Leveraging Gradient Boosting for Improved Anti-HIV Activity Prediction

Rekia Kadari[1], Abderrazeq Ajradi[2], Moufida Touhami[3], Berkane Ariche[1], Ali Rahmouni[1]
[1] Modeling and Calculation Methods Laboratory, Tahar Moulay University of Saida. Saida, Algeria
[2] Chemistry Department, Tahar Moulay University of Saida. Saida, Algeria
[3] Process Engineering Department, Tahar Moulay University of Saida. Saida, Algeria
Corresponding author e-mail: rekia.kadari@univ-saida.dz / rekiakadari@gmail.com

Received: 01/07/2023; Accepted: 07/10/2023; Published: 14/10/2023

Abstract
QSAR modelling is a widely used method that aims at learning relationships between input structures and output bioactivity data in order to make accurate predictions of bioactivities based on data structure. Prediction of the anti-HIV activity has been one of the most important tasks in chemical sciences where dominant approaches based on machine learning methods have been proposed. In this paper, we present a machine learning approach based on Gradient Boosting Regressor (GBR) to improve the performance of the HEPT anti-HIV activity prediction. The study was carried out with the estimation of the anti-HIV activity of a large set of 107 HEPT compounds using five quantum molecular descriptors. We evaluate our model on test and over all datasets, and in both cases we achieve state-of-the-art results.
Keywords: QSAR, HEPT derivatives, Anti-HIV, GBR, Descriptors, Prediction.

Tob Regul Sci. ™ 2023 ;9(2): 263-277
DOI: doi.org/10.18001/TRS.9.2.18

1. Introduction

Acquired Immunodeficiency Syndrome Disease (AIDS) epidemic is a threat to the health of the population worldwide caused by the Human Immunodeficiency Virus (HIV) [1]. 38,4 million people globally were living with HIV in 2021 [2]. To overcome the problem, there is an enormous interest by the medical and scientific community where current combination therapy named highly active antiretroviral therapy has been optimized. 28,7 million people were accessing antiretroviral therapy in 2021 [2]. This treatment, which inhibit multiple viral replication cycle, reduced remarkably AIDS-related mortality. Non-nucleoside reverse transcriptase inhibition is an essential component of this treatment. To date more than 50 structurally diverse classes of compounds have been reported as non-nucleoside reverse transcriptase inhibitors [3–10].

The first discovered non-nucleoside reverse transcriptase inhibitor is 1-[(2-hydroxyethoxy) methyl]-6 (phenylthio) thymine (HEPT) which exhibited moderate bioactivity and selectivity [11]. Since numerous HEPT derivatives have been designed and synthetized to optimize their anti-VIH-1 bioactivity [12–29]. The HEPT derivatives common structure is of thymine molecule structure where different substituents are changed. Due to HIV-1 drug-resistant mutant emergence, the development of novel non-nucleoside reverse transcriptase inhibitors have to be

continued [30]. This process can be aided and accelerated by using efficient Quantitative Structure Activity Relationship (QSAR) models.

QSAR modeling surrounds an important class of computational chemistry problems. Using such numerical means can facilitate and benefit several applications such as assessing the efficiency and general toxicity of drugs. QSAR is an important task that can be cast as a regression problem and can be formulated as follow: given a data set of structure-derived features of compounds, the QSAR model aims to relate the set of descriptors of each compound to its biological activity [31, 32].

QSAR was first proposed by Cros in 1863 for water solubility based on the toxicity of primary aliphatic alcohols [33] and then applied to compounds possessing anti-HIV activity since 1991 [34]. The area of QSAR modeling for the anti-HIV inhibition has been enriched over the last few decades by the contribution from several researchers. Several sophisticated machine learning algorithms have been developed.

Statistical machine learning techniques have been applied to several brunches. Much progress has been made to advance the state-of-the-art on QSAR modeling for the anti-HIV task. In literature, Contributions mainly come from two research directions. One is the choice of descriptors which contain information about each compound that is very important to the task [35]. The second line of research is the selection of the appropriate statistical method [31, 36–38]. Several QSAR studies have carried out on HIV-1 non-nucleoside reverse transcriptase inhibition of HEPT derivatives [39–50]. This section provides a brief mini review of some of them.

The first work was proposed by Luco [39] which applied two machine learning algorithms, partial least squares (PLS) and multiple linear regression (MLR) with a set of 10 hydrophobic and geometric descriptors. Douali et al, [40] were the first to explore artificial Neural Nets (ANN) for the estimation of the anti-HIV activity with a set of eight structural and physicochemical descriptors on 80 HEPT derivatives.

Additionally, in the same way as Douali et al., [40]; Shaik et al., [41] used Neural networks with a different set of descriptors. Shaik et al., [41] used 4 topological descriptors and a dataset of 107 compounds. Moreover, Shaik et al., [41] also used MLR and another most widely used state-of-the-art machine learning technique; Support Vector Machine (SVM). The ANN model shows the highest results on both train, test and overall data. In recent years, an MLR based model have been proposed by Rahmouni et al., [42]. The proposed model was performed using 60 HEPT derivatives with the help of 9 quantum descriptors.

The achieved experimental results confirm the effectiveness of all the proposed machine learning techniques for the anti-HIV bioactivity of HEPT derivatives prediction. The main characteristics of the proposed QSAR models in the literature are summarized in table 1.

Table 1: Brief description of the state of the art proposed models.

