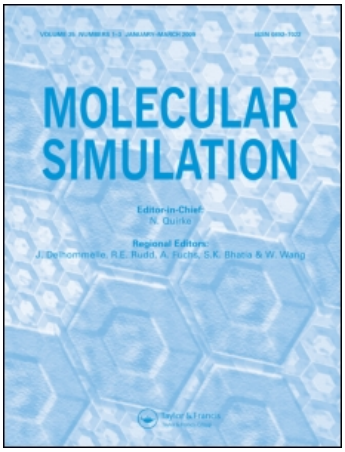

Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

2D autocorrelation modelling of the anti-HIV HEPT analogues using multiple linear regression approaches

Amir Najafi[a]; Soheil Sobhan Ardakani[b]
[a] Islamic Azad University, Young Researchers Club, Hamedan, Iran [b] Department of Environment, Islamic Azad University, Hamedan, Iran

Online publication date: 28 January 2011

To cite this Article Najafi, Amir and Sobhan Ardakani, Soheil(2011) '2D autocorrelation modelling of the anti-HIV HEPT analogues using multiple linear regression approaches', Molecular Simulation, 37: 1, 72 — 83
To link to this Article: DOI: 10.1080/08927022.2010.520134
URL: http://dx.doi.org/10.1080/08927022.2010.520134

Molecular Simulation
Vol. 37, No. 1, January 2011, 72–83

Taylor & Francis
Taylor & Francis Group

2D autocorrelation modelling of the anti-HIV HEPT analogues using multiple linear regression approaches

Amir Najafi[a]* and Soheil Sobhan Ardakani[b]

[a]*Islamic Azad University, Young Researchers Club, Hamedan Branch, Hamedan, Iran;* [b]*Department of Environment, Islamic Azad University, Hamedan Branch, Hamedan, Iran*

(*Received 22 June 2010; final version received 24 August 2010*)

A quantitative structure–anti HIV-1 activity relationship study has been applied in a series of 1-[2-hydroxyethoxy-methyl]-6-(phenylthio)thymine] analogues acting as non-nucleoside reverse transcriptase inhibitors. The relevant 2D autocorrelation descriptors for deriving a quantitative relation between the anti-HIV activity and structural properties were selected by the multiple linear regression approach. Analysis of the resulting model revealed a correlation coefficient and a root mean square error of 0.859 and 0.503, respectively. The predictive ability of the model indicates that this model can be used for virtual library screening of databases for novel potent anti-HIV agents.

Keywords: HEPT analogues; anti-HIV activity; 2D autocorrelation descriptors; QSAR

1. Introduction

The human immunodeficiency virus (HIV) is a lentivirus that quickly compromises the immune system of its host [1]. Since the first recognition of acquired immunodeficiency syndrome (AIDS) in the early 1980s, much progress has been made in understanding the pathogenesis, treatment and opportunistic infections associated with HIV infection [2]. Twenty-five years later, HIV infection has become a worldwide epidemic; more than 25 million people have died and more than 40 million people are infected [3]. AIDS is caused by the depletion of helper T-lymphocytes through infection by the human immunodeficiency virus type 1 (HIV-1) and human immunodeficiency virus type 2 (HIV-2). HIV-1 and HIV-2 are retroviruses, requiring a reverse transcriptase (RT) to convert viral RNA into proviral DNA that can be inserted into the host [4]. Disorders of the central nervous system (CNS) in patients with HIV-1 infection are often associated with severe morbidity and mortality. Seizures are relatively common manifestations of HIV infection itself and of its several complications involving the brain (see [5] and references therein). HIV-1 protease inhibitors have thus become a major target for anti-AIDS drug design [6]. HIV-1 belongs to the Lentiviridae subfamily of retroviruses; it invades the CNS shortly after primary infection. The virus does not infect neurons, but infects the microglia, macrophages and astrocytes within the CNS. HIV infection originates from the migration of infected microcytes into the CNS via the blood–brain barrier [2]. Combination therapy with antiretroviral agents has proven an effective strategy for delaying disease progression and prolonging life.[1] Because of this important role in HIV production, a large number of compounds have been developed to target various sites on RT [7–11] and the crystal structure of HIV-1 RT has been determined [12].

Among the developing anti-HIV agents, 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thyamine (HEPT) analogues are a very well-known class. HEPT is a non-nucleoside RT inhibitor (NNRTI) with potent anti-HIV-1 activity at nanomolar concentrations [13]. The synthesis and various studies of HEPT analogues have been performed by Tanaka et al. [14–17].

Over the past several decades, the quantitative structure–activity/property relationships (QSAR/QSPR) have become an alternative powerful theory for the description and prediction of properties of complex molecular systems in different environments. The QSAR/QSPR approach proceeds from the assumption of the one-to-one correspondence between any physical affinity, chemical affinity or biological activity of a chemical compound and its molecular structure. Linear or nonlinear statistic methods such as multiple linear regression (MLR), principal component regression, partial least squares (PLS), different types of artificial neural networks (ANN), genetic algorithms (GA), support vector machines, etc. can be selected in the development of a mathematical relationship between the structural descriptors and biological effects. In most cases, it is more convenient that a linear relationship between activity/property and structural descriptors is considered.

Although there are several targets and different antiretroviral therapies, drug resistance emerges quickly because of mutation at the transcription phase. This necessitates

*Corresponding author. Email: najafi@iauh.ac.ir; am.najafi@yahoo.com

ISSN 0892-7022 print/ISSN 1029-0435 online

DOI: 10.1080/08927022.2010.520134
http://www.informaworld.com

QSAR studies for developing good predictive models for ligands acting on different anti-HIV targets. Barreca et al. [18] have designed, synthesised and performed structure–activity relationships and molecular modelling studies on 2,3-diaryl-1,3-thiazolidin-4-ones as RT inhibitors. These authors have performed computational studies to delineate the ligand–RT interactions and to probe the binding of the ligands to HIV-1 RT. Rawal et al. [19] have studied a series of 4-thiazolidines as selective inhibitors of the HIV-RT enzyme and correlated the inhibitory activity with physico-chemical properties using statistically significant QSAR models with good predictive ability. A structure-based design of non-nucleoside RT has been proposed by Mao et al. [20] to explore the lowering of binding affinity with changes in binding pocket shape, volume and chemical properties of residue mutations. Tintori et al. [21] combined an electron–ion interaction potential technique with molecular modelling approaches for the identification of new HIV-1 integrase inhibitors. Bhhataraj and Garg [22] investigated the effect of hydrophobicity on the design of 4-hydroxy-5,6-dihydro-pyran-2-ones as a new class of emerging HIV-1 protease inhibitors. Kellenberger et al. [23] have screened about 1.6 million commercially available compounds against the CCR5 model by sequential filters (drug-likeness, 2D pharmacophore, 3D docking and scaffold clustering) and 10 compounds were detected of having binding affinity to CCR5. Mandal and Roy [24] have performed QSAR modelling of anti-HIV compounds of different chemical series acting on different targets. CoMFA and CoMSIA and docking studies have been performed by Buolamwini and Assefa [25] on conformationally restrained cinnamoyl HIV-1 integrase inhibitors to explore a binding mode at the active site.

In the present work, the 2D autocorrelation pool was used for encoding structural information of HEPT analogues and development of the linear model for the prediction of anti-HIV activity of these compounds. The general structure of HEPT analogues is shown in Figure 1. This study may help designing new analogues with a better biological profile.

2. Materials and method

2.1 Software

A Pentium IV personal computer (CPU at 2.6 MB) under the Windows XP operating system was used. Molecular modelling and geometry optimisation were employed by HyperChem (version 7.0, Hypercube Inc., http://www.hyper.com). Dragon software [26] was employed for the calculation of theoretical molecular descriptors. SPSS software (version 13.0, SPSS Inc., http://www.spss.com) was used for MLR analysis. Other statistics calculations were also performed in the MATLAB (version 7.0, MathWorks Inc., http://www.mathworks.com) environment.

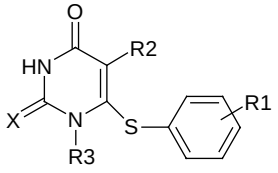

Figure 1. General structure of HEPT analogues.

2.2 2D autocorrelation pool

Three spatial autocorrelation vectors were employed for modelling:

Broto–Moreau's autocorrelation coefficients [27]

$$\mathrm{ATS}_w l = \sum_{i=1}^{A}\sum_{j=1}^{A}\delta_{ij} w_i w_j; \tag{1}$$

Moran's indices [28]

$$\mathrm{MATS}_w l = \frac{N}{2L}\frac{\sum_{ij}\delta_{ij}(w_i - \bar{w})(w_j - \bar{w})}{\sum_i (w_i - \bar{w})^2}; \tag{2}$$

Geary's coefficient [29]

$$\mathrm{GATS}_w l = \frac{(N-1)}{4L}\frac{\sum_{ij}(w_i - \bar{w})(w_j - \bar{w})}{\sum_i (w_i - \bar{w})}, \tag{3}$$

where $\mathrm{ATS}_w l$, $\mathrm{MATS}_w l$ and $\mathrm{GATS}_w l$ are Broto–Moreau's autocorrelation coefficient, Moran's index and Geary's coefficient at spatial lag l, respectively; w_i and w_j are the values of any atomic property of atoms i and j, respectively; $\bar{w}$ is the average value of the property; L is the number of non-zero values in the sum; N is the number of atoms in the molecule and $\delta(l, d_{ij})$ is a Dirac delta function defined as

$$\delta(l, d_{ij}) = \left\{ \begin{array}{lll} 1, & \text{if} & d_{ij} = l \\ 0, & \text{if} & d_{ij} \neq l \end{array} \right\}, \tag{4}$$

where d_{ij} is the topological distance or spatial lag between atoms i and j.

In general the 2D autocorrelation descriptors explain how the considered property is distributed along the topological structure. The most important factor in interpreting them in the model is the topological distance once weighted equally. The computation of these descriptors involves the summations of different autocorrelation functions corresponding to the different fragment lengths and leads to different autocorrelation vectors corresponding to the lengths of the structural fragments [30]. Also, a weighting component in terms of a physico-chemical property has been embedded in this descriptor. As a result, these descriptors address the topology of the structure or

Table 1. Molecular structures of the HEPT analogues and their experimental and predicted values of the anti-HIV activity.

Analogues	R1	R2	R3	X	Experimental anti-HIV	Predicted MLR
1	2-Me	Me	$CH_2OCH_2CH_2OH$	O	4.15	3.92
2[a]	2-NO_2	Me	$CH_2OCH_2CH_2OH$	O	3.85	4.58
3	2-OMe	Me	$CH_2OCH_2CH_2OH$	O	4.72	5.32
4	3-Me	Me	$CH_2OCH_2CH_2OH$	O	5.59	5.52
5[a]	3-Et	Me	$CH_2OCH_2CH_2OH$	O	5.57	5.21
6	3-*t*-Bu	Me	$CH_2OCH_2CH_2OH$	O	4.92	4.85
7	3-CF_3	Me	$CH_2OCH_2CH_2OH$	O	4.35	4.95
8	3-F	Me	$CH_2OCH_2CH_2OH$	O	5.48	4.79
9	3-Cl	Me	$CH_2OCH_2CH_2OH$	O	4.89	5.20
10[a]	3-Br	Me	$CH_2OCH_2CH_2OH$	O	5.24	5.31
11	3-I	Me	$CH_2OCH_2CH_2OH$	O	5.00	5.31
12	3-NO_2	Me	$CH_2OCH_2CH_2OH$	O	4.47	4.86
13[a]	3-OH	Me	$CH_2OCH_2CH_2OH$	O	4.09	4.82
14	3-OMe	Me	$CH_2OCH_2CH_2OH$	O	4.66	5.18
15[a]	3,5-Me_2	Me	$CH_2OCH_2CH_2OH$	O	6.59	6.59
16	3,5-Cl_2	Me	$CH_2OCH_2CH_2OH$	O	5.89	5.43
17	3,5-Me_2	Me	$CH_2OCH_2CH_2OH$	S	6.66	6.54
18[a]	3-COOMe	Me	$CH_2OCH_2CH_2OH$	O	5.10	5.64
19	3-COMe	Me	$CH_2OCH_2CH_2OH$	O	5.14	5.86
20	3-CN	Me	$CH_2OCH_2CH_2OH$	O	5.00	5.26
21	H	$CH_2CH{=}CH_2$	$CH_2OCH_2CH_2OH$	O	5.60	5.39
22	H	Et	$CH_2OCH_2CH_2OH$	S	6.96	6.25
23	H	Pr	$CH_2OCH_2CH_2OH$	S	5.00	5.84
24	H	*i*-Pr	$CH_2OCH_2CH_2OH$	S	7.23	7.39
25	3,5-Me_2	Et	$CH_2OCH_2CH_2OH$	S	8.11	7.66
26	3,5-Me_2	*i*-Pr	$CH_2OCH_2CH_2OH$	S	8.30	8.56
27[a]	3,5-Cl_2	Et	$CH_2OCH_2CH_2OH$	S	7.37	7.03
28	H	Et	$CH_2OCH_2CH_2OH$	O	6.92	6.18
29[a]	H	Pr	$CH_2OCH_2CH_2OH$	O	5.47	5.66
30	H	*i*-Pr	$CH_2OCH_2CH_2OH$	O	7.20	7.21
31[a]	3,5-Me_2	Et	$CH_2OCH_2CH_2OH$	O	7.89	7.61
32	3,5-Me_2	*i*-Pr	$CH_2OCH_2CH_2OH$	O	8.57	8.41
33	3,5-Cl_2	Et	$CH_2OCH_2CH_2OH$	O	7.85	6.98
34	4-Me	Me	$CH_2OCH_2CH_2OH$	O	3.66	4.41
35[a]	H	Me	$CH_2OCH_2CH_2OH$	O	5.15	4.79
36	H	Me	$CH_2OCH_2CH_2OH$	S	6.01	4.72
37	H	I	$CH_2OCH_2CH_2OH$	O	5.44	4.85
38[a]	H	$CH{=}CH_2$	$CH_2OCH_2CH_2OH$	O	5.69	5.78
39	H	CH=CHPh	$CH_2OCH_2CH_2OH$	O	5.22	5.22
40	H	CH_2Ph	$CH_2OCH_2CH_2OH$	O	4.37	4.86
41	H	$CH{=}CPh_2$	$CH_2OCH_2CH_2OH$	O	6.07	5.78
42	H	Me	$CH_2OCH_2CH_2Me$	O	5.06	5.46
43[a]	H	Me	$CH_2OCH_2CH_2Ac$	O	5.17	5.04
44	H	Me	$CH_2OCH_2CH_2OCOPh$	O	5.12	5.65
45[a]	H	Me	CH_2OCH_2Me	O	6.48	5.49
46	H	Me	$CH_2OCH_2CH_2Cl$	O	5.82	5.68
47	H	Me	$CH_2OCH_2CH_2N_3$	O	5.24	5.61
48	H	Me	$CH_2OCH_2CH_2F$	O	5.96	5.03
49	H	Me	$CH_2OCH_2CH_2Me$	O	5.48	5.46
50	H	Me	CH_2OCH_2Ph	O	7.06	6.02
51	H	Et	CH_2OCH_2Me	O	7.72	6.91
52	H	Et	CH_2OCH_2Me	S	7.58	7.05
53	3,5-Me_2	Et	CH_2OCH_2Me	O	8.24	8.35
54	3,5-Me_2	Et	CH_2OCH_2Me	S	8.30	8.51
55	H	Et	CH_2OCH_2Ph	O	8.23	7.23
56	3,5-Me_2	Et	CH_2OCH_2Ph	O	8.55	8.47
57	H	Et	CH_2OCH_2Ph	S	8.09	7.67
58[a]	3,5-Me_2	Et	CH_2OCH_2Ph	S	8.14	8.87
59[a]	H	*i*-Pr	CH_2OCH_2Me	O	7.99	7.98
60	H	*i*-Pr	CH_2OCH_2Ph	O	8.51	8.19
61	H	*i*-Pr	CH_2OCH_2Me	S	7.89	8.27
62	H	*i*-Pr	CH_2OCH_2Ph	S	8.14	8.74

Table 1 – *continued*

Analogues	R1	R2	R3	X	Experimental anti-HIV	Predicted MLR
63[a]	H	Me	CH_2OMe	O	5.68	5.44
64	H	Me	CH_2OBu	O	5.33	5.73
65	H	Me	Et	O	5.66	6.61
66	H	Me	Bu	O	5.92	6.04
67[a]	3,5-Cl_2	Et	CH_2OCH_2Me	S	7.89	7.99
68	H	Et	CH_2O-*i*-Pr	S	6.66	7.01
69[a]	H	Et	CH_2O-*c*-Hex	S	5.79	6.47
70	H	Et	CH_2OCH_2-*c*-Hex	S	6.45	6.18
71[a]	H	Et	$CH_2OCH_2C_6H_4$(4-Me)	S	7.11	7.24
72	H	Et	$CH_2OCH_2C_6H_4$(4-Cl)	S	7.92	7.54
73	H	Et	$CH_2OCH_2CH_2Ph$	S	7.04	7.25
74[a]	3,5-Cl_2	Et	CH_2OCH_2Me	S	8.13	7.99
75	H	Et	CH_2O-*i*-Pr	O	6.47	6.83
76	H	Et	CH_2O-*c*-Hex	O	5.40	5.95
77	H	Et	CH_2OCH_2-*c*-Hex	O	6.35	5.95
78	H	Et	$CH_2OCH_2CH_2Ph$	O	7.02	6.97
79	H	*c*-Pr	CH_2OCH_2Me	S	7.02	7.37
80	H	*c*-Pr	CH_2OCH_2Me	O	7.00	7.13

[a] Test examples.

parts thereof in association with a selected physico-chemical property. The number of consecutively connected edges considered in its computation was called the autocorrelation vector of lag *n* (corresponding to the number of edges in the unit fragment). Autocorrelation vectors were calculated at spatial lags *l* ranging from 1 to 8. The physico-chemical property considered in the four different weighting schemes was: atomic mass (m), atomic van der Waals volume (v), atomic Sanderson electronegativity (e) and atomic polarisability (p). The autocorrelation descriptors are denoted by the scheme: type of descriptor–spatial lag–weighting property; for instance, GATS5p is the Geary autocorrelation of lag 5 weighted by atomic polarisabilities. The 2D autocorrelation descriptors have been successfully applied in recent QSAR studies [31–37].

2.3 Activity data and descriptor generation

A data-set containing 80 HEPT analogues was used in this study. These compounds were first synthesised by Tanaka et al. [14–17] and various QSAR studies were performed in previous works [38–44]. The chemical structures and inhibitory activities used in this study are shown in Figure 1 and Table 1. The biological evaluation of these compounds was made by using one numerical indicator for activity, pEC50, negative logarithm of molar concentration of a drug required to achieve 50% protection of MT-4 cells against the cytopathic effect of HIV-1.

HyperChem software was used to draw the chemical structure of the molecules. AM1 semi-empirical quantum chemical calculation was used to optimise the 3D geometry of the molecules. The geometry optimisation was preceded by the Polak–Rebiere algorithm until the root mean square gradient reached 0.01. Afterwards, the calculation of weighted Broto–Moreau, Moran and Geary 2D autocorrelation vectors was carried out at spatial lags ranging from 1 to 8. Four different weighting properties including atomic masses (m), atomic van der Waals volumes (v), atomic Sanderson electronegativities (e) and atomic polarisabilities (p) have been used. Since we calculated three types of autocorrelation descriptors (Moran's index, Geary's coefficient and Broto–Moreau's autocorrelation coefficient) weighted by four atomic properties at eight lags, a total of 96 ($3 \times 4 \times 8$) descriptors for each compound were computed.

Table 2. 2D autocorrelation descriptors used in this study.

No.	Symbol	Definition
1	MATS5e	Moran autocorrelation–lag 5/weighted by atomic Sanderson electronegativities
2	MATS5p	Moran autocorrelation–lag 5/weighted by atomic polarisabilities
3	ATS8v	Broto–Moreau autocorrelation of a topological structure–lag 8/weighted by atomic van der Waals volumes
4	MATS1v	Moran autocorrelation–lag 1/weighted by atomic van der Waals
5	ATS8e	Broto–Moreau autocorrelation of a topological structure–lag 8/weighted by atomic Sanderson electronegativities
6	MATS8e	Moran autocorrelation–lag 8/weighted by atomic Sanderson electronegativities

Table 3. Data of the selected descriptors used in this study for the HEPT analogues.

No.	MATS5e	MATS5p	ATS8v	MATS1v	ATS8e	MATS8e	No.	MATS5e	MATS5p	ATS8v	MATS1v	ATS8e	MATS8e
1	0.007	−0.057	0.249	−0.022	1.027	−0.242	41	−0.146	−0.034	0.312	−0.018	1.023	−0.161
2	−0.107	−0.058	0.274	−0.002	1.114	0.090	42	−0.044	−0.123	0.299	−0.057	1.012	−0.223
3	0.004	−0.163	0.300	−0.065	1.028	−0.127	43	0.040	−0.014	0.315	−0.094	1.026	−0.104
4	−0.030	−0.148	0.315	−0.022	1.018	−0.094	44	−0.056	−0.037	0.337	−0.050	1.026	−0.119
5	0.000	−0.070	0.342	−0.022	1.021	−0.123	45	−0.046	−0.045	0.274	−0.061	0.997	−0.104
6	0.037	0.038	0.386	−0.022	1.025	−0.170	46	−0.036	−0.125	0.309	−0.056	1.016	−0.118
7	−0.055	−0.161	0.331	−0.020	1.090	−0.124	47	−0.041	−0.150	0.299	−0.059	1.020	−0.149
8	−0.056	−0.059	0.288	−0.021	1.092	0.113	48	−0.029	−0.039	0.280	−0.061	1.027	−0.111
9	−0.054	−0.068	0.320	−0.015	1.066	0.045	49	−0.044	−0.123	0.299	−0.057	1.012	−0.223
10	−0.050	−0.063	0.338	−0.010	1.054	−0.005	50	−0.060	−0.125	0.386	−0.021	1.022	−0.205
11	−0.039	−0.042	0.355	−0.006	1.032	−0.108	51	−0.195	−0.052	0.290	−0.057	0.988	−0.081
12	−0.058	−0.107	0.323	−0.002	1.080	−0.060	52	−0.230	0.044	0.301	−0.052	0.979	−0.081
13	−0.052	−0.091	0.294	0.017	1.068	0.080	53	−0.195	−0.289	0.350	−0.052	0.988	0.015
14	−0.014	−0.015	0.324	−0.065	1.067	0.056	54	−0.235	−0.182	0.366	−0.048	0.981	0.008
15	−0.051	−0.326	0.340	−0.022	1.015	−0.046	55	−0.199	−0.124	0.383	−0.020	1.011	−0.171
16	−0.073	−0.077	0.357	−0.008	1.110	0.180	56	−0.201	−0.339	0.415	−0.018	1.007	−0.058
17	−0.053	−0.204	0.360	−0.018	1.004	−0.038	57	−0.234	−0.027	0.406	−0.019	0.994	−0.143
18	−0.032	−0.147	0.339	−0.060	1.051	−0.101	58	−0.240	−0.235	0.440	−0.017	0.993	−0.044
19	−0.024	−0.171	0.328	−0.062	1.030	−0.083	59	−0.309	−0.057	0.302	−0.054	0.981	−0.068
20	−0.041	−0.074	0.338	−0.015	1.037	−0.117	60	−0.310	−0.123	0.380	−0.019	1.002	−0.147
21	−0.088	−0.046	0.313	−0.019	1.019	−0.183	61	−0.366	0.029	0.311	−0.050	0.974	−0.071
22	−0.181	0.034	0.309	−0.017	0.998	−0.095	62	−0.368	−0.035	0.400	−0.018	0.988	−0.127
23	−0.127	0.019	0.316	−0.018	1.000	−0.139	63	−0.043	−0.200	0.238	−0.065	1.005	−0.148
24	−0.299	0.021	0.317	−0.018	0.990	−0.078	64	−0.083	−0.009	0.325	−0.054	1.010	−0.184
25	−0.181	−0.188	0.369	−0.018	0.996	−0.022	65	−0.167	−0.236	0.235	−0.017	0.979	−0.051
26	−0.284	−0.176	0.376	−0.018	0.990	−0.011	66	−0.113	−0.067	0.276	−0.016	0.977	−0.071
27	−0.211	0.009	0.392	−0.005	1.069	0.196	67	−0.242	0.021	0.389	−0.044	1.053	0.325
28	−0.167	−0.064	0.296	−0.022	1.009	−0.115	68	−0.222	0.131	0.324	−0.050	0.977	−0.054
29	−0.121	−0.069	0.300	−0.022	1.015	−0.186	69	−0.171	0.197	0.395	−0.018	0.990	−0.139
30	−0.270	−0.066	0.306	−0.022	1.000	−0.093	70	−0.128	0.112	0.381	−0.017	0.990	−0.176
31	−0.163	−0.296	0.352	−0.022	1.005	−0.028	71	−0.192	0.020	0.411	−0.018	0.995	−0.143
32	−0.253	−0.272	0.361	−0.022	0.998	−0.015	72	−0.217	−0.009	0.415	−0.018	1.004	−0.101
33	−0.200	−0.081	0.370	−0.010	1.082	0.229	73	−0.200	0.036	0.390	−0.018	0.988	−0.097
34	0.004	0.043	0.316	−0.022	1.025	−0.183	74	−0.242	0.021	0.389	−0.044	1.053	0.325
35	−0.033	−0.060	0.282	−0.022	1.023	−0.152	75	−0.186	0.051	0.315	−0.054	0.984	−0.058
36	−0.027	0.052	0.297	−0.015	1.009	−0.124	76	−0.139	0.130	0.373	−0.019	1.007	−0.178
37	−0.074	−0.111	0.341	−0.003	1.061	−0.300	77	−0.139	0.130	0.373	−0.019	1.007	−0.178
38	−0.146	−0.034	0.312	−0.018	1.023	−0.161	78	−0.168	−0.052	0.371	−0.019	0.999	−0.100
39	−0.059	0.047	0.401	0.012	1.034	−0.208	79	−0.287	0.055	0.322	−0.022	0.980	−0.087
40	−0.007	0.051	0.393	0.028	1.017	−0.191	80	−0.244	−0.031	0.311	−0.024	0.988	−0.085

Table 4. Correlation matrix of the six selected descriptors.

	MATS5e	MATS5p	ATS8v	MATS1v	ATS8e	MATS8e
MATS5e	1.000					
MATS5p	−0.072	1.000				
ATS8v	−0.329	0.072	1.000			
MATS1v	−0.080	0.092	0.339	1.000		
ATS8e	0.471	−0.061	−0.047	0.228	1.000	
MATS8e	−0.264	−0.099	0.137	0.016	0.417	1.000

Descriptors with constant or near-to-constant values were discarded. Table 2 presents the notation and a short description of the molecular descriptors used to generate the QSAR models.

3. Results and discussion

3.1 Variable selection

A data matrix was generated with the spatial autocorrelation vectors calculated for each compound. Afterwards, dimensionality reduction methods were employed for selecting the most relevant vector components for building the QSAR models. To decrease the redundancy existing in the descriptor data matrix, the correlation of descriptors with each other and with the activity (pEC50) of the molecules was examined, and collinear descriptors (i.e. $r > 0.95$) were detected. Among the collinear descriptors, one with the highest correlation with activity was retained, and the others were removed from the data matrix. The stepwise MLR procedure based on the forward selection and backward elimination methods was used for inclusion or rejection of descriptors in the screened models. Table 3 shows the data of the descriptors used in this study. The correlation matrix of the descriptors used in this study is given in Table 4. Inspection of these results shows that all the values deviate from unity noticeably, so there is no significant correlation between the six independent variables.

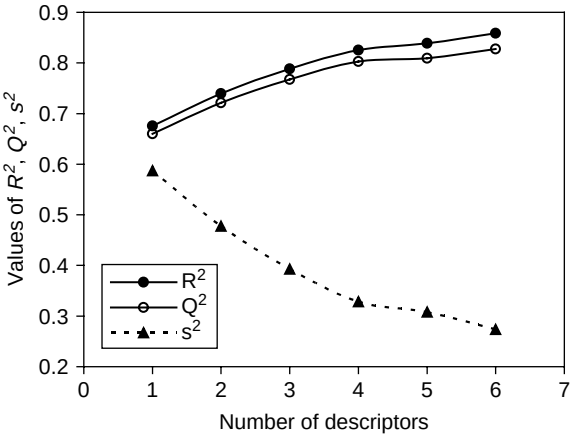

Figure 2. Influences of the number of descriptors on the correlation coefficient (R^2), the CV correlation coefficient (Q^2) and the square of standard error (s^2) of the regression model.

3.2 MLR analysis

By using the stepwise multiple regression method, the models were developed for 80 HEPT analogues. The correlations performed for the whole set provided the optimal equations for different numbers of descriptors in the range of 1–6. Figure 2 shows the plots of R^2, Q^2 and s^2 (squared standard deviation) as a function of the number of variables in the regression model. It suggests that the best one-descriptor model with the highest impact

Table 5. Comparison between some works on the same set of HEPT analogues.

Reference	Model	R^2	Number of descriptors	Descriptors used
[38]	MLR	0.900	9	Structural descriptors and physico-chemical variables
	PLS	0.889	9	
[39]	MLR	0.811	6	Topological, geometric, electronic and physico-chemical
	ANN	0.919	6	
[40]	MLR	0.830	4	Pertinent descriptors
	ANN	0.852	4	
[41]	ANN	0.977	7	Several descriptors used in previous works
[42]	CP-ANN	0.875	11	Structural descriptors
[43]	MLR	0.862	6	Quantum chemical, topological and several descriptors used in previous works
[44]	MLR	0.841	5	Physico-chemical constants, topological and structural and several descriptors used in previous works
This study	GA-MLR	0.859	6	2D autocorrelation descriptors

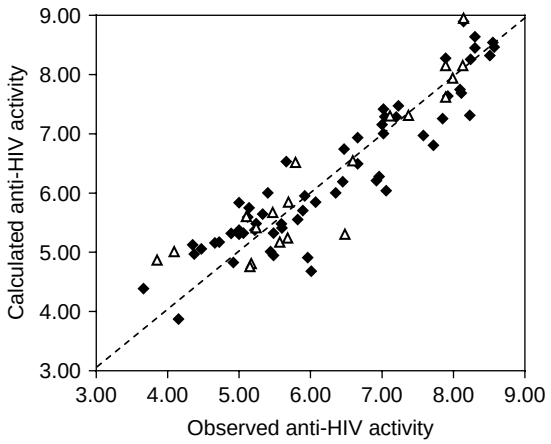

Figure 3. Plot of the anti-HIV activity predicted by MLR for training set (■) and test set (Δ) against the experimental values.

is MATS5e, which is defined as the Moran autocorrelation vector weighted by atomic Sanderson electronegativities, representing the topological substructure of sizes 5 in the HEPT molecule. Subsequent addition of variables produces monotonously increasing values of R^2 and Q^2 and decreasing values of s^2, and the break point is not clearly defined. We decided to select the best model to be the one having the smallest number of parameters and outliers. The model with six descriptors is as follows:

$$\begin{aligned} \text{pEC50} =& 14.163(\pm 3.177) - 7.502(\pm 0.883)\text{MATS5e} \\ & - 3.009(\pm 0.557)\text{MATS5p} \\ & + 9.103(\pm 1.495)\text{ATS8v} \\ & - 7.839(\pm 3.127)\text{MATS1v} \\ & - 11.921(\pm 2.924)\text{ATS8e} \\ & + 2.287(\pm 0.716)\text{MATS8e} \end{aligned}$$

$$N = 80,\ R^2 = 0.859,\ \text{SE} = 0.523,\ \text{RMS} = 0.503,$$
$$F = 73.829,\ Q^2_{\text{LOO}} = 0.828. \quad (5)$$

Some statistical parameters such as squared correlation coefficients (R^2), standard error of estimation (SE), RMS and Fisher statistic ratio (F) are given in Equation (5). As can be seen, the MLR model has good statistical quality with low prediction error. The quality of the QSAR-derived regression models given in Equation (5) is compared with those reported in the literature [38–44], which is shown in Table 5. The model given by Equation (5) is superior in prediction and count and type of descriptors used to those reported in the literature.

The robustness of the model and its predictive ability for the anti-HIV activity were evaluated by both leave-one-out cross-validation (LOO-CV) and external validation (EV) procedures.

Table 6. Anti-HIV activity predicted for the odd- and even-number samples.

	R^2_{RS}	RMS_{RS}	R^2_{HO}	RMS_{HO}
Odd samples	0.897	0.432	0.844	0.557
Even samples	0.856	0.510	0.801	0.638

Most of the QSAR modelling methods implement the leave-one-out (LOO) or leave-some-out (LSO) cross-validation (CV) procedure. The outcome from the CV procedure is cross-validated R^2 (Q^2_{LOO} or Q^2_{LSO}), which is used as a criterion of both robustness and predictive ability of the model. Cross-validated squared correlation coefficient Q^2_{LOO} is calculated according to the formula:

$$Q^2 = 1 - \frac{\sum(Y_{\text{obs}} - Y_{\text{cal}})^2}{\sum(Y_{\text{obs}} - \bar{Y})^2}. \quad (6)$$

In Equation (6), $\bar{Y}$ means average activity value of the entire data-set, while Y_{obs} and Y_{cal} represent observed and estimated activity values, respectively. Often, a high Q^2 value ($Q^2 > 0.5$) is considered as a proof of high predictive ability of the model [45]. Furthermore, the difference between model R^2 and LOO Q^2 should not exceed 0.3 [46,47]. The CV result ($Q^2 = 0.827$) shows that the model is surely stable and has good predictive ability.

In order to estimate the predictive power of the MLR model, an EV test was performed by splitting the data into two sub-samples, with one being used to fit (training set) and the other to test (test set) [48]. The models are generated based on training-set compounds, and predictive capacity of the models is judged based on the predictive R^2 values. Selection of the training-set compounds is significantly important in QSAR analysis. One most widely used method for dividing a data-set into training and test sets is mere random selection. Another frequently used method is based on the activity sampling in which the data could be ranked according to the magnitude of biological response, and the sorted data-set is divided into odd- and even-number groups. The QSPR models were fitted to the odd- and even-number samples separately, and the resulted fitness was assessed by applying the QSPR models to both samples. To compare the estimation abilities of the models, two statistical parameters, namely RMS and R^2, were calculated. The same data-set (i.e. 'training set') that was already used to fit the model was employed to determine resubstitution parameters, i.e. RMS_{RS} and R^2_{RS}, and to determine holdout parameters, i.e. RMS_{HO} and R^2_{HO}, for the other data-set which was not

Table 7. Statistical qualities of the model.

R^2	Q^2	R^2_{pred}	$r^2_{m(\text{test})}$	$r^2_{m(\text{LOO})}$	$r^2_{m(\text{overall})}$
0.855	0.811	0.848	0.802	0.789	0.791

Table 8. Results of the virtual screening.

Analogues	R1	R2	R3	X	Predicted MLR
1	3,5-$(Cl)_2$	Et	Ph	O	8.97
2	3,5-$(Cl)_2$	Et	Ph	S	9.11
3	2-Me	Me	Ph	O	5.14
4	2-Me	Me	Ph	S	5.21
5	3-Me	Me	Ph	O	6.94
6	3-Me	Me	Ph	S	6.96
7	4-Me	Me	Ph	O	5.60
8	4-Me	Me	Ph	S	5.51
9	3,5-$(Me)_2$	Me	Ph	O	8.04
10	3,5-$(Me)_2$	Me	Ph	S	8.12
11	3,5-$(Me)_2$	Et	Ph	O	9.10
12	3,5-$(Me)_2$	Et	Ph	S	9.52
13	3,5-$(Me)_2$	Me	CH_2Ph	O	8.04
14	3,5-$(Me)_2$	Me	CH_2Ph	S	8.12
15	3,5-$(Me)_2$	Et	CH_2Ph	O	9.10
16	3,5-$(Me)_2$	Et	CH_2Ph	S	9.52
17	3,5-$(Me)_2$	Pr	CH_2Ph	O	8.46
18	3,5-$(Me)_2$	Pr	CH_2Ph	S	8.87
19	3,5-$(Me)_2$	*i*-Pr	CH_2Ph	O	9.94
20	3,5-$(Me)_2$	*i*-Pr	CH_2Ph	S	10.68
21	3,5-$(Me)_2$	*i*-Pr	CH_2CH_2Ph	O	9.48
22	3,5-$(Me)_2$	*i*-Pr	CH_2CH_2Ph	S	10.31
23	3,5-$(Me)_2$	*i*-Pr	OCH_2Ph	O	8.79
24	3,5-$(Me)_2$	*i*-Pr	OCH_2Ph	S	9.02
25	3,5-$(Me)_2$	*i*-Pr	OCH_2Me	O	8.61
26	3,5-$(Me)_2$	*i*-Pr	OCH_2Me	S	8.56
27	3,5-$(Me)_2$	*i*-Pr	CH_2OMe	O	9.38
28	3,5-$(Me)_2$	*i*-Pr	CH_2OMe	S	9.66
29	3,5-$(Me)_2$	*i*-Pr	CH_2OEt	O	9.17
30	3,5-$(Me)_2$	*i*-Pr	CH_2OEt	S	9.47
31	3,5-$(Me)_2$	*i*-Pr	CH_2OPh	O	9.42
32	3,5-$(Me)_2$	*i*-Pr	CH_2OPh	S	9.97
33	3,5-$(Me)_2$	*i*-Pr	$CH_2OCH_2CH_2OH$	O	8.41
34	3,5-$(Me)_2$	*i*-Pr	$CH_2OCH_2CH_2OH$	S	8.56
35	3,5-$(Me)_2$	*i*-Pr	$CH_2OCH_2CH_2OMe$	O	8.88
36	3,5-$(Me)_2$	*i*-Pr	$CH_2OCH_2CH_2OMe$	S	9.07
37	3,5-$(Me)_2$	Et	CH_2OEt	O	8.35
38	3,5-$(Me)_2$	Et	CH_2OEt	S	8.51
39	4-Ph	Et	CH_2Ph	O	6.55
40	4-Ph	Et	CH_2Ph	S	6.92
41	4-Ph	Et	Ph	O	7.31
42	4-Ph	Et	Ph	S	7.70
43	2-NH_2	*i*-Pr	CH_2Ph	O	6.53
44	2-NH_2	*i*-Pr	CH_2Ph	S	7.00
45	3-NH_2	*i*-Pr	CH_2Ph	O	8.09
46	3-NH_2	*i*-Pr	CH_2Ph	S	8.67
47	4-NH_2	*i*-Pr	CH_2Ph	O	7.10
48	4-NH_2	*i*-Pr	CH_2Ph	S	7.60
49	2-OH	*i*-Pr	CH_2Ph	O	6.30
50	2-OH	*i*-Pr	CH_2Ph	S	6.49
51	3-OH	*i*-Pr	CH_2Ph	O	7.82
52	3-OH	*i*-Pr	CH_2Ph	S	8.13
53	4-OH	*i*-Pr	CH_2Ph	O	6.94
54	4-OH	*i*-Pr	CH_2Ph	S	7.20
55	2-F	*i*-Pr	CH_2Ph	O	6.38
56	2-F	*i*-Pr	CH_2Ph	S	6.42
57	3-F	*i*-Pr	CH_2Ph	O	7.68
58	3-F	*i*-Pr	CH_2Ph	S	7.78
59	4-F	*i*-Pr	CH_2Ph	O	7.03
60	4-F	*i*-Pr	CH_2Ph	S	7.16
61	2-Cl	*i*-Pr	CH_2Ph	O	7.33
62	2-Cl	*i*-Pr	CH_2Ph	S	7.61

Table 8 – *continued*

Analogues	R1	R2	R3	X	Predicted MLR
63	3-Cl	*i*-Pr	CH_2Ph	O	8.48
64	3-Cl	*i*-Pr	CH_2Ph	S	8.97
65	4-Cl	*i*-Pr	CH_2Ph	O	7.58
66	4-Cl	*i*-Pr	CH_2Ph	S	8.01
67	2-Br	*i*-Pr	CH_2Ph	O	7.81
68	2-Br	*i*-Pr	CH_2Ph	S	8.30
69	3-Br	*i*-Pr	CH_2Ph	O	8.77
70	3-Br	*i*-Pr	CH_2Ph	S	9.48
71	4-Br	*i*-Pr	CH_2Ph	O	7.85
72	4-Br	*i*-Pr	CH_2Ph	S	8.47
73	3,5-$(NH_2)_2$	*i*-Pr	CH_2Ph	O	8.63
74	3,5-$(NH_2)_2$	*i*-Pr	CH_2Ph	S	9.19
75	3,5-$(OH)_2$	*i*-Pr	CH_2Ph	O	7.84
76	3,5-$(OH)_2$	*i*-Pr	CH_2Ph	S	7.95
77	3,5-$(Br)_2$	*i*-Pr	CH_2Ph	O	9.48
78	3,5-$(Br)_2$	*i*-Pr	CH_2Ph	S	10.17
79	3,5-$(Cl)_2$	*i*-Pr	CH_2Ph	O	8.85
80	3,5-$(Cl)_2$	*i*-Pr	CH_2Ph	S	9.16
81	3,5-$(F)_2$	*i*-Pr	CH_2Ph	O	7.24
82	3,5-$(F)_2$	*i*-Pr	CH_2Ph	S	7.13

involved in the fitting. The resubstitution statistical parameters of the samples base their predictions on the regression fitted to those samples, and holdout statistical parameters base their predictions on the regression fitted to the other samples.

In the EV procedure, 20 HEPT analogues are randomly selected and eliminated from the data-set as unknown test samples. Then, the training set generated using remaining 60 samples and anti-HIV activities of eliminated samples are predicted using the MLR model. The test examples are marked with the superscript letter 'a' in Table 1. A model based on training set is as follows:

$$\begin{aligned}\text{pEC50} =& 11.394(\pm 3.651) - 8.152(\pm 1.035)\text{MATS5e} \\ &- 2.929(\pm 0.708)\text{MATS5p} \\ &+ 9.464(\pm 1.933)\text{ATS8v} \\ &- 5.379(\pm 4.047)\text{MATS1v} \\ &- 9.312(\pm 3.415)\text{ATS8e} \\ &+ 2.248(\pm 0.933)\text{MATS8e}\end{aligned}$$

$$N = 60,\ R^2 = 0.855,\ \text{SE} = 0.539,\ \text{RMS} = 0.511,$$
$$\text{REP} = 8.095,\ Q^2 = 0.811,\ F = 52.162. \quad (7)$$

The R^2_{pred} and RMS for the test set are 0.848 and 0.544, respectively. The plot of the experimental vs. predicted values for the above presented model is shown in Figure 3.

Statistical parameters obtained according to odd–even EV are summarised in Table 6. As can be seen, in the odd- and even-number samples, the resubstitution and holdout RMS are very similar, indicating that the same sample and the other sample predictions are equally precise.

For a better external predictive potential of the model, a modified r^2 [$r^2_{m(\text{test})}$] was introduced by the following equation [49]:

$$r^2_{m(\text{test})} = r^2 \times \left(1 - \sqrt{r^2 - r_0^2}\right). \quad (8)$$

In Equation (8), r_0^2 is the squared correlation coefficient between the observed and predicted values of the test-set compounds with the intercept set to zero. The value of $r^2_{m(\text{test})}$ should be greater than 0.5 for an acceptable model. Two other variants [50] of the r^2_m parameter, $r^2_{m(\text{LOO})}$ and $r^2_{m(\text{overall})}$, were also calculated. The parameter $r^2_{m(\text{overall})}$ is based on the prediction of both training-set (LOO prediction) and test-set compounds. It was shown [50] that $r_{m(\text{LOO}_2)}$ and $r^2_{m(\text{test})}$ penalise a model more strictly than Q^2 and R^2_{pred}, respectively. The statistical quality of the model is shown in Table 7.

3.3 Virtual library construction and screening

Virtual screening of chemical databases and virtual libraries is now a well-established method for researchers to identify new therapeutic agents in the drug discovery process. Computational approaches can thus significantly reduce the cost, time and labour required to synthesise and screen large libraries, as well as to enhance the success rate in lead compound generation [51].

In order to identify novel potent compounds, the developed model is considered as good tools for a virtual library [46] screening when the descriptor values, calculated for the molecules belonging to virtual libraries that are shown in Tables 8 and 9.

Table 9. Data of the selected descriptors used in this study for virtual library screening.

No.	MATS5e	MATS5p	ATS8v	MATS1v	ATS8e	MATS8e	No.	MATS5e	MATS5p	ATS8v	MATS1v	ATS8e	MATS8e
1	−0.321	−0.113	0.396	0.012	1.032	0.372	42	−0.257	−0.015	0.383	0.024	0.972	−0.064
2	−0.323	−0.102	0.415	0.012	1.023	0.320	43	−0.256	−0.021	0.351	0.050	0.998	−0.227
3	−0.080	−0.050	0.305	0.012	0.999	−0.238	44	−0.325	0.071	0.362	0.049	0.987	−0.231
4	−0.051	−0.022	0.321	0.012	0.983	−0.224	45	−0.319	−0.071	0.376	0.050	0.994	0.059
5	−0.132	−0.155	0.358	0.012	0.986	−0.039	46	−0.411	0.025	0.386	0.049	0.986	0.054
6	−0.117	−0.120	0.374	0.012	0.977	−0.048	47	−0.268	0.017	0.372	0.050	0.988	−0.106
7	−0.086	0.054	0.353	0.012	0.992	−0.147	48	−0.344	0.105	0.381	0.049	0.980	−0.100
8	−0.056	0.070	0.365	0.012	0.983	−0.163	49	−0.224	−0.023	0.357	0.043	1.008	−0.223
9	−0.155	−0.345	0.385	0.012	0.986	0.008	50	−0.260	0.070	0.367	0.043	0.998	−0.226
10	−0.157	−0.291	0.403	0.011	0.978	−0.007	51	−0.292	−0.049	0.370	0.043	1.009	0.139
11	−0.298	−0.310	0.373	0.011	0.975	0.035	52	−0.350	0.046	0.380	0.043	1.000	0.121
12	−0.351	−0.259	0.388	0.011	0.969	0.024	53	−0.233	−0.008	0.368	0.043	0.993	−0.073
13	−0.155	−0.345	0.385	0.012	0.986	0.008	54	−0.276	0.084	0.377	0.043	0.985	−0.057
14	−0.157	−0.291	0.403	0.011	0.978	−0.007	55	−0.199	−0.025	0.366	0.011	1.017	−0.208
15	−0.298	−0.310	0.373	0.011	0.975	0.035	56	−0.217	0.067	0.375	0.011	1.008	−0.212
16	−0.351	−0.259	0.388	0.011	0.969	0.024	57	−0.262	−0.026	0.368	0.011	1.021	0.167
17	−0.231	−0.302	0.389	0.011	0.982	−0.041	58	−0.296	0.067	0.377	0.011	1.012	0.137
18	−0.260	−0.254	0.406	0.010	0.971	−0.022	59	−0.203	−0.031	0.366	0.011	0.997	−0.049
19	−0.413	−0.283	0.366	0.011	0.968	0.051	60	−0.227	0.062	0.375	0.011	0.988	−0.030
20	−0.509	−0.234	0.378	0.010	0.962	0.042	61	−0.265	−0.047	0.388	0.011	1.003	−0.197
21	−0.348	−0.257	0.398	0.010	0.978	0.020	62	−0.318	0.043	0.397	0.010	0.995	−0.210
22	−0.432	−0.221	0.415	0.010	0.969	0.040	63	−0.319	−0.040	0.397	0.011	1.006	0.117
23	−0.299	−0.282	0.395	0.006	0.994	−0.070	64	−0.393	0.050	0.410	0.010	0.997	0.107
24	−0.308	−0.240	0.413	0.006	0.984	−0.069	65	−0.272	0.032	0.387	0.011	0.990	−0.070
25	−0.278	−0.400	0.313	−0.022	0.984	−0.058	66	−0.332	0.115	0.396	0.010	0.982	−0.053
26	−0.281	−0.340	0.327	−0.019	0.978	−0.089	67	−0.295	−0.052	0.400	0.010	0.996	−0.178
27	−0.309	−0.376	0.341	−0.052	0.986	0.004	68	−0.371	0.026	0.409	0.009	0.988	−0.194
28	−0.368	−0.265	0.357	−0.048	0.979	−0.008	69	−0.338	−0.041	0.414	0.010	0.998	0.069
29	−0.296	−0.266	0.36	−0.049	0.983	0.021	70	−0.432	0.038	0.428	0.009	0.989	0.067
30	−0.354	−0.17	0.374	−0.045	0.977	0.014	71	−0.301	0.059	0.399	0.010	0.987	−0.081
31	−0.31	−0.271	0.427	−0.018	0.997	−0.011	72	−0.383	0.129	0.408	0.009	0.979	−0.068
32	−0.372	−0.18	0.447	−0.017	0.986	0.01	73	−0.372	−0.135	0.388	0.086	1.007	0.182
33	−0.253	−0.272	0.361	−0.022	0.998	−0.015	74	−0.466	−0.035	0.399	0.085	0.999	0.163
34	−0.284	−0.176	0.376	−0.018	0.99	−0.011	75	−0.316	−0.073	0.377	0.075	1.036	0.261
35	−0.265	−0.186	0.376	−0.081	0.998	0.004	76	−0.357	0.023	0.388	0.074	1.027	0.205
36	−0.299	−0.1	0.393	−0.075	0.99	0.001	77	−0.354	−0.051	0.465	0.009	1.016	0.201
37	−0.195	−0.289	0.35	−0.052	0.988	0.015	78	−0.440	0.017	0.485	0.009	1.007	0.182
38	−0.235	−0.182	0.366	−0.048	0.981	0.008	79	−0.320	−0.052	0.432	0.010	1.032	0.252
39	−0.127	0.003	0.409	0.026	0.989	−0.127	80	−0.375	0.036	0.448	0.010	1.022	0.209
40	−0.179	0.083	0.419	0.024	0.981	−0.116	81	−0.237	−0.025	0.373	0.011	1.063	0.256
41	−0.221	−0.035	0.374	0.026	0.979	−0.066	82	−0.252	0.067	0.383	0.011	1.052	0.182

The QSAR analysis of the HEPT compounds indicated that there are three important zones to promote biological activity. (a) R1 substitute zone of the primary structure with substituted phenylthio groups; (b) R2 substitute zone of the primary structure with substitutes with relatively small volumes and (c) R3 substitute zone of the primary structure.

The effect of substitution at R1 was too complex to analyse. Generally, disubstitution with methyl or halogen or hydroxyl or amino gave better activities than the corresponding mono-substitution. The bulky *i*-Pr group at substitution R2 showed better activity than *n*-Pr, Et and Me substituents that occupy less space. The introduction of a phenyl substituent at R3 resulted in improved activities. Introduction of an alkyl or alkoxy spacer between the phenyl substituent and the ring nitrogen at R3 reduced the activities. Moreover, the introduction of sulphur (X = S) in almost all cases resulted in better activity than the oxygen (X = O) analogues. Finally, virtual screening identified three attractive compounds (analogues 20, 22 and 78; Table 8) that have high-quality activities (10.22, 10.48 and 10.15, respectively) and these deserve further study.

4. Conclusions

In the present work, a quantitative structure–property relationship approximation using a multiple linear correlation approach was developed to predict RT inhibition of HEPT analogues acting as NNRTIs. 2D autocorrelation space was employed, obtained from different weighting schemes, viewed as an adaptive descriptor space, containing topological information able to capture structural complexity. The 2D autocorrelation descriptors appeared to capture sufficient structural detail to yield very useful results in modelling biological properties. Our results corroborate that the employment of 2D autocorrelation descriptors is extremely useful in QSAR studies giving simple correlations between the molecular structures and biological activities. We have selected six variables from a pool of 96 descriptors to obtain a multilinear QSAR equation by the MLR method. The predictive ability and robustness of the models were examined by LOO cross-validation and EV, and, in general, satisfactory results were obtained. We expect this model to be useful in conjunction with the experimental methods for filtering likely HEPT analogues from chemical libraries and virtual chemical databases to identify new potential and selective compounds.

Acknowledgements

This work was supported by the Research Council of Jihad Daneshgahi. The authors would like to acknowledge Prof. Kunal Roy for his informative discussions pertaining to regression analysis, and the referees for their excellent suggestions to improve the quality of the paper.

Note

1. Guidelines for the use of antiretroviral agents in HIV infected adults and adolescents. Panel on Clinical Practice for Treatment of HIV Infection convened by the Department of Health and Human Services (DHHS) and the Henry J. Kaiser Family Foundation, 2002 (http://www.hivatis.org).

References

[1] D.H. Gabuzda and M.S. Hirsch, *Neurologic manifestations of infection with human immunodeficiency virus clinical features and pathogenesis*, Ann. Intern. Med. 107 (1987), pp. 383–391.
[2] L.L. Barton and N.R. Friedman, *The Neurological Manifestations of Pediatric Infectious Diseases and Immunodeficiency Syndromes*, Humana Press, Totowa, NJ, 2008.
[3] UNAIDS, *Report on the global AIDS epidemic*, XV International AIDS Conference, Bangkok, 2004.
[4] R.F. Schinazi, *Competitive inhibitors of human immunodeficiency virus reverse transcriptase*, Perspect. Drug Discov. Des. 1 (1993), pp. 151–180.
[5] R.K. Garg, *HIV infection and seizures*, Post. Grad. Med. J. 75 (1999), pp. 387–390.
[6] M. Fernandez and J. Caballero, *Modeling of activity of cyclic urea HIV-1 protease inhibitors using regularized-artificial neural networks*, Bioorg. Med. Chem. 14 (2006), pp. 280–294.
[7] M. Baba, Z. Debyser, S. Shigeta, and E. DeClercq, *Highly potent and selective inhibition of human immunodeficiency virus type I (HIV-I) by the HIV-I-specific reverse transcriptase inhibitors*, Drugs Fut. 17 (1992), pp. 891–897.
[8] E. DeClercq, *Trends in drug development for the treatment of AIDS compounds interfering with the initial stages of the HIV replicative cycle*, Eur. J. Pharm. Sci. 2 (1994), pp. 4–6.
[9] E. DeClercq, *Toward improved anti-HIV chemotherapy: Therapeutic strategies for intervention with HIV infections*, J. Med. Chem. 38 (1995), pp. 2491–2517.
[10] H. Mitsuya, R. Yarchoan, and S. Broder, *Molecular targets for AIDS therapy*, Science 249 (1990), pp. 1533–1544.
[11] J.A. Sandberg and W. Slikker, *Developmental pharmacology and toxicology of anti-HIV therapeutic agents: Dideoxynucleosides*, FASEB J. 9 (1995), pp. 1157–1163.
[12] A. Jacobo-Molina, J. Ding, R.G. Nanni, A.D. Clark, X. Lu, C. Tantillo, R.L. Williams, G. Kamer, A.L. Ferris, P. Clerck, A. Hizi, S. Huges, and E. Arnold, *Crystal structure of human immunodeficiency virus type 1 reverse transcriptase complexed with double-stranded DNA at 3.0 Å resolution shows bent DNA*, Proc. Natl Acad. Sci. USA 90 (1993), pp. 6320–6324.
[13] T. Miyasaka, H. Tanaka, M. Baba, H. Hayakawa, R.T. Walker, J. Balzarini, and E. DeClercq, *A novel lead for specific anti-HIV-1 agents: 1-[(2-Hydroxyethoxy) methyl]-6-(phenylthio) thymine*, J. Med. Chem. 32 (1989), pp. 2507–2509.
[14] 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. DeClercq, *A new class of HIV-1-specific 6-substituted acyclouridine derivatives: Synthesis and anti-HIV-1 activity of 5- or 6-substituted analogues of 1-[(2-hydroxyethoxy) methyl]-6-(phenylthio)thymine (HEPT)*, J. Med. Chem. 34 (1991), pp. 349–357.
[15] H. Tanaka, M. Baba, M. Ubasawa, H. Takashima, K. Sekiya, I. Nitta, S. Shigeta, R.T. Walker, and T. Miyasaka, *Synthesis and antiviral activity of deoxy analogs of 1-[(2-hydroxyethoxy)-methyl]-6-(phenylthio)thymine (HEPT) as potent and selective anti-HIV-1 agents*, J. Med. Chem. 34 (1991), pp. 1394–1399.
[16] H. Tanaka, H. Takashima, M. Ubasawa, K. Sekiya, I. Nitta, M. Baba, S. Shigeta, R.T. Walker, E. DeClercq, and T. Miyasaka, *Structure–activity relationships of 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine analogs: Effect of substitutions at the C-6 phenyl ring and at the C-5 position on anti-HIV-1 activity*, J. Med. Chem. 35 (1992), pp. 337–345.
[17] H. Tanaka, H. Takashima, M. Ubasawa, K. Sekiya, I. Nitta, M. Baba, S. Shigeta, R.T. Walker, E. DeClercq, and T. Miyasaka, *Synthesis and antiviral activity of deoxy analogs of 1-[(2-*

hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT) as potent and selective anti-HIV-1 agents, J. Med. Chem. 35 (1992), pp. 4713–4719.

[18] M.L. Barreca, J. Balzarini, A. Chimirri, E. De Clercq, L. De Luca, H. De Höltje, M. Höltje, A.M. Monforte, P. Monforte, C. Pannecouque, A. Rao, and M. Zappalà, *Design, synthesis, structure–activity relationships, and molecular modeling studies of 2,3-diaryl-1,3-thiazolidin-4-ones as potent anti-HIV agents*, J. Med. Chem. 45 (2002), pp. 5410–5413.

[19] R.K. Rawal, Y.S. Prabhakar, S.B. Katti, and E. De Clercq, *2-(Aryl)-3-furan-2-ylmethyl-thiazolidin-4-ones as selective HIV-RT inhibitors*, Bioorg. Med. Chem. 13 (2005), pp. 6771–6776.

[20] C. Mao, E.A. Sudbeck, T.K. Venkatachalam, and F.M. Uckun, *Structure-based design of non-nucleoside reverse transcriptase inhibitors of drug-resistant human immunodeficiency virus*, Antivir. Chem. Chemother. 10 (1999), pp. 233–240.

[21] C. Tintori, F. Manetti, N. Veljkovic, V. Perovic, J. Vercammen, S. Hayes, S. Massa, M. Witvrouw, Z. Debyser, V. Veljkovic, and M. Botta, *Novel virtual screening protocol based on the combined use of molecular modeling and electron–ion interaction potential techniques to design HIV-1 integrase inhibitors*, J. Chem. Inf. Model. 47 (2007), pp. 1536–1544.

[22] B. Bhhataraj and R. Garg, *From SAR to comparative QSAR: Role of hydrophobicity in the design of 4-hydroxy-5,6-dihydropyran-2-ones HIV-1 protease inhibitors*, Bioorg. Med. Chem. 13 (2005), pp. 4078–4084.

[23] E. Kellenberger, J.Y. Springael, M. Parmentier, M.H. Haas, J.L. Galzi, and D. Rognan, *Identification of nonpeptide CCR5 receptor agonists by structure-based virtual screening*, J. Med. Chem. 50 (2007), pp. 1294–1303.

[24] A.S. Mandal and K. Roy, *Predictive QSAR modeling of HIV reverse transcriptase inhibitor TIBO derivatives*, Eur. J. Med. Chem. 44 (2009), pp. 1509–1524.

[25] J.K. Buolamwini and H. Assefa, *CoMFA and CoMSIA 3D QSAR and docking studies on conformationally-restrained cinnamoyl HIV-1 integrase inhibitors: Exploration of a binding mode at the active site*, J. Med. Chem. 45 (2002), pp. 841–852.

[26] R. Todeschini, Milano Chemometrics and QSAR Group; software available at http://www.talete.mi.it

[27] G. Moreau and P. Broto, *The autocorrelation of a topological structure: A new molecular descriptor*, Nouv. J. Chim. 4 (1980), pp. 359–360.

[28] P.A.P. Moran, *Notes on continuous stochastic phenomena*, Biometrika 37 (1950), pp. 17–23.

[29] R.F. Geary, *The contiguity ratio and statistical mapping*, Statistician 5 (1954), pp. 115–141.

[30] P. Broto, G. Moreau, and C. Vandycke, *Molecular structures: Perception, autocorrelation descriptor and SAR studies*, Eur. J. Med. Chem. 19 (1984), pp. 66–70.

[31] M. Fernández and J. Caballero, *Modeling of activity of cyclic urea HIV-1 protease inhibitors using regularized-artificial neural networks*, J. Bioorg. Med. Chem. 14 (2006), pp. 280–294.

[32] J. Caballero, M. Garriga, and M. Fernández, *2D autocorrelation modeling of the negative inotropic activity of calcium entry blockers using Bayesian-regularized genetic neural networks*, Bioorg. Med. Chem. 14 (2006), p. 3330.

[33] M. Fernández and J. Caballero, *Bayesian-regularized genetic neural networks applied to the modeling of non-peptide antagonists for the human luteinizing hormone-releasing hormone receptor*, J. Mol. Graph. Model. 25 (2006), pp. 410–422.

[34] M. Fernández, J. Caballero, and A. Tundidor-Camba, *Linear and nonlinear QSAR study of N-hydroxy-2-[(phenylsulfonyl)amino]acetamide derivatives as matrix metalloproteinase inhibitors*, Bioorg. Med. Chem. 14 (2006), pp. 4137–4150.

[35] J. Caballero, A. Tundidor-Camba, and M. Fernández, *Modeling of the inhibition constant (Ki) of some cruzain ketone-based inhibitors using 2D spatial autocorrelation vectors and data-diverse ensembles of Bayesian-regularized genetic neural networks*, QSAR Comb. Sci. 26 (2007), pp. 27–40.

[36] M. Fernández and J. Caballero, *QSAR modeling of matrix metalloproteinase inhibition by N-hydroxy-α-phenylsulfonylacetamide derivatives*, J. Bioorg. Med. Chem. 15 (2007), pp. 6298–6310.

[37] L.S. Urra, M.P. Gonzáleza, and M. Teijeira, *2D-autocorrelation descriptors for predicting cytotoxicity of naphthoquinone ester derivatives against oral human epidermoid carcinoma*, J. Bioorg. Med. Chem. 15 (2007), pp. 3565–3571.

[38] 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.

[39] 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.

[40] H. Bazoui, M. Zahouily, S. Boulajaaj, S. Sebti, and D. Zakarya, *QSAR for anti-HIV activity of HEPT derivatives*, SAR QSAR Environ. Res. 13 (2002), pp. 567–577.

[41] 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.

[42] M.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*, Chem. Intel. Lab. Syst. 83 (2006), pp. 91–98.

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

[44] A. Afantitis, G. Melagraki, H. Sarimveