Taylor & Francis
Taylor & Francis Group

SAR and QSAR in Environmental Research

ISSN: 1062-936X (Print) 1029-046X (Online) Journal homepage: https://www.tandfonline.com/loi/gsar20

4D-QSAR study of HEPT derivatives by electron conformational–genetic algorithm method

L. Akyüz , E. Sarıpınar , E. Kaya & E. Yanmaz

SAR and QSAR in Environmental Research
Vol. 23, Nos. 5–6, July–September 2012, 409–433

Taylor & Francis
Taylor & Francis Group

4D-QSAR study of HEPT derivatives by electron conformational–genetic algorithm method

L. Akyüz[a], E. Sarıpınar[a]*, E. Kaya[a] and E. Yanmaz[b]

[a]*Department of Chemistry, Erciyes University, Kayseri, Turkey;* [b]*Department of Chemistry, Altınoluk Vacational College, Balıkesir Universty, Balıkesir, Turkey*

(*Received 9 July 2011; in final form 4 November 2011*)

In this work, the EC–GA method, a hybrid 4D-QSAR approach that combines the electron conformational (EC) and genetic algorithm optimization (GA) methods, was applied in order to explain pharmacophore (Pha) and predict anti-HIV-1 activity by studying 115 compounds in the class of 1-[(2-hydroxyethoxy)-methyl]-6-(phenylthio) thymine (HEPT) derivatives as non-nucleoside reverse transcriptase inhibitors (NNRTIs). The series of NNRTIs were partitioned into four training and test sets from which corresponding quantitative structure–activity relationship (QSAR) models were constructed. Analysis of the four QSAR models suggests that the three models generated from the training and test sets used in previous works yielded comparable results with those of previous studies. Model 4, the data set of which was partitioned randomly into two training and test sets with 11 descriptors, including electronical and geometrical parameters, showed good statistics both in the regression ($r^2_{training} = 0.867$, $r^2_{test} = 0.923$) and cross-validation ($q^2 = 0.811$, $q^2_{ext1} = 0.909$, $q^2_{ext2} = 0.909$) for the training set of 80 compounds and the test set of 27 compounds. The prediction of the anti-HIV-1 activity of HEPT compounds by means of the EC–GA method allowed for a quantitatively consistent QSAR model. In addition, eight novel compounds never tested experimentally have been designed theoretically using model 4.

Keywords: HEPT; EC–GA; genetic algorithm; pharmacophore; 4D-QSAR

1. Introduction

Acquired immunodeficiency syndrome (AIDS) is a fatal disorder for which no completely successful chemotherapy has been developed so far. AIDS is the result of a chronic persistent infection by the human retrovirus, human immunodeficiency virus (HIV) [1]. Reverse transcriptase (RT) is a key enzyme which is responsible for the process of HIV-1 replication. Recently, different types of studies have been conducted to achieve a better understanding of the mechanisms of HIV-1 replication, and several classes of compounds have been synthesized and tested as highly specific inhibitors of HIV-1 for AIDS therapy [2–5]. One of the most potent, selective and widespread inhibitors displaying high activity against HIV-1 reverse transcriptase (HIV-1RT) is 1-[(2-hydroxyethoxy)-methyl]-6-(phenylthio) thymine (HEPT). It was first synthesized by Tanaka et al. [2–5] as a non-nucleoside reverse transcriptase inhibitor (NNRTI) that is necessary for the catalytic formation of proviral DNA from viral RNA [6–8]. HEPT derivatives, whose

*Corresponding author. Email: emin@erciyes.edu.tr

ISSN 1062–936X print/ISSN 1029–046X online
© 2012 Taylor & Francis
http://dx.doi.org/10.1080/1062936X.2012.665082
http://www.tandfonline.com

the basic skeleton is shown in Table 1, have potent anti-HIV-1 activity at nanomolar concentration [5].

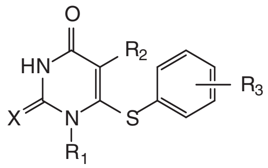
Quantitative structure–activity relationship (QSAR) modelling methods provide an effective means for investigating the relationship between the chemical structure of molecules and their biological action, during the development of novel drug candidates [9–11]. QSAR methods describe the mathematical relationship between the structural descriptors and biological activity of chemical compounds. Significant developments have occurred over the last two decades in these methods [12], with the aim of obtaining a fuller understanding of the relationship between biological activity and the structure of compounds. Recently, higher-dimensional QSAR techniques have been developed, such as 3D-QSAR and nD-QSAR methods. The primary goal of these techniques is to establish a correlation between biological activities of a series of compounds with the structural properties of each molecule, such as steric demand, lipophilicity and electrostatic interactions. The main differences between these techniques are the types of the structural parameters used within the mathematical approaches developed to predict biological activity. Since it is known that the three-dimensional features of molecules govern biological activity, 3D-QSAR methods are especially informative in demonstrating a 3D model of how structural changes affect biological activities. An advantage of the 3D-QSAR method is that it takes into account the 3D structures of molecules and is applicable to sets of structurally varied compounds. However, each compound is represented by a single bioactive conformation in 3D-QSAR methods; the other molecule conformers are not analysed, and the lowest energy of the conformation is used to generate the QSAR model.

The four-dimensional quantitative structure–activity relationship (4D-QSAR) approach, which includes the concepts of conformational flexibility and alignment freedom, was developed by Hopfinger et al. [13] as an extension of 3D-QSAR methodology for the representation of each compound by an ensemble of conformations. 4D-QSAR models are similar to 3D models but, when compared, the ligands of both the training set and test set are provided as an ensemble of conformations, instead of one fixed conformation [14]. Since the active conformer is often not the lowest-energy conformer, the 4D-QSAR approach used in this work is based on the generation of a conformational ensemble profile describing each molecule instead of the lowest-energy conformer. The relatively small energy differences between conformers can result in significant variations in electronic structure. Therefore, the 4D-QSAR approach used in the present work takes into account the Boltzmann populations and the dynamics of the conformational changes of all compounds in order to understand the effects of all energetically stable conformers on the biological activity [15]. In this study, not only the lowest-energy conformation but also all reasonable conformers were used in order to reveal the pharmacophore and predict the bioactivity.

Recently, several anti-HIV-1 QSAR studies have also been carried out in a series of HEPT derivatives acting as NNRTIs by research groups using different techniques, such as simple multiple linear regression (MLR) [16], artificial neural networks (ANNs) [17,18], holographic (HQSAR) methods, genetic algorithm-based partial least squares regression (GAPLS) [19], computational docking [20], principal component analysis (PCA) [21], stepwise discriminant analysis (SDA) [21], comparative molecular field analysis (CoMFA) [22] and supervised stochastic resonance [23].

The first attempt to establish the structure–activity relationships for HEPT derivatives was made by Tanaka et al. in 1992. In their research, the active and inactive regions of

Table 1. Chemical molecular structures: A^{exp}, experimental activity taken from reference 25 and A^{calc}, calculated activity for training and test sets according to Model 4 which is the best model for 115 HEPT derivatives as NNRTIs using EC–GA method. *CN* is the number of conformers in the table. All of the conformers in the table were used to reveal the pharmacaphore and to predict activity.



No	*R3*	*R2*	*R1*	*X*	*CN*	A^{exp}	A^{calc}
1[a, b, c]	2-Me	Me	$CH_2OCH_2CH_2OH$	O	12	4.150	4.554
2[a, b, c]	2-NO_2	Me	$CH_2OCH_2CH_2OH$	O	9	3.850	3.752
3[a, b, c]	2-OMe	Me	$CH_2OCH_2CH_2OH$	O	9	4.720	4.098
4[a, c]	3-Me	Me	$CH_2OCH_2CH_2OH$	O	15	5.590	4.833
5[a, b, c]	3-Et	Me	$CH_2OCH_2CH_2OH$	O	10	5.570	4.991
6[a, b, c]	3-t-Bu	Me	$CH_2OCH_2CH_2OH$	O	11	4.920	5.168
7[a, b, c]	3-CF_3	Me	$CH_2OCH_2CH_2OH$	O	12	4.350	4.858
8[a, b, c]	3-F	Me	$CH_2OCH_2CH_2OH$	O	9	5.480	4.742
9[a, b, c]	3-Cl	Me	$CH_2OCH_2CH_2OH$	O	13	4.890	4.926
10[a, b, c]	3-Br	Me	$CH_2OCH_2CH_2OH$	O	9	5.240	4.829
11[a, b, c]	3-I	Me	$CH_2OCH_2CH_2OH$	O	8	5.000	4.738
12[a, c]	3-NO_2	Me	$CH_2OCH_2CH_2OH$	O	9	4.470	4.603
13[a, b, c]	3-OH	Me	$CH_2OCH_2CH_2OH$	O	9	4.090	4.671
14[a, b]	3-OMe	Me	$CH_2OCH_2CH_2OH$	O	12	4.660	4.824
15[a, b, c]	3,5-Me_2	Me	$CH_2OCH_2CH_2OH$	O	17	6.590	6.285
16[a, b]	3,5-Cl_2	Me	$CH_2OCH_2CH_2OH$	O	6	5.890	5.951
17[a, b, c]	3,5-Me_2	Me	$CH_2OCH_2CH_2OH$	S	10	6.660	6.243
18[a, b, c]	3-COOMe	Me	$CH_2OCH_2CH_2OH$	O	13	5.100	4.792
19[a, b]	3-COMe	Me	$CH_2OCH_2CH_2OH$	O	6	5.140	4.790
20[a]	3-CN	Me	$CH_2OCH_2CH_2OH$	O	12	5.000	4.575
21[a, b, c]	H	$CH_2CH{=}CH_2$	$CH_2OCH_2CH_2OH$	O	9	5.600	5.630
22[a, b]	H	Et	$CH_2OCH_2CH_2OH$	S	14	6.960	6.161
23[a, c]	H	Pr	$CH_2OCH_2CH_2OH$	S	8	5.000	5.752
24[a, b, c]	H	i-Pr	$CH_2OCH_2CH_2OH$	S	8	7.230	6.885
25[a, b, c]	3,5-Me_2	Et	$CH_2OCH_2CH_2OH$	S	12	8.110	7.666
26[a, b]	3,5-Me_2	i-Pr	$CH_2OCH_2CH_2OH$	S	6	8.300	8.666
27[a, b, c]	3,5-Cl_2	Et	$CH_2OCH_2CH_2OH$	S	8	7.370	8.244
28[a, b, c]	H	Et	$CH_2OCH_2CH_2OH$	O	9	6.920	5.770
29[a, b, c]	H	Pr	$CH_2OCH_2CH_2OH$	O	9	5.470	5.116
30[a, b, c]	H	i-Pr	$CH_2OCH_2CH_2OH$	O	8	7.200	6.309
31[a, b, c]	3,5-Me_2	Et	$CH_2OCH_2CH_2OH$	O	12	7.890	6.950
32[a, b, c]	3,5-Me_2	i-Pr	$CH_2OCH_2CH_2OH$	O	7	8.570	8.570
33[a, b]	3,5-Cl_2	Et	$CH_2OCH_2CH_2OH$	O	10	7.850	7.570
34[a, b, c]	4-Me	Me	$CH_2OCH_2CH_2OH$	O	10	3.660	4.533
35[a, b, c]	H	Me	$CH_2OCH_2CH_2OH$	O	11	5.150	4.696
36[a, b, c]	H	Me	$CH_2OCH_2CH_2OH$	S	9	6.010	5.224
37[a, b]	H	I	$CH_2OCH_2CH_2OH$	O	6	5.440	5.653
38[a, b, c]	H	$CH{=}CH_2$	$CH_2OCH_2CH_2OH$	O	10	5.690	6.208
39[a, b, c]	H	CH=CHPh	$CH_2OCH_2CH_2OH$	O	8	5.220	6.047
40[a, b, c]	H	CH_2Ph	$CH_2OCH_2CH_2OH$	O	13	4.370	5.516

(*continued*)

Table 1. Continued.