Reference	N molecules	N Descriptors	Type of Descriptors	Model	R2
Ref [41]	84	4	Topological descriptors	MLR ANN SVM	0.799 0.825 0.817
Ref [40]	80	8	Structural and physicochemical descriptors	ANN PLS	0.958 0.944
Ref [39]	107 79	10	Hydrophobic and geometric descriptors	MLR PLS	0.951 0.943
Ref [42]	60	9	Fukui Indices	MLR	0.815

In this paper, we introduce a GBR-based approach for modeling the anti-HIV activity. The main contributions of the present work are as follows: (1) We present a QSAR approach to predict anti-HIV activity with a set of five quantum descriptors. (2) We report results on a large dataset containing 107 compounds and show that the approach outperforms state-of-the-art methods. (3) We report results that only simple statistical techniques are weak in estimating the activity while boosting the regression dramatically improves the prediction performance.
The paper is organized as follows. In section 2 we introduce the gradient boosting method. Next, Section 3 gives details about our GBR QSAR model and experiments setup. In addition, Section 4 provides experimental results with a comparison against previous works for both statistical method and the set of descriptors. Finally, in Section 5 we conclude the paper.

2. Gradient Boosting

We present an overview of the GBR approach in this section.

Gradient Boosting (GB) [51–53] is a machine learning technique primarily used for regression tasks. The method produces more accurate prediction models based on the boosting principle where several weak learners are added to form a strong learner.

The main idea behind the method as the name suggests is to generate sequentially models during the learning process where each tries to correct its predecessor. First, base simple trees with single root nodes are constructed. Then, subsequent trees are built from errors of the previous tree. After each iteration, each data sample is given a weight based on its prediction. The more often a data sample residual is large (prediction error), the more important it becomes. The trees are scaled by using the learning rate. The goal is to minimize an objective function:

$$O(x) = \sum_{t} l(\hat{y}_i, y_i) + \sum_{t} \Omega(f_t)$$

Where:

- $l(\hat{y}_i, y_i)$is the loss function
- $\Omega(f_t)$ is the regularization function

The subsequent trees are combined with the preceding trees to predict the response. The process is repeated until the model prediction stops improving or the maximum number of trees is reached. The overall model becomes a stronger predictor.

3. Materials and Methods

In this section, we present our approach for developing an Anti-HIV inhibition QSAR model using the Gradient Boosting Regression (GBR) method. We provide detailed information about the development of our GBR model to predict the anti-HIV bioactivity of HEPT derivatives.

3.1. Dataset

3.1.1. Target function

The anti-HIV activity (log (1/EC50)) for 107 HEPT derivatives were collected from the previously published literature [39–42, 54]. The general base structure of these compounds is shown by figure 1. This figure shows also the atomic numbering used in this work.

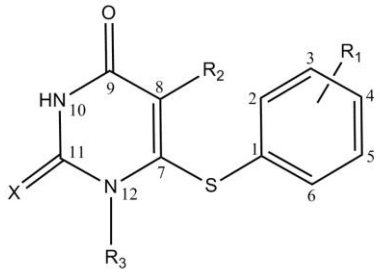

Figure 1. HEPT derivatives general structure and atomic numbering

The structural details as well as anti-HIV activity (log (1/EC50)) are reported in Table 2. HEPT derivatives have been obtained by varying thymine base structure substituents in order to improve their bioactivity [12–19]. Nature, number and position of substituent influence on HEPT derivatives anti-HIV have been investigated and quantified. From Table 2, it can be noted the cooperative effects of different aspects of substituents. It can also be noted that a little substituent structure change can induce large bioactivity modification. No linear correlation between substituent structure and bioactivity was identified.

Table 2: Chemical Structures and Observed Anti-HIV-Activities of HEPT Derivatives.

N° Compounds	R1	R2	R3	X	Anti-HIV ACTIVITY
1	2-Me	Me	CH2OCH2CH2OH	O	4.15
2	2-NO2	Me	CH2OCH2CH2OH	O	3.85
3	2-OMe	Me	CH2OCH2CH2OH	O	4.72
4	3-Me	Me	CH2OCH2CH2OH	O	5.59
5	3-Et	Me	CH2OCH2CH2OH	O	5.57
6	3-t-Bu	Me	CH2OCH2CH2OH	O	4.92
7	3-CF3	Me	CH2OCH2CH2OH	O	4.35
8	3-F	Me	CH2OCH2CH2OH	O	5.48
9	3-Cl	Me	CH2OCH2CH2OH	O	4.89
10	3-Br	Me	CH2OCH2CH2OH	O	5.24
11	3-I	Me	CH2OCH2CH2OH	O	5.00
12	3-NO2	Me	CH2OCH2CH2OH	O	4.47
13	3-OH	Me	CH2OCH2CH2OH	O	4.09
14	3-OMe	Me	CH2OCH2CH2OH	O	4.66
15	3,5-Me2	Me	CH2OCH2CH2OH	O	6.59
16	3,5-Cl2	Me	CH2OCH2CH2OH	O	5.89

17	3,5-Me2	Me	CH2OCH2CH2OH	S	6.66
18	3-COOMe	Me	CH2OCH2CH2OH	O	5.10
19	3-COMe	Me	CH2OCH2CH2OH	O	5.14
20	3-CN	Me	CH2OCH2CH2OH	O	5.00
21	H	CH2CH=CH2	CH2OCH2CH2OH	O	5.60
22	H	Et	CH2OCH2CH2OH	S	6.96
23	H	Pr	CH2OCH2CH2OH	S	5.00
24	H	i-Pr	CH2OCH2CH2OH	S	7.23
25	3,5-Me2	Et	CH2OCH2CH2OH	S	8.11
26	3,5-Me2	i-Pr	CH2OCH2CH2OH	S	8.30
27	3,5-Cl2	Et	CH2OCH2CH2OH	S	7.37
28	H	Et	CH2OCH2CH2OH	O	6.92
29	H	Pr	CH2OCH2CH2OH	O	5.47
30	H	i-Pr	CH2OCH2CH2OH	O	7.20
31	3,5-Me2	Et	CH2OCH2CH2OH	O	7.89
32	3,5-Me2	i-Pr	CH2OCH2CH2OH	O	8.57
33	3,5-Cl2	Et	CH2OCH2CH2OH	O	7.85
34	4-Me	Me	CH2OCH2CH2OH	O	3.66
35	H	Me	CH2OCH2CH2OH	O	5.15
36	H	Me	CH2OCH2CH2OH	S	6.01
37	H	I	CH2OCH2CH2OH	O	5.44
38	H	CH=CH2	CH2OCH2CH2OH	O	5.69
39	H	CH=CHPh	CH2OCH2CH2OH	O	5.22
40	H	CH2Ph	CH2OCH2CH2OH	O	4.37
41	H	CH=CPh2	CH2OCH2CH2OH	O	6.07
42	H	Me	CH2OCH2CH2OMe	O	5.06
43	H	Me	CH2OCH2CH2OAc	O	5.17
44	H	Me	CH2OCH2CH2OCOPh	O	5.12
45	H	Me	CH2OCH2Me	O	6.48
46	H	Me	CH2OCH2CH2Cl	O	5.82

Continued on next page

Table 2 – Continued from previous page

N° Compounds	R1	R2	R3	X	Anti-HIV ACTIVITY
47	H	Me	CH2OCH2CH2N3	O	5.24
48	H	Me	CH2OCH2CH2F	O	5.96
49	H	Me	CH2OCH2CH2Me	O	5.48
50	H	Me	CH2OCH2Ph	O	7.06
51	H	Et	CH2OCH2Me	O	7.72
52	H	Et	CH2OCH2Me	S	7.58
53	3,5-Me2	Et	CH2OCH2Me	O	8.24
54	3,5-Me2	Et	CH2OCH2Me	S	8.30
55	H	Et	CH2OCH2Ph	O	8.23
56	3,5-Me2	Et	CH2OCH2Ph	O	8.55
57	H	Et	CH2OCH2Ph	S	8.09
58	3,5-Me2	Et	CH2OCH2Ph	S	8.14
59	H	i-Pr	CH2OCH2Me	O	7.99
60	H	i-Pr	CH2OCH2Ph	O	8.51
61	H	i-Pr	CH2OCH2Me	S	7.89
62	H	i-Pr	CH2OCH2Ph	S	8.14
63	H	Me	CH2OMe	O	5.68

64	H	Me	CH2OBu	O	5.33
65	H	Me	Et	O	5.66
66	H	Me	Bu	O	5.92
67	3,5-Cl2	Et	CH2OCH2Me	S	7.89
68	H	Et	CH2O-i-Pr	S	6.66
69	H	Et	CH2O-c-Hex	S	5.79
70	H	Et	CH2OCH2-c-Hex	S	6.45
71	H	Et	CH2OCH2C6H4(4-Me)	S	7.11
72	H	Et	CH2OCH2C6H4(4-Cl)	S	7.92
73	H	Et	CH2OCH2CH2Ph	S	7.04
74	3,5-Cl2	Et	CH2OCH2Me	O	8.13
75	H	Et	CH2O-i-Pr	O	6.47
76	H	Et	CH2O-c-Hex	O	5.40
77	H	Et	CH2OCH2-c-Hex	O	6.35
78	H	Et	CH2OCH2CH2Ph	O	7.02
79	H	c-Pr	CH2OCH2Me	S	7.02
80	H	c-Pr	CH2OCH2Me	O	7.00
81	H	Me	CH2OCH2CH2OC5H11-n	O	<4.46
82	2-Cl	Me	CH2OCH2CH2OH	O	<3.89
83	3-CH2OH	Me	CH2OCH2CH2OH	O	<3.53
84	4-F	Me	CH2OCH2CH2OH	O	<3.60
85	4-Cl	Me	CH2OCH2CH2OH	O	<3.60
86	4-NO2	Me	CH2OCH2CH2OH	O	<3.72
87	4-CN	Me	CH2OCH2CH2OH	O	<3.60
88	4-OH	Me	CH2OCH2CH2OH	O	<3.56
89	4-OMe	Me	CH2OCH2CH2OH	O	<3.60
90	4-COMe	Me	CH2OCH2CH2OH	O	<3.96
91	3-COOH	Me	CH2OCH2CH2OH	O	<3.45
92	3-CONH2	Me	CH2OCH2CH2OH	O	<3.51

Continued on next page

Table 2 – Continued from previous page

N° Compounds	**R1**	**R2**	**R3**	**X**	**Anti-HIV ACTIVITY**
93	H	COOMe	CH2OCH2CH2OH	O	<5.18
94	H	CONHPh	CH2OCH2CH2OH	O	<4.74
95	H	SPh	CH2OCH2CH2OH	O	<4.68
96	H	CCH	CH2OCH2CH2OH	O	<4.74
97	H	CC-Ph	CH2OCH2CH2OH	O	<5.47
98	3-NH2	Me	CH2OCH2CH2OH	O	<3.60
99	H	COCHMe2	CH2OCH2CH2OH	O	<4.92
100	H	COPh	CH2OCH2CH2OH	O	<4.89
101	H	CCMe	CH2OCH2CH2OH	O	<4.72
102	H	F	CH2OCH2CH2OH	O	<4.00
103	H	Cl	CH2OCH2CH2OH	O	<4.52
104	H	Br	CH2OCH2CH2OH	O	<4.70
105	H	Me	CH2OCH2CH2OCH2Ph	O	<4.70
106	H	Me	H	O	<3.60
107	H	Me	Me	O	<3.82

