BT6320

Protein Interaction: Computational Techniques QSAR

Shreeharsha G Bhat | BE21B037
Department of Biotechnology
Indian Institute of Technology Madras

TABLE OF CONTENTS

QSAR study of FDA approved drugs for HIV protease inhibitors

Introduction:

The US Department of Health and Human Services in 1984 declared that the retrovirus Human immunodeficiency virus (HIV) was the cause of the immunodeficiency syndrome AIDS. As of 2024, the estimated number of people infected have been recorded at 88.4 million with a death toll of 42.3 million according to the Joint United Nations Program on HIV/AIDS (UNAIDS) [1]. The development of multiple therapeutic agents since the disease first surfaced have targeted the various stages of the HIV life cycle [2]. This has helped transform the deadly infection to a manageable ailment. The set of targets for antiviral therapy development for HIV include the reverse transcriptase, protease and integrase. This paper focuses on the current set of clinically approved drugs that have been developed for protease inhibitors. These drugs include Saquinavir, Ritonavir, Indinavir, Nelfinavir , Amprenavir, Lopinavir , Atazanavir, Tipranavir, Darunavir and Telinavir, which is a Phase 2 drug. These drugs block the activity of the HIV protease enzyme, and this inhibition leads to the virus being unable to cleave the polyproteins. Polyproteins are what allow the virus to produce new and mature viruses. By preventing the maturation of HIV, the inhibitors reduce the viral load in the body. These drugs are typically used as a part of antiretroviral therapy (ART) in combination with other drugs. This approach does not cure HIV, but does help in allowing the person to live longer and healthier [3].

Literature Review:

QSAR stands for Quantitative Structure-Activity Relationship. The QSAR study relies on the principle that the deviation in biological response or activity of a series of compounds can be accounted for by variation in their structure properties [4]. By building models and by calculating minimum energy conformations we are able to calculate the descriptors, these descriptor values for a particular drug can be obtained using databases like ChemDes, PaDEL, BlueDesc, RDKit, E-DRAGON and many others. In this study we use ChemDes for obtaining data. ChemDes is an integrated web-based platform for molecular descriptor and fingerprint computation [5]. This study focuses on 2D QSAR models, and finding two molecular descriptors that best define the 10 FDA (US Food and drug administration) approved protease inhibitors. The binding affinity or activity of a drug can be measured in several ways, most commonly IC50, Ki and EC50 are used. IC50 is a measure of how much a drug or compound is required to inhibit a biological process by 50%. Ki or the inhibition constant is the measure of how strong an inhibitor is at blocking an enzyme's activity, and EC50 refers to the concentration of a drug which induces a response halfway between the baseline and maximum after specified exposure time. To perform QSAR analysis we use the IC50 values of all the 10 drugs. ChEMBL is a manually curated database of bioactive molecules with drug-like properties [6]. ChEMBL is used to find the SMILES representation and IC50 values of all the drugs for HIV protease inhibitors

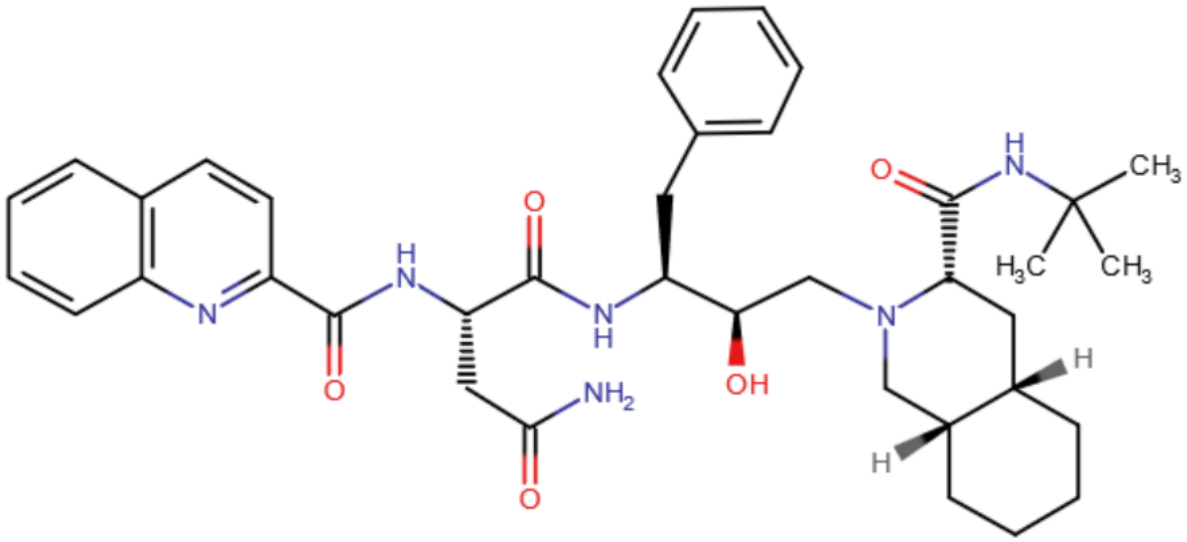
as the target (ID: CHEMBL243). The 10 FDA approved drugs are listed below with their CHEMBL IDs and their corresponding activity (IC50 value)

Sl.no	Drug	Mol.Formula	Mol.wt	ChEMBL ID	IC50 Value
1	Saquinavir	C38H50N6O5	670.86	CHEMBL114	0.23
2	Ritonavir	C37H48N6O5S2	720.96	CHEMBL163	34
3	Indinavir	C36H47N5O4	613.80	CHEMBL115	0.56
4	Nelfinavir	C32H45N3O4S	567.80	CHEMBL584	12
5	Amprenavir	C25H35N3O6S	505.64	CHEMBL116	30
6	Lopinavir	C37H48N4O5	628.81	CHEMBL729	25
7	Atazanavir	C38H52N6O7	704.87	CHEMBL1163	4
8	Tipranavir	C31H33F3N2O5S	602.68	CHEMBL222559	30
9	Darunavir	C27H37N3O7S	547.67	CHEMBL1323	3.5
10	Telinavir	C33H44N6O5	604.75	CHEMBL322241	6.3

1) Saquinavir:

First approved in 1995, Saquinavir is a small molecule whose structure in the SMILES format is CC(C)(C)NC(=O)[C@@H]1C[C@@H]2CCCC[C@@H]2CN1C[C@@H](O)[C@H](Cc1ccccc1)NC(=O)[C@H](CC(N)=O)NC(=O)c1ccc2ccccc2n1

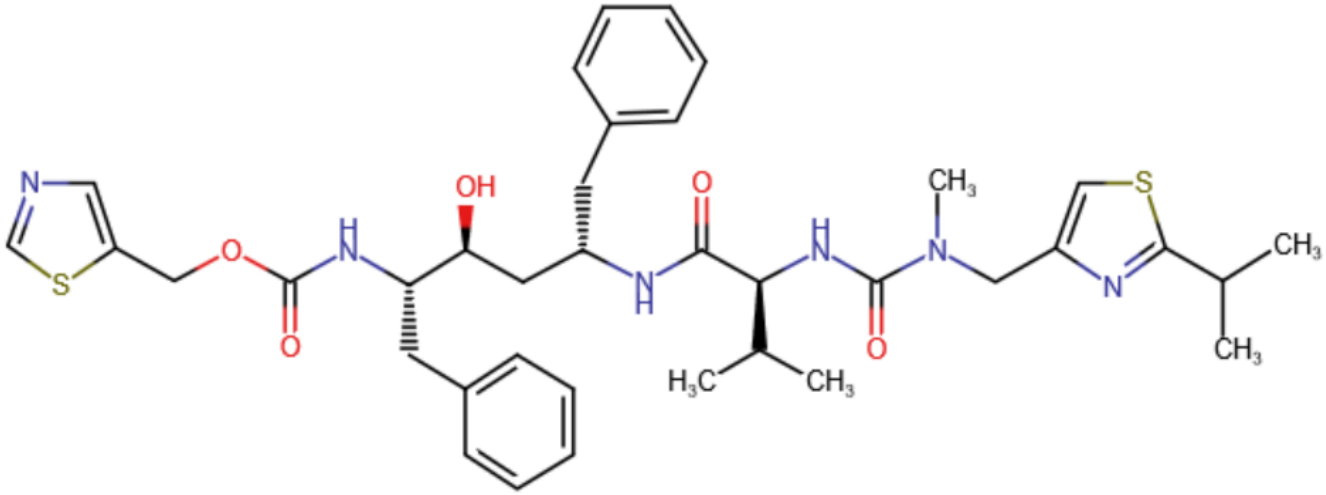
The structure of this small molecule is given below



2) Ritonavir:

First approved in 1996, Ritonavir is a small molecule whose structure in the SMILES format is CC(C)c1nc(CN(C)C(=O)N[C@H](C(=O)N[C@@H](Cc2ccccc2)C[C@H](O)[C@H](Cc2ccccc2)NC(=O)OCc2cncs2)C(C)C)cs1

