A nonlinear QSAR Study Using Oscillating Search and SVM as an Efficient Algorithm to Model the Inhibition of Reverse Transcriptase by HEPT Derivatives

[1,2]Ahmed Allali*, [1]Fouad Ferkous, [1,3]Khaireddine Kraim, [1]Youcef Saihi, [4,5]Mohammed Brahimi
[6]Faouzi Zaiz and [1]Ouassila Attoui-Yahia
[1]*Laboratory of Applied Organic Chemistry, Chemistry Department, Badji Mokhtar Annaba University, Algeria.*
[2]*Biology Sciences Department, Hamma Lakhdar El-Oued University, Algeria.*
[3]*High School of Technological Education (ENSET) Skikda, Algeria.*
[4]*Department of Computer Science, USTHB University, Algiers, Algeria.*
[5]*Department of Computer Science, Mohamed El Bachir El Ibrahimi University, Bordj Bou Arreridj, Algeria.*
[6]*Computer Sciences Department; Hamma Lakhdar El-Oued University, Algeria.*
ahmed-allali@univ-eloued.dz*

(Received on 22nd November 2016, accepted in revised form 21st August 2017)

Summary: Quantitative structure-activity relationships were constructed for 107 inhibitors of *HIV*-1 reverse transcriptase that are derivatives of 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT). A combination of a support vector machine *(SVM*) and oscillating search (OS) algorithms for feature selection was adopted to select the most appropriate descriptors. The application was optimized to obtain an SVM model to predict the biological activity EC_{50} of the HEPT derivatives with a minimum number of descriptors (SpMax4_Bh (e) MLOGP MATS5m) and high values of R^2 and Q^2 (0.8662, 0.8769). The statistical results showed good correlation between the activity and three best descriptors were included in the best SVM model. The values of R^2 and Q^2 confirmed the stability and good predictive ability of the model. The SVM technique was adequate to produce an effective *QSAR* model and outperformed those in the literature and the predictive stages for the inhibitory activity of reverse transcriptase by HEPT derivatives.

Keywords*:* QSAR, Features Selection, SVM, HEPT, HIV.

Introduction

Viruses are infectious agents that reproduce in the intracellular environment of a host cell. Viruses can be classified according to the nature of their genomes (DNA or RNA) [1]. Reverse transcriptase (RT), the major target of antiviral chemotherapy for AIDS/HIV [2] and a key multifunctional enzyme in the life cycle of human immunodeficiency virus type-1 (HIV-1). HIV-1 is the causative agent of AIDS (Acquired immunodeficiency syndrome) [3] and features high-molecular-weight RNA that encodes an enzyme (reverse transcriptase) [4] that allows the transcription of the viral RNA into pro-viral DNA. This pro-viral DNA can then integrate into the genome of the host cell [5]. HIV-1 and -2 are the most common and most pathogenic retroviruses and are responsible for the onset of AIDS.[6] HIV infection is a chronic infection that affects host cells carrying the CD4 receptor (T cells) [7]. Once integrated, the retrovirus replication cycle starts, resulting in the production of new viral particles. This replication results in the slow death of infected cells.

The disappearance of CD4 lymphocytes leads to immunodeficiency, thereby inducing the occurrence stage of AIDS [8]. Inhibition of reverse transcriptase remains the ideal way to combat this type of retrovirus by blocking the viral replication cycle. The beneficial effects of reverse transcriptase inhibition have attracted the attention of scientists and pharmaceutical companies to develop and enrich this therapeutic class. Several molecules have been marketed as anti-HIV drugs, including Combivir, Kivexa, Trizivir, Tenofovir, and Efavirenz [9].

Modeling methods have undergone progressive development in the fields of pharmaceutical chemistry and drug design to study enzyme inhibition in the absence of detailed information on the underlying mechanism. Among these methods, the use of quantitative structure-activity relationships (QSAR) requires further progress [10]. QSAR studies have become essential in pharmaceutical chemistry and drug design, particularly when the availability of samples is limited or experimental measurements are dangerous, time consuming and expensive [11]. Technically, this method is based on four pillars: the basic structure, physicochemical parameters (molecular descriptors), descriptor selection method, and learning algorithm. The use of this modeling approach is the heart of our

*To whom all correspondence should be addressed.

work, particularly in the development of stable and robust QSAR models that can effectively predict inhibitory activity against reverse transcriptase and thus contribute to the development of new molecules that inhibit this enzyme. Here, we conducted a nonlinear QSAR study of reverse transcriptase inhibitors based on a set of molecules derived from 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT) (Fig 1).

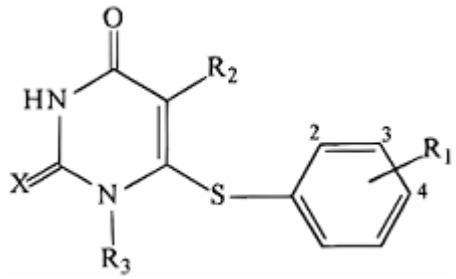


Fig. 1: General structure of HEPT derivatives.

Experimental

The data set used in this QSAR study consisted of 107 compounds selected from the literature [12] (Table-1). All 2D structures (Molecule.mol) were downloaded from the ChemSpider database [13] and were preoptimized using the Molecular Mechanics Force Field method (MM^+) included in HyperChem [14]. Finally, optimization was performed utilizing the semi-empirical Austin Method (AM1) with a root mean square gradient of 0.001 kcal.mol^{-1}.

The molecular descriptors were computed using Dragon software [15], which calculated the parameters of all compounds (a total of 4,885 different molecular descriptors). The descriptors obtained were analyzed to remove constant and near constant variables to reduce redundancy in the descriptor data matrix. The correlation was examined to eliminate highly correlated descriptors ($R > 0.90$). Finally, 686 molecular descriptors remained.

The EC_{50} values were scaled (from 0 to 1) to reduce the skewness of the data set and were used for subsequent *QSAR* analysis as dependent variables. In our study, to build the oscillating search-support vector machine (*OS-SVM*) models, we combined *SVM* with the OS algorithm for feature selection as the objective function. *SVM*, which was developed by Vapnik [16] as a novel algorithm of machine learning, is increasingly popular due to its many striking features and superior empirical performance. *SVM* was first introduced for solving pattern recognition problems and measure the quality of subset features, particularly R^2 and Q^2, which are calculated using the trained *SVM* model. The variable selection flowchart with the algorithm research oscillating support vector machine (*OS-SVM*) is shown in Fig 2.