3.1.2. Descriptors

Quantum molecular descriptors have the advantage that they can be calculated for hypothetical molecule. The bioactivity of drugs depends mainly on its interactions with active site. Good representation of these interactions needs rigorous descriptions of the drug charges distribution. Approximative charge distribution representation can be expressed by global and local properties of electronic density. Such electronic density indices have been successfully used to develop QSAR models [42, 56]. In the same way as Rahmouni et al. [42], four global descriptors which are Ionization Potential, Electron Affinity, Softness, Global Electrophilicity index and the local descriptors which is Dual Fukui function Δfk. This function has been calculated at five sites which are: C7, C8, S, O, N12 atom (figure 1). As can be seen from Table 2 and figure 1, these are the thymine structure sites where the substituent variation experimentally has been done. All quantum descriptors have been calculated at density functional theory level using the functional B3LYP and 6-311g + (d,p) atomic orbital basis set [58–62]. The used neutral molecular geometries have been optimized at the same level of theory. These geometries have been used in calculations of charged species. All quantum calculations have been carried out using Gaussian 09 package [63]. Fukui function Δfk have been evaluated using Natural Population Analysis (NPA) [64]. The results for 83 compound have been previously published [42, 57]. It should be noted that variations in all calculated descriptors can be correlated with the anti-HIV bioactivity of the HEPT derivatives.

3.1.3. Descriptors Selection

A feature or descriptor is an independent measurable variable of the process being observed. The set of features used to train machine learning models have a huge influence on the model's performance. Thus, it is indispensable to carefully select only those features that contribute meaningful information when used as input to a linear model.

Feature selection aims to choose a small subset of the relevant features from the original features by removing irrelevant, redundant, or noisy features. Feature selection usually can lead to better learning performance, higher learning accuracy, lower computational cost, and better model interpretability.

Instead of choosing all calculated descriptors which is our source data, we concentrate on identifying the appropriate attributes to anti-HIV activity. To choose the most optimum descriptors we use univariate selection which measures by KBest from sklearn [55] using score function chisquared. The calculation will help us find out the most important attributes, below are the top 5 suitable descriptors for prediction:

- Ionization Potential (I)
- C7 site dual descriptor
- C8 site dual descriptor
- S site dual descriptor
- N12 site dual descriptor

3.1.4. GBR QSAR Model development

After the descriptors calculations and the selection of the relevant ones steps are performed, we work towards training and validating the model's performance.

Our dataset has been divided into two subsets; a training set for generating QSAR model and a test set to evaluate how well our trained model performs on unseen data. Using the train_test_split function from scikit-learn's model _selection module [55], we randomly split the total set of compounds into 20 percent test data (22 compounds) and 80 percent training data (85 compounds). Note that the train_test_split function already shuffles the datasets before splitting. We use a fixed random_state to ensure that our results are reproducible.

Implementation was done in python [65, 66] using the version 0.22.1 of the scikit-learn machine Learning Library [55, 67] and all models were trained on MacBook Pro laptop with 2.5 GHz Intel Core i7 CPU, 16 GB RAM.

4. Results and Discussion

In this section, we report details and the evaluation of the performance of our proposed GBR model for HEPT derivatives and the anti-HIV activity prediction for both train and test sets. We also perform experiments on the overall data, the results are discussed below.

We tuned the hyper-parameters then trained the models. The results are reported with the best model, which is selected by the performance on the train set. The final chosen parameters are reported in Table 3.

Table 3. Best parameters obtained with gradient boosting.

Hyper-parameter	value
No. Estimator	600
Learning Rate	0.02
Max Depth	25
Max Features	Log2
Min Sample Leaf	1
Min Sample Split	25

Table 4 reports the results of evaluation of our GBR model on the datasets. We compare our model with baseline existing models including ANN models proposed by Douali et al., [40] and Shaik et al., [41] , the MLR models by Shaik et al., [39, 41, 42, 54], the PLS model proposed by [39] and the SVM model by Shaik et al., [41].

Table 4. Performance of our proposed model.

	N° Compounds	N° Descriptors	Sets	R2	MAE	MSE	RMSE
GBR	85	5	Train	0.9923	0.1052	0.0174	0.1322
	22		Test	0.5858	0.7714	0.8666	0.9309
	107		Over All	0.9147	0.2422	0.1920	0.4382

As shown in Table 4, our results demonstrate that the coefficient of determination (R^2) of our model is more informative and truthful than state of the art proposed models. Our model also get results better (or close to) than neural networks models based on ANN architectures which show that boosting the predictions is very important for the task.

The performance of the GBR model is also assessed by using statistical error metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root MSE (RMSE). The statistical summary of errors is also reported in Table 4 for training, test and overall data.

The statistical evaluation reflects that the developed model manifested close agreement between experimental and predicted results. In comparison, the GBR model surpassed the accuracy of the state of the art models, yielding the highest R^2 and the lowest MAE and RMSE.

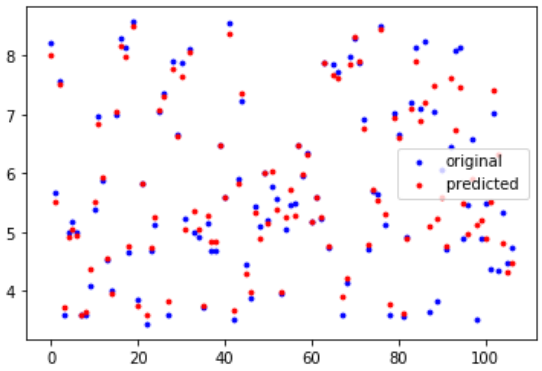

Figure 2. Plot of the experimental and model predicted values of the anti-HIV activity of HEPT derivatives in overall datasets.