No	*R3*	*R2*	*R1*	*X*	*CN*	A^{exp}	A^{calc}
41[a, b]	H	$CH{=}CPh_2$	$CH_2OCH_2CH_2OH$	O	12	6.070	6.018
42[a, b, c]	H	Me	$CH_2OCH_2CH_2OMe$	O	10	5.060	5.214
43[a, c]	H	Me	$CH_2OCH_2CH_2OAc$	O	15	5.170	5.176
44[a, b]	H	Me	$CH_2OCH_2CH_2OCOPh$	O	12	5.120	5.420
45[a, b, c]	H	Me	CH_2OCH_2Me	O	9	6.480	5.794
46[a, b, c]	H	Me	$CH_2OCH_2CH_2Cl$	O	10	5.820	5.327
47[a, c]	H	Me	$CH_2OCH_2CH_2N_3$	O	12	5.240	6.241
48[a, b, c]	H	Me	$CH_2OCH_2CH_2F$	O	9	5.960	5.811
49[a]	H	Me	$CH_2OCH_2CH_2Me$	O	10	5.480	5.395
50[a, b]	H	Me	CH_2OCH_2Ph	O	7	7.060	5.820
51[a, b]	H	Et	CH_2OCH_2Me	O	10	7.720	6.751
52[a, b, c]	H	Et	CH_2OCH_2Me	S	7	7.580	6.849
53[a, c]	3,5-Me_2	Et	CH_2OCH_2Me	O	10	8.240	8.461
54[a, c]	3,5-Me_2	Et	CH_2OCH_2Me	S	11	8.300	8.697
55[a, b, c]	H	Et	CH_2OCH_2Ph	O	10	8.230	6.712
56[a, b,]	3,5-Me_2	Et	CH_2OCH_2Ph	O	8	8.550	7.953
57[a, b, c]	H	Et	CH_2OCH_2Ph	S	7	8.090	7.101
58[a, b, c]	3,5-Me_2	Et	CH_2OCH_2Ph	S	9	8.140	8.427
59[a, b, c]	H	i-Pr	CH_2OCH_2Me	O	8	7.990	7.412
60[a, b, c]	H	i-Pr	CH_2OCH_2Ph	O	10	8.510	7.729
61[a, c]	H	i-Pr	CH_2OCH_2Me	S	8	7.890	8.068
62[a, c]	H	i-Pr	CH_2OCH_2Ph	S	8	8.140	7.996
63[a, b, c]	H	Me	CH_2OMe	O	9	5.680	5.125
64[a, b, c]	H	Me	CH_2OBu	O	16	5.330	5.455
65[a, b, c]	H	Me	Et	O	8	5.660	5.251
66[a, b]	H	Me	Bu	O	9	5.920	5.343
67[a, c]	3,5-Cl_2	Et	CH_2OCH_2Me	S	10	7.890	8.490
68[a]	H	Et	CH_2O-i-Pr	S	13	6.660	6.677
69[a, b, c]	H	Et	CH_2O-c-Hex	S	13	5.790	6.631
70[a, b]	H	Et	CH_2OCH_2-c-Hex	S	13	6.450	6.162
71[a, b, c]	H	Et	$CH_2OCH_2C_6H_4$(4-Me)	S	10	7.110	7.084
72[a, b, c]	H	Et	$CH_2OCH_2C_6H_4$(4-Cl)	S	10	7.920	7.104
73[a, b, c]	H	Et	$CH_2OCH_2CH_2Ph$	S	12	7.040	6.851
74[a]	3,5-Cl_2	Et	CH_2OCH_2Me	O	10	8.130	7.938
75[a, c]	H	Et	CH_2O-i-Pr	O	10	6.470	6.240
76[a, b, c]	H	Et	CH_2O-c-Hex	O	11	5.400	6.189
77[a, b, c]	H	Et	CH_2OCH_2-c-Hex	O	13	6.350	6.158
78[a, b]	H	Et	$CH_2OCH_2CH_2Ph$	O	14	7.020	6.871
79[a, c]	H	c-Pr	CH_2OCH_2Me	S	13	7.020	7.694
80[a, b, c]	H	c-Pr	CH_2OCH_2Me	O	17	7.000	7.191
81[b, c]	H	Me	$CH_2OCH_2CH_2OC_5H_{11}$-n	O	13	4.460	4.290
82[b, c]	2-Cl	Me	$CH_2OCH_2CH_2OH$	O	22	3.890	3.601
83[b, c]	3-CH_2OH	Me	$CH_2OCH_2CH_2OH$	O	20	3.530	3.661
84[b, c]	4-F	Me	$CH_2OCH_2CH_2OH$	O	19	3.600	3.861
85	4-Cl	Me	$CH_2OCH_2CH_2OH$	O	24	3.600	3.835
86[b, c]	4-NO_2	Me	$CH_2OCH_2CH_2OH$	O	26	3.720	3.695
87[c]	4-CN	Me	$CH_2OCH_2CH_2OH$	O	29	3.600	3.902
88[b, c]	4-OH	Me	$CH_2OCH_2CH_2OH$	O	19	3.560	3.845
89	4-OMe	Me	$CH_2OCH_2CH_2OH$	O	13	3.600	4.062
90[b, c]	4-COMe	Me	$CH_2OCH_2CH_2OH$	O	8	3.960	3.706
91[b]	3-COOH	Me	$CH_2OCH_2CH_2OH$	O	15	3.450	3.594

(*continued*)

Table 1. Continued.

No	*R3*	*R2*	*R1*	*X*	*CN*	A^{exp}	A^{calc}
92[b, c]	3-$CONH_2$	Me	$CH_2OCH_2CH_2OH$	O	7	3.510	3.547
93[b, c]	H	COOMe	$CH_2OCH_2CH_2OH$	O	5	5.180	4.865
94[b, c]	H	CONHPh	$CH_2OCH_2CH_2OH$	O	4	4.740	4.795
95[b, c]	H	SPh	$CH_2OCH_2CH_2OH$	O	14	4.680	5.451
96[c]	H	CCH	$CH_2OCH_2CH_2OH$	O	8	4.740	4.743
97[c]	H	CCPh	$CH_2OCH_2CH_2OH$	O	10	5.470	4.896
98	3-NH_2	Me	$CH_2OCH_2CH_2OH$	O	28	3.600	3.643
99[c]	H	$COCHMe_2$	$CH_2OCH_2CH_2OH$	O	7	4.920	4.585
100	H	COPh	$CH_2OCH_2CH_2OH$	O	6	4.890	4.638
101[c]	H	CCMe	$CH_2OCH_2CH_2OH$	O	25	4.720	4.704
102[b]	H	F	$CH_2OCH_2CH_2OH$	O	18	4.000	4.030
103[b]	H	Cl	$CH_2OCH_2CH_2OH$	O	18	4.520	3.774
104[b]	H	Br	$CH_2OCH_2CH_2OH$	O	25	4.700	4.696
105[c]	H	Me	$CH_2OCH_2CH_2OCH_2Ph$	O	15	4.700	4.071
106[c]	H	Me	H	O	7	3.600	2.863
107[b]	H	Me	Me	O	7	3.820	3.423
108[d]	2-Me	Me	$CH_2OCH_2CH_2SH$	O	10	NA	5.635
109[d]	2-Me	Me	$CH_2OCH_2CH_2SH$	S	12	NA	6.163
110[d]	2-Me	Me	CH_2OCH_2COH	O	18	NA	6.721
111[d]	2-Me	Me	CH_2OCH_2COH	S	2	NA	7.030
112[d]	2-Me	*c*-Hex	$CH_2OCH_2CH_2OH$	O	4	NA	6.671
113[d]	2-Me	*c*-Hex	$CH_2OCH_2CH_2OH$	S	11	NA	8.191
114[d]	2- *c*-Hex	Me	$CH_2OCH_2CH_2OH$	O	13	NA	5.039
115[d]	2- *c*-Hex	Me	$CH_2OCH_2CH_2OH$	S	5	NA	6.391

Notes: [a]: training set compounds in model 2
[b]: training set compounds in model 3
[c]: training set compounds in model 4
[d]: novel compounds never tested experimentally designed using Model 4.
NA: Not available
compounds not marked [a], [b] or [c] appear in test set for regarding model.

these compounds were shown in these studies [5]. Duda-Seiman et al. [16] performed research on a large series of HEPT compounds using the MTD (minimal topological difference) and HyperChem molecular modelling methods. However, they did not use a validation method to obtain a predictive QSAR model. To create a good statistical model requires an available data set to be divided into training sets and test sets [16]. Hannongbua et al. [22] performed a CoMFA study to describe QSARs, particularly to investigate the steric and electrostatic interactions for HEPT derivatives. Kireev et al. [24] performed a 3D-QSAR study including 87 HEPT derivatives using the MLR method. Luco and Ferretti [25] developed a QSAR based on MLR and PLS methods using topological descriptors in order to construct the relationship between the physicochemical parameters and biological activity of 107 HEPT derivatives. These authors concluded that PLS is a better method than MLR for evolving data, and the PLS method has better predictive power for representing models. In many cases, the PLS and MLR methods exhibit some limitations and give poor statistical results, especially when the relationship between dependent and independent variables is so complex that it can not be emulated by a simple linear relationship [25]. The correlation coefficient (*r*) and cross-validated correlation coefficient (q) values for these compounds are given in the literature [25]. Bak and Polanski [26] applied both Hopfinger's 4D-QSAR and self-organizing mapping-4D-QSAR (SOM-4D-QSAR), which are the self-organizing neural network versions of traditional 4D-QSAR, for the investigation of the antiviral activity of HEPT derivatives. These authors compared these methods for their performance in predicting the QSAR model. Both methods yielded comparable results according to cross-validated regression coefficient (q^2) values. However, they did not explain the regression coefficients ($r^2_{training}$ or r^2_{test}) and external validated regression coefficients (q^2_{ext1} or q^2_{ext2}). A high q^2 value for the training set has often been considered as a sufficient criterion of QSAR model accuracy. However, a high q^2 does not automatically imply the high predictive power of the model. There is no relationship between internal and external predictivity [27]: high internal predictivity may result in low external predictivity and vice versa. This effect is called the 'Kubinyi paradox' [28]. The overall picture which emerges from these QSAR studies shows that hydrophobic, electronic and steric characteristics of the compounds have predominant roles in the anti-HIV-1 activity of HEPT derivatives.

The electron conformational (EC) method by Bersuker and co-workers, presented as a QSAR method, is aimed at searching rules for predicting different activities based on the pharmacophores found previously by specific EC calculations [29]. For this purpose, a nonlinear mathematical model which defines the relationship between bioactivity and the parameters was presented for bioactivity prediction using one conformer for compounds. This EC method has been recently applied to a variety of problems such as screening rice blast activity inhibitors, angiotensin-converting enzyme inhibitors, group I metabotropic glutamate receptor agonists, inhibitors of human breast carcinoma, guanidino- and aminoguanidinopropionic acid analogues [29–36].

Genetic algorithm (GA) is a heuristic search method used for identifying optimal solutions to a problem where the possible solution space is too large to be exhaustively enumerated. GA has been widely used for feature optimization in QSAR models for variable selection [37–39]. The purpose of variable selection is to select the variables significantly contributing to prediction and to discard other variables by a fitness function [37–39].

The EC and GA methods, namely electron conformational and genetic algorithm method (EC–GA) [40–43], were combined for identifying pharmacophore and predicting anti-HIV-1 activity by studying 107 compounds in the class of HEPT derivatives as NNRTIs [25]. In the EC–GA method which incorporates conformational and alignment freedom, the conformers, heavily populated at room temperature, are taken into account by using Boltzmann weighting for pharmacophore identification and bioactivity predictions for a series of compounds that have the same type of biological activity. The EC–GA method is categorized as a useful ligand-based QSAR method. In the EC–GA method, as in all QSAR methods used to design a novel drug, the molecular structure is also represented with the physicochemical and structural properties of the molecules. These are usually represented by a set of descriptors, with the assumption that the molecule's activity is related to the values of these descriptors in some way. In this method, the optimal subset of these chemical descriptors is selected from a molecular descriptor pool to obtain a statistically robust model. For selecting a subset of relevant descriptors and building the optimal QSAR model, the GA optimization method is used in the EC–GA method. The leave-one-out cross-validation method is used in order to explore the reliability of the EC–GA method by dividing data into two groups: one used to train the model and the other to validate it. The results obtained in this way allowed us to perform computer-based screening of eight new compounds (never tested before) and to predict theoretically the novel molecular structures 108–115 with statistically significant anti-HIV-1 activities as prospective candidates for future experimental studies.