Table-1: Chemical Structures for Experimental. and Calculated Anti-HIV-Activities of $EC_{50.}$

ID	SMILES	Status	EC50 -Exp	EC50-Cal
1	Cc2ccccc2SC1=C(C)C(=O)NC(=O)N1COCCO	train	4.15	4.63
2	[O-][N+](=O)c2ccccc2SC1=C(C)C(=O)NC(=O)N1COCCO	train	3.85	4.89
3	COc2ccccc2SC1=C(C)C(=O)NC(=O)N1COCCO	train	4.72	5.16
4	Cc1cc(ccc1)SC2=C(C)C(=O)NC(=O)N2COCCO	train	5.59	5.65
5	CCc1cc(ccc1)SC2=C(C)C(=O)NC(=O)N2COCCO	train	5.57	5.20
6	CC(C)(C)c1cc(ccc1)SC2=C(C)C(=O)NC(=O)N2COCCO	test	4.92	4.70
7	FC(F)(F)c1cc(ccc1)SC2=C(C)C(=O)NC(=O)N2COCCO	train	4.35	4.51
8	Fc1cccc(c1)SC2=C(C)C(=O)NC(=O)N2COCCO	train	5.48	4.99
9	Clc1cccc(c1)SC2=C(C)C(=O)NC(=O)N2COCCO	test	4.89	5.12
10	Brc1cccc(c1)SC2=C(C)C(=O)NC(=O)N2COCCO	train	5.24	5.09
11	Ic1cccc(c1)SC2=C(C)C(=O)NC(=O)N2COCCO	train	5.00	5.09
12	[O-][N+](=O)c1cc(ccc1)SC2=C(C)C(=O)NC(=O)N2COCCO	train	4.47	4.41
13	Oc1cc(ccc1)SC2=C(C)C(=O)NC(=O)N2COCCO	train	4.09	4.56
14	COc1cc(ccc1)SC2=C(C)C(=O)NC(=O)N2COCCO	train	4.66	4.10
15	Cc1cc(cc(C)c1)SC2=C(C)C(=O)NC(=O)N2COCCO	train	6.59	7.02
16	Clc1cc(cc(Cl)c1)SC2=C(C)C(=O)NC(=O)N2COCCO	test	5.89	5.42
17	Cc1cc(cc(C)c1)SC2=C(C)C(=O)NC(=S)N2COCCO	train	6.66	6.78
18	O=C(OC)c1cc(ccc1)SC2=C(C)C(=O)NC(=O)N2COCCO	train	5.10	4.44
19	CC(=O)c1cc(ccc1)SC2=C(C)C(=O)NC(=O)N2COCCO	train	5.14	3.98
20	N#Cc1cc(ccc1)SC2=C(C)C(=O)NC(=O)N2COCCO	train	5.00	4.27
21	C=CCC2=C(Sc1ccccc1)N(COCCO)C(=O)NC2=O	train	5.60	5.34
22	CCC2=C(Sc1ccccc1)N(COCCO)C(=S)NC2=O	train	6.96	6.06
23	CCCC2=C(Sc1ccccc1)N(COCCO)C(=S)NC2=O	train	5.00	5.78
24	CC(C)C2=C(Sc1ccccc1)N(COCCO)C(=S)NC2=O	train	7.23	7.04
25	Cc1cc(cc(C)c1)SC2=C(CC)C(=O)NC(=S)N2COCCO	train	8.11	7.64
26	Cc1cc(cc(C)c1)SC2=C(C(=O)NC(=S)N2COCCO)C(C)C	train	8.30	8.31
27	Clc1cc(cc(Cl)c1)SC2=C(CC)C(=O)NC(=S)N2COCCO	train	7.37	6.53
28	CCC2=C(Sc1ccccc1)N(COCCO)C(=O)NC2=O	train	6.92	6.22
29	CCCC2=C(Sc1ccccc1)N(COCCO)C(=O)NC2=O	train	5.47	5.80
30	CC(C)C2=C(Sc1ccccc1)N(COCCO)C(=O)NC2=O	test	7.20	7.30

Table-1 continues. . .