Figure 2 illustrates the real activity values and the estimated ones for the GBR model on the 107 compounds. As we can see from Figure 2, a good agreement between the predicted and the measured values of the activity is observed. The experimental and the calculated anti-HIV are reported in Table 5.

Table 5. Experimental and calculated values of the anti-HIV activity

N° Compounds	Anti-HIV Activity	Calculated Anti-HIV Activity	Residual
1	4.15	4.224395	0.074395
2	3.85	3.758439	-0.091561
3	4.72	4.760461	0.040461
4	5.59	5.595656	0.005656
5	5.57	5.385797	-0.184203
6	4.92	5.039858	0.119858
7	4.35	6.316261	1.966261
8	5.48	5.287556	-0.192444
9	4.89	5.210121	0.320121
10	5.24	5.267761	0.027761
11	5.00	4.959957	-0.040043
12	4.47	4.325951	-0.144049
13	4.09	4.376748	0.286748

14	4.66	4.763285	0.103285
15	6.59	5.912189	-0.677811
16	5.89	5.928105	0.038105
17	6.66	6.625048	-0.034952
18	5.10	4.897999	-0.202001
19	5.14	5.2481	0.1081
20	5.00	4.90885	-0.09115
21	5.60	5.593229	-0.006771
22	6.96	6.833767	-0.126233
23	5.00	5.363776	0.363776
24	7.23	7.355596	0.125596

Continued on next page

Table 5 – Continued from previous page

N° Compounds	Anti-HIV Activity	Calculated Anti-HIV Activity	Residual
25	8.11	8.056116	-0.053884
26	8.30	8.169909	-0.130091
27	7.37	7.299992	-0.070008
28	6.92	6.762607	-0.157393
29	5.47	5.718068	0.248068
30	7.20	7.104644	-0.095356
31	7.89	7.900826	0.010826
32	8.57	8.508361	-0.061639
33	7.85	7.663453	-0.186547
34	3.66	5.105185	1.445185
35	5.15	5.294746	0.144746
36	6.01	6.010196	0.000196
37	5.44	5.338469	-0.101531
38	5.69	5.722825	0.032825
39	5.22	5.143688	-0.076312
40	4.37	5.527927	1.157927
41	6.07	5.608641	-0.461359
42	5.06	5.246104	0.186104
43	5.17	5.055092	-0.114908
44	5.12	5.320029	0.200029
45	6.48	6.485013	0.005013
46	5.82	5.830899	0.010899
47	5.24	5.060602	-0.179398
48	5.96	5.978538	0.018538
49	5.48	4.886623	-0.593377
50	7.06	7.073937	0.013937
51	7.72	7.613487	-0.106513
52	7.58	7.505393	-0.074607
53	8.24	7.196055	-1.043945
54	8.30	8.325924	0.025924
55	8.23	7.999344	-0.230656
56	8.55	8.378993	-0.171007

57	8.09	6.745106	-1.344894
58	8.14	7.46304	-0.67696
59	7.99	7.84448	-0.14552
60	8.51	8.44929	-0.06071
61	7.89	7.650892	-0.239108
62	8.14	7.913083	-0.226917
63	5.68	5.509005	-0.170995
64	5.33	4.82577	-0.50423
65	5.66	5.53596	-0.12404
66	5.92	5.84298	0.07702
67	7.89	7.883271	-0.006729
68	6.66	6.614121	-0.045879
69	5.79	6.026258	0.236258
70	6.45	7.630187	1.180187
Continued on next page			
Table 5 – Continued from previous page			
N° Compounds	Anti-HIV Activity	Calculated Anti-HIV Activity	Residual
71	7.11	6.883806	-0.226194
72	7.92	7.790105	-0.129895
73	7.04	7.479501	0.439501
74	8.13	7.997541	-0.132459
75	6.47	6.473724	0.003724
76	5.40	5.507018	0.107018
77	6.35	6.324032	-0.025968
78	7.02	6.952559	-0.067441
79	7.02	7.411836	0.391836
80	7.00	7.041514	0.041514
81	<4.46	4.287657	-0.172343
82	<3.89	3.981287	0.091287
83	<3.53	5.122133	1.592133
84	<3.60	3.656126	0.056126
85	<3.60	3.910322	0.310322
86	<3.72	3.763059	0.043059
87	<3.60	3.779849	0.179849
88	<3.56	3.63339	0.07339
89	<3.60	3.819368	0.219368
90	<3.96	3.989068	0.029068
91	<3.45	3.603296	0.153296
92	<3.51	3.686436	0.176436
93	<5.18	5.175317	-0.004683
94	<4.74	4.76652	0.02652
95	<4.68	4.841471	0.161471
96	<4.74	4.479576	-0.260424
97	<5.47	4.985387	-0.484613
98	<3.60	3.601492	0.001492
99	<4.92	4.890554	-0.029446

100	<4.89	5.48845	0.59845
101	<4.72	4.795568	0.075568
102	<4.00	3.966014	-0.033986
103	<4.52	4.560081	0.040081
104	<4.70	4.740508	0.040508
105	<4.70	4.837693	0.137693
106	<3.60	3.717912	0.117912
107	<3.82	5.242484	1.422484

5. Conclusion

In this paper, we have described an approach for anti-HIV activity prediction. The proposed approach uses five quantum descriptors. We performed experiments on 107 HEPT derivatives with the coefficient of determination as an evaluation metric. Experiment results show that our approach achieves state-of-the-art performances.

Further work in this area could be done in several directions. There are some possible extensions that must been taken into consideration as immediate goals, and we think that they should be studied such as the exploration of deep learning algorithms. Moreover, the enlargement of the dataset would be very advantageous to build more accurate models. As one long-term goal, it would be beneficial and useful to conduct a deep study concerning the most important and potent descriptors to the task.