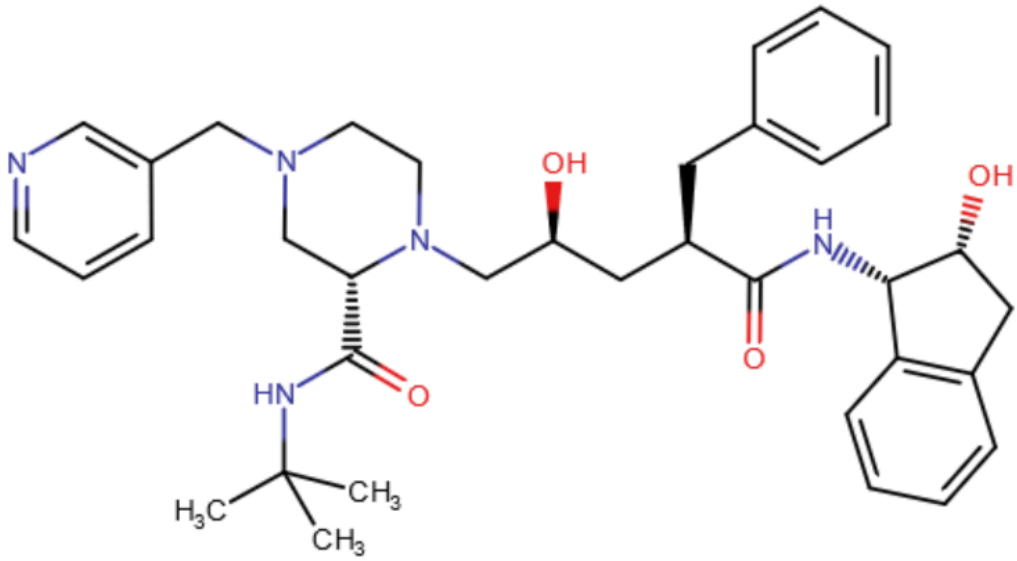
The structure of Ritonavir is given below



3) Indinavir

First approved in 1996, Indinavir is a small molecule whose structure in the SMILES format is CC(C)(C)NC(=O)[C@@H]1CN(Cc2cccnc2)CCN1C[C@@H](O)C[C@@H](Cc1ccccc1)C(=O)N[C@H]1c2ccccc2C[C@H]1O

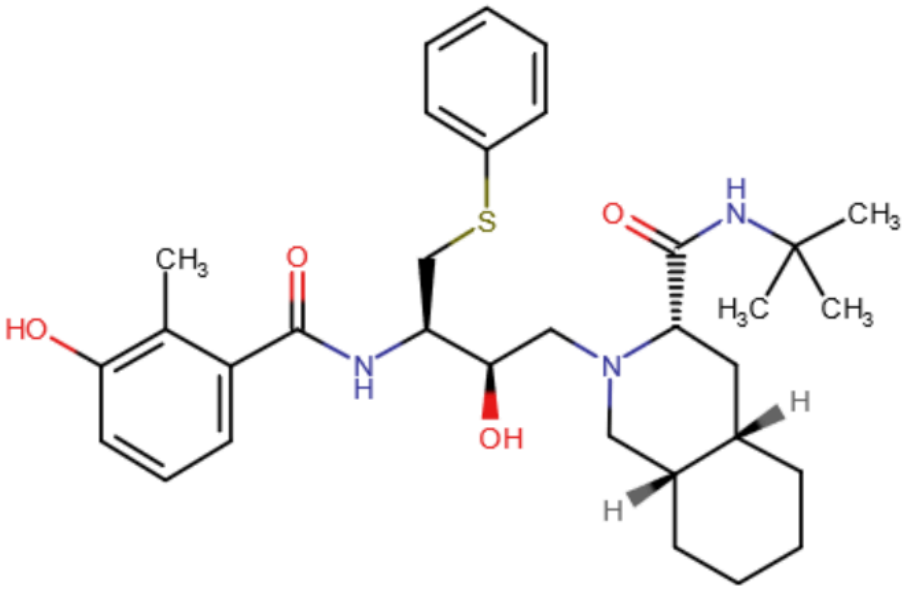
The structure of Indinavir is given below



4) Nelfinavir

First approved in 1997, Nelfinavir is a small molecule whose structure in the SMILES format is Cc1c(O)cccc1C(=O)N[C@@H](CSc1ccccc1)[C@H](O)CN1C[C@H]2CCCC[C@H]2C[C@H]1C(=O)NC(C)(C)C

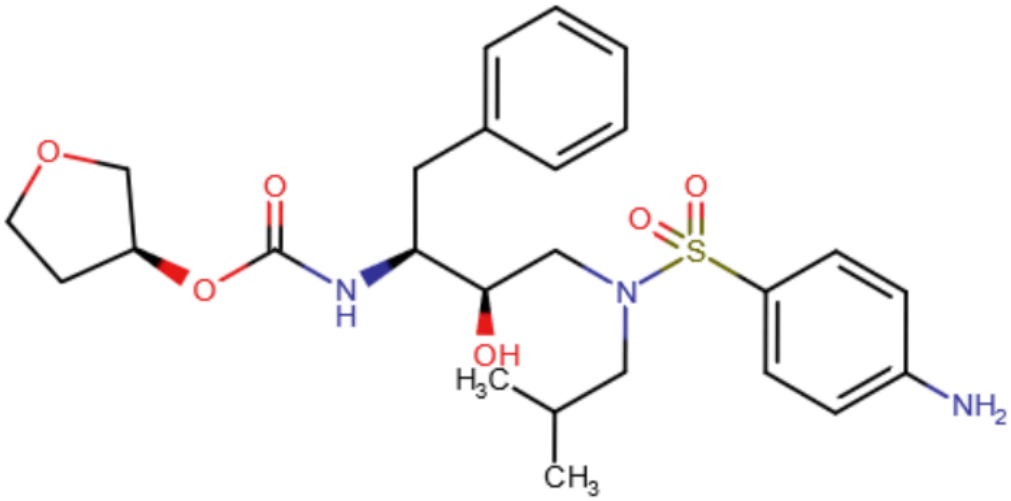
The structure of Nelfinavir is given below



5) Amprenavir

First approved in 1999, Amprenavir is a small molecule whose structure in the SMILES format is CC(C)CN(C[C@@H](O)[C@H](Cc1ccccc1)NC(=O)O[C@H]1CCOC1)S(=O)(=O)c1ccc(N)cc1

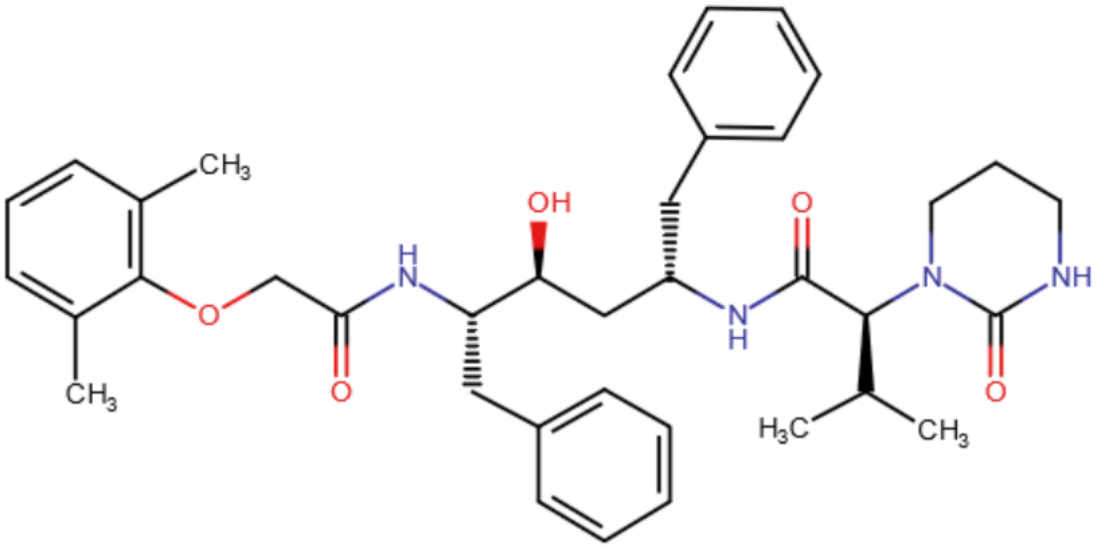
The structure of Amprenavir is given below



6) Lopinavir

First approved in 2000, Lopinavir is a small molecule whose structure in the SMILES format is Cc1cccc(C)c1OCC(=O)N[C@@H](Cc1ccccc1)[C@@H](O)C[C@H](Cc1ccccc1)NC(=O)[C@H](C(C)C)N1CCCNC1=O

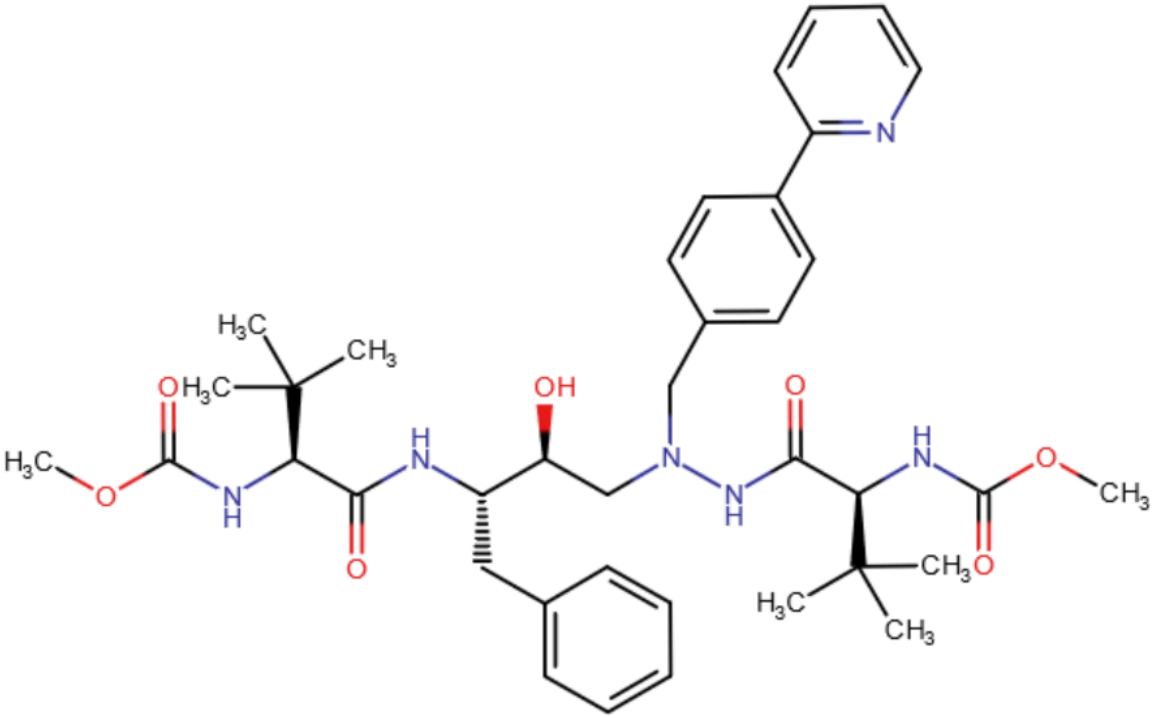
The structure of Lopinavir is given below