In our previous studies, this method was successfully performed for a 4D-QSAR procedure to identify the pharmacophore for benzotriazines as Src inhibitors, the anti-cancer activity of *N*-morpholino triaminotriazine derivatives, penicillins and 1,4-dihydropyridines as calcium channel antagonists as well as to make a quantitative prediction of activity [40–43].

The aim of the research is to explain the pharmacophore and to predict anti-HIV-1 activity of HEPT derivatives as NNRTIs. Below, we compare the performance of the EC–GA method on HEPT compounds with those of MLR, ANNs, HQSAR, GAPLS, SSR (supervised stochastic resonance) and CoMFA analyses reported in the literature [16–26]. In most of these studies, molecules in the compound series corresponding to one fixed conformation were employed for model building and bioactivity prediction. Although these methods are popular 3D QSAR methods, they do not always lead to reliable predictions because of several internal problems: (i) identification of the pharmacophore features of active conformation; (ii) consideration of the conformation of molecules; and (iii) external prediction ability of models. It has been shown that the EC–GA method is useful for overcoming these difficulties in structure–property studies of HEPT derivatives and the other series of compounds [44].

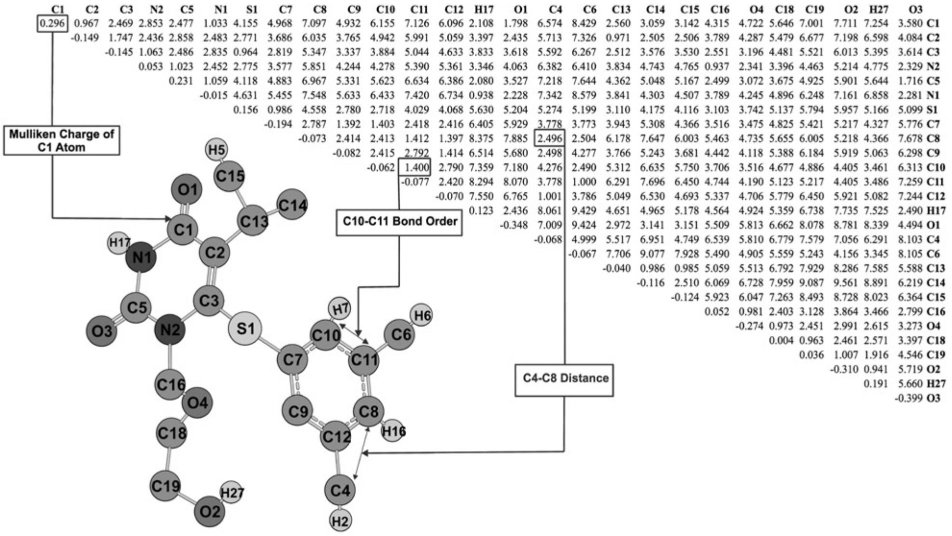
2. Materials and methods

4D-QSAR analysis using the EC–GA method was carried out on a series of 107 HEPT derivatives to reveal the pharmacophore and to predict anti-HIV-1 activity. The chemical structures of the HEPT derivatives are shown in Table 1 and the experimental activity data (IC50, micromol) are taken from literature [25].

The newly developed EC–GA method is described in more detail elsewhere [40–43]. The computational part of the EC–GA method consists of the following steps: (1) calculation of conformational and quantum-chemical analysis; (2) formation of the electron conformational matrices of congruity (ECMC) for each conformer of all compounds; (3) multiple intercomparison of the ECMC between themselves and activity feature (pharmacophore) selection; (4) preparation of the molecular descriptors; (5) selection of the best subset of parameter combinations which contribute mostly to activity using the GA method; and (6) implementation of robust statistical methods to predict the model's power.

Quantum mechanical calculation and conformational analysis for HEPT derivatives have been performed from SPARTAN 08 software using the parametric model number 3 (PM3) method [45]. Because lower energy conformers are responsible for biological activity much more than higher energy conformers, the conformers were seperated from heavily populated conformers (smaller than 1.5 $\mathrm{kcal\,mol^{-1}}$) at room temperature using Boltzmann weighting [29].

In the EC–GA method, the properties of a molecule in its interaction with the bioreceptor are described by a set of electronic and geometric features presented in terms of elements of the ECMC. Figure 1 illustrates an example of the ECMC calculated for compound 32 for its lowest-energy conformer. The ECMC is a 3D square matrix of the order *nxn* (*n* is the number of atoms in the molecule) and it is symmetric with respect to



C1	C2	C3	N2	C5	N1	S1	C7	C8	C9	C10	C11	C12	H17	O1	C4	C6	C13	C14	C15	C16	O4	C18	C19	O2	H27	O3	
0.296	0.967	2.469	2.853	2.477	1.033	4.155	4.968	7.097	4.932	6.155	7.126	6.096	2.108	1.798	6.574	8.429	2.560	3.059	3.142	4.315	4.722	5.646	7.001	7.711	7.254	3.580	C1
	-0.149	1.747	2.436	2.858	2.483	2.771	3.686	6.035	3.765	4.942	5.991	5.059	3.397	2.435	5.713	7.326	0.971	2.505	2.506	3.789	4.287	5.479	6.677	7.198	6.598	4.084	C2
		-0.145	1.063	2.486	2.835	0.964	2.819	5.347	3.337	3.884	5.044	4.633	3.833	3.618	5.592	6.267	2.512	3.576	3.530	2.551	3.196	4.481	5.521	6.013	5.395	3.614	C3
			0.053	1.023	2.452	2.775	3.577	5.851	4.244	4.278	5.390	5.361	3.346	4.063	6.382	6.410	3.834	4.743	4.765	0.937	2.341	3.396	4.463	5.214	4.775	2.329	N2
				0.231	1.059	4.118	4.883	6.967	5.331	5.623	6.634	6.386	2.080	3.527	7.218	7.644	4.362	5.048	5.167	2.499	3.072	3.675	4.925	5.901	5.644	1.716	C5
					-0.015	4.631	5.455	7.548	5.633	6.433	7.420	6.734	0.938	2.228	7.342	8.579	3.841	4.303	4.507	3.789	4.245	4.896	6.248	7.161	6.858	2.281	N1
						0.156	0.986	4.558	2.780	2.718	4.029	4.068	5.630	5.204	5.274	5.199	3.110	4.175	4.116	3.103	3.742	5.137	5.794	5.957	5.166	5.099	S1
							-0.194	2.787	1.392	1.403	2.418	2.416	6.405	5.929	3.778	3.773	3.943	5.308	4.366	3.516	3.475	4.825	5.421	5.217	4.327	5.776	C7
								-0.073	2.414	2.413	1.412	1.397	8.375	7.885	2.496	2.504	6.178	7.647	6.003	5.463	4.735	5.655	6.005	5.218	4.366	7.678	C8
									-0.082	2.415	2.792	1.414	6.514	5.680	2.498	4.277	3.766	5.243	3.681	4.442	4.118	5.388	6.184	5.919	5.063	6.298	C9
										-0.062	1.400	2.790	7.359	7.180	4.276	2.490	5.312	6.635	5.750	3.706	3.516	4.677	4.886	4.405	3.461	6.313	C10
											-0.077	2.420	8.294	8.070	3.778	1.000	6.291	7.696	6.450	4.744	4.190	5.123	5.217	4.405	3.486	7.259	C11
												-0.070	7.550	6.765	1.001	3.786	5.049	6.530	4.693	5.337	4.706	5.779	6.450	5.921	5.082	7.244	C12
													0.123	2.436	8.061	9.429	4.651	4.965	5.178	4.564	4.924	5.359	6.738	7.735	7.525	2.490	H17
														-0.348	7.009	9.424	2.972	3.141	3.151	5.509	5.813	6.662	8.078	8.781	8.339	4.494	O1
															-0.068	4.999	5.517	6.951	4.749	6.539	5.810	6.779	7.579	7.056	6.291	8.103	C4
																-0.067	7.706	9.077	7.928	5.490	4.905	5.559	5.243	4.156	3.345	8.105	C6
																	-0.040	0.986	0.985	5.059	5.513	6.792	7.929	8.286	7.585	5.588	C13
																		-0.116	2.510	6.069	6.728	7.959	9.087	9.561	8.891	6.219	C14
																			-0.124	5.923	6.047	7.263	8.493	8.728	8.023	6.364	C15
																				0.052	0.981	2.403	3.128	3.864	3.466	2.799	C16
																					-0.274	0.973	2.451	2.991	2.615	3.273	O4
																						0.004	0.963	2.461	2.571	3.397	C18
																							0.036	1.007	1.916	4.546	C19
																								-0.310	0.941	5.719	O2
																									0.191	5.660	H27
																										-0.399	O3

Figure 1. Electron-conformational matrices of congruity (ECMC) of the lowest energy conformer of the reference compound (compound 32) with the highest activity in the series of HEPT derivatives. The diagonal elements represent to the Mulliken charges while the non-diagonal elements are bond orders for chemically bonded pairs of atoms and interatomic distances for non-bonded pairs. Hydrogen atoms bonded to carbon atoms are excluded in the ECMC for simplicity.

diagonal elements. The diagonal elements of the ECMC seen in Figure 1 refer to Mulliken atomic charges. The non-diagonal elements are either the bond orders for chemically bonded pairs of atoms or the interatomic distance for non-bonded atoms [46,47]. In this way, the ECMC contains both geometric and electronic parameters. ECMs of congruity have been constructed from the data of conformational analysis and the electronic structure calculation of molecules in a compound series to reveal pharmacophore atoms. In this study, 1233 ECMs of congruity, corresponding to the conformers of the HEPT derivatives, were constructed using EMRE software [40–44].

In this work, the ECMC of the most active compound chosen as a template was compared with other ECMs of congruity to find the pharmacophore, within given tolerances, and identified by the electron conformational submatrix of activity (ECSA) which represents the pharmacophore. The pharmacophore, which can also be described as a group essential for activity, is defined as a specific three-dimensional arrangement of functional groups that are found in active molecules. To begin the identification of pharmacophore groups, the most active compound was chosen as a template molecule, and 55 compounds with values of $-\log(EC_{50}) \geq 5.47$ were classified as high-activity compounds and 52 molecules with values of $-\log(EC_{50}) < 5.47$ were considered as low-activity compounds [34,35,46–49]. The ECMC of the template molecule, which is the lowest-energy conformer of the compound with the highest activity, is compared with all other ECMs of congruity within given tolerances to reveal the ECSA [40]. In general, tolerance values existing in compounds of high activity are lower than those existing in compounds of low activity.

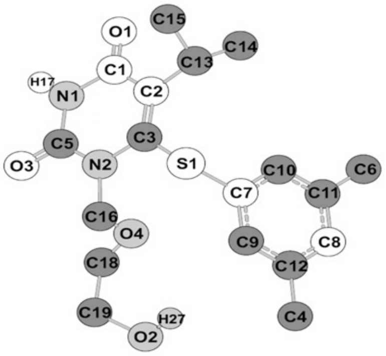
The tolerance evaluation procedure starts with smaller (small) values. Then, it is increased until an ECSA with the smallest tolerance values is reached in all the active compounds. By gradually changing the limits of tolerance for diagonal and non-diagonal matrix elements, the tolerance limits of the ECSA are obtained [29–33]. Two criteria are used to determine the pharmacophore. The first (P_α) only demonstrates the probability of pharmacophore presence in active compounds; the other (α_a) shows the possibility of pharmacophore presence in inactive compounds. They are given by the following equations [34,35,46–49]: $P_\alpha = (c_1 + 1)/(c_1 + c_2 + 2)$, $\alpha_a = (c_1 * c_4 - c_2 * c_3)/(b_1 * b_2 * b_3 * b_4)^{1/2}$, where c_1 (53 compounds) and c_2 (2 compounds) are the numbers of the molecules that are inclusive and non-inclusive, respectively, as the features of activity in the active compounds; c_3 (two compounds) and c_4 (50 compounds) have the same meaning for low-active compounds; b_1 and b_2 are the numbers of the molecules in the class of active and low-active compounds; $b_3 = c_1 + c_3$; $b_4 = c_2 + c_4$ [46]. Accordingly, the group is determined as the ECSA existing in all of the compounds and having the minimum tolerance value and maximum (P_α) and (α_a) values. Under these circumstances, the probabilistic estimation values are high enough, $P_\alpha = 0.9474$, $\alpha_a = 0.9252$.