References

[1] Weiss, R. A. (1993). How does HIV cause AIDS?. Science, 260(5112), 1273-1279.
[2] https://www.unaids.org/en.
[3] NAMASIVAYAM, Vigneshwaran, VANANGAMUDI, Murugesan, KRAMER, Victor G., et al. The journey of HIV-1 non-nucleoside reverse transcriptase inhibitors (NNRTIs) from lab to clinic. Journal of medicinal chemistry, 2018, vol. 62, no 10, p. 4851-4883.
[4] Fauci, 2003 : Fauci, A. S. (2003). HIV and AIDS: 20 years of science. Nature medicine, 9(7), 839-843.
[5] Whiteside, 2016 : Whiteside, A. (2016). HIV & AIDS: A very short introduction. Oxford University Press.
[6] Piot et al., 2001 : Piot, P., Bartos, M., Ghys, P. D., Walker, N., & Schwartländer, B. (2001). The global impact of HIV/AIDS. Nature, 410(6831), 968-973.
[7] Chulin 2020 : Zhuang, C., Pannecouque, C., De Clercq, E., & Chen, F. (2020). Development of nonnucleoside reverse transcriptase inhibitors (NNRTIs): our past twenty years. Acta Pharmaceutica Sinica B, 10(6), 961-978.
[8] Clercq, E. D. (2007). The design of drugs for HIV and HCV. Nature reviews Drug discovery, 6(12), 1001-1018.
[9] Martins, S., Ramos, M. J., & Fernandes, P. A. (2008). The current status of the NNRTI family of antiretrovirals used in the HAART regime against HIV infection. Current medicinal chemistry, 15(11), 1083-1095.
[10] Zhan, P., Chen, X., Li, D., Fang, Z., De Clercq, E., & Liu, X. (2013). HIV-1 NNRTIs: structural diversity, pharmacophore similarity, and impliations for drug design. Medicinal research reviews, 33(S1), E1-E72.
[11] Miyasaka, T., Tanaka, H., Baba, M., Hayakawa, H., Walker, R. T., Balzarini, J., & De Clercq, E. (1989). A novel lead for specific anti-HIV-1 agents: 1-[(2 hydroxyethoxy) methyl]-6-(phenylthio) thymine. Journal of medicinal chemistry, 32(12), 2507-2509.
[12] TANAKA, Hiromichi, BABA, Masanori, HAYAKAWA, Hiroyuki, et al. A new class of HIV-1 specific 6-substituted acyclouridine derivatives: synthesis and anti HIV-1 activity of 5-or 6-substituted analogs of 1-[(2-hydroxyethoxy) methyl]-6-(phenylthio) thymine (HEPT). Journal of medicinal chemistry, 1991, vol. 34, no 1, p. 349-357.

[13] TANAKA, Hiromichi, BABA, Masanori, UBASAWA, Masaru, et al. Synthesis and anti-HIV activity of 2-, 3-, and 4-substituted analogs of 1-[(2-hydroxyethoxy) methyl]-6-(phenylthio) thymine (HEPT). Journal of medicinal chemistry, 1991, vol. 34, no 4, p. 1394-1399.

[14] TANAKA, Hiromichi, TAKASHIMA, Hideaki, UBASAWA, Masaru, et al. Structure-activity relationships of 1-[(2-hydroxyethoxy) methyl]-6-(phenylthio) thymine analogs: effect of substitutions at the C-6 phenyl ring and at the C-5 position on anti-HIV-1 activity. Journal of Medicinal Chemistry, 1992, vol. 35, no 2, p. 337-345.

[15] TANAKA, Hiromichi, TAKASHIMA, Hideaki, UBASAWA, Masaru, et al. Synthesis and antiviral activity of deoxy analogs of 1-[(2-hydroxyethoxy) methyl]-6-(phenylthio) thymine (HEPT) as potent and selective anti-HIV-1 agents. Journal of medicinal chemistry, 1992, vol. 35, no 25, p. 4713-4719.

[16] TANAKA, Hiromichi, MIYASAKA, Tadashi, SEKIYA, Kouichi, et al. Synthesis of some analogues Of 1-[(2-hydroxyethoxy)-methyl]-6-(phenylthio) thymine (HEPT) which have different types of acyclic structures. Nucleosides & nucleotides, 1992, vol. 11, no 2-4, p. 447-456.

[17] TANAKA, Hiromichi, BABA, Masanori, TAKAHASHI, Emiko, et al. Design and synthesis of regioisomeric analogues of a specific anti-HIV-1 agent 1-[(2-hydroxyethoxy) methyl]-6-(phenylthio) thymine (HEPT). Nucleosides, Nucleotides & Nucleic Acids, 1994, vol. 13, no 1-3, p. 155-162.

[18] TANAKA, Hiromichi, TAKASHIMA, Hideaki, UBASAWA, Masaru, et al. Synthesis and antiviral activity of 6-benzyl analogs of 1-[(2-hydroxyethoxy) methyl]-5-(phenylthio) thymine (HEPT) as potent and selective anti-HIV-1 agents. Journal of medicinal chemistry, 1995, vol. 38, no 15, p. 2860-2865.

[19] TANAKA, Hiromichi, HAYAKAWA, Hiroyuki, HARAGUCHI, Kazuhiro, et al. HEPT: from an investigation of lithiation of nucleosides towards a rational design of non-nucleoside reverse transcriptase inhibitors of HIV-1. In : Advances in Antiviral Drug Design. Elsevier, 1999. p. 93-144.