7) Atazanavir

First approved in 2003, Atazanavir is a small molecule whose structure in the SMILES format is COC(=O)N[C@H](C(=O)N[C@@H](Cc1ccccc1)[C@@H](O)CN(Cc1ccc(-c2ccccn2)cc1)NC(=O)[C@@H](NC(=O)OC)C(C)(C)C)C(C)(C)C

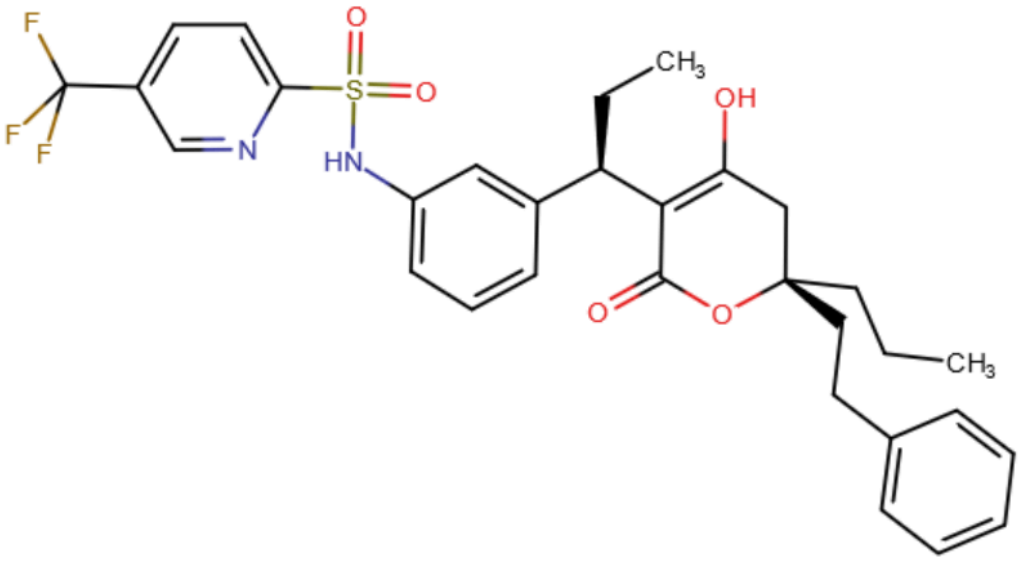
The structure of Atazanavir is given below



8) Tipranavir

First approved in 2005, Tipranavir is a small molecule whose structure in the SMILES format is CCC[C@@]1(CCc2ccccc2)CC(O)=C([C@H](CC)c2cccc(NS(=O)(=O)c3ccc(C(F)(F)F)cn3)c2)C(=O)O1

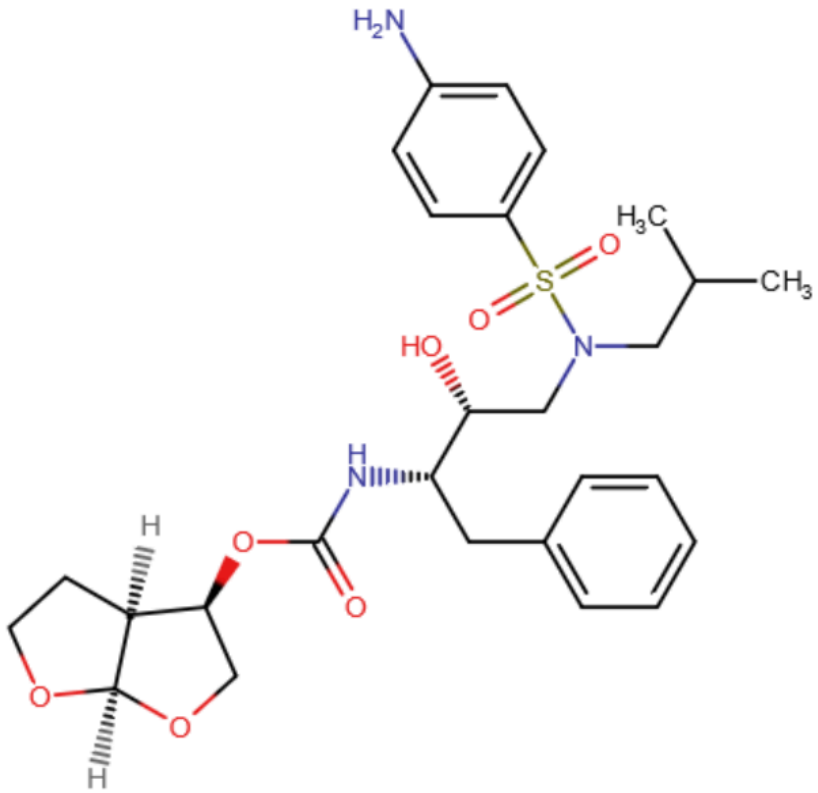
The structure of Tipranavir is given below



9) Darunavir

First approved in 2006, Darunavir is a small molecule whose structure in the SMILES format is CC(C)CN(C[C@@H](O)[C@H](Cc1ccccc1)NC(=O)O[C@H]1CO[C@H]2OCC[C@H]21)S(=O)(=O)c1ccc(N)cc1

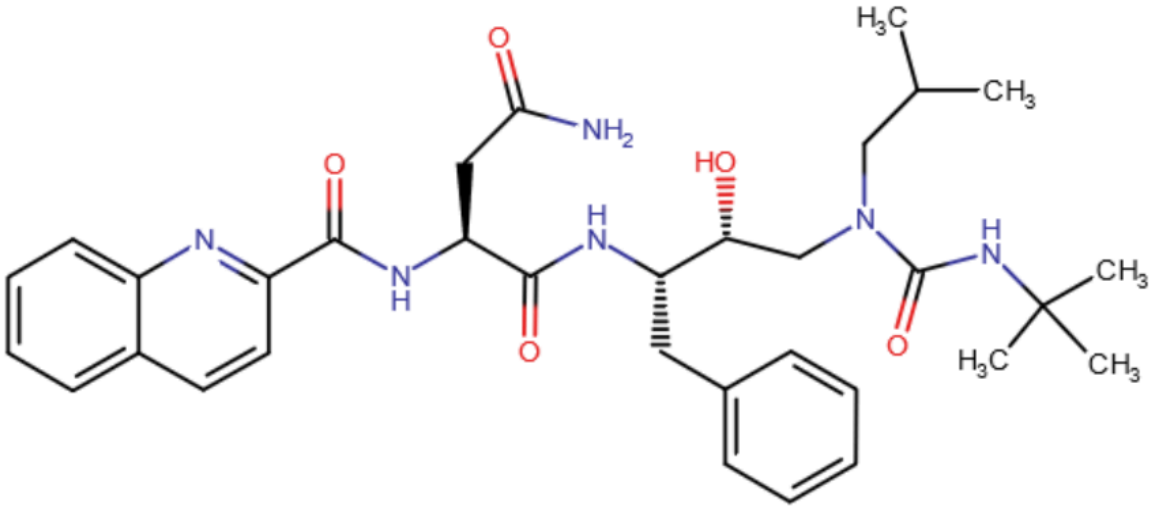
The structure of Darunavir is given below



10) Telinavir

Currently in Phase 2, Telinavir is a small molecule whose structure in the SMILES format is CC(C)CN(C[C@@H](O)[C@H](Cc1ccccc1)NC(=O)[C@H](CC(N)=O)NC(=O)c1ccc2ccccc2n1)C(=O)NC(C)(C)C

The structure of Telinavir is given below



The activity of all these drugs can be found using the ChEMBL database.

Methodology:

ChemDes:

We use ChemDes Chemopy descriptor calculator to get all the descriptor values for all the 10 drugs. The input taken to calculate the molecular descriptors is in the SMILES format which have been provided above, we calculated the 1D and 2D descriptors as part of this computational study. The descriptors that Chemopy uses are as follows :

1) Constitutional descriptors (30)
2) Connectivity descriptors (44)
3) Basak descriptors (21)
4) Topology descriptors (35)
5) Kappa descriptors (7)
6) Burden descriptors (64)
7) E-state descriptors (245)
8) Moran autocorrelation descriptors(32)
9) Geary autocorrelation descriptors(32)
10) Molecular property descriptors (6)
11) Moreau-Broto autocorrelation descriptors(32)
12) Charge descriptors (25)
13) MOE-type descriptors (60)

QSAR:

The data is split into test and train, we use the current FDA approved drugs (Drug 1-9 from table) and use telinavir to test the accuracy of the linear regression model. The model used is a multiple linear regression model with 2 descriptors the final workflow is as follows

Step 1: Creating a training Dataset with 9 drugs, corresponding descriptor and IC50 values

Step 2: Find the correlation matrix to find the two descriptors with high absolute correlation with IC50. We set a cut-off for minimum correlation with IC50 for a particular descriptor based on the maximum correlation value

Step 3: Find two descriptors which pass the cutoff correlation and also are not correlated highly with each other

Step 4: Perform Linear regression with two independent variables with IC50 value as the output.