31	Cc1cc(cc(C)c1)SC2=C(CC)C(=O)NC(=O)N2COCCO	train	7.89	7.93
32	Cc1cc(cc(C)c1)SC2=C(C(=O)NC(=O)N2COCCO)C(C)C	train	8.57	8.64
33	Clc1cc(cc(Cl)c1)SC2=C(CC)C(=O)NC(=O)N2COCCO	test	7.85	6.50
34	Cc1ccc(cc1)SC2=C(C)C(=O)NC(=O)N2COCCO	test	3.66	4.13
35	CC2=C(Sc1ccccc1)N(COCCO)C(=O)NC2=O	test	5.15	4.75
36	CC2=C(Sc1ccccc1)N(COCCO)C(=S)NC2=O	test	6.01	4.77
37	IC2=C(Sc1ccccc1)N(COCCO)C(=O)NC2=O	train	5.44	5.09
38	C=CC2=C(Sc1ccccc1)N(COCCO)C(=O)NC2=O	train	5.69	5.84
39	O=C3NC(=O)N(COCCO)C(Sc1ccccc1)=C3/C=C/c2ccccc2	train	5.22	5.73
40	O=C3NC(=O)N(COCCO)C(Sc1ccccc1)=C3Cc2ccccc2	train	4.37	4.87
41	O=C4NC(=O)N(COCCO)C(Sc1ccccc1)=C4\C=C(/c2ccccc2)c3ccccc3	train	6.07	5.91
42	CC2=C(Sc1ccccc1)N(COCCOC)C(=O)NC2=O	train	5.06	4.61
43	CC(=O)OCCOCN2C(Sc1ccccc1)=C(C)C(=O)NC2=O	train	5.17	4.77
44	CC3=C(Sc1ccccc1)N(COCCOC(=O)c2ccccc2)C(=O)NC3=O	train	5.12	6.11
45	CC2=C(Sc1ccccc1)N(COCC)C(=O)NC2=O	train	6.48	5.02
46	CC2=C(Sc1ccccc1)N(COCCCl)C(=O)NC2=O	train	5.82	5.71
47	CC2=C(Sc1ccccc1)N(COCC\N=[N+]=[N-])C(=O)NC2=O	train	5.24	4.58
48	CC2=C(Sc1ccccc1)N(COCCF)C(=O)NC2=O	train	5.96	5.43
49	CC2=C(Sc1ccccc1)N(COCCC)C(=O)NC2=O	train	5.48	5.48
50	CC3=C(Sc1ccccc1)N(COCc2ccccc2)C(=O)NC3=O	train	7.06	5.98
51	CCC2=C(Sc1ccccc1)N(COCC)C(=O)NC2=O	train	7.72	6.54
52	CCC2=C(Sc1ccccc1)N(COCC)C(=S)NC2=O	test	7.58	6.42
53	Cc1cc(cc(C)c1)SC2=C(CC)C(=O)NC(=O)N2COCC	train	8.24	8.32
54	Cc1cc(cc(C)c1)SC2=C(CC)C(=O)NC(=S)N2COCC	train	8.30	8.03
55	CCC3=C(Sc1ccccc1)N(COCc2ccccc2)C(=O)NC3=O	train	8.23	7.38
56	Cc1cc(cc(C)c1)SC3=C(CC)C(=O)NC(=O)N3COCc2ccccc2	train	8.55	9.02
57	CCC3=C(Sc1ccccc1)N(COCc2ccccc2)C(=S)NC3=O	test	8.09	7.26
58	Cc1cc(cc(C)c1)SC3=C(CC)C(=O)NC(=S)N3COCc2ccccc2	test	8.14	8.75
59	CC(C)C2=C(Sc1ccccc1)N(COCC)C(=O)NC2=O	train	7.99	7.66
60	CC(C)C3=C(Sc1ccccc1)N(COCc2ccccc2)C(=O)NC3=O	test	8.51	8.42
61	CC(C)C2=C(Sc1ccccc1)N(COCC)C(=S)NC2=O	train	7.89	7.42
62	CC(C)C3=C(Sc1ccccc1)N(COCc2ccccc2)C(=S)NC3=O	train	8.14	8.20
63	COCN2C(Sc1ccccc1)=C(C)C(=O)NC2=O	train	5.68	5.40
64	CC2=C(Sc1ccccc1)N(COCCCC)C(=O)NC2=O	train	5.33	5.37
65	O=C2NC(=O)C(C)=C(Sc1ccccc1)N2CC	train	5.66	7.36
66	CC2=C(Sc1ccccc1)N(CCCC)C(=O)NC2=O	train	5.92	6.45
67	Clc1cc(cc(Cl)c1)SC2=C(CC)C(=O)NC(=S)N2COCC	test	7.89	6.92
68	CC(C)OCN2C(Sc1ccccc1)=C(CC)C(=O)NC2=S	train	6.66	6.19
69	CCC3=C(Sc1ccccc1)N(COC2CCCCC2)C(=S)NC3=O	train	5.79	7.94
70	CCC3=C(Sc1ccccc1)N(COCC2CCCCC2)C(=S)NC3=O	train	6.45	6.24
71	Cc3ccc(COCN2C(Sc1ccccc1)=C(CC)C(=O)NC2=S)cc3	train	7.11	7.02
72	Clc3ccc(COCN2C(Sc1ccccc1)=C(CC)C(=O)NC2=S)cc3	train	7.92	7.45
73	CCC3=C(Sc1ccccc1)N(COCCc2ccccc2)C(=S)NC3=O	train	7.04	7.26
74	Clc1cc(cc(Cl)c1)SC2=C(CC)C(=O)NC(=O)N2COCC	train	8.13	6.86
75	CC(C)OCN2C(Sc1ccccc1)=C(CC)C(=O)NC2=O	test	6.47	6.23
76	CCC3=C(Sc1ccccc1)N(COC2CCCCC2)C(=O)NC3=O	train	5.40	6.53
77	CCC3=C(Sc1ccccc1)N(COCC2CCCCC2)C(=O)NC3=O	train	6.35	6.19
78	CCC3=C(Sc1ccccc1)N(COCCc2ccccc2)C(=O)NC3=O	train	7.02	6.83
79	O=C2NC(=S)N(COCC)C(Sc1ccccc1)=C2C3CC3	train	7.02	6.93
80	O=C2NC(=O)N(COCC)C(Sc1ccccc1)=C2C3CC3	train	7.00	6.05
81	CC2=C(Sc1ccccc1)N(COCCOCCCCC)C(=O)NC2=O	train	4.46	5.07
82	Clc2ccccc2SC1=C(C)C(=O)NC(=O)N1COCCO	test	3.89	4.90
83	OCc1cc(ccc1)SC2=C(C)C(=O)NC(=O)N2COCCO	train	3.53	4.54
84	Fc1ccc(cc1)SC2=C(C)C(=O)NC(=O)N2COCCO	train	3.60	4.14
85	Clc1ccc(cc1)SC2=C(C)C(=O)NC(=O)N2COCCO	train	3.60	3.85
86	[O-][N+](=O)c1ccc(cc1)SC2=C(C)C(=O)NC(=O)N2COCCO	test	3.72	4.46
87	N#Cc1ccc(cc1)SC2=C(C)C(=O)NC(=O)N2COCCO	train	3.60	4.02
88	Oc1ccc(cc1)SC2=C(C)C(=O)NC(=O)N2COCCO	train	3.56	3.50
89	COc1ccc(cc1)SC2=C(C)C(=O)NC(=O)N2COCCO	test	3.60	3.68
90	CC(=O)c1ccc(cc1)SC2=C(C)C(=O)NC(=O)N2COCCO	train	3.96	4.05
91	O=C(O)c1cc(ccc1)SC2=C(C)C(=O)NC(=O)N2COCCO	test	3.45	3.93
92	NC(=O)c1cc(ccc1)SC2=C(C)C(=O)NC(=O)N2COCCO	train	3.51	3.60
93	O=C(OC)C2=C(Sc1ccccc1)N(COCCO)C(=O)NC2=O	train	5.18	4.30
94	O=C(Nc1ccccc1)C3=C(Sc2ccccc2)N(COCCO)C(=O)NC3=O	train	4.74	4.65
95	O=C3NC(=O)N(COCCO)C(Sc1ccccc1)=C3Sc2ccccc2	test	4.68	5.15
96	C#CC2=C(Sc1ccccc1)N(COCCO)C(=O)NC2=O	train	4.74	5.44
97	O=C3NC(=O)N(COCCO)C(Sc1ccccc1)=C3C#Cc2ccccc2	test	5.47	5.27
98	Nc1cc(ccc1)SC2=C(C)C(=O)NC(=O)N2COCCO	train	3.60	4.85
99	CC(C)C(=O)C2=C(Sc1ccccc1)N(COCCO)C(=O)NC2=O	train	4.92	4.84
100	O=C(C2=C(Sc1ccccc1)N(COCCO)C(=O)NC2=O)c3ccccc3	train	4.89	4.62
101	CC#CC2=C(Sc1ccccc1)N(COCCO)C(=O)NC2=O	train	4.72	4.96
102	FC2=C(Sc1ccccc1)N(COCCO)C(=O)NC2=O	train	4.00	4.62
103	ClC2=C(Sc1ccccc1)N(COCCO)C(=O)NC2=O	train	4.52	4.70
104	BrC2=C(Sc1ccccc1)N(COCCO)C(=O)NC2=O	train	4.70	4.92
105	CC3=C(Sc1ccccc1)N(COCCOCc2ccccc2)C(=O)NC3=O	train	4.70	5.56
106	O=C2NC(Sc1ccccc1)=C(C)C(=O)N2	test	3.60	5.08
107	O=C2NC(=O)N(C)C(Sc1ccccc1)=C2C	train	3.82	5.63