In this paper, the ECSA that is common for all of the active molecules contains eight atoms consisting of C1, C2, S1, C7, C8, H17, O1 and O3 in all of the compounds. The resulting ECSA, which represents the pharmacophore for the HEPT derivatives and its tolerance values for the compounds of both high and low activity, is given in Table 2. The pharmacophore atoms are shown in white in Table 2. The first submatrix, which demonstrates the lowest-energy conformer of the template molecule, corresponds to the pharmacophore group. The second submatrix in Table 2 shows the tolerance values for 55 compounds with high activity, and the third submatrix shows the tolerance values for 52 compounds with low activity. The fourth submatrix is obtained without limitation on tolerance values of the pharmacophore group; the maximum tolerance values are calculated for all conformers of all compounds, and the submatrix shows the tolerance values for the 1233 conformers of the 107 compounds. As seen in Table 2, the tolerance values in compounds with high activity are usually lower than those found in compounds with low activity. For example, the tolerance values of the distance between the H17 and O3 atoms for higher and less active compounds are ±0.2112 and ±2.0871, and the charge values of the C8 atom are ±0.0522 and ±0.3372, respectively.

Two of the eight atoms in the pharmacophore group are oxygen atoms, O1 and O3, and one is a sulphur atom, S1. The highest positive charge is concentrated in the C1 atom, which is part of the carbonyl group. The S1 and H17 atoms have positive charges, and the other atoms have negative charges. C1, C2 and O1 form a rigid plane, whereas the positions of the O3, C7 and C8 atoms are highly flexible. The flexible position of the C8 atom of the pharmacophore can also be seen directly in the tolerance values for the corresponding distances in the ECSA matrix. The C8 atom, which is a distal atom, has different distances from the other pharmacophore atoms. First of all, the positioning flexibilty of all atoms in the overall pharmacophore structure is taken into account in the EC–GA method, and then the relationships between the flexibility and activity of the compounds are used for bioactivity predictions. The pharmacophore analysis explains that the pharmacophore group, containing the C1, C2, S1, C7, C8, H17, O1 and O3 atoms, is a principal component of activity within the specifics of the drug–receptor interaction mechanism for HEPT derivatives.

A 3D pharmacophore is defined as an arrangement of molecular features in space that are required for a desired biological activity. However, the concept of pharmacophore does not explain why different compounds with the same pharmacophore have quite different activities. A pharmacophore is a necessary, but not sufficient, condition for

Table 2. (a) ECSA (pharmacophore) of reference compound (compound 32) for HEPT derivatives; (b) tolerance matrix of ECSA for 55 compounds with high activity; (c) tolerance matrix of ECSA for 52 compounds with low activity; (d) tolerance values for all conformers (1233). Pharmacophore atoms are shown in white. P_α and α_a values of pharmacophore were found as 0.9474 and 0.9252, respectively.



C1	*C2*	*S1*	*C7*	*C8*	*H17*	*O1*	*O3*	*Pha Atoms*
(a) ECSA of reference compound (Pharmacophore group)								
0.2955	0.9674	4.1549	4.9677	7.0969	2.1078	1.7980	3.5799	C1
	−0.1487	2.7711	3.6858	6.0353	3.3966	2.4349	4.0838	C2
		0.1555	0.9864	4.5580	5.6303	5.2036	5.0987	S1
			−0.1936	2.7869	6.4046	5.9286	5.7759	C7
				−0.0733	8.3751	7.8847	7.6781	C8
					0.1231	2.4362	2.4898	H17
						−0.3479	4.4944	O1
							−0.3989	O3
(b) Tolerance matrix of ECSA for 55 compounds with high activity								
±0.0425	±0.0153	±0.0822	±0.3875	±0.6701	±0.0824	±0.0323	±0.4177	C1
	±0.1344	±0.0690	±0.3970	±0.6016	±0.0648	±0.0597	±0.4616	C2
		±0.0326	±0.0095	±0.0393	±0.1431	±0.1276	±0.4152	S1
			±0.0412	±0.0220	±0.3485	±0.5038	±0.6197	C7
				±0.0522	±0.5789	±0.9110	±0.8722	C8
					±0.0134	±0.0354	±0.2112	H17
						±0.0219	±0.3781	O1
							±0.0917	O3
(c) Tolerance matrix of ECSA for 52 compounds with low activity								
±0.0294	±0.0198	±0.1343	±0.3422	±0.7381	±0.0084	±0.0220	±0.7490	C1
	±0.1928	±0.1693	±0.4211	±0.6655	±0.0485	±0.0475	±0.4569	C2
		±0.0517	±0.0251	±0.0282	±0.0517	±0.2248	±2.0413	S1
			±0.0688	±0.0252	±0.1743	±0.4686	±2.0541	C7
				±0.3372	±0.3858	±1.0919	±1.8240	C8
					±0.0128	±0.0860	±2.0871	H17
						±0.0280	±1.0506	O1
							±0.2976	O3
(d) Tolerance matrix of ECSA for 1233 conformations of 107 compounds								
±0.0425	±0.0211	±0.1438	±0.3875	±0.7744	±0.0825	±0.0407	±0.7490	C1
	±0.2048	±0.1786	±0.4450	±0.6877	±0.0660	±0.0605	±0.4650	C2
		±0.0647	±0.0261	±0.0424	±0.1448	±0.2347	±2.0422	S1
			±0.0782	±0.0259	±0.3979	±0.5127	±2.0541	C7
				±0.3372	±0.6986	±1.1296	±1.8240	C8
					±0.0138	±0.0884	±2.0871	H17
						±0.0322	±1.0506	O1
							±0.2976	O3

the ligand to interact at the receptor site. Therefore, other factors, such as auxilary groups (AG), anti-pharmacophore shielding groups (APS), electronic properties and three-dimensional space, which are linked to biological activity, must be considered [29]. APS and AG parameters, which affect the activity of molecules in all molecular systems with a pharmacophore, exist there. Both AG and APS can be described by means of geometrical, electronic and physicochemical parameters. The effect of AG and APS is determined by introducing the function S that is the sum of all these effects and is given in short as follows [29]:

$$S_{ni} = \sum_{j=1}^{N} \kappa_j a_{ni}^{(j)} \tag{1}$$

where $a_{ni}^{(j)}$ are the molecular parameters describing the jth kind of APS or AG groups in the ith conformation of the nth molecule, N is the number of selected APS and AG parameters and the κ_j are variational parameters. A formula of activity is obtained for all conformations using the function S and taking into account the Boltzmann weighting of each conformation as a function of its energy and temperature, and $\kappa_{j,}$ variational constants are calculated from the formula that was developed by Bersuker et al., given below [29–36].

$$A_n = A_l \frac{\sum_{i=1}^{m_l} e^{-E_{li}/RT}}{\sum_{i=1}^{m_n} e^{-E_{ni}/RT}} \frac{\sum_{i=1}^{m_n} \delta_{ni}[Pha]e^{-S_{ni}}e^{-E_{ni}/RT}}{\sum_{i=1}^{m_l} \delta_{li}[Pha]e^{-S_{li}}e^{-E_{li}/RT}} \tag{2}$$

where δ is a kind of Dirac δ function. It is equal to Pha 1 when pharmacophore is present. It is equal to Pha 0 when pharmacophore is absent. A_n and A_l stand for the numerical values of activity of the nth compound and the reference compound, respectively. E_{li} is the relative energy of the ith conformation of the reference compound (in kcal mol^{-1}). E_{li} is the relative energy of the ith conformation of the nth compound (in kcal mol^{-1}) and R (kcal mol^{-1} K^{-1}) is the gas constant. T is the temperature in Kelvin.

The choice of $a_{ni}^{(j)}$ parameters and the determination of κ_j variational constants in Equation (1) are important components in this method. The unknown coefficients κ_j in the S function can be found by performing a least square minimization ($\sum_n |A_n^{calc} - A_n^{exp}|^2$) for all the compounds in training set. This procedure is carried out using Matlab software in conjunction with the optimization function lsqnonlin, which is a general nonlinear least squares fitting algorithm to fit the data. If the κ_j variational constants are known, bioactivity prediction for novel compounds may be possible. The numbers "κ_j" $= 1, 2, \ldots, N$, obtained in this way characterize the weights of each kind of the $a_{ni}^{(j)}$ parameters in the overall APS/AG influence [29].

In classical QSAR analysis, it is important to select the best subset descriptors from a large pool of descriptors. In this study, 300 different molecular descriptors including topological, geometrical and thermodynamical parameters were prepared and calculated for each conformer of 107 compounds using EMRE software [40]. Generally, selecting a proper subset of descriptors from a large descriptor pool is difficult, and is one of the most important steps in the QSAR modelling process. For selection of the most important descriptors, the GA technique was used [50]. In this QSAR study, the GA codes were written in Matlab by the authors. GA randomly creates subsets of chromosomes with the input parameters for the QSAR model. The lsqnonlin function within the statistics toolbox in MATLAB [51] was used to obtain κ_j values by numerically solving the system of differential equations for the best subset of variables.

The GA method represents a probable solution of a given problem by means of bit strings. It is optimized towards better solutions by applying genetic operators such as selection, mutation and crossover. The first step in the GA method is to create a population (population size = 500) of N individuals (feature subsets). Each individual encodes the same number of randomly chosen descriptors, and the fitness of each individual in this generation (generations = 500) is determined. In our study, further increasing the population and generation size from 500 to 1000 did not create a significant improvement but involved much longer computational time. The compounds in the dataset were divided into two: training sets (80) and test sets (27). The test set is not used during training but serves to test the predictive ability of final models. In this study, the predictive residual sum of squares (*PRESS*) is also taken as the fitness measure. Next, a fraction of children of the next generation is produced by crossover (crossover probability = 0.850, elite count = 2) and the rest by mutation (mutation probability = 0.015) from the parents on the basis of their scaled fitness scores. The new offspring contains characteristics from one or both parents, and is evaluated for fitness. The cycle continues for a predetermined number of generations, or until the results do not change continuously for a specified number of generations. Model parameters (kappa indices κ_j) are coded by chromosomes as integer numbers. Each parent is submitted to the lsqnonlin function to calculate the κ_j values of model parameters. The lsqnonlin function iterates κ_j values over the Equation (1). After determining the κ_j values and the most important parameters for the HEPT series, and in order to explore the reliability of the proposed method, the leave-one-out (LOO) cross-validation method is used. The validation of the developed model is a very important task in the building of the predictive QSAR model. For this reason, model validation is performed by internal validation and external validation techniques [52–56].

PRESS is a standard index to measure the accuracy of a modelling method based on the LOO cross-validation technique. PRESS is defined as the sum of the squared differences between calculated and experimental values of activity and can be written as:

$$PRESS_N = \sum_{n=1}^{N} |A_n^{exp} - A_n^{calc}|^2 \tag{3}$$

where A_n^{calc} is the value of the activity that is calculated in the LOO cross-validation model and A_n^{exp} is the value of experimental activity taken from the experimental data. The predictive performance of the model is measured by q^2, which is the cross-validation regression coefficient that is given as follows:

$$q^2 = 1 - \frac{\sum_{n=1}^{N} |A_n^{exp} - A_n^{calc}|^2}{\sum_{n=1}^{N} |A_n^{exp} - \bar{A}_n^{exp}|^2} \equiv 1 - \frac{PRESS}{SSY} \tag{4}$$

where N is the total number of training compounds in the data set. SSY is the sum of squares of deviations of the experimental values (A_n^{exp}) from their mean ($\bar{A}_n^{exp}$). A_n^{exp} and $\bar{A}_n^{exp}$ are the experimental and averaged experimental activity values of the dependent variable, respectively. The smaller the *PRESS* is, the better the model's accuracy is. Its value must be less than the SSY value for the model to be statistically significant.