Step 5: Find the error the model makes with the test data

Results:

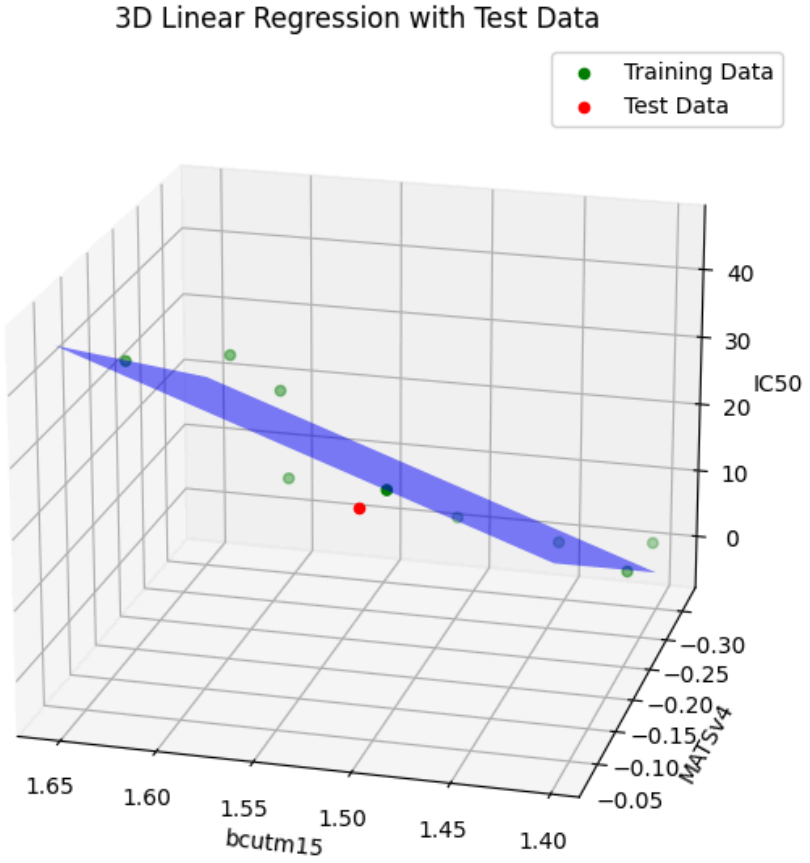
The correlation between IC50 and all descriptors has the maximum value of 0.78. The cutoff for correlation is hence kept at 0.7. Measuring all cross correlation terms we find that the two descriptors that are highly correlated with activity and are not themselves correlated are

Sl. no	Descriptor	Correlation with IC50
1	bcutm15	0.77
2	MATSv4	0.73

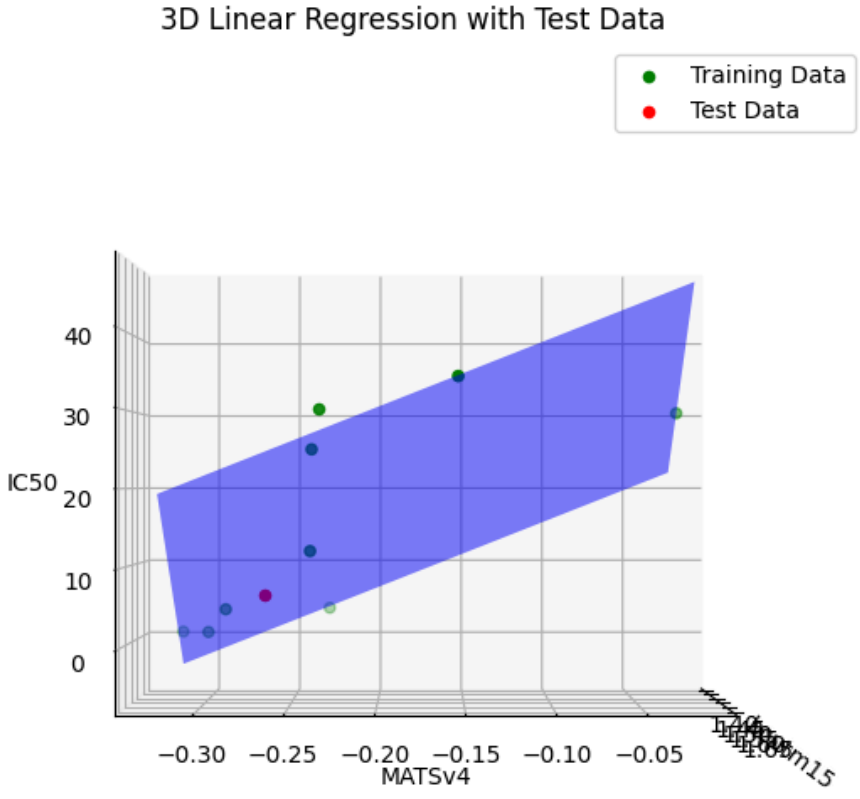
The cross-correlation between the two descriptors being **0.27**

Model is tested with the test data to give a RMSE of **8.75**

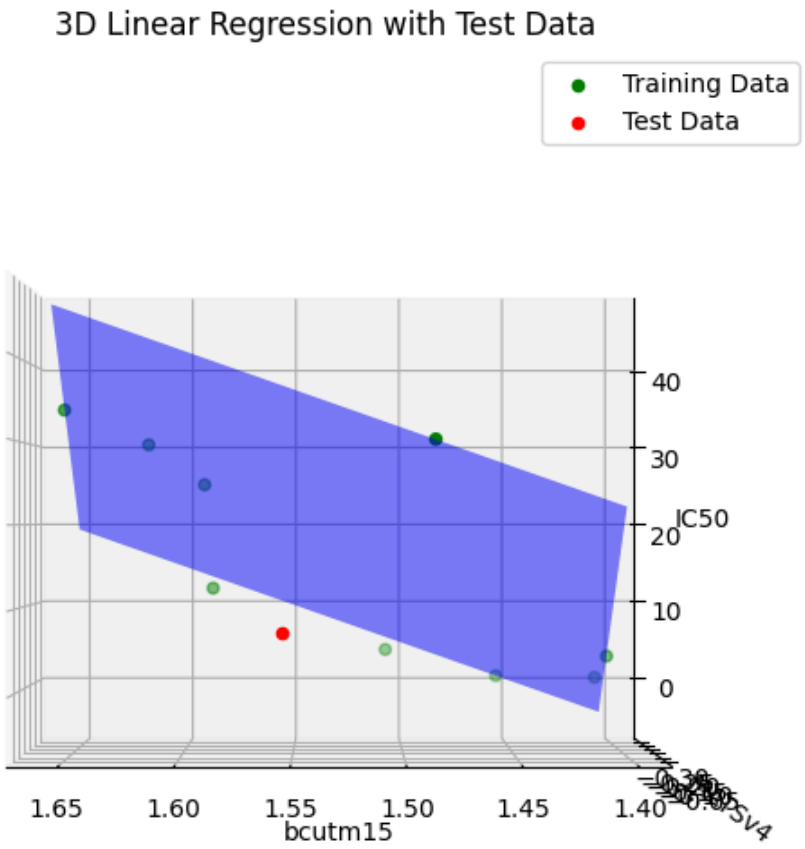
The regression model is plotted with the training data in red and test data in green


3D Linear Regression with Test Data
Training Data
Test Data
IC50
40
30
20
10
0
−0.30
−0.25
−0.20
−0.15
−0.10
−0.05
MATSv4
1.65
1.60
1.55
1.50
1.45
1.40
bcutm15

X-axis :


3D Linear Regression with Test Data
Training Data
Test Data
IC50
40
30
20
10
0
−0.30
−0.25
−0.20
−0.15
−0.10
−0.05
MATSv4

Y-axis :



Discussion:

The descriptors themselves have a moderately high correlation with the activity and are independent with respect to each other. The descriptors are bcutm15 and MATSv4 in the CHEMDES database.

Bcutm15:

This is a molecular descriptor derived from the Burden Matrix in QSAR modeling. BCUT stands for Burden CAS Unzaled Transformation and they represent the class of descriptors that encode electronic and geometric properties of the molecule . The **m** stands for molecular property like charge or polarizability and the number 15 represents the specific eigenvalue in the ranked list of eigenvalues of the Burden Matrix. The burden matrix encodes the atomic contributions and connectivity information of the drug/molecule.

MATSv4:

This is a Moran autocorrelation descriptor where MAT stands for Molecular Autocorrelation of Topological Structure. This descriptor captures spatial distribution of molecular properties across different atom pairs. This is done by measuring properties like atomic masses, van der waals volumes and Electronegativity and their distribution across atoms in the molecule as a function of atomic distance. The **v** here stands for van der Waals volume, which is the measure of the size of atoms in the

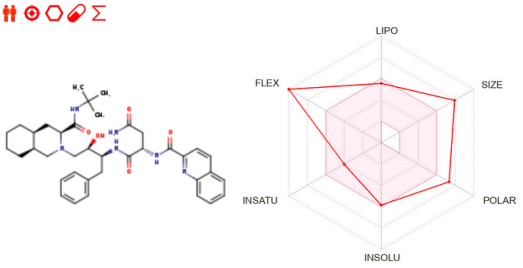
molecules. The number 4 represents the topological distance i.e, the bond separation between the atoms being considered.

We hence observe that for the HIV protease inhibitors the drugs we have chosen are such that their structure and specific atomic charges are responsible majorly for their activity. This provides us crucial information on how to model new drugs such that the activity is correspondingly higher.