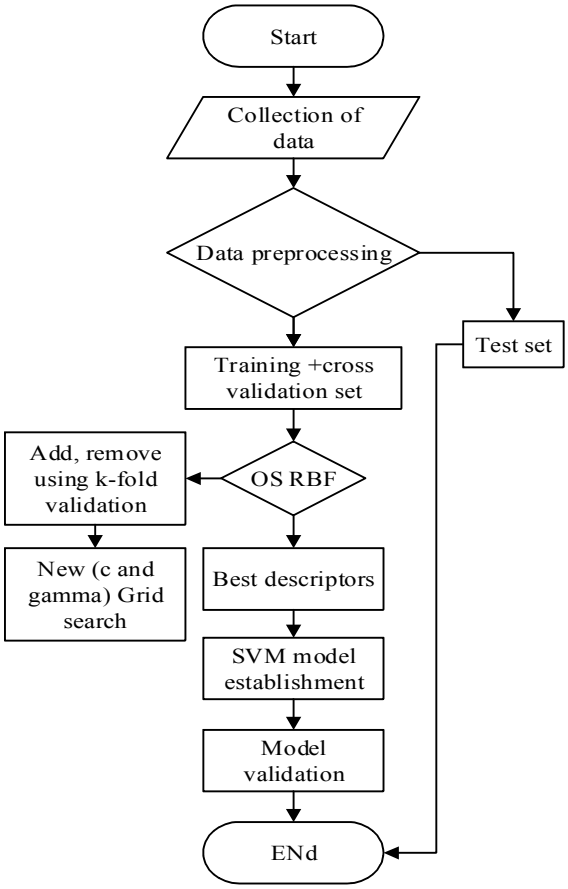
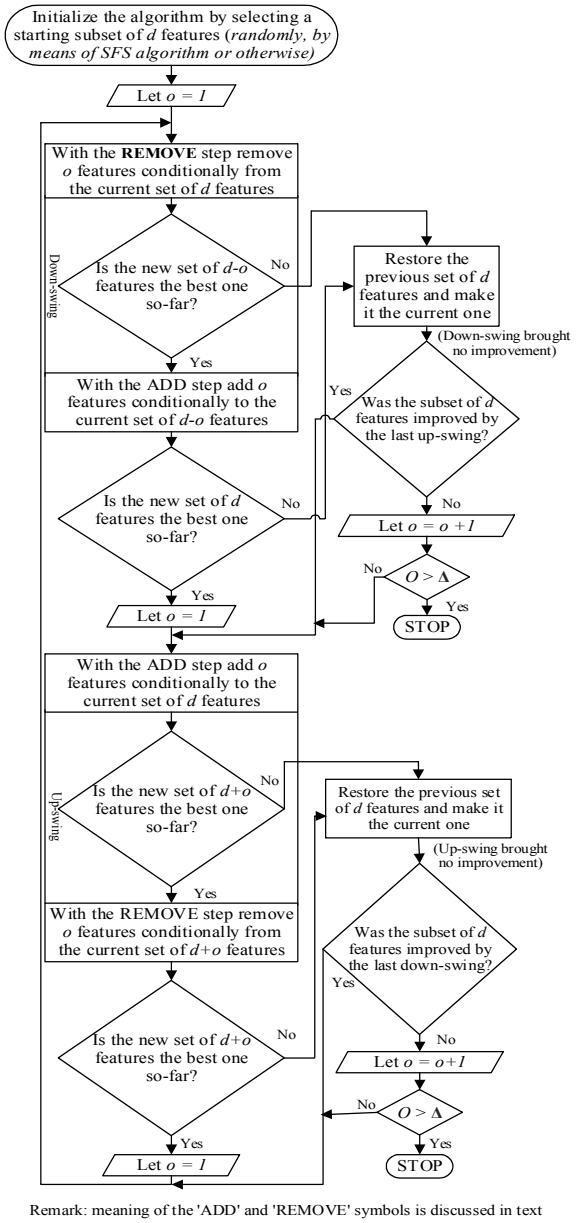


Fig. 2: System flowchart.

C and Gamma are the parameters for a nonlinear support vector machine (SVM) with a Gaussian (RBF) radial basis function kernel

The OS algorithm begins by initializing the main variables (o) and selecting an initial subset of d features using a random, SFS algorithm or another algorithm. Next, the algorithm launches a loop of two main phases: down-swing and up-swing. The down-swing phase starts with a REMOVE step that removes conditionally o features from the current set of d features; if the new set of d-o is the best, conditionally, an ADD step is used to add a set of o features to the current set of d-o features; otherwise, the previous state is restored to launch an up-swing phase. If the ADD step shows improvement, o is reinitialized to 1, and an up-swing phase is launched; otherwise, the previous state is restored, and o is incremented by one and then proceeds to an up-swing. After a restoring step, if the subset shows improvement, an up-swing is started; otherwise, o is incremented by one. If the condition $o < \Delta$ is true, the process is stopped; otherwise, an up-swing is launched. Additionally, the up-swing phase uses the same ADD / REMOVE steps but in the reverse order. The ADD step adds conditionally o features to the current set; if the new set of d + o features is the best so far, a REMOVE step begins to remove conditionally o features from the current set of d + o features; otherwise, the previous set of d features is restored (Fig 3). Thereafter, the algorithm tests the quality of the d features; if the quality is the best, o is reinitialized to 1, and another downswing is started; otherwise, the previous set is restored to determine if the subset is improved by the last down-swing. If improvement is observed, another down-swing is launched; otherwise, o is incremented by one. Finally, if $o < \Delta$, then the processes are stopped; otherwise, a down-swing phase is launched [17].



