discovery. *Journal of Asian Natural Products Research*, **15**, 764-788. https://doi.org/10.1080/10286020.2013.798314

[18] de Souza Neto, L.R., Moreira-Filho, J.T., Neves, B.J., Maidana, R.L.B.R., Guimarães, A.C.R., Furnham, N., *et al.* (2020) In Silico Strategies to Support Fragment-To-Lead Optimization in Drug Discovery. *Frontiers in Chemistry*, **8**, Article 93. https://doi.org/10.3389/fchem.2020.00093

[19] Merchant, H.A. (2021) Basic Pharmacokinetics. In: Batchelor, H., Ed., *Biopharmaceutics*: *From Fundamentals to Industrial Practice*, Wiley. https://doi.org/10.1002/9781119678366.ch2

[20] van den Anker, J., Reed, M.D., Allegaert, K. and Kearns, G.L. (2018) Developmental Changes in Pharmacokinetics and Pharmacodynamics. *The Journal of Clinical Pharmacology*, **58**, S10-S25. https://doi.org/10.1002/jcph.1284

[21] Jones, L.H. and Bunnage, M.E. (2017) Applications of Chemogenomic Library Screening in Drug Discovery. *Nature Reviews Drug Discovery*, **16**, 285-296. https://doi.org/10.1038/nrd.2016.244

[22] Abdel-Ilah, L., Veljović, E., Gurbeta, L. and Badnjević, A. (2017) Applications of QSAR Study in Drug Design. *International Journal of Engineering Research & Technology*, **6**, 582-587.

[23] Lewis, R.A. and Wood, D. (2014) Modern 2D QSAR for Drug Discovery. *WIREs Computational Molecular Science*, **4**, 505-522. https://doi.org/10.1002/wcms.1187

[24] Aparoy, P., Kumar Reddy, K. and Reddanna, P. (2012) Structure and Ligand Based Drug Design Strategies in the Development of Novel 5-LOX Inhibitors. *Current Medicinal Chemistry*, **19**, 3763-3778. https://doi.org/10.2174/092986712801661112

[25] Fan, J., Fu, A. and Zhang, L. (2019) Progress in Molecular Docking. *Quantitative Biology*, **7**, 83-89. https://doi.org/10.1007/s40484-019-0172-y

[26] Bai, X., Yang, Z., Zhu, M., Dong, B., Zhou, L., Zhang, G., *et al.* (2017) Design and

Synthesis of Potent HIV-1 Protease Inhibitors with (S)-Tetrahydrofuran-Tertiary Amine-Acetamide as P2-ligand: Structure-Activity Studies and Biological Evaluation. *European Journal of Medicinal Chemistry*, **137**, 30-44. https://doi.org/10.1016/j.ejmech.2017.05.024

[27] Anthony-Cahill, S. and Walsh, E.J. (1999) Book & Media Reviews. *Journal of Chemical Education*, **76**, 905-906. https://doi.org/10.1021/ed076p905

[28] Spessard, G.O. (1998) ACD Labs/LogP dB 3.5 and ChemSketch 3.5. *Journal of Chemical Information and Computer Sciences*, **38**, 1250-1253. https://doi.org/10.1021/ci980264t

[29] Chen, H., Kogej, T. and Engkvist, O. (2018) Cheminformatics in Drug Discovery, an Industrial Perspective. *Molecular Informatics*, **37**, Article ID: 1800041. https://doi.org/10.1002/minf.201800041

[30] Weininger, D. (1988) SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *Journal of Chemical Information and Computer Sciences*, **28**, 31-36. https://doi.org/10.1021/ci00057a005

[31] Pires, D.E.V., Blundell, T.L. and Ascher, D.B. (2015) PKCSM: Predicting Small-Molecule Pharmacokinetic and Toxicity Properties Using Graph-Based Signatures. *Journal of Medicinal Chemistry*, **58**, 4066-4072. https://doi.org/10.1021/acs.jmedchem.5b00104

[32] Lombardo, F., Shalaeva, M.Y., Tupper, K.A. and Gao, F. (2001) ElogD$_{oct}$: A Tool for Lipophilicity Determination in Drug Discovery. 2. Basic and Neutral Compounds. *Journal of Medicinal Chemistry*, **44**, 2490-2497. https://doi.org/10.1021/jm0100990

[33] Pecoraro, B., Tutone, M., Hoffman, E., Hutter, V., Almerico, A.M. and Traynor, M. (2019) Predicting Skin Permeability by Means of Computational Approaches: Reliability and Caveats in Pharmaceutical Studies. *Journal of Chemical Information and Modeling*, **59**, 1759-1771. https://doi.org/10.1021/acs.jcim.8b00934

[34] Doniger, S., Hofmann, T. and Yeh, J. (2002) Predicting CNS Permeability of Drug Molecules: Comparison of Neural Network and Support Vector Machine Algorithms. *Journal of Computational Biology*, **9**, 849-864. https://doi.org/10.1089/10665270260518317

[35] Lee, D.K., In, J. and Lee, S. (2015) Standard Deviation and Standard Error of the Mean. *Korean Journal of Anesthesiology*, **68**, 220-223. https://doi.org/10.4097/kjae.2015.68.3.220

[36] Hughes, J., Rees, S., Kalindjian, S. and Philpott, K. (2011) Principles of Early Drug Discovery. *British Journal of Pharmacology*, **162**, 1239-1249. https://doi.org/10.1111/j.1476-5381.2010.01127.x

[37] Pavlov, A., Takuchev, N. and Georgieva, N. (2012) Drug Design by Regression Analyses of Newly Synthesized Derivatives of 8-Quinolinol. *Biotechnology & Biotechnological Equipment*, **26**, 164-169. https://doi.org/10.5504/50yrtimb.2011.0030

Abbreviations

HIV	Human Immunodeficiency Virus
AIDS	Acquired Immune Deficiency Syndrome
ADME	Absorption Distribution Metabolism Excretion
QSAR	Quantitative Structure Activity Relationship
MLR	Multiple Linear Regression
IC50	Inhibitory Concentrations
pkCSM	Predicting small-molecule pharmacokinetic properties using graph
R_2	Correlation Coefficient
SMILES	Simplified Molecular Input Line Entry System
Q2	Regression Coefficient