To analyze further we check the ADME properties of each of the 10 drugs. We use the web server SWIZZ ADME for obtaining the information [7]

1) Saquinavir

Molecule 1



SMILES NC(=O)C[C@@H](C(=O)N[C@H]([C@@H](CN1C[C@H]2CCCC[C@H]2C[C@H]1C(=O)NC(C)(C)C)O)Cc1ccccc1)NC(=O)c1ccc2c(n1)cccc2

Physicochemical Properties	
Formula	C38H50N6O5
Molecular weight	670.84 g/mol
Num. heavy atoms	49
Num. arom. heavy atoms	16
Fraction Csp3	0.50
Num. rotatable bonds	16
Num. H-bond acceptors	7
Num. H-bond donors	5
Molar Refractivity	192.87
TPSA	166.75 Å²

Lipophilicity	
Log $P_{o/w}$ (iLOGP)	2.15
Log $P_{o/w}$ (XLOGP3)	4.24
Log $P_{o/w}$ (WLOGP)	2.71
Log $P_{o/w}$ (MLOGP)	1.40
Log $P_{o/w}$ (SILICOS-IT)	3.84
Consensus Log $P_{o/w}$	2.87

Water Solubility	
Log *S* (ESOL)	-5.86
Solubility	9.34e-04 mg/ml ; 1.39e-06 mol/l
Class	Moderately soluble
Log *S* (Ali)	-7.45
Solubility	2.36e-05 mg/ml ; 3.53e-08 mol/l
Class	Poorly soluble
Log *S* (SILICOS-IT)	-8.54
Solubility	1.91e-06 mg/ml ; 2.85e-09 mol/l
Class	Poorly soluble

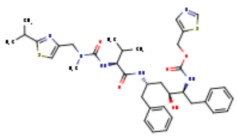
Pharmacokinetics	
GI absorption	Low
BBB permeant	No
P-gp substrate	Yes
CYP1A2 inhibitor	No
CYP2C19 inhibitor	No
CYP2C9 inhibitor	No
CYP2D6 inhibitor	No
CYP3A4 inhibitor	Yes
Log K_p (skin permeation)	-7.38 cm/s

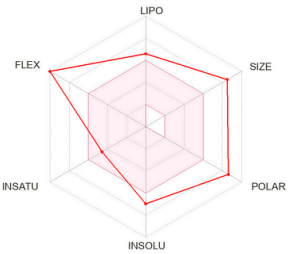
Druglikeness	
Lipinski	No; 2 violations: MW>500, NorO>10
Ghose	No; 3 violations: MW>480, MR>130, #atoms>70
Veber	No; 2 violations: Rotors>10, TPSA>140
Egan	No; 1 violation: TPSA>131.6
Muegge	No; 3 violations: MW>600, TPSA>150, Rotors>15
Bioavailability Score	0.17

Medicinal Chemistry	
PAINS	0 alert
Brenk	0 alert
Leadlikeness	No; 3 violations: MW>350, Rotors>7, XLOGP3>3.5
Synthetic accessibility	5.94

2) Ritonavir

Molecule 1





SMILES O=C(N[C@H]([C@H](C[C@H](Cc1ccccc1)NC(=O)[C@H](C(C)C)NC(=O)N(Cc1csc(n1)C(C)C)C)O)Cc1ccccc1)OCc1cncs1

Physicochemical Properties	
Formula	C37H48N6O5S2
Molecular weight	720.94 g/mol
Num. heavy atoms	50
Num. arom. heavy atoms	22
Fraction Csp3	0.43
Num. rotatable bonds	22
Num. H-bond acceptors	7
Num. H-bond donors	4
Molar Refractivity	197.82
TPSA	202.26 Å²

Lipophilicity	
Log $P_{o/w}$ (iLOGP)	5.18
Log $P_{o/w}$ (XLOGP3)	6.05
Log $P_{o/w}$ (WLOGP)	5.60
Log $P_{o/w}$ (MLOGP)	1.80
Log $P_{o/w}$ (SILICOS-IT)	7.33
Consensus Log $P_{o/w}$	5.19

Water Solubility	
Log *S* (ESOL)	-6.99
Solubility	7.29e-05 mg/ml ; 1.01e-07 mol/l
Class	Poorly soluble
Log *S* (Ali)	-10.08
Solubility	6.04e-08 mg/ml ; 8.38e-11 mol/l
Class	Insoluble
Log *S* (SILICOS-IT)	-10.02
Solubility	6.87e-08 mg/ml ; 9.53e-11 mol/l
Class	Insoluble

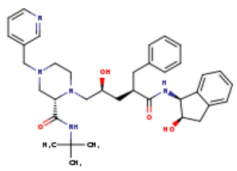
Pharmacokinetics	
GI absorption	Low
BBB permeant	No
P-gp substrate	Yes
CYP1A2 inhibitor	No
CYP2C19 inhibitor	No
CYP2C9 inhibitor	No
CYP2D6 inhibitor	No
CYP3A4 inhibitor	Yes
Log K_p (skin permeation)	-6.40 cm/s

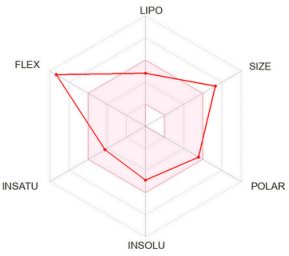
Druglikeness	
Lipinski	No; 2 violations: MW>500, NorO>10
Ghose	No; 4 violations: MW>480, WLOGP>5.6, MR>130, #atoms>70
Veber	No; 2 violations: Rotors>10, TPSA>140
Egan	No; 1 violation: TPSA>131.6
Muegge	No; 4 violations: MW>600, XLOGP3>5, TPSA>150, Rotors>15
Bioavailability Score	0.17

Medicinal Chemistry	
PAINS	0 alert
Brenk	0 alert
Leadlikeness	No; 3 violations: MW>350, Rotors>7, XLOGP3>3.5
Synthetic accessibility	6.45

3) Indinavir

Molecule 1





SMILES O[C@H](CN1CCN(C[C@H]1C(=O)NC(C)(C)C)Cc1cccnc1)C[C@H](C(=O)N[C@@H]1[C@H](O)Cc2c1cccc2)Cc1ccccc1

Physicochemical Properties	
Formula	C36H47N5O4
Molecular weight	613.79 g/mol
Num. heavy atoms	45
Num. arom. heavy atoms	18
Fraction Csp3	0.47
Num. rotatable bonds	14
Num. H-bond acceptors	7
Num. H-bond donors	4
Molar Refractivity	182.62
TPSA	118.03 Å²

Lipophilicity	
Log $P_{o/w}$ (iLOGP)	4.05
Log $P_{o/w}$ (XLOGP3)	2.92
Log $P_{o/w}$ (WLOGP)	1.63
Log $P_{o/w}$ (MLOGP)	1.33
Log $P_{o/w}$ (SILICOS-IT)	3.97
Consensus Log $P_{o/w}$	2.78

Water Solubility	
Log *S* (ESOL)	-4.86
Solubility	8.53e-03 mg/ml ; 1.39e-05 mol/l
Class	Moderately soluble
Log *S* (Ali)	-5.06
Solubility	5.35e-03 mg/ml ; 8.71e-06 mol/l
Class	Moderately soluble
Log *S* (SILICOS-IT)	-8.44
Solubility	2.25e-06 mg/ml ; 3.67e-09 mol/l
Class	Poorly soluble

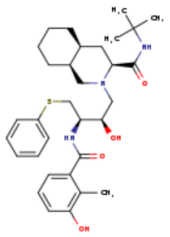
Pharmacokinetics	
GI absorption	High
BBB permeant	No
P-gp substrate	Yes
CYP1A2 inhibitor	No
CYP2C19 inhibitor	No
CYP2C9 inhibitor	No
CYP2D6 inhibitor	No
CYP3A4 inhibitor	No
Log K_p (skin permeation)	-7.97 cm/s

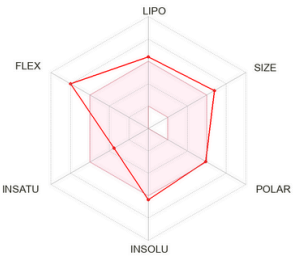
Druglikeness	
Lipinski	Yes; 1 violation: MW>500
Ghose	No; 3 violations: MW>480, MR>130, #atoms>70
Veber	No; 1 violation: Rotors>10
Egan	Yes
Muegge	No; 1 violation: MW>600
Bioavailability Score	0.55

Medicinal Chemistry	
PAINS	0 alert
Brenk	0 alert
Leadlikeness	No; 2 violations: MW>350, Rotors>7
Synthetic accessibility	5.60

4) Nelfinavir

Molecule 1





SMILES O[C@@H]([C@@H](NC(=O)c1cccc(c1C)O)CSc1ccccc1)CN1C[C@H]2CCCC[C@H]2C[C@H]1C(=O)NC(C)(C)C

Physicochemical Properties	
Formula	C32H45N3O4S
Molecular weight	567.78 g/mol
Num. heavy atoms	40
Num. arom. heavy atoms	12
Fraction Csp3	0.56
Num. rotatable bonds	12
Num. H-bond acceptors	5
Num. H-bond donors	4
Molar Refractivity	166.17
TPSA	127.20 Å²

Lipophilicity	
Log $P_{o/w}$ (iLOGP)	4.24
Log $P_{o/w}$ (XLOGP3)	5.67
Log $P_{o/w}$ (WLOGP)	4.37
Log $P_{o/w}$ (MLOGP)	3.20
Log $P_{o/w}$ (SILICOS-IT)	4.56
Consensus Log $P_{o/w}$	4.41

Water Solubility	
Log *S* (ESOL)	-6.36
Solubility	2.47e-04 mg/ml ; 4.34e-07 mol/l
Class	Poorly soluble
Log *S* (Ali)	-8.11
Solubility	4.45e-06 mg/ml ; 7.83e-09 mol/l
Class	Poorly soluble
Log *S* (SILICOS-IT)	-7.26
Solubility	3.11e-05 mg/ml ; 5.48e-08 mol/l
Class	Poorly soluble