Remark: meaning of the 'ADD' and 'REMOVE' symbols is discussed in text

Fig. 3: OS Oscillating Search Flowchart [17].

During each evaluation of a descriptor set, the SVM model parameter estimation uses the grid search algorithm provided with LibSVM [18]. For learning, we used a nonlinear *SVM* with an *RBF* kernel to optimize the cost function. All experiments were conducted on a PC with 4.0GB of RAM, a 2.53-GHz processor, Manjaro Linux OS [19], and GNU Octave script [20]. The program was developed in our laboratory based on a set of M-files.

The dataset of 107 molecules was first split randomly into a training and cross-validation set (TSET) (85 molecules: 80% training data) to develop the QSAR models and a Test Set (22 molecules: 20% TestSet) to validate the QSAR models. The latter dataset did not undergo any treatment during the development of the QSAR models and was reserved for testing the reliability of the models only.

Results and Discussion

Variable selection is a key step in the process of developing a *QSAR* model. Many selection algorithms (statistics or heuristics) have been developed for this purpose. Success with the use of genetic algorithms for this purpose has been reported in the literature. In this work, we used another descriptor selection algorithm.

A good descriptor selection algorithm and a high number of cases (all data) are recommended to establish a statistically reliable *QSAR* model. The *QSAR* model was developed using theoretical descriptors calculated from a sample of 107 molecules derived from HEPT that have been tested as inhibitors of HIV-1 reverse transcriptase (Table-2).

The application of the OS selection algorithm on the set of descriptors allowed us to obtain a series of optimal models of different sizes. Table-2 presents the series of optimum models obtained in different sizes ranging from 1 to 5 model descriptors and their validation using the coefficients of determination R and Q, which ranged from 0.5738 to 0.9046 and 0.6534 to 0.9020, respectively.

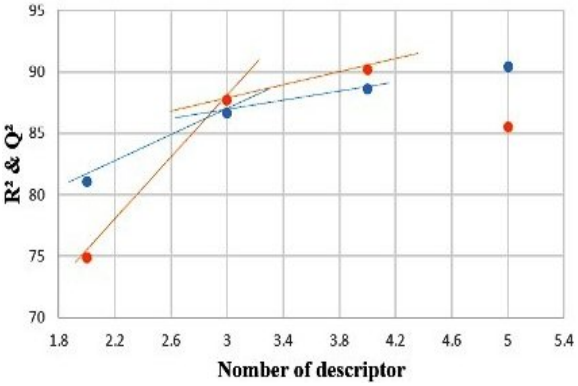


Fig. 4: Optimal number of descriptors for the optimal model.

Determining the number of descriptors in the *QSAR* model is important to prevent, to some extent, chance correlations between the descriptors that compose the model. To determine the optimal number of descriptors in the *QSAR* model, we used a simple method: the breakpoint. We consecutively set several models of different sizes ranging from 1 to 5 descriptors in the model. The optimal model corresponding to the breakpoint was obtained by analyzing the graphical representation of the coefficients of determination R^2 and Q^2 validation based on the number of descriptors (Fig 4). This graph reveals that the breakpoint [21] corresponds to the use of 3 descriptors. As a result, the optimal model is composed of three descriptors (SpMax4_Bh €; MATS5m and MLOGP).

This model has a high value of the coefficient of determination, which explains the good correlation between the descriptors and inhibitory activity. The collinearity problem between the descriptors included in the final model *QSAR* was tested by examining the correlation matrix by calculating the correlation coefficient for all possible combinations of the three pairs of descriptors. High values of the correlation coefficient $R > 0.9$ correspond to strong correlations among the model's descriptors. The results are summarized in Table-3.

Table-2: Series of optimum models obtained in different dimensions

n	Descriptors	*train-R^2*	*train-MSE*	*test-Q^2*	*test-MSE*
1	MATS5e	0.5738	0.0389	0.6534	0.0225
2	MATS5m:MLOGP	0.8105	0.0148	0.7486	0.0294
3	SpMax4_Bh(e):MATS5m:MLOGP	0.8662	0.0107	0.8769	0.0142
4	SpMax8_Bh(m):MATS5m:MLOGP:TDB07s	0.8859	0.0092	0.9020	0.0085
5	P1s:R4e:MATS5m:MATS1p:MLOGP	0.9046	0.0078	0.8534	0.0144

Table-3: Correlation matrix among the three descriptors of the model.

	SpMax4_Bh(e)	MLOGP	MATS5m
MLOGP	0.356		
MATS5m	-0.059	-0.333	
EC_{50} exp	0.301	0.71	-0.742

To generalize the implemented model to other data, the model was first tested and validated using different methods of internal and external validation. *QSAR* model validation is required to estimate its reliability. In this study, we used the randomization test method [22] (Table-4). The results of the randomization test shown in Table-4 confirm clearly the robustness and stability of our model; the statistical parameters of our model are higher than those obtained by the models generated randomly and show that our model is not due to chance. Thus, this model is effective and capable of modeling and predicting the inhibitory activity of reverse transcriptase by HEPT derivatives.