The external predictive ability of a QSAR model is an important feature, especially if the model is to be used for the prediction of untested compounds. External validation is used for predicting the activity values of chemical structures that are in the same chemical domain as the training set but not used in the creation of the correlation in the training set. A QSAR model is developed from data that are obtained from the training set which is applied to the test set in order to explain the predictive ability of the model. Two different expressions used for quantifying the external prediction capability of QSAR models are discussed by Schüüremann et al. [53].

The average experimental value of training compounds and the average experimental value of test compounds are, respectively, used in q^2_{ext1} and in q^2_{ext2}. These expressions are defined as follows:

$$q^2_{ext1} = 1 - \frac{\sum_{n=1}^{N} |A^{exp}_{test} - A^{calc}_{test}|^2}{\sum_{n=1}^{N} |A^{exp}_{test} - \bar{A}^{exp}_{training}|^2} \tag{5}$$

$$q^2_{ext2} = 1 - \frac{\sum_{n=1}^{N} |A^{exp}_{n_{test}} - A^{calc}_{n_{test}}|^2}{\sum_{n=1}^{N} |A^{exp}_{n_{test}} - \bar{A}^{exp}_{n_{training}}|^2} \tag{6}$$

where N is the number of tested molecules and $A^{exp}_{n_{test}}$ is the experimental activity of the nth molecule in the test set. $A^{calc}_{n_{test}}$ is the calculated activity of test set without using the left-out compound in the model building. $\bar{A}^{exp}_{n_{training}}$ and $\bar{A}^{exp}_{n_{test}}$ are the average of experimental activities of training and test set, respectively.

The predictive ability of our 4D-QSAR model was evaluated by the LOO cross-validation method. In many cases, the LOO cross-validated regression coefficient (q^2) and regression coefficient ($r^2_{training}$) are taken as an evidence of the high predictive ability of QSAR models. In addition to a high q^2 and $r^2_{training}$, a reliable model should also be characterized by high r^2_{test}, q^2_{ext1} and q^2_{ext2} for the test set of the molecules that were not used to develop the models. To obtain a statistically significant QSAR model, there should not be a large difference between the $r^2_{training}$ and q^2 values and, in addition, an external set should be used for the predictive ability of the QSAR model [57].

3. Results and discussion

In this study, QSAR models were generated using four different training sets and then validated using the corresponding test sets; thus four independent models (models 1–4) could be obtained to evaluate both the robustness and the predictive ability of the models. Models 1–3 were constructed using the training set and test set compounds given in previous studies [25,26], but the training set and test set compounds in model 4 were randomly selected for comparison with the other models. These models were used to make a comparison with the model under discussion. For comparison, correlation, cross-validation and external validation coefficients were used for each model. The results are summarized in Table 3. By using sets containing different compounds and working in different ways, the form of the training set was found to have a significant impact on the predictive ability of the models.

In order to show the performance of these models and to obtain the optimum number of descriptors, the EC–GA method was applied to determine the anti-HIV-1 activity of the

Table 3. Comparison of four models according to statistical parameters to predict anti-HIV activity of HEPT derivatives. Calculated activity values according to statistical data obtained from Model 4 are given in Table 1.

	Training set	*Test set*	$r^2_{training}$	q^2	r^2_{test}	q^2_{ext1}	q^2_{ext2}
Model 1	107	–	0.887	0.868	–	–	–
Model 2	80	27	0.861	0.815	0.568	0.500	−5.562
Model 3	80	27	0.850	0.797	0.855	0.718	0.715
Model 4	80	27	0.867	0.811	0.923	0.909	0.909

Notes: model 1: all molecules were used as training set;
model 2: compounds 1–80 marked with [a] in Table 1 and compounds 81–107 were used as training set and test sets, respectively;
model 3: compounds marked with [b] in table 1 were used as training set;
model 4: compounds selected randomly and marked with [c] in Table 1 were used as training set.

HEPT derivatives. Since the optimum number of variables is not known in advance, several runs are needed to examine the relationship between the predictive power of a model (q^2) and the number of descriptors selected. Consequently, by using coefficient κ_j which was obtained from the training set by the lsqnonlin function, the activities of the test compounds were calculated using Equation (2). The test compounds are not included in the model generation for all of the models. The most active molecule, 32, was used as a template for alignment in the models. In order to demonstrate the predictive power and accuracy of the EC–GA method, the four models developed in this work were compared with those obtained with other QSAR approaches reported in the literature for the same data sets on HEPT compounds. We demonstrated that our models, which are constructed using the newly developed EC–GA method, were described by similar or better statistics and predictive power as compared with the other QSAR models. Thus, it has been proven that this approach was a powerful alternative to more popular QSAR methods.

In Luco et al.'s study [25], the activity of the compounds that were used as a training set in model 1 was estimated by applying the coefficients derived by PLS ($r^2 = 0.891$ and $q^2 = 0.866$) and MLR ($r^2 = 0.900$ and $q^2 = 0.745$) statistical methodology from all molecules in the compound series. Note that these models were constructed without using test set-predicted values to validate the model. In our study, the statistical quality of model 1, as depicted in Table 3, was also determined by $r^2_{training}$ and q^2. Model 1, which was developed by 12 descriptors (model 1's descriptors and κ_j values are given in Table 4), had the following statistics: $q^2 = 0.868$, $r^2_{training} = 0.887$. This seems adequate for comparing the results given in the previous work [25] with those of our study. The reggression coefficient ($r^2_{training}$) between the experimental and predicted activities calculated with the EC–GA method was rarely lower, but it implied the model's high predictive power with a higher cross-validation regression coefficient (q^2) than in the previous work. We can assume that model 1, which was generated with the EC–GA method, outperforms those given in the literature in terms of predictive ability. Moreover, model 1 has an advantage because of the nonlinear character of the relationship between variables and biological activities obtained by the EC–GA method.

In order to apply the EC–GA modelling method to derive Models 2 and 3, the data set of 107 was split into training (80 compounds) and test sets (27 compounds) as in Bak and Polanski's study [26]. They used Hopfinger's 4D-QSAR and SOM-4D-QSAR methods for

Table 4. Optimum 12 molecular parameters selected with GA and κ_j values used in activity calculation for HEPT derivatives in Model 1.

$a_n^{(j)}$	*Molecular parameters*	κ_j *values*
$a^{(1)}$	Distance between C1 and S1	−4.001
$a^{(2)}$	Electrostatic charge of the farthest atom (H27) bonded to N2	0.176
$a^{(3)}$	Atomic valency of the farthest atom (C15) bonded to C2	−13.602
$a^{(4)}$	Distance between C2 and N1	16.991
$a^{(5)}$	Distance between C2 and the farthest atom (C15) bonded to C2	−0.825
$a^{(6)}$	Distance between C8 and N2	−0.038
$a^{(7)}$	Distance between C8 and the farthest atom (H16) bonded to C8	−0.772
$a^{(8)}$	Distance between O1 and C2+ van der Waals radius of C2 atom	−16.527
$a^{(9)}$	Distance between C1 and the farthest atom (H7) bonded to C10+ van der Waals radius of H7 atom	0.023
$a^{(10)}$	Angle O3-C5-N1	−0.002
$a^{(11)}$	Polarizability (gamma) (au) YYYY	2×10^{-5}
$a^{(12)}$	Polarizability (gamma) (au) XXZZ	1.5×10^{-5}

Table 5. Optimum 12 molecular parameters selected with GA and κ_j values used in activity calculation for HEPT derivatives in Model 2.

$a_n^{(j)}$	*Molecular parameters*	κ_j *values*
$a^{(1)}$	Mulliken charge of C3 atom	−2.787
$a^{(2)}$	Atomic valency of the farthest atom (H27) bonded to N2	−0.026
$a^{(3)}$	Distance between C2 and N1	10.768
$a^{(4)}$	Distance between C2 and O3	17.690
$a^{(5)}$	Distance between S1 and the farthest atom (H16) bonded to C8	−0.455
$a^{(6)}$	Distance between C8 and H atom bonded to N1 (H17)	0.100
$a^{(7)}$	Distance between C8 and N2	−0.041
$a^{(8)}$	Distance between C8 and the farthest atom (H27) bonded to N2	0.089
$a^{(9)}$	Distance between O1 and C2+ van der Waals radius of C2 atom	−29.707
$a^{(10)}$	Distance between O1 and the farthest atom (H2) bonded to C12+ van der Waals radius of H2 atom	−0.059
$a^{(11)}$	Distance between C1 and O3+ van der Waals radius of O3 atom	0.026
$a^{(12)}$	Angle between line of C5–O3 atoms and C8–C7–O1 plane	−0.183

the investigation of the structure–activity relationships of a HEPT series. The training set compounds of Models 2 and 3 are marked with a and b, respectively, in Table 1. The cross-validated q^2 values of 4D-QSAR and SOM-4D-QSAR models ranged from 0.76–0.98 for the same training and test set compounds in model 2. However, test set statistics were not used in these methods. Therefore it did not provide any information about the predictive performance of the method. In another work, Heravi et al. [18] used both ANN and MLR techniques for the same data set. While the q^2 values for MLR and ANN were found as 0.605 and ranged from 0.525–0.954, the r^2 values were found to be 0.811 and 0.919, respectively (but no q^2_{ext1} and q^2_{ext2} values). In our study, the predictive ability of model 2 was determined by $r^2_{training}$, q^2, r^2_{test}, q^2_{ext1} and q^2_{ext2} values, as seen from Table 3. Model 2 was developed using 12 descriptors, which are given in Table 5. Although the training set had good statistical prediction, contrary to expectations, the external set showed much worse statistical prediction for model 2. In this model, the training set

Table 6. Optimum 12 molecular parameters selected with GA and κ_j values used in activity calculation for HEPT derivatives in Model 3.

$a_n^{(j)}$	*Molecular parameters*	κ_j *values*
$a^{(1)}$	Electrostatic charge of the farthest atom (H27) bonded to N2	0.482
$a^{(2)}$	Distance between O3 and C12	–0.077
$a^{(3)}$	Distance between O3 and the farthest atom (H7) bonded to C10	0.107
$a^{(4)}$	Distance between C1and the farthest atom (H6) bonded to C11+ van der Waals radius of H6 atom	0.025
$a^{(5)}$	Distance between C8 and the farthest atom (H7) bonded to C10 + van der Waals radius of H7 atom	0.126
$a^{(6)}$	Distance between C8 and the farthest atom (H2) bonded to C12 + van der Waals radius of H2 atom	–0.215
$a^{(7)}$	Distance between C2–C7–C8 plane and the farthest atom bonded to C11	0.045
$a^{(8)}$	Distance between S1–O4–O1 plane and the farthest atom bonded to C11+ van der Waals radius of H6 atom	–0.025
$a^{(9)}$	Angle O1–C1–C2	–0.009
$a^{(10)}$	Angle C2–N2–the farthest atom (H27) bonded to N2	0.004
$a^{(11)}$	Angle C1–C2–the farthest atom bonded to C2 (C15) (radian)	0.320
$a^{(12)}$	log P, partition coefficient	–0.029

showed a good fit with $r^2_{training} = 0.861$ and $q^2 = 0.815$, but the test set did not have high correlation coefficients, relatively. Model 2 gave a negative q^2_{ext2} value (−5.562) and lower r^2_{test} (0.568) for the test set.