[20] Sun, G. F., Chen, X. X., Chen, F. E., Wang, Y. P., De Clercq, E., Balzarini, J., & Pannecouque, C. (2005). Nonnucleoside HIV-1 reverse-transcriptase inhibitors, part 5. Synthesis and anti-HIV-1 activity of novel 6-naphthylthio HEPT analogues. Chemical & pharmaceutical bulletin, 53(8), 886– 892. https://doi.org/10.1248/cpb.53.886

[21] Puig-De-La-Bellacasa, R., Giménez, L., Pettersson, S., Pascual, R., Gonzalo, E., Esté, J. A., ... & Teixidó, J. (2012). Diverse combinatorial design, synthesis and in vitro evaluation of new HEPT analogues as potential non-nucleoside HIV-1 reverse transcription inhibitors. European journal of medicinal chemistry, 54, 159-174.

[22] NOVIKOV, M. S. et GEISMAN, A. N. Methods of synthesis of 6-substituted uracil derivatives–the structural base of antiviral agents. Chemistry of Heterocyclic Compounds, 2014, vol. 49, p. 1426-1450.

[23] Hao, Q., Ling, X., Pannecouque, C., De Clercq, E., & Chen, F. (2023). Linker optimization of HEPT derivatives as potent non-nucleoside HIV-1 reverse transcriptase inhibitors: From S= O to CHOR. Chinese Chemical Letters, 34(4), 107663.

[24] Hao, Q., Wang, S., Huang, W., Zhang, Y., Pannecouque, C., De Clercq, E., & Chen, F. (2022). Structure-Based design of [(2-Hydroxyethoxy) methyl]-6-(phenylthio)-thymine derivatives as nonnucleoside HIV-1 reverse transcriptase Inhibitors: From HEPTs to Sulfinyl-substituted HEPTs. Bioorganic Chemistry, 126, 105880.

[25] Zhou, R. L., Ju, Z., Pannecouque, C., De Clercq, E., Wang, S., & Chen, F. E. (2023). Structure guided design of novel HEPT analogs with enhanced potency and safety: From Isopropyl-HEPTs to Cyclopropyl-HEPTs. European Journal of Medicinal Chemistry, 246, 114939.

[26] Zhuang, C., Pannecouque, C., De Clercq, E., & Chen, F. (2020). Development of non-nucleoside reverse transcriptase inhibitors (NNRTIs): our past twenty years. Acta Pharmaceutica Sinica B, 10(6), 961-978.

[27] Chen, W., Zhan, P., Wu, J., Li, Z., & Liu, X. (2012). The development of HEPT-type HIV nonnucleoside reverse transcriptase inhibitors and its implications for DABO family. Current pharmaceutical design, 18(27), 4165-4186.

[28] Chen, W., Zhan, P., Wu, J., Li, Z., & Liu, X. (2012). The development of HEPT-type HIV nonnucleoside reverse transcriptase inhibitors and its implications for DABO family. Current pharmaceutical design, 18(27), 4165-4186.

[29] Zhuang, C., Pannecouque, C., De Clercq, E., & Chen, F. (2020). Development of non-nucleoside reverse transcriptase inhibitors (NNRTIs): our past twenty years. Acta Pharmaceutica Sinica B, 10(6), 961-978.

[30] Hao, Q. Q., Chen, X. M., Pannecouque, C., De Clercq, E., Wang, S., & Chen, F. E. (2023). Structure-directed linker optimization of novel HEPTs as non nucleoside inhibitors of HIV-1 reverse transcriptase. Bioorganic Chemistry, 133, 106413.

[31] Toplak, M., Močnik, R., Polajnar, M., Bosnić, Z., Carlsson, L., Hasselgren, C., ... & Stalring, J. (2014). Assessment of machine learning reliability methods for quantifying the applicability domain of QSAR regression models. Journal of chemical information and modeling, 54(2), 431-441.

[32] Dudek, A. Z., Arodz, T., & Gálvez, J. (2006). Computational methods in developing quantitative structure-activity relationships (QSAR): a review. Combinatorial chemistry & high throughput screening, 9(3), 213-228.

[33] Cros, A. F. A. (1863). Action de l'alcool amylique sur l'organisme (Doctoral dissertation, Faculté de médecine de Strasbourg).

[34] Castro, E. A., Torrens, F., Toropov, A. A., Nesterov, I. V., & Nabiev, O. M. (2004). QSAR modeling ANTI-HIV-1 activities by optimization of correlation weights of local graph invariants. Molecular Simulation, 30(10), 691-696.

[35] Khan, A. U. (2016). Descriptors and their selection methods in QSAR analysis: paradigm for drug design. Drug discovery today, 21(8), 1291-1302.

[36] Liu, Y. (2005, December). Drug design by machine learning: ensemble learning for qsar modeling. In Fourth International Conference on Machine Learning and Applications (ICMLA'05) (pp. 7-pp). IEEE.

[37] Lo, Y. C., Rensi, S. E., Torng, W., & Altman, R. B. (2018). Machine learning in chemoinformatics and drug discovery. Drug discovery today, 23(8), 1538-1546.

[38] Muratov, E. N., Bajorath, J., Sheridan, R. P., Tetko, I. V., Filimonov, D., Poroikov, V., ... & Tropsha, A. (2020). QSAR without borders. Chemical Society Reviews, 49(11), 3525-3564.

[39] Luco, J. M., & Ferretti, F. H. (1997). QSAR based on multiple linear regression and PLS methods for the anti-HIV activity of a large group of HEPT derivatives. Journal of chemical information and computer sciences, 37(2), 392-401.

[40] Douali, L., Villemin, D., Zyad, A., & Cherqaoui, D. (2004). Artificial neural networks: Non-linear QSAR studies of HEPT derivatives as HIV-1 reverse transcriptase inhibitors. Molecular Diversity, 8, 1-8.