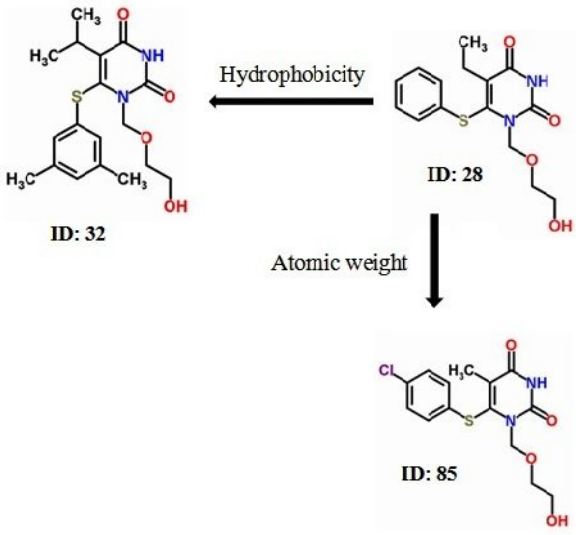
Table-4: Randomization Test

ID	R^2	Q^2	MSE
Our model	0.866217	0.87691	0.014164
1	0.009884	-0.59519	0.13714
2	0.00794	-0.61329	0.138696
3	0.014342	-0.97974	0.170201
4	0.014641	-0.55733	0.133885
5	0.00618	-0.6317	0.140279
6	0.000026	-0.77865	0.152913
7	0.001367	-0.83449	0.157713
8	0.000002	-0.77197	0.152338
9	0.006046	-0.6332	0.140408
10	0.000513	-0.80935	0.155552

The high value of the coefficient of determination from external validation, Q^2ext = 0.8769, confirms the excellent predictive power of our model. This high value also demonstrates that the SVM technique is adequate to produce an effective *QSAR* model for modeling and predicting the inhibitory activity of the HEPT derivatives against HIV-1 reverse transcriptase. The influence of the three descriptors on the inhibitory activity of HIV-1 reverse transcriptase can be interpreted chemically. Every biological activity is directly linked to the molecular forms of chemicals (electronic, geometric, and constitutional); it is thus possible to determine the factors responsible factors for the observed inhibitory effects of the compounds by interpreting the descriptors collected in the resulting model.

As shown in Fig 6, the two descriptors (MLOGP and MATS5m) significantly influence the biological magnitude of the inhibition of HIV-1 reverse transcriptase. The first descriptor (MLOGP) correlates positively with biological response (EC50) and describes the hydrophobicity of molecules and their influence on the inhibition of HIV-1 reverse transcriptase. The second descriptor (MATS5m) correlates negatively with the biological response (EC_{50}), indicating greater inhibitory activity. Mats5m is weighted by atomic weight: molecules consisting of more hetero activity are best to promote hydrogen bond formation and a stable ligand-enzyme complex.

The Scheme 1 demonstrates clearly the effect of the two descriptors, MLOGP and MATS5m, on the inhibitory activity of transcriptase enzyme, where the substitution of compound 28 (EC50 = 6.92) with a chloride group on the aromatic ring was significantly improved its activity (Compound 85, EC50 = 3.60), contrary to the increase in hydrophobicity of the same compound by adding methyl groups has lowered its activity (Compound 32, EC50 = 8,57). The third descriptor (SpMax4_Bh (e)) has a low correlation and high distribution for all molecules (Fig 5).



Scheme-1: Two descriptors MLOGP and MATS5m.

Table-5: Definition of the selected descriptors.

Descriptors	Definition
MLOGP	(Moriguchi octanol water partition coefficient) descriptor choice belonging to the family of molecular properties.
MATS5m	(Moran Autocorrelation of Lag 5 weighted by mass) 2D block family autocorrelation.
SpMax4_Bh (e)	(largest eigenvalue n.4 of the Burden matrix weighted by Sanderson Electronegativity) of the Burden eigenvalue family.

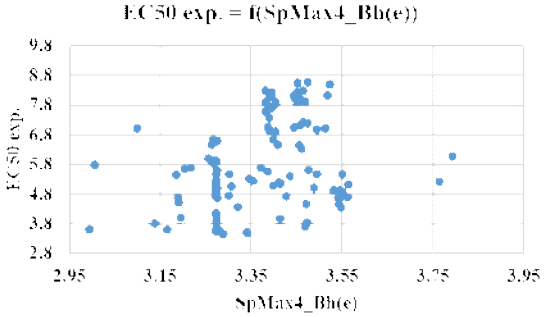


Fig. 5: Graphical representation of EC50 (Experimental) = f(SpMax4_Bh(e)).

The effects of all molecules on biological activity are clearly explained by the two MLOGP and MATS5m descriptors (Fig 6). The results also revealed a small influence of the third descriptor, SpMax4_Bh(e) (Fig 5), on the inhibitory activity of HIV-1 reverse transcriptase for all molecules (Fig 6).

We aimed to compare the results obtained using this model and those obtained from previous analyses of the same set of molecules and the same biological magnitude (Table-6). Based on this representation, the model obtained by the application of OS-SVM is superior to previous models in two aspects: the strong correlation between the value of EC_{50} and the descriptors (R^2 = 0.8662), the nature of the selected descriptors and information carried by these selected descriptors, and, finally, the high predictive power of Q^2ext (0.8769).

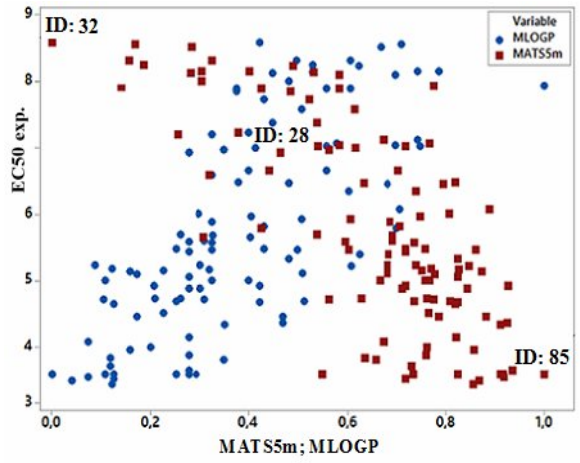


Fig. 6: Effects of MLOGP and MATS5M on the experimental EC50values.

We graphed the calculated EC_{50} values as a function of the experimental EC_{50} values (Fig 7).

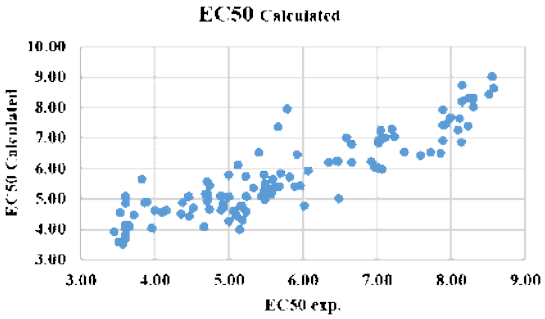


Fig. 7: Calculated and experimental anti-HIV-1 activity values for the test set.

Table-6: Comparison of studies using the same set of HEPT derivatives.