Some of the statistics of the 4D-QSAR, SOM-4D-QSAR, ANN and MLR methods can be expected to be better than, though still comparable with, those obtained for model 2 to reliably predict the modelled property for the entire universe of chemicals. Many authors consider a high q^2 value (for instance, $q^2 > 0.5$) as an indicator, or even as ultimate proof, that the model is highly predictive. Indeed, according to the current OECD guidelines [58], high q^2 cannot be a single parameter to imply the predictive ability of a model. Thus, a high q^2 value, alone, is insufficient proof that a QSAR model shows a high predictive power. It has been shown that the only way to estimate the true predictive power of a model is to test it using an external test set. Therefore, goodness-of-fit and robustness, and the predictivity of a model are represented by internal performance and external validation, respectively [59]. Model 2 explains the importance of q^2_{ext1} and q^2_{ext2} for the predictive abilities of QSAR models, and proved that high q^2 values do not automatically imply a model's high predictive ability; external validation is the only way to 'determine' the true predictive power of a QSAR model.

In other training set compounds that correspond with b in Table 1 and our model 3, the values of q^2 obtained by Bak and Polanski [26] using Hopfinger's 4D-QSAR and SOM-4D-QSAR methods ranged from 0.23–0.77 and 0.60–0.71, respectively, but external values were not given. In our study, model 3 yielded high predictive correlation coefficients ($q^2 = 0.797$, $q^2_{ext1} = 0.718$ and $q^2_{ext2} = 0.715$) and high fitted correlation coefficients ($r^2_{training} = 0.850$ and $r^2_{test} = 0.855$). Model 3 was developed using 12 descriptors which are given in Table 6. Results obtained from model 3 show that EC–GA gives slightly better values for q^2, $r^2_{training}$, q^2_{ext1} and q^2_{ext2} than the other methods. This is because in our method a conformational ensemble of compounds according to Boltzmann weighting is used instead of a single representative structure. It is important to note that the prediction

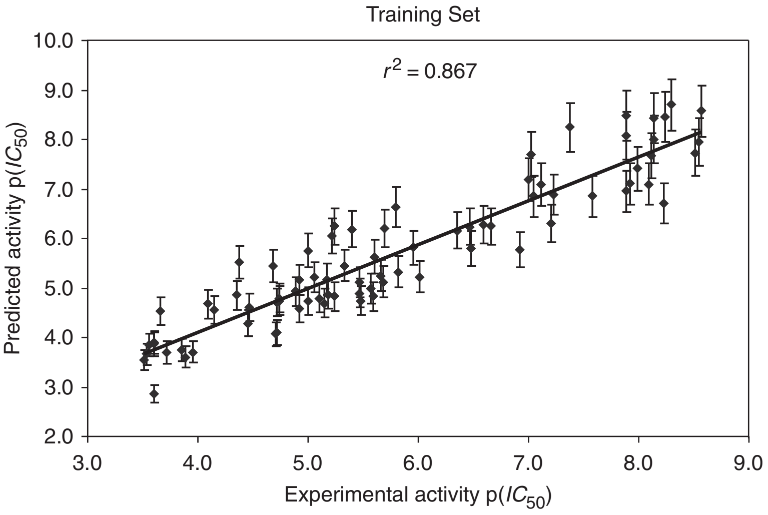


Figure 2. Actual vs. predicted and experimental anti-HIV-1 activity values for the training set obtained by Model 4 with 11 variables, using the EC–GA method.

ability estimated by EC–GA methods for this analysis is at least as good as the results obtained with other QSAR methods originally used on the same data sets.

An additional study using a training/test set protocol was performed in order to better estimate the results of the EC–GA method. In model 4, 18 molecules (marked with *c* in Table 1) were randomly partitioned into a training set of 107 HEPT derivatives and a test set of 27 compounds which were used to validate the QSAR models. The best model was selected based on the best value of the cross-validated coefficient q^2 and the regression coefficient $r^2_{training}$ for the training set, and the external validation coefficients q^2_{ext1} and q^2_{ext2} and the regression coefficient r^2_{test} for the test set. All the models were compared with the other models; model 4 had a considerably high $r^2_{training}$, q^2, q^2_{ext1} and q^2_{ext2}. In this case, the best model was determined as model 4 with the highest cross-validated coefficicent $q^2 = 0.811$ and the regression coefficients $r^2_{training} = 0.867$ and $r^2_{test} = 0.923$ for the training set and test set, respectively. For the test set, both of the external validation coefficients q^2_{ext1} and q^2_{ext2} were as high as 0.909. It was proven that model 4 was a statistically significant model and had high predictive power. This was a demonstration that model 4 was not obtained by chance correlation. Hence, we successfully developed an externally validated QSAR model for predicting anti-HIV-1 activity for HEPT derivatives. The plots of the predicted activities versus experimental values of anti-HIV-1 activity are shown individually for the training and test set in Figures 2 and 3, respectively.

To determine the best subset of descriptors in the best model, the HEPT derivatives were calculated for a range of 1–14 parameters. To obtain the optimum number of descriptors, r^2 and q^2 values were presented in a graph against the number of descriptors as seen in Figure 4. According to the q^2 values, the results indicated that 7–14 parameters are acceptable. As seen from the graph, the model reaches a stable condition after 11 variables, and any new additional variable is unnecessary. Therefore, the best model was found as the 11-parameter model involving charges, distance, angle and dipole (X) (X component of the dipole moment) using the EC–GA method. The predictive performances of the generated QSAR models using the reduced set of descriptors are shown in Table 7.

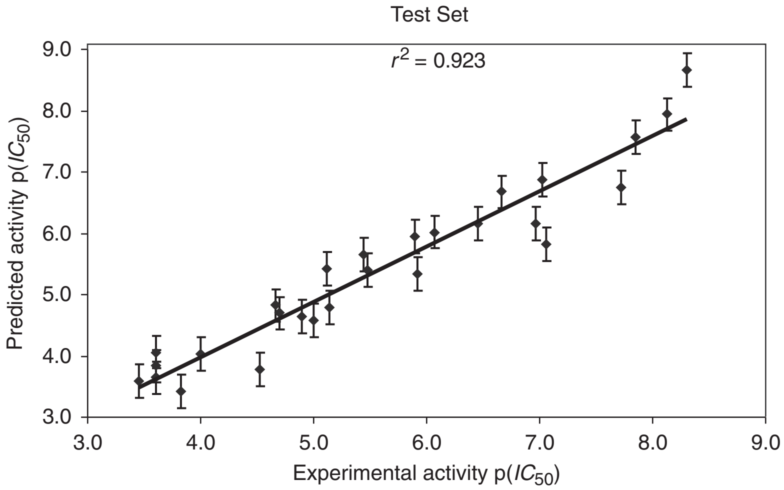


Figure 3. Actual vs. predicted and experimental anti-HIV-1 activity values for the test set obtained by Model 4 with 11 variables, using the EC–GA method.

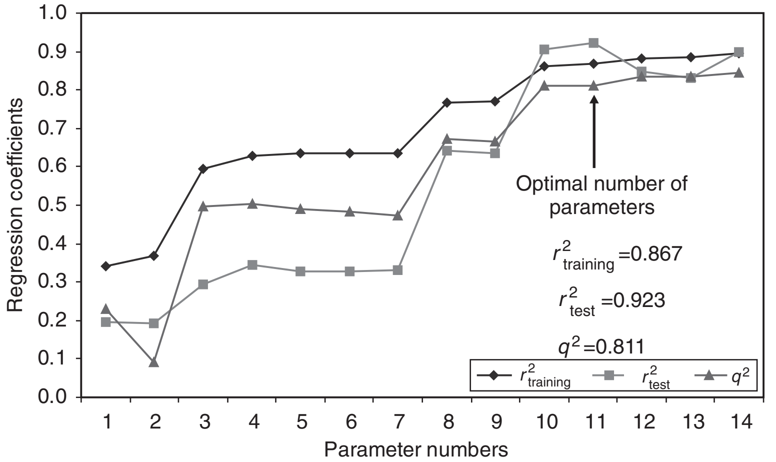


Figure 4. Plot of $r^2_{training}$, r^2_{test} and cross-validated q^2 against the parameter numbers. The model reaches a stable condition after 11 variables, and any new variable added is unnecessary.

In Table 7, the $a^{(1)}$, $a^{(5)}$, $a^{(6)}$ and $a^{(7)}$ parameters, which are corresponding interatomic distances employed to take into account the influence of their limited flexibility on activity, are related to the overall shape of the molecules. We employed the difference in distance between all pairs of the atoms in all of the conformers of molecules. We found that the distances of the C2-O1, C2-N1, C8-N2 and C8-the farthest atom (H16) bonded to C8 atom were strongly influenced by such conformational changes. The first parameter, which gives the largest negative contribution in the change of activity as the κ_j value, $a^{(1)}$, is the distance between C2 and O1, which are pharmacophoric atoms. Distance descriptors, which are defined by molecular conformation, have been proven to be useful in the construction of QSAR models and in the prediction of important properties of the active conformation [60]. Distance descriptors are helpful in quantifying their inter and

Table 7. Optimum 11 molecular parameters selected with GA and κ_j values used in activity calculation for HEPT derivatives in Model 4.

$a_n^{(j)}$	*Molecular parameters*	κ_j *values*
$a^{(1)}$	Distance between C2 and O1	−15.370
$a^{(2)}$	Mulliken charge of C2	−2.530
$a^{(3)}$	Electrostatic charge of the farthest atom (C15) bonded to C2	0.153
$a^{(4)}$	Electrostatic charge of the farthest atom (H27) bonded to N2	0.190
$a^{(5)}$	Distance between C2 and N1	13.281
$a^{(6)}$	Distance between C8 and N2	2.494
$a^{(7)}$	Distance between C8 and the farthest atom (H16) bonded to C8	−0.025
$a^{(8)}$	Distance between C7 and the farthest atom (H5) bonded to C2+ van der Waals radius of H5 atom	−0.628
$a^{(9)}$	Angle between line of C5–O3 atoms and C8–C7–O1 plane	−0.224
$a^{(10)}$	Angle O3–C5–N2	−0.002
$a^{(11)}$	Dipole (X), X component of the dipole moment	0.015

intramolecular 3D interactions between ligand and bioreceptor, and these descriptors are related to the ability of the ligand molecule to fit into its site in the receptor.

The second parameter, $a^{(2)}$, is the Mulliken charge of the C3 atom of the thymine ring, and its parameter value is changeable in all molecules. $a^{(3)}$ and $a^{(4)}$ are electronic parameters of the electrostatic charges of the farthest atom (C15) bonded to C2 and the farthest atom (H27) bonded to N2, respectively. The electrostatic charges of the C15 and H27 atoms directly affect activity. The atomic charges at the C2 and C3 positions, representing the steric interaction of the substituents of the compounds, play an important role in the model, particularly in the relationships between the atomic charge and the nature of the substituent on the atom. Positive electrostatic charges located near the substituent, which is attached to the C2 position of the thymine ring, showed favourable positive charges. It can be understood that the C3 atom with negative Mulliken charges may reduce the binding affinity of all the molecule conformers in the central area where the increase of negative charge is; the C15 and H27 atoms with positive electrostatic charges in the more distant area, where the increase of the positive charge is, may decrease binding affinity, too. $a^{(8)}$ is the distance between C7-H5 excluding hydrogen as the farthest atom van der Waals radius of the H5 atom (Figure 5). The van der Waals atomic radius is one of the criteria used for determining whether atoms are bonded to one another [61]. Therefore, the van der Waals radius is one of the most important descriptors for describing the interaction with the receptors. $a^{(9)}$ is the angle between the line of the C5-O3 atoms and the C8-C7-O1 plane. $a^{(10)}$ is the angle (radian) between the O3-C5-N2 atoms. $a^{(11)}$ is the dipole (X) which is the X component of the dipole moment. The dipole (X) is one of the most important molecular descriptors for predicting activity [62]. Dipole (X) is applicable in reciprocal format: its contribution is negative and positive for some of the conformers in the HEPT series in predicting biological activity.

Activity depends exponentially on S ($A \sim e^{-S}$). It shows the APS parameter, and providing that the product of the parameter and kappa values is positive and vice versa, it also shows the AG. $a^{(2)}$, $a^{(3)}$, $a^{(4)}$, $a^{(5)}$ and $a^{(8)}$ are APS parameters. $a^{(1)}$, $a^{(6)}$, $a^{(7)}$, $a^{(9)}$ and $a^{(10)}$ are AG parameters. $a^{(11)}$ has positive or negative values in different conformers of the same compound; it is not only AG, but also APS.