[41] Shaik, B., Zafar, T., & Agrawal, V. K. (2013). Estimation of anti-HIV activity of HEPT analogues using MLR, ANN, and SVM techniques. International journal of medicinal chemistry, 2013.

[42] Rahmouni, A., Touhami, M., & Benaissa, T. (2017). Fukui Indices as QSAR Model Descriptors: The Case of the Anti-HIV Activity of 1-2-[(Hydroxyethoxy) Methyl]-6-(Phenylthio) Thymine Derivatives. International Journal of Chemoinformatics and Chemical Engineering (IJCCE), 6(2), 31-44.

[43] Riahi, S., Pourbasheer, E., Dinarvand, R., Ganjali, M. R., & Norouzi, P. (2009). Quantitative Structure–Activity Relationship Study on the Anti-HIV-1 Activity of Novel 6-Naphthylthio HEPT Analogs. Chemical biology & drug design, 74(2), 165-172.

[44] Ji, L., Chen, F. E., Feng, X. Q., De Clercq, E., Balzarini, J., & Pannecouque, C. (2006). Nonnucleoside HIV-1 reverse transcriptase inhibitors, part 7. Synthesis, antiviral activity, and 3D-QSAR investigations of novel 6-(1-naphthoyl) HEPT analogues. Chemical and pharmaceutical bulletin, 54(9), 1248-1253.

[45] Mao, Y., Li, Y., Hao, M., Zhang, S., & Ai, C. (2012). Docking, molecular dynamics and quantitative structure-activity relationship studies for HEPTs and DABOs as HIV-1 reverse transcriptase inhibitors. Journal of molecular modeling, 18, 2185-2198.

[46] Canizares-Carmenate, Y., Castillo-Garit, J. A., & Machado, Y. H. (2019). Chemometric study of antiretroviral activities of HEPT derivatives using Multiple Linear Regression.

[47] Bazoui, H., Zahouily, M., Boulajaaj, S., Sebti, S., & Zakarya, D. (2002). QSAR for anti-HIV activity of HEPT derivatives. SAR and QSAR in Environmental Research, 13(6), 567-577.

[48] Duda-Seiman, C., Duda-Seimana, D., Putz, M. V., & Ciubotariub, D. (2007). QSAR modelling of anti-HIV activity with HEPT derivatives. Digest Journal of Nanomaterials and Biostructures, 2(2), 207-219.

[49] Toropova, A. P., Toropov, A. A., Veselinović, J. B., Miljković, F. N., & Veselinović, A. M. (2014). QSAR models for HEPT derivates as NNRTI inhibitors based on Monte Carlo method. European journal of medicinal chemistry, 77, 298-305.
[50] Hosny, A., Ashton, M., Gong, Y., & McGarry, K. (2020). The development of a predictive model to identify potential HIV-1 attachment inhibitors. Computers in Biology and Medicine, 120, 103743.
[51] Mason et al. 1999: Mason, L., Baxter, J., Bartlett, P., & Frean, M. (1999). Boosting algorithms as gradient descent. Advances in neural information processing systems, 12.
[52] Natekin, Alexey, and Alois Knoll. "Gradient boosting machines, a tutorial." Frontiers in neurorobotics 7 (2013): 21.
[53] Ridgeway, Greg. "The state of boosting." Computing science and statistics (1999): 172-181.
[54] Hansch, C.; Zhang, L. QSAR of HIV Inhibitors.Bioorg. Med. Chem. Lett. 1992, 2, 1165-1169.
[55] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.
[56] Djeradi, H., Rahmouni, A., & Cheriti, A. (2014). Antioxidant activity of flavonoids: a QSAR modeling using Fukui indices descriptors. Journal of molecular modeling, 20(10), 2476. https://doi.org/10.1007/s00894-014-2476-1
[57] Touhami, M. (2019). Modélisation de l'activité biologique de composés hétérocycles. Doctoral dissertation. University of Saida. Algeria
[58] Parr, R. G., & Yang, W. (1989). Density-Functional Theory of Atoms and Molecules New York: Oxford Univ. Press.
[59] McLean, A. D., & Chandler, G. S. (1980). Contracted Gaussian basis sets for molecular calculations. I. Second row atoms, Z= 11–18. The Journal of Chemical Physics, 72(10), 5639–5648.
[60] Krishnan, R. B. J. S., Binkley, J. S., Seeger, R., & Pople, J. A. (1980). Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions. The Journal of Chemical Physics, 72(1), 650–654.
[61] Clark, T., Chandrasekhar, J., Spitznagel, G. W., & Schleyer, P. V. R. (1983). Efficient diffuse function-augmented basis sets for anion calculations. III. The 3-21+ G basis set for first-row elements, Li–F. Journal of ComputationalChemistry, 4(3), 294–301.
[62] Frisch, M. J., Pople, J. A., & Binkley, J. S. (1984). Self-consistent molecular orbital methods 25. Supplementary functions for Gaussian basis sets. The Journal of Chemical Physics, 80(7), 326.
[63] Frisch, M. J. E. A., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., ... & Nakatsuji, H. (2009). Gaussian 09, revision D. 01.
[64] Hirshfeld, F. L. (1977). Bonded-atom fragments for describing molecular charge densities. Theoretical Chemistry Accounts: Theory, Computation, and Modeling. Theoretica Chimica Acta, 44(2), 129–138.
[65] Martelli, A., Ravenscroft, A. M., Holden, S., & McGuire, P. (2023). Python in a Nutshell. " O'Reilly Media, Inc.".
[66] Lutz, M. (2001). Programming python. " O'Reilly Media, Inc.".
[67] Kramer, O., & Kramer, O. (2016). Scikit-learn. Machine learning for evolution strategies, 45-53.