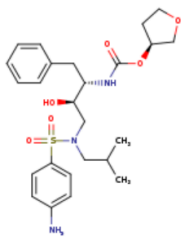
Pharmacokinetics	
GI absorption	Low
BBB permeant	No
P-gp substrate	Yes
CYP1A2 inhibitor	No
CYP2C19 inhibitor	Yes
CYP2C9 inhibitor	No
CYP2D6 inhibitor	No
CYP3A4 inhibitor	Yes
Log K_p (skin permeation)	-5.74 cm/s

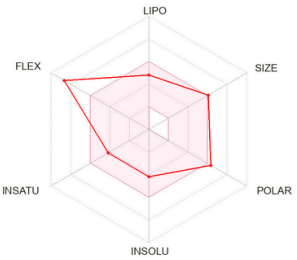
Druglikeness	
Lipinski	Yes; 1 violation: MW>500
Ghose	No; 3 violations: MW>480, MR>130, #atoms>70
Veber	No; 1 violation: Rotors>10
Egan	Yes
Muegge	No; 1 violation: XLOGP3>5
Bioavailability Score	0.55

Medicinal Chemistry	
PAINS	0 alert
Brenk	0 alert
Leadlikeness	No; 3 violations: MW>350, Rotors>7, XLOGP3>3.5
Synthetic accessibility	5.58

5) Amprenavir

Molecule 1





SMILES CC(CN(S(=O)(=O)c1ccc(cc1)N)C[C@H]([C@H](Cc1ccccc1)NC(=O)O[C@H]1CCOC1)O)C

Physicochemical Properties	
Formula	C25H35N3O6S
Molecular weight	505.63 g/mol
Num. heavy atoms	35
Num. arom. heavy atoms	12
Fraction Csp3	0.48
Num. rotatable bonds	13
Num. H-bond acceptors	7
Num. H-bond donors	3
Molar Refractivity	133.62
TPSA	139.57 Å²

Lipophilicity	
Log $P_{o/w}$ (iLOGP)	3.46
Log $P_{o/w}$ (XLOGP3)	2.89
Log $P_{o/w}$ (WLOGP)	3.49
Log $P_{o/w}$ (MLOGP)	1.14
Log $P_{o/w}$ (SILICOS-IT)	1.78
Consensus Log $P_{o/w}$	2.55

Water Solubility	
Log *S* (ESOL)	-4.19
Solubility	3.25e-02 mg/ml ; 6.44e-05 mol/l
Class	Moderately soluble
Log *S* (Ali)	-5.48
Solubility	1.67e-03 mg/ml ; 3.30e-06 mol/l
Class	Moderately soluble
Log *S* (SILICOS-IT)	-5.55
Solubility	1.42e-03 mg/ml ; 2.81e-06 mol/l
Class	Moderately soluble

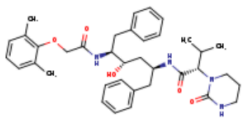
Pharmacokinetics	
GI absorption	Low
BBB permeant	No
P-gp substrate	Yes
CYP1A2 inhibitor	No
CYP2C19 inhibitor	No
CYP2C9 inhibitor	No
CYP2D6 inhibitor	Yes
CYP3A4 inhibitor	Yes
Log K_p (skin permeation)	-7.33 cm/s

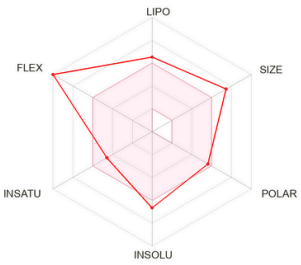
Druglikeness	
Lipinski	Yes; 1 violation: MW>500
Ghose	No; 2 violations: MW>480, MR>130
Veber	No; 1 violation: Rotors>10
Egan	No; 1 violation: TPSA>131.6
Muegge	Yes
Bioavailability Score	0.55

Medicinal Chemistry	
PAINS	0 alert
Brenk	1 alert: aniline
Leadlikeness	No; 2 violations: MW>350, Rotors>7
Synthetic accessibility	4.98

6) Lopinavir

Molecule 1





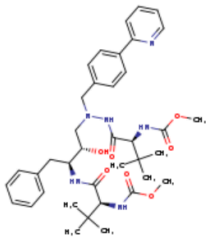
SMILES O=C(N[C@H]([C@H](C[C@H](Cc1ccccc1)NC(=O)[C@@H](N1CCCNC1=O)C(C)C)O)Cc1ccccc1)COc1c(C)cccc1C

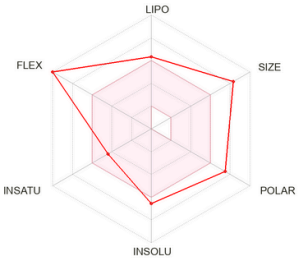
Physicochemical Properties	
Formula	C37H48N4O5
Molecular weight	628.80 g/mol
Num. heavy atoms	46
Num. arom. heavy atoms	18
Fraction Csp3	0.43
Num. rotatable bonds	17
Num. H-bond acceptors	5
Num. H-bond donors	4
Molar Refractivity	187.92
TPSA	120.00 Å²
Lipophilicity	
Log $P_{o/w}$ (iLOGP)	3.75
Log $P_{o/w}$ (XLOGP3)	5.92
Log $P_{o/w}$ (WLOGP)	3.57
Log $P_{o/w}$ (MLOGP)	2.93
Log $P_{o/w}$ (SILICOS-IT)	6.02
Consensus Log $P_{o/w}$	4.44

Water Solubility	
Log *S* (ESOL)	-6.64
Solubility	1.45e-04 mg/ml ; 2.31e-07 mol/l
Class	Poorly soluble
Log *S* (Ali)	-8.21
Solubility	3.84e-06 mg/ml ; 6.10e-09 mol/l
Class	Poorly soluble
Log *S* (SILICOS-IT)	-10.05
Solubility	5.57e-08 mg/ml ; 8.85e-11 mol/l
Class	Insoluble
Pharmacokinetics	
GI absorption	High
BBB permeant	No
P-gp substrate	Yes
CYP1A2 inhibitor	No
CYP2C19 inhibitor	Yes
CYP2C9 inhibitor	No
CYP2D6 inhibitor	No
CYP3A4 inhibitor	Yes
Log K_p (skin permeation)	-5.93 cm/s
Druglikeness	
Lipinski	Yes; 1 violation: MW>500
Ghose	No; 3 violations: MW>480, MR>130, #atoms>70
Veber	No; 1 violation: Rotors>10
Egan	Yes
Muegge	No; 3 violations: MW>600, XLOGP3>5, Rotors>15
Bioavailability Score	0.55
Medicinal Chemistry	
PAINS	0 alert
Brenk	0 alert
Leadlikeness	No; 3 violations: MW>350, Rotors>7, XLOGP3>3.5
Synthetic accessibility	5.67

7) Atazanavir

Molecule 1





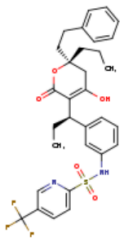
SMILES COC(=O)N[C@@H](C(C)(C)C)C(=O)NN(Cc1ccc(cc1)c1ccccn1)C[C@@H]([C@H](Cc1ccccc1)NC(=O)[C@H](C(C)(C)C)NC(=O)OC)O

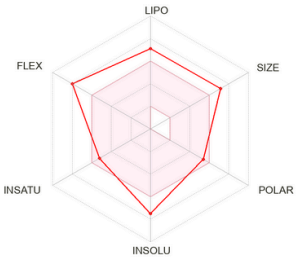
Physicochemical Properties	
Formula	C38H52N6O7
Molecular weight	704.86 g/mol
Num. heavy atoms	51
Num. arom. heavy atoms	18
Fraction Csp3	0.45
Num. rotatable bonds	22
Num. H-bond acceptors	9
Num. H-bond donors	5
Molar Refractivity	193.76
TPSA	171.22 Å²
Lipophilicity	
Log $P_{o/w}$ (iLOGP)	4.56
Log $P_{o/w}$ (XLOGP3)	5.60
Log $P_{o/w}$ (WLOGP)	4.06
Log $P_{o/w}$ (MLOGP)	1.76
Log $P_{o/w}$ (SILICOS-IT)	4.11
Consensus Log $P_{o/w}$	4.02

Water Solubility	
Log *S* (ESOL)	-6.55
Solubility	2.00e-04 mg/ml ; 2.84e-07 mol/l
Class	Poorly soluble
Log *S* (Ali)	-8.96
Solubility	7.77e-07 mg/ml ; 1.10e-09 mol/l
Class	Poorly soluble
Log *S* (SILICOS-IT)	-9.46
Solubility	2.47e-07 mg/ml ; 3.50e-10 mol/l
Class	Poorly soluble
Pharmacokinetics	
GI absorption	Low
BBB permeant	No
P-gp substrate	Yes
CYP1A2 inhibitor	No
CYP2C19 inhibitor	No
CYP2C9 inhibitor	No
CYP2D6 inhibitor	No
CYP3A4 inhibitor	Yes
Log K_p (skin permeation)	-6.62 cm/s
Druglikeness	
Lipinski	No; 2 violations: MW>500, NorO>10
Ghose	No; 3 violations: MW>480, MR>130, #atoms>70
Veber	No; 2 violations: Rotors>10, TPSA>140
Egan	No; 1 violation: TPSA>131.6
Muegge	No; 4 violations: MW>600, XLOGP3>5, TPSA>150, Rotors>15
Bioavailability Score	0.17
Medicinal Chemistry	
PAINS	0 alert
Brenk	1 alert: more_than_2_esters
Leadlikeness	No; 3 violations: MW>350, Rotors>7, XLOGP3>3.5
Synthetic accessibility	6.24

8) Tipranavir

Molecule 1





SMILES CCC[C@@]1(CCc2ccccc2)CC(=C(C(=O)O1)[C@@H](c1cccc(c1)NS(=O)(=O)c1ccc(cn1)C(F)(F)F)CC)O