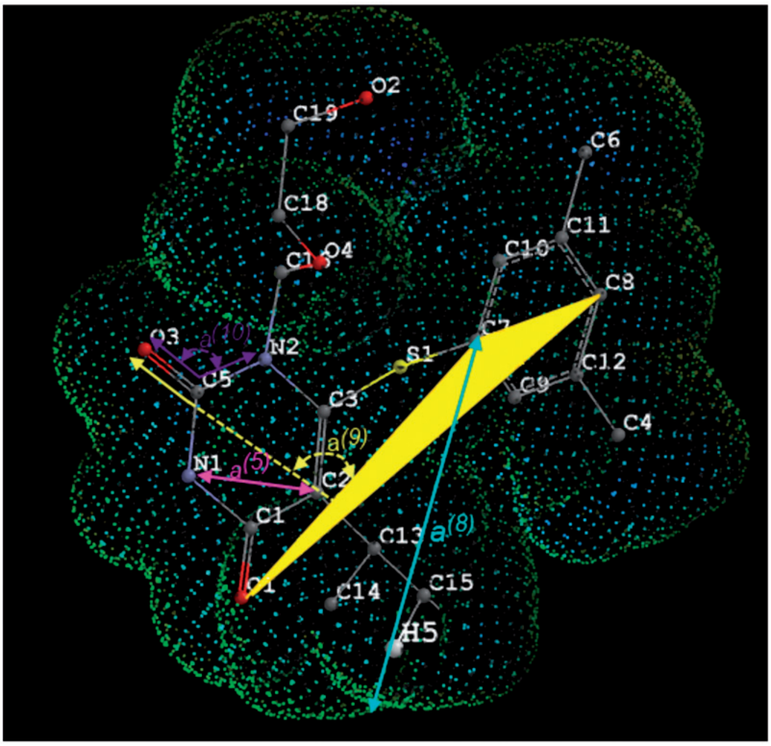


Figure 5. Van der Waals surface of the reference compound. $a^{(8)}$ is the angle between C7-C8-O1 plane shown in yellow and O3-C5 line. $a^{(9)}$ indicates C7-H5 distance+ van der Waals radius of H5 atom. $a^{(5)}$ is distance between C2 and N1 atoms.

The E statistical technique [63,64] was used to determine the role of selected molecular parameters on anti-HIV-1 activity. The statistical E value is defined as follows [34]:

$$E = \frac{PRESS_P}{PRESS_{P-1}} \tag{7}$$

The E statistical method is similar to the LOO cross-validation method. Each molecular parameter was omitted and 10 molecular parameters were evaluated from within the 11 molecular parameters. Therefore, the effect each of the molecular parameters was determined for the HEPT series. The E, q^2, $r^2_{training}$, r^2_{test}, q^2_{ext1} and q^2_{ext2} values, which are displayed in Table 8, were calculated to determine which variables affect the performance of the generated model 4. The performance of the model affected $a^{(10)}$ much more than the other variables in terms of O3-C5-N2. Both the E value (0.542) and q^2 value (0.704) are the lowest for the case of negligence of $a^{(10)}$ variables. Moreover, considering the r^2_{test}, q^2_{ext1} and q^2_{ext2} values, we can see in Table 8 that they are also low. Therefore, $a^{(10)}$ is the most important variable for this model. The second variable most affecting the model is $a^{(8)}$, which represents the distance between C7-H5+ van der Waals radius of the H5 atom, and when this variable is neglected the values of E, q^2, r^2_{test}, q^2_{ext1} and q^2_{ext2} decrease. Considering the case of the negligence of $a^{(5)}$, the E and q^2 values do not decrease much, but a significant decline is observed in the r^2_{test}, q^2_{ext1} and q^2_{ext2} values.

Table 8. E, q^2, $r^2_{training}$, r^2_{test}, q^2_{ext1} and q^2_{ext2} values showing the contribution of each descriptor to model performance for anti-HIV-1 activity of HEPT derivatives. q^2_{ext1} and q^2_{ext2} are external validations in the leave-one-out cross-validation for 11 parameters.

Index	E	$r^2_{training}$	q^2	r^2_{test}	q^2_{ext1}	q^2_{ext2}
$a^{(1)}$	0.663	0.793	0.883	0.715	0.882	0.880
$a^{(2)}$	0.948	0.856	0.828	0.801	0.786	0.781
$a^{(3)}$	0.940	0.854	0.890	0.799	0.879	0.877
$a^{(4)}$	0.910	0.868	0.868	0.793	0.859	0.855
$a^{(5)}$	0.915	0.745	0.794	0.738	0.635	0.628
$a^{(6)}$	0.978	0.862	0.913	0.807	0.889	0.887
$a^{(7)}$	0.995	0.860	0.895	0.810	0.869	0.866
$a^{(8)}$	0.638	0.778	0.609	0.704	0.436	0.424
$a^{(9)}$	0.981	0.862	0.896	0.808	0.873	0.866
$a^{(10)}$	0.542	0.770	0.665	0.652	0.570	0.560
$a^{(11)}$	0.996	0.864	0.895	0.810	0.863	0.860

Individually, the $a^{(2)}$, $a^{(3)}$, $a^{(6)}$, $a^{(7)}$ and $a^{(11)}$ parameters have only a small effect on the model. Using the developed EC–GA model we found that 11 parameters, which include electronic and geometric characteristics, are the most important for affecting the activity of HEPT derivatives.

The EC–GA method, which was successfully used in a 4D-QSAR study employing four independent models, improve model selection and predictivity. The results from our study clearly show that electronic and, in particular, geometric parameters are of prime importance for determining the anti-HIV-1 activity of the HEPT derivatives under study. The QSAR models were validated both internally and externally. External validation should be seen as a useful supplement to internal validation, rather than as a superior alternative. Whenever additional data sets with compounds of unknown activity are available, it is preferable that QSAR models be externally validated.

At last, eight novel compounds (see Table 1, compounds 108–115) never tested experimentally have been designed theoretically to predict the anti-HIV-1 activity of compounds before their synthesis. These new derivatives were used as external test set and their predictive pIC50 values checked based on our model 4. Eleven parameters and κ_j values obtained from model 4 entered the final activity formula (Equation (2)) in order to predict the anti-HIV-1 activity of eight novel compounds. Calculated activity values are shown in Table 1. Because of the absence of experimental values for new compounds, statistical evaluations have not been performed for these compounds. However, according to the EC–GA method, these compounds are expected to demonstrate statistically significant anti-HIV-1 activity under experimental conditions.

4. Conclusion

This study provided statistical interpretations of the activity predictions of HEPT derivatives investigated to reveal pharmacophore and to predict anti-HIV-1 activity using a 4D-QSAR method called EC–GA that combines the electron conformational and genetic algorithm methods. The goal of the EC–GA method was not only to explain the relationships between molecular descriptors and anti-HIV-1 activity, but also to describe the pharmacophore group using the conformational flexibility of the HEPT compounds. However, a conformational ensemble of compounds according to Boltzmann weighting was also used instead of a single representative structure to predict anti-HIV-1 activity. Four independent models were constructed using four different training and test sets to evaluate both the robustness and the predictive ability of the models and to compare the results obtained from this study with previous works. Internal and external validations were used to explain the goodness-of-fit and robustness, and the predictivity of a model. Finally, the results of model 2 which had a negative q^2_{ext2} value and high q^2 value, emphasized that external validation is essential to interprete the predictive power of QSAR models. Based on both internally and externally validated results, we concluded that the best model for the prediction of the anti-HIV-1 activity of HEPT derivatives was model 4. The investigated activity of the HEPT derivatives proved to be of electrostatic, geometric and topological nature according to the model 4 results. It depended on compound charges, van der Waals radius of atoms, and distance between of two atoms in the model 4. The EC–GA method provided reliable and valid model in terms of statistical characterization and LOO analyses.

Acknowledgements

This project was financially supported by the Scientific Technical Research Council of Turkey (TUBITAK, Grant No. 105T396) and the Research Fund of Erciyes University (BAP, Project Number: FBA-06-04). The authors would you like to thank Mustafa Yıldırım and Fatih Çopur for their valuable suggestions.

References

[1] R.A. Koup, V.J. Merluzzi, K.D. Hargrave, J. Adems, and K.J. Grozinger, *Inhibition of human immunodeficiency virus type 1 replication by the dipyridocliazepinone*, Infect. Dıs. 163 (1991), pp. 966–970.

[2] H. Tanaka, M. Baba, H. Hayakawa, T. Sakamaki, T. Miyasaka, M. Ubasawa, H. Takashima, K. Sekiya, I. Nitta, S. Shigeta, R. T. Walker, J. Balzarini, and E. De Clercq, *A new class of HIV-1 specific 6-substituted acyclouridine derivatives: Synthesis and anti-HIV activity of 5- or 6-substituted analogues of 1-[(2-hydroxyethoxy)-methyl]-6-(phenylthio)thymine (HEPT)*, J. Med. Chem. 34 (1991), pp. 349–357.

[3] H. Tanaka, M. Baba, M. Ubasawa, H. Takashima, K. Sekiya, I. Nitta, S. Shigeta, R.T. Walker, and T. Miyasaka, *Synthesis and anti-HIV activity of 2-, 3-, and 4-substituted analogues of 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT)*, J. Med. Chem. 34 (1991), pp. 1394–1399.

[4] H. Tanaka, H. Takashima, M. Ubasawa, K. Sekiya, I. Nitta, M. Baba, S. Shigeta, T.R. Walker, E. De Clercq, and T. Miyasaka, *Structure-activity relationships of 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)-thymine analogues: Effect of substitutions at the C-6 phenyl ring and the C-5 position on anti-HIV-1 activity*, J. Med. Chem. 35 (1992), pp. 337–345.

[5] H. Tanaka, H. Takashima, M. Ubasawa, K. Sekiya, I. Nitta, M. Baba, S. Shigeta, T.R. Walker, E. DeClercq, and T. Miyasaka, *Synthesis and antiviral activity of deoxy analogues of 1-[(2-hydroxyethoxy)-methyl]-6-(phenylthio) thymine (HEPT) as potent and selective anti-HIV agents*, J. Med. Chem. 35 (1992), pp. 4713–4719.

[6] D. Warnke, J. Barreto, and Z. Temesgen, *Antiretroviral drugs*, J. Clin. Pharmacol. 47 (2007), pp. 1570–1579.

[7] E. De Clercq, *Mini-review: The role of non-nucleoside reverse transcriptase inhibitors (NNRTIs) in the therapy of HIV-1 infection*, Antiviral Res. 38 (1998), pp. 153–179.

[8] E. De Clercq, *Non-nucleoside reverse transcriptase ınhibitors (NNRTIs) for the treatment of human immunodeficiency virus type 1 (HIV-1) infections: Strategies to overcome drug resistance development*, Med. Res. Rev. 16 (1996), pp. 125–157.

[9] A. Tropsha and A. Golbraikh, *Predictive QSAR modeling workflow, model applicability domains, and virtual screening*, Curr. Pharm. Des. 13 (2007), pp. 3494–3504.

[10] A. Tropsha, *Best practices for QSAR model development, validation, and exploitation*, Mol. Inf. 29 (2010), pp. 476–488.

[11] A. Tropsha, *Burger's Medicinal Chemistry, Drug Discovery and Development*, Wiley, 2010, pp. 505.

[12] J. Polanski, *Receptor dependent multidimensional QSAR for modeling drug- receptor interactions*, Curr. Med Chem. 16 (2009), pp. 3243–3257.

[13] A.J. Hopfinger, S. Wang, J.S. Tokarski, B. Jin, M. Albuquerque, P.J. Madhav, and C. Duraiswami, *Construction of 3D-QSAR models using the 4D-QSAR analysis formalism*, J. Am. Chem. Soc. 119 (1997), pp. 10509–10524.

[14] M.A. Lill and A. Vedani, *Combining 4D pharmacophore generation and multidimensional QSAR: Modeling ligand binding to the bradykinin B2 receptor*, J. Chem. Inf. Model. 46 (2006), pp. 2135–2145.