	Model	K	R^2	Q^2	Ref.
1	MLR	9	0.9	0.745	[12]
	PLS	9	0.889	0.8599	
2	MLR	5	0.815	0.783	[24]
3	MLR	4	0.83	0.7	[25]
	FFNN	4	0.852	0.81	
4	FFNN	7	0.977	0.862	[26]
5	RBFN	11	0.927	0.855	[27]
6	MLR	5	0.904	0.827	[28]
7	MLR	8	0.837	0.837	[29]
	BPNN	8	0.867	0.841	
	SVM	8	0.863	0.856	
8	MLR	4	0.799	0.597	[30]
	ANN	4	0.825	0.671	
	SVM	4	0.817	0.561	
9	SPA-UVE-PLS	9	0.84	0.8	[31]
10	MLR-ACO	7	0.86	0.85	[32]
11	SVM-RBF	6	0.945	0.881	[33]
12	MLR	6	0.811	0.778	[34]
	FFNN	6	0.919	0.919	
13	MLR	9	0.949	0.745	[35]
14	CP neural network	11	0.875	/	[36]
15	RBFNs	11	0.927	0.925	[37]
16	Montecarlo	7	0.8818	0.9243	[38]
17	OS-SVM	3	0.8662	0.8789	This work

K: Number of descriptors per model
MLR: Multiple Linear Regressions
PLS: partial least squares
FFNN: Feed Forward Neural Network
RBFN: Radial Basis Function Network
BPNN: Back Propagation Neural Network
SVM: Support Vector Machine
ANN: Artificial Neural Network
SPA-UVE-PLS: SPA Successive Projection Algorithm, UVE Uninformative Variable Elimination
MLR-ACO: Ant Colony Optimization
CP neural network : Counter-Propagation Neural Network
RBFNs: Radial Basis Function Networks

Conclusion

In this study, we conducted a quantitative analysis of the inhibitory activity-structure relationship for HIV-1 reverse transcriptase using a set of molecules derived from HEPT. A non-linear model was developed using the SVM method. The statistical results showed good correlation between the activity and three molecular descriptors opened in the built model. The high values of the determination

coefficients from internal and external validation clearly confirmed the stability, robustness and good predictive ability of the built model. These results demonstrate that the SVM technique is adequate to produce an effective *QSAR* model that is capable of modeling and predicting the inhibitory activity of HEPT derivatives against reverse transcriptase.

Acknowledgments

This work was generously supported by the General Directorate for Scientific Research and Technological Development (DGRSDT), Algerian Ministry of Scientific Research. The authors would like to dedicate this work to the memory of their colleague Azzouzi Abdelkader.

References

1. D. C. Boettiger, T. Sudjaritruk, R. Nallusamy, P. Lumbiganon, S. Rungmaitree, R. Hansudewechakul, N. Kumarasamy, T. Bunupuradah, V. Saphonn, K. H. Truong, Non-Nucleoside Reverse Transcriptase Inhibitor–Based Antiretroviral Therapy in Perinatally HIV-Infected, Treatment-Naïve Adolescents in Asia, *J. Adolesc. Health*, **58**, 451 (2016).
2. S. Butini, S. Gemma, M. Brindisi, G. Borrelli, I. Fiorini, A. Samuele, A. Karytinos, M. Facchini, A. Lossani, S. Zanoli, G. Campiani, E. Novellino, F. Focher, G. Maga, Enantioselective binding of second generation pyrrolobenzoxazepinones to the catalytic ternary complex of HIV-1 RT wild-type and L100I and K103N drug resistant mutants, *Bioorg. Med. Chem. Lett*., **21**, 3935 (2011).
3. S. Butini, M. Brindisi, S. Cosconati, L. Marinelli, G. Borrelli, S. S. Coccone, A. Ramunno, G. Campiani, E. Novellino, S. Zanoli, A. Samuele, G. Giorgi, A. Bergamini, M. D. Mattia, S. Lalli, B. Galletti, S. Gemma, G. Maga, Specific Targeting of Highly Conserved Residues in the HIV-1 Reverse Transcriptase Primer Grip Region. 2. Stereoselective Interaction to Overcome the Effects of Drug Resistant Mutations, *J. Med. Chem*., **52**, 1224 (2009).
4. A. Hameed, M. I. Abdullah, E. Ahmed, A. Sharif, A. Irfan, S. Masood, Anti-HIV cytotoxicity enzyme inhibition and molecular docking studies of quinoline based chalcones as potential non-nucleoside reverse transcriptase inhibitors (NNRT), *Bioorg. Chem*., **65**, 175 (2016).
5. E. Herzig, A. Hizi, The importance of glutamine 294 that affects the ribonuclease H activity of the reverse transcriptase of HIV-2 to viral replication, Virology, **483**, 13 (2015).
6. W.-G. Lee, K. M. Frey, R. Gallardo-Macias, K. A. Spasov, A. H. Chan, K. S. Anderson, W. L. Jorgensen, Discovery and crystallography of bicyclic arylaminoazines as potent inhibitors of HIV-1 reverse transcriptase, Bioorg. *Med. Chem. Lett*., **25**, 4824 (2015).
7. M. H. Manyeruke, T. O. Olomola, S. Majumder, S. Abrahams, M. Isaacs, N. Mautsa, S. Mosebi, D. Mnkandhla, R. Hewer, H. C. Hoppe, Synthesis and evaluation of 3-hydroxy-3-phenylpropanoate ester–AZT conjugates as potential dual-action HIV-1 Integrase and Reverse Transcriptase inhibitors, Biorg. Med. Chem., **23**, 7521 (2015).
8. A. Massarotti, A. Coluccia, An in-silico approach aimed to clarify the role of Y181C and K103N HIV-1 reverse transcriptase mutations versus Indole Aryl Sulphones, *J. Mol. Graphics Model.,* **63**, 49 (2016).
9. R. V. Patel, S. W. Park, Pyrroloaryls and pyrroloheteroaryls: Inhibitors of the HIV fusion/attachment, reverse transcriptase and integrase, *Biorg. Med. Chem.,* **23**, 5247 (2015).
10. F. S. Sarfo, M. A. Sarfo, D. Chadwick, Incidence and risk factors for neuropsychiatric events among Ghanaian HIV patients on long-term non-nucleoside reverse transcriptase inhibitor-based therapy, *eNeurologicalSci,* **3**, 21 (2016).
11. R. J. Visalli, H. Ziobrowski, K. R. Badri, J. J. He, X. Zhang, S. R. Arumugam, H. Zhao, Ionic derivatives of betulinic acid exhibit antiviral activity against herpes simplex virus type-2 (HSV-2), but not HIV-1 reverse transcriptase, *Bioorg. Med. Chem. Lett.,* **25**, 3168 (2015).
12. J. M. Luco, F. H. Ferretti, QSAR based on multiple linear regression and PLS methods for the anti-HIV activity of a large group of HEPT derivatives, *J. Chem. Inf. Comput. Sci*., **37**, 392 (1997).
13. ChemSpider., accessed. March 2016; http://www.chemspider.com
14. HyperChem(TM) Professional 8.0.10, Hypercube, Inc., 1115 NW 4th Street, Gainesville, Florida 32601, USA
15. Dragon v.6 software for molecular descriptors, Talete srl Copyright © 2013 Milano – Italy.
16. V. Vapnik, S. E. Golowich, A. Smola, Support vector method for function approximation, regression estimation, and signal processing, *Adv. Neural Inf. Process. Syst*., 281 (1997).
17. P. Somol, P. Pudil Oscillating search algorithms for feature selection. In Pattern Recognition, 2000. Proceedings. 15th International

Conference on, 2000; IEEE: 2000; Vol. **2**; pp 406

18. C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Transactions on Intelligent Systems and Technology (TIST), **2**, 27 (2011).
19. Manjaro Xfce 15.12 Linux (2016) Roland Singer, Guillaume Benoit, Philip Müller Free software licenses https://sourceforge.net/projects/manjarolinux
20. John W. Eaton, David Bateman, Søren Hauberg, Rik Wehbring (2015). GNU Octave version 4.0.0 manual: a high-level interactive language for numerical computations. CreateSpace Independent Publishing Platform. ISBN 1441413006, URL http://www.gnu.org/software/octae/doc/interpreter,
21. K. Kraim, D. Khatmi, Y. Saihi, F. Ferkous, M. Brahimi, Quantitative structure activity relationship for the computational prediction of α-glucosidase inhibitory*, Chemometr. Intell. Lab. Sys*., **97**, 118 (2009)
22. C. Rücker, G. Rücker, M. Meringer, y-Randomization and its variants in QSPR/QSAR, *J. Chem. Inf.* Model., **47**, 2345 (2007).
23. R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, WILEY-VCH Verlag GmbH, p. 652 (2008).
24. R. Garg, S. P. Gupta, H. Gao, M. S. Babu, A. K. Debnath, C. Hansch, Comparative quantitative structure-activity relationship studies on anti-HIV drugs, *Chem. Rev*., **99**, 3525 (1999).
25. H. Bazoui, M. Zahouily, D. Zakarya, S. Sebti, S. Boulajaaj, Structure-Activity Anti-Hiv-1 Relationships Study Of A Series Of Hept, *Phys. Chem. News,* **6**, 135 (2002).
26. L. Douali, D. Villemin, D. Cherqaoui, Neural networks: accurate nonlinear QSAR model for HEPT derivatives, *J. Chem. Inf. Comput. Sci*., **43**, 1200 (2003).
27. A. Malek-Khatabi, M. Kompany-Zareh, S. Gholami, S. Bagheri, Replacement based non-linear data reduction in radial basis function networks QSAR modeling, Chemometrics Intellig. Lab. Syst., **135**, 157 (2014).
28. A. Afantitis, G. Melagraki, H. Sarimveis, P. A. Koutentis, J. Markopoulos, O. Igglessi-Markopoulou, A novel simple QSAR model for the prediction of anti-HIV activity using multiple linear regression analysis, *Mol. Divers*., **10**, 405 (2006).
29. N. J. Pancholi, S. Gupta, N. Sapre, N. S. Sapre, Design of novel leads: ligand based computational modeling studies on non-nucleoside reverse transcriptase inhibitors (NNRTIs) of HIV-1, Mol. *Biosyst.,* **10**, **313** (2014).
30. B. Shaik, T. Zafar, V. K. Agrawal, Estimation of anti-HIV activity of HEPT analogues using MLR, ANN, and SVM techniques, *Int. J. Med. Chem.,* **2013** (2013).
31. N. Omidikia, M. Kompany-Zareh, Uninformative variable elimination assisted by Gram–Schmidt orthogonalization/successive projection algorithm for descriptor selection in QSAR, Chemometrics *Intellig. Lab. Syst.,* **128**, 56 (2013).
32. V. Zare‐shahabadi, F. Abbasitabar, Application of ant colony optimization in development of models for prediction of anti‐HIV‐1 activity of HEPT derivatives, *J. Comput. Chem*., **31**, 2354 (2010).
33. R. Darnag, A. Schmitzer, Y. Belmiloud, D. Villemin, A. Jarid, A. Chait, M. Seyagh, D. Cherqaoui, QSAR studies of HEPT derivatives using support vector machines, QSAR & Combinatorial Science, **28**, 709 (2009).
34. M. Jalali‐Heravi, A. Kyani, Comparison of Shuffling‐Adaptive Neuro Fuzzy Inference System (Shuffling‐ANFIS) with Conventional ANFIS as Feature Selection Methods for Nonlinear Systems, QSAR & Combinatorial Science, **26**, 1046 (2007).
35. C. Duda-Seiman, D. Duda-Seiman, M. Putz, D. Ciubotariu, QSAR modelling of anti-HIV activity with HEPT derivatives, *Digest J. Nanomat. Biostruct*, **2**, 207 (2007).
36. M. Arakawa, K. Hasegawa, K. Funatsu, QSAR study of anti-HIV HEPT analogues based on multi-objective genetic programming and counter-propagation neural network*, Chemometrics Intellig. Lab. Syst*., **83**, 91 (2006).
37. Y. Akhlaghi, M. Kompany‐Zareh, Application of radial basis function networks and successive projections algorithm in a QSAR study of anti‐HIV activity for a large group of HEPT derivatives, *J. Chemometrics,* **20**, 1 (2006).
38. A. P. Toropova, A. A. Toropov, J. B. Veselinović, F. N. Miljković, A. M. Veselinović, QSAR models for HEPT derivates as NNRTI inhibitors based on Monte Carlo method, *Eur. J. Med. Chem*., **77**, 298 (2014).