Physicochemical Properties	
Formula	C31H33F3N2O5S
Molecular weight	602.66 g/mol
Num. heavy atoms	42
Num. arom. heavy atoms	18
Fraction Csp3	0.35
Num. rotatable bonds	12
Num. H-bond acceptors	9
Num. H-bond donors	2
Molar Refractivity	153.80
TPSA	113.97 Å²

Lipophilicity	
Log $P_{o/w}$ (iLOGP)	3.48
Log $P_{o/w}$ (XLOGP3)	6.97
Log $P_{o/w}$ (WLOGP)	9.37
Log $P_{o/w}$ (MLOGP)	3.74
Log $P_{o/w}$ (SILICOS-IT)	6.55
Consensus Log $P_{o/w}$	6.02

Water Solubility	
Log *S* (ESOL)	-7.49
Solubility	1.94e-05 mg/ml ; 3.22e-08 mol/l
Class	Poorly soluble
Log *S* (Ali)	-9.18
Solubility	4.01e-07 mg/ml ; 6.65e-10 mol/l
Class	Poorly soluble
Log *S* (SILICOS-IT)	-10.86
Solubility	8.24e-09 mg/ml ; 1.37e-11 mol/l
Class	Insoluble

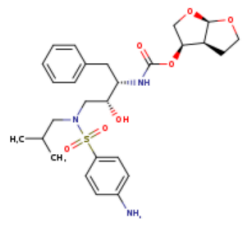
Pharmacokinetics	
GI absorption	Low
BBB permeant	No
P-gp substrate	Yes
CYP1A2 inhibitor	No
CYP2C19 inhibitor	Yes
CYP2C9 inhibitor	No
CYP2D6 inhibitor	Yes
CYP3A4 inhibitor	Yes
Log K_p (skin permeation)	-5.03 cm/s

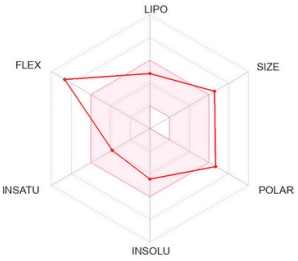
Druglikeness	
Lipinski	Yes; 1 violation: MW>500
Ghose	No; 4 violations: MW>480, WLOGP>5.6, MR>130, #atoms>70
Veber	No; 1 violation: Rotors>10
Egan	No; 1 violation: WLOGP>5.88
Muegge	No; 2 violations: MW>600, XLOGP3>5
Bioavailability Score	0.56

Medicinal Chemistry	
PAINS	0 alert
Brenk	0 alert
Leadlikeness	No; 3 violations: MW>350, Rotors>7, XLOGP3>3.5
Synthetic accessibility	5.29

9) Darunavir

Molecule 1





SMILES CC(CN(S(=O)(=O)c1ccc(cc1)N)C[C@H]([C@H](Cc1ccccc1)NC(=O)O[C@H]1CO[C@@H]2[C@H]1CCO2)O)C

Physicochemical Properties	
Formula	C27H37N3O7S
Molecular weight	547.66 g/mol
Num. heavy atoms	38
Num. arom. heavy atoms	12
Fraction Csp3	0.52
Num. rotatable bonds	13
Num. H-bond acceptors	8
Num. H-bond donors	3
Molar Refractivity	142.20
TPSA	148.80 Å²

Lipophilicity	
Log $P_{o/w}$ (iLOGP)	3.30
Log $P_{o/w}$ (XLOGP3)	2.94
Log $P_{o/w}$ (WLOGP)	3.46
Log $P_{o/w}$ (MLOGP)	1.18
Log $P_{o/w}$ (SILICOS-IT)	1.46
Consensus Log $P_{o/w}$	2.47

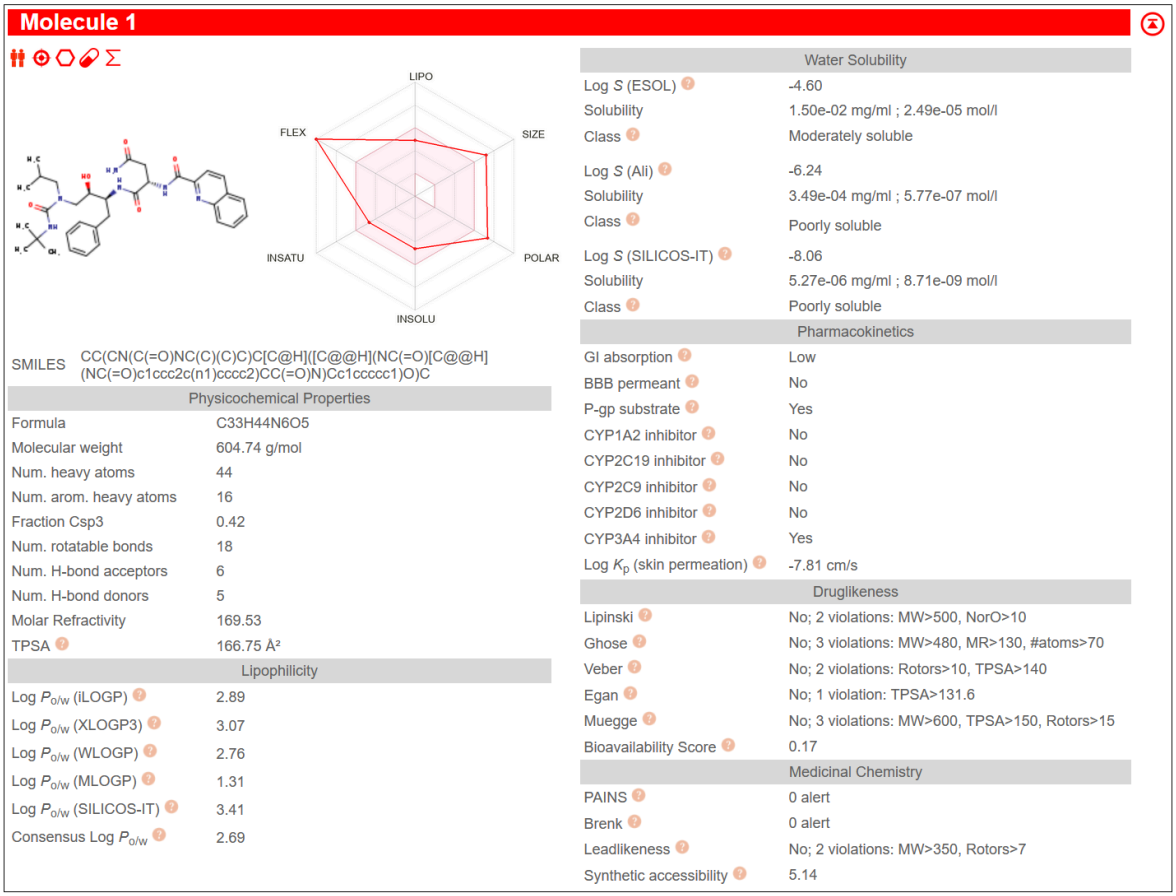
Water Solubility	
Log *S* (ESOL)	-4.46
Solubility	1.88e-02 mg/ml ; 3.44e-05 mol/l
Class	Moderately soluble
Log *S* (Ali)	-5.73
Solubility	1.03e-03 mg/ml ; 1.88e-06 mol/l
Class	Moderately soluble
Log *S* (SILICOS-IT)	-5.33
Solubility	2.55e-03 mg/ml ; 4.66e-06 mol/l
Class	Moderately soluble

Pharmacokinetics	
GI absorption	Low
BBB permeant	No
P-gp substrate	Yes
CYP1A2 inhibitor	No
CYP2C19 inhibitor	No
CYP2C9 inhibitor	No
CYP2D6 inhibitor	No
CYP3A4 inhibitor	Yes
Log K_p (skin permeation)	-7.55 cm/s

Druglikeness	
Lipinski	Yes; 1 violation: MW>500
Ghose	No; 3 violations: MW>480, MR>130, #atoms>70
Veber	No; 2 violations: Rotors>10, TPSA>140
Egan	No; 1 violation: TPSA>131.6
Muegge	Yes
Bioavailability Score	0.55

Medicinal Chemistry	
PAINS	0 alert
Brenk	1 alert: aniline
Leadlikeness	No; 2 violations: MW>350, Rotors>7
Synthetic accessibility	5.67

10) Telinavir



We observe that all the drug molecules rank really high on flexibility with each of the drugs ranging from moderately soluble to insoluble in nature. The test drug which is currently in Phase II trials is extremely flexible and polar.

Conclusion:

The QSAR study performed on the HIV protease inhibitor drugs showed that the drugs flexibility and polar properties are highly correlated with the drug activity. This study was performed with only two descriptors and was performed using the data acquired from the CHEMDES server. For further studies using the E-dragon server, which contains close to 2000 descriptors would serve as a better alternative. The study was also restricted to 2D QSAR, hence performing 3D QSAR can be under further scope of this study. The current literature available on the drugs activity is diverse for every particular drug, and the IC50 values range widely for a few. The test error for the model gave a RMSE of 8.75. Both the descriptors themselves are positively correlated with the activity of the drug molecules. Currently Telinavir is undergoing Phase II trials for FDA approval, and based on this study we find that telinavir performs similarly to all the current drugs that are available in the market.

References:

[1] "HIV." *World Health Organization (WHO)*, https://www.who.int/data/gho/data/themes/hiv-aids.