[15] T. Pavlov, M. Todorov, G. Stoyanova, P. Schmieder, H. Aladjov, R. Serafimova, and O. Mekenyan, *Conformational coverage by a genetic algorithm: Saturation of conformational space*, J. Chem. Inf. Model. 47 (2007), pp. 851–863.

[16] C. Duda-Seiman, D. Duda-Seiman, M.V. Putz, and D. Ciubotariub, *QSAR modelling of anti-HIV activity with HEPT derivatives*, Digest J. Nano. Biostruct. 2 (2007), pp. 207–219.

[17] L. Douali, D. Villemin, and D. Cherqaoui, *Neural networks: Accurate nonlinear QSAR model for HEPT derivatives*, J. Chem. Inf. Comput. Sci. 43 (2003), pp. 1200–1207.

[18] M. Jalali-Heravi and F. Parastar, *Use of artificial neural networks in a QSAR study of anti-HIV activity for a large group of HEPT derivatives*, J. Chem. Inf. Comput. Sci. 40 (2000), pp. 147–154.

[19] M. Arakawa, K. Hasegawa, and K. Funatsu, *QSAR study of anti-HIV HEPT analogues based on multi-objective genetic programming and counter-propagation neural network*, Chemom. Intel. Lab. Syst. 83 (2006), pp. 91–98.

[20] D. Dana Weekes and G.B. Fogel, *Evolutionary optimization, backpropagation, and data preparation issues in QSAR modeling of HIV inhibition by HEPT derivatives*, BioSystems 72 (2003), pp. 149–158.

[21] C.N. Alves, J.C. Pinheiroa, A.J. Camargob, M.M.C. Ferreirac, and A.B.F. Da Silva, *A structure–activity relationship study of HEPT-analog compounds with anti-HIV activity*, J. Mol. Struct: Theochem 530 (2000), pp. 39–47.

[22] S. Hannongbua, K. Nivesanond, L. Lawtrakul, P. Pungpo, and P. Wolschann, *3D-quantitative structure-activity relationships of HEPT derivatives as HIV-1 reverse transcriptase inhibitors, based on ab-initio calculations*, J. Chem. İnf. Comput. Sci. 41 (2001), pp. 848–855.

[23] W. Guo, X. Hu, N. Chu, and C. Yin, *Quantitative structure-activity relationship studies on HEPTs by supervised stochastic resonance*, Bioorg. Med. Chem. Lett. 16 (2006), pp. 2855–2859.

[24] D.B. Kireev, J.R. Chrétien, D.S. Grierson, and C. Monneret, *A 3D QSAR study of a series of HEPT analogues: The influence of conformational mobility on HIV-1 reverse transcriptase inhibition*, J. Med. Chem. 40 (1997), pp. 4257–4264.

[25] J.M. Luco and 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 (1997), pp. 392–401.

[26] A. Bak and J. Polanski, *4D-QSAR study on anti-HIV HEPT analogues*, Bioorg. Med. Chem. 14 (2006), pp. 273–279.

[27] H. Kubinyi, F.A. Hamprecht, and T. Mietzner, *Three-dimensional quantitative similarity-activity relationships (3D QSAR) from SEAL similarity matrices*, J. Med. Chem. 41 (1998), pp. 2553–2564.

[28] J.H. Van Drie, *Pharmacophore discovery: A critical review, in Computational Medicinal Chemistry for Drug Discovery*, Curr. Pharm. Des. 9 (2003), pp. 1649–1664.

[29] I.B. Bersuker, *Pharmacophore identification and quantitative bioactivity prediction using the electron-conformational method*, Curr. Pharm. Des. 9 (2003), pp. 1575–1606.

[30] I.B. Bersuker, S. Bahceci, J.E. Boggs, and R.S. Pearlman, *A novel electron-conformational approach to molecular modeling for QSAR by identification of pharmacophore and anti-pharmacophore shielding*, SAR QSAR Environ. Res. 10 (1998), pp. 157–173.

[31] I.B. Bersuker, S. Bahceci, J.E. Boggs, and R.S. Pearlman, *An electron- conformational method of identification of pharmacophore and anti- pharmacophore shielding: Application to rice blast activity*, J. Comp. Aided. Mol. Des. 13 (1999), pp. 419–434.

[32] I.B. Bersuker, S. Bahceci, and J.E. Boggs, *Improved electron-conformational method of pharmacophore identification and bioactivity prediction. application to angiotensin converting enzyme inhibitors*, J. Chem. Inf. Comput. Sci. 40 (2000), pp. 1363–1376.

[33] E. Rosines, I.B. Bersuker, and J.E. Boggs, *Pharmacophore identification and bioactivity prediction for group i metabotropic glutamate receptor agonists by the Electron-Conformational QSAR method*, Quant. Struct. Act. Relat. 20 (2001), pp. 327–334.

[34] Al H. Makkouk, I.B. Bersuker, and J.E. Boggs, *Quantitative drug activity prediction for inhibitors of human breast carcinoma*, Int. J. Pharm. Med. 18 (2004), pp. 81–89.

[35] A.V. Marenich, P.H. Yong, I.B. Bersuker, and J.E. Boggs, *Quantitative antidiabetic activity prediction for the class of guanidino- and aminoguanidinopropionic acid analogs based on electron-conformational studies*, J. Chem. Inf. Model. 48 (2008), pp. 556–568.

[36] I.B. Bersuker, *QSAR without arbitrary descriptors: The electron-conformational Method*, J. Comp. Aid. Mol. Des. 22 (2008), pp. 423–430.

[37] K. Hasegawa, Y. Miyashita, and K. Funatsu, *Strategy for variable selection in qsar studies: ga-based pls analysis of calcium channel antagonists*, J. Chem. Inf. Comput. Sci. 37 (1997), pp. 306–310.

[38] M. Fernandez, J. Caballero, L. Fernandez, and A. Sarai, *Genetic algorithm optimization in drug design QSAR: Bayesian-regularized genetic neural networks (BRGNN) and genetic algorithm-optimized support vectors machines (GA-SVM)*, Mol. Divers. (2010) DOI 10.1007/s11030-010-9234-9.

[39] S.J. Cho and M.A. Hermsmeier, *Genetic algorithm guided selection: Variable selection and subset selection*, J. Chem. Inf. Comput. Sci. 42 (2002), pp. 927–936.

[40] E. Sarıpınar, N. Geçen, K. Şahin, and E. Yanmaz, *Pharmacophore identification and bioactivity prediction for triaminotriazine derivatives by electron conformational-genetic algorithm QSAR method*, Eur. J. Med. Chem. 45 (2010), pp. 4157–4168.

[41] E. Yanmaz, E. Sarıpınar, K. Şahin, N. Geçen, and F. Çopur, *4D-QSAR analysis and pharmacophore modeling: Electron conformational-genetic algorithm approach for penicillins*, Bioorg. Med. Chem. 19 (2011), pp. 2199–2210.

[42] K. Şahin, E. Sarıpınar, E. Yanmaz, and N. Geçen, *Quantitative bioactivity prediction and pharmacophore identification for benzotriazines derivatives by electron conformational-genetic algorithm QSAR method*, SAR QSAR Environ. Res. 22 (2011), pp. 217–238.

[43] N. Geçen, E. Saripinar, E. Yanmaz, and K. Sahin, *Application of electron conformational–genetic algorithm approach to 1,4-dihydropyridines as calcium channel antagonists: Pharmacophore identification and bioactivity prediction*, J. Mol. Model. DOI 10.1007/s00894-011-1024-5.

[44] L. Akyüz, *Determination of the active groups of Anti-HIV effective HEPT, TIBO, thiazolidine and DABO derivatives using electron conformational-genetic algorithm method as a new 4D-QSAR method*, Ph.D. Diss., Erciyes University, 2011.

[45] Spartan 08 for Windows, Macintosh and Linux; Tutorial and User's Guide; Wavefunction, Inc. 2006.

[46] E. Saripinar, Y. Guzel, S. Patat, I. Yildirim, Y. Akcamur, and A.S. Dimoglo, *Electron-topological investigation of the structure-antitubercular activity relationship of thiosemicarbazone derivatives*, Arzneımıttel-Forschung/Drug Research 46 (1996), pp. 824–828.

[47] A.S. Dimoglo, N.M. Shvets, I.V. Tetko, and D.J. Livingstone, *Electronic-topological investigation of thestructure – acetylcholinesterase inhibitor activity relationship in the series of n-benzylpiperidine derivatives*, Quant. Struct.-Act. Relat. 20 (2001), pp. 31–45.

[48] Y. Güzel, E. Sarıpınar, and I. Yıldırım, *Electron-topological (ET) investigation of structure-antagonist activity of a series of dibenzo[a,d]cycloalkenimines*, J. Mol. Struc: Theochem. 418 (1997), pp. 83–91.

[49] A.S. Dimoglo, P.F. Vlad, N.M. Shvets, M.N. Coltsa, Y. Güzel, M. Saracoglu, E. Sarıpınar, and S. Patat, *Electronic-topological investigations of the relationships between chemical structure and ambergris odor*, New J. Chem. 19 (1995), pp. 1217–1226.

[50] J.H. Holland, *Adaptation in artificial and natural systems,* University of Michigan Press, Ann Arbor, Michigan, 1975.

[51] MATLAB (ver 7.0), The MathWorks Inc, 3 Apple Hill Drive, Natick, MA 01760-2098.

[52] P. Gramatica, *Principles of QSAR models validation: Internal and external*, QSAR Comb. Sci. 26 (2007), pp. 694–701.

[53] G. Schüürmann, R.U. Ebert, J. Chen, B. Wang, and R. Kuhne, *External validation and prediction employing the predictive squared correlation coefficients test set activity mean vs training set activity mean*, J. Chem. Inf. Model. 48 (2008), pp. 2140–2145.

[54] S.V. Damme and P. Bultınck, *Software news and update a new computer program for QSAR-analysis: ARTE-QSAR*, J. Comput. Chem. 28 (2007), pp. 1924–1928.

[55] V. Consonni, D. Ballabio, and R. Todeschini, *Comments on the definition of the Q2 parameter for QSAR validation*, J. Chem. Inf. Model. 49 (2009), pp. 1669–1678.

[56] A. Golbraikh, M. Shen, Z. Xiao, Y.D. Xiao, K.H. Lee, and A. Tropsha, *Rational selection of training and test sets for the development of validated QSAR models*, J. Comput.-Aided Mol. Des. 17 (2003), pp. 241–253.

[57] S.O. Podunavac-Kuzmanović, D.D. Cvetković, and D. Barna, *QSAR analysis of 2-amino or 2-methyl-1-substituted benzimidazoles against Pseudomonas aeruginosa*, J. Int. Mol. Sci. 10 (2009), pp. 1670–1682.

[58] Organisation for Economic Co-operation and Development. *Guidance document on the validation of (quantitative) structure-activity relationship [(Q)SAR] models. OECD Series on Testing and Assessment 69*. OECD Document ENV/JM/MONO(2007)2, 2007.

[59] A. Golbraikh and J. Tropsha, *Beware of q2!*, J. Mol. Graph. Mod. 20 (2002), pp. 269–276.

[60] M. Dervarics, F. Ötvös, and T.A. Martinek, *Development of a chiralitysensitive flexibility descriptor for 3 + 3D-QSAR*, J. Chem. Inf. Model. 46 (2006), pp. 1431–1438.

[61] R.A. Klein, *Modified van der Waals atomic radii for hydrogen bonding based on electron density topology*, Chem. Phys. Lett. 425 (2006), pp. 128–133.

[62] D.M. Patel and N.M. Patel, *QSAR analysis of aminoquinoline analogues as MCH1 receptor antagonist*, J. Sci. Res. 1 (2009), pp. 594–605.

[63] D. Livingstone, *Data Analysis for Chemists*, Oxford University Press, Oxford, 1995.

[64] S. Wold, *Cross-validatory estimation of the number of components in factor and principal components models*, Technometrics 20 (1978), pp. 397–405.