[2] LiverTox: Clinical and Research Information on Drug-Induced Liver Injury [Internet]. Bethesd (MD): National Institute of Diabetes and Digestive and Kidney Diseases; 2012-. Protease Inhibitors (HIV) [Updated 2017 Sep 1]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK548893/

[3] Recent Progress in the Development of HIV-1 Protease Inhibitors for the Treatment of HIV/AIDS Arun K. Ghosh, Heather L. Osswald, and Gary Prato Journal of Medicinal Chemistry 2016 59 (11), 5172-5208 DOI: 10.1021/acs.jmedchem.5b01697

[4] Roy, K., Kar, S., & Das, R. N. (2014). Introduction to 3D-QSAR. *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*, 291-317. https://doi.org/10.1016/B978-0-12-801505-6.00008-9

[5] Dong, J., Cao, DS., Miao, HY. et al. ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation. J Cheminform 7, 60 (2015). https://doi.org/10.1186/s13321-015-0109-z

[6] Zdrazil B, Felix E, Hunter F, Manners EJ, Blackshaw J, Corbett S, de Veij M, Ioannidis H, Lopez DM, Mosquera JF, Magarinos MP, Bosc N, Arcila R, Kizilören T, Gaulton A, Bento AP, Adasme MF, Monecke P, Landrum GA, Leach AR. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Res. 2024 Jan 5;52(D1):D1180-D1192. doi: 10.1093/nar/gkad1004. PMID: 37933841; PMCID: PMC10767899.

[7] SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. Sci. Rep. (2017) 7:42717.

Website Links:

ChEMBL : https://www.ebi.ac.uk/chembl/
CHEMDES : http://www.scbdd.com/chemopy_desc/index/
SwissADME: http://www.swissadme.ch/

Human immunodeficiency virus type 1 protease ChEMBL :
https://www.ebi.ac.uk/chembl/web_components/explore/target/CHEMBL243#LigandEfficiencies
Drugs and their activities:

1) Saquinavir: https://www.ebi.ac.uk/chembl/web_components/explore/compound/CHEMBL114
2) Ritonavir: https://www.ebi.ac.uk/chembl/web_components/explore/compound/CHEMBL163
3) Indinavir: https://www.ebi.ac.uk/chembl/web_components/explore/compound/CHEMBL115
4) Nelfinavir: https://www.ebi.ac.uk/chembl/web_components/explore/compound/CHEMBL584

5) Amprenavir: https://www.ebi.ac.uk/chembl/web_components/explore/compound/CHEMBL116
6) Lopinavir: https://www.ebi.ac.uk/chembl/web_components/explore/compound/CHEMBL729
7) Atazanavir: https://www.ebi.ac.uk/chembl/web_components/explore/compound/CHEMBL1163
8) Tipranavir: https://www.ebi.ac.uk/chembl/web_components/explore/compound/CHEMBL222559
9) Darunavir: https://www.ebi.ac.uk/chembl/web_components/explore/compound/CHEMBL1323
10) Telinavir: https://www.ebi.ac.uk/chembl/web_components/explore/compound/CHEMBL322241

Appendix:

The dataset used for testing and training can be obtained in this google drive link.
The python script used to plot and create the regression model is attached here.

Code:

```
import pandas as pd
import numpy as np
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
from mpl_toolkits.mplot3d import Axes3D
import plotly.graph_objects as go
import warnings
warnings.filterwarnings("ignore")

# Training Data
df1 = pd.read_csv('/content/Amprenavir.csv')
df2 = pd.read_csv("/content/Indinavir.csv")
df3 = pd.read_csv("/content/Nelfinavir.csv")
df4 = pd.read_csv("/content/Ritonavir.csv")
df5 = pd.read_csv("/content/Saquinavir.csv")
df6 = pd.read_csv("/content/Atazanavir.csv")
df7 = pd.read_csv("/content/darunavir.csv")
df8 = pd.read_csv("/content/Lopinavir.csv")
df9 = pd.read_csv("/content/Tipranavir.csv")

# Test Data
df10 = pd.read_csv("/content/telinavir.csv")
# Create the Training Data
Train = pd.concat([df1,df2,df3,df4,df5,df6,df7,df8,df9], axis=0)
Train.reset_index(drop=True, inplace=True)
Train.insert(0, 'Drug',
['Amprenavir','Indinavir','Nelfinavir','Ritonavir','Saquinavir', 'Atazanavir',
'darunavir','Lopinavir','Tipranavir'])
```

```
Train['IC50'] =  [0.23, 34, 0.56, 12, 30, 4, 3.5,25,30]
print(Train)

Test = pd.concat([df10])
Test.reset_index(drop=True,inplace=True)
Test.insert(0,'Drug',['telinavir'])
Test['IC50'] = [6.3]
print(Test)

# Correlation between IC50 and every Descriptor
numerical_data = Train.drop(columns=['Drug'])
corr_matrix = numerical_data.corr()
ic50_sorted = corr_matrix['IC50'].sort_values(ascending=False)
ic50_clean = ic50_sorted.dropna()

# Sorted Correlation to observe the highest
print('Sorted Correlation')
print(ic50_clean[1:])

# Cross correlation between every descriptors
cross_corr = []

for i in range(1,len(ic50_clean)):
    for j in range(i,len(ic50_clean)):
        # Each of their individual correlation with the activity should be > 0.7
        if abs(ic50_clean[i]) > 0.7 and abs(ic50_clean[j]) > 0.7:
            correlation = Train[ic50_clean.index[i]].corr(Train[ic50_clean.index[j]])
            cross_corr.append([i,j,correlation])

# Least cross correlation between the descriptors
min_index = min(range(len(cross_corr)), key=lambda i: abs(cross_corr[i][2]))
min_corr = cross_corr[min_index]

print('Final Choice')
print(f'Descriptor A : {ic50_clean.index[min_corr[0]]}, correlation with activity :
{ic50_clean[min_corr[0]]}')
print(f'Descriptor B : {ic50_clean.index[min_corr[1]]}, correlation with activity :
{ic50_clean[min_corr[1]]}')
print(f'The cross-correlation between the two descriptors being {min_corr[2]}')
```

```
# Interactive Plot
A = ic50_clean.index[min_corr[0]]
B = ic50_clean.index[min_corr[1]]
```

```
QSAR_Train = Train[['Drug',A,B,'IC50']]

# QSAR_Train.to_csv('QSAR_Train.csv')   # If needed to save

df_test = Test[['Drug',A,B,'IC50']]
X = QSAR_Train[[A,B]]
y = QSAR_Train['IC50']

# Create the linear regression model
model = LinearRegression()
model.fit(X, y)

X_test = df_test[[A,B]]
y_test = df_test['IC50']
y_test_pred = model.predict(X_test)

mse = mean_squared_error(y_test, y_test_pred)
print(f"Root Mean Squared Error on test data: {mse**0.5}")

# Generate predictions for plotting
A_range = np.linspace(X[A].min(), X[A].max(), 10)
B_range = np.linspace(X[B].min(), X[B].max(), 10)
A_grid, B_grid = np.meshgrid(A_range, B_range)
IC50_pred_grid = model.predict(np.c_[A_grid.ravel(),
B_grid.ravel()]).reshape(A_grid.shape)

# Create the 3D interactive plot
fig = go.Figure()

# Plot training data points in green
fig.add_trace(go.Scatter3d(x=QSAR_Train[A], y=QSAR_Train[B],
z=QSAR_Train['IC50'],mode='markers',marker=dict(size=5, color='green'),
    name='Training Data'))

# Plot test data points in red
fig.add_trace(go.Scatter3d(x=df_test[A], y=df_test[B],
z=df_test['IC50'],mode='markers',marker=dict(size=5, color='red'),
    name='Test Data'))

# Plot regression plane
fig.add_trace(go.Surface(x=A_range, y=B_range,
z=IC50_pred_grid,colorscale='Blues',opacity=0.7,name='Regression Plane'
))
```

```
# Set plot titles and labels
fig.update_layout(
    title="3D Interactive Linear Regression with Test Data",
    scene=dict(
        xaxis_title=A,
        yaxis_title=B,
        zaxis_title='IC50'
    )
)

# Show the plot
fig.show()
```

```
# Non-interactive Plot
A = ic50_clean.index[min_corr[0]]
B = ic50_clean.index[min_corr[1]]

QSAR_Train = Train[['Drug',A,B,'IC50']]
df_test = Test[['Drug',A,B,'IC50']]



X = QSAR_Train[[A,B]]
y = QSAR_Train['IC50']

X_test = df_test[[A, B]]
y_test_actual = df_test['IC50']



model = LinearRegression()
model.fit(X, y)

# Predict IC50 for the test data
y_test_pred = model.predict(X_test)
mse_test = mean_squared_error(y_test_actual, y_test_pred)
print(f"Root Mean Squared Error for test data: {mse_test**0.5}")

# Generate predictions for the training data for plotting
A_range = np.linspace(X[A].min(), X[A].max(), 10)
B_range = np.linspace(X[B].min(), X[B].max(), 10)
A_grid, B_grid = np.meshgrid(A_range, B_range)
IC50_pred_train = (model.coef_[0] * A_grid) + (model.coef_[1] * B_grid) +
model.intercept_

# 3D Plotting
```

```
fig = plt.figure(figsize=(10, 7))
ax = fig.add_subplot(111, projection='3d')


ax.scatter(QSAR_Train[A], QSAR_Train[B], QSAR_Train['IC50'], color='green',
label='Training Data')
ax.scatter(df_test[A], df_test[B], y_test_actual, color='red', label='Test Data')
ax.plot_surface(A_grid, B_grid, IC50_pred_train, color='blue', alpha=0.5,
rstride=100, cstride=100)

# Set labels
ax.set_xlabel(A)
ax.set_ylabel(B)
ax.set_zlabel('IC50')
ax.set_title('3D Linear Regression with Test Data')
ax.grid(True)
ax.view_init(elev=20, azim=105)  # (0,0) for X axis and (0,90) for Y axis

plt.legend()
plt.show()
```

Acknowledgements:

I would like to thank the course faculty Prof.Mukesh Doble, for his invaluable inputs throughout the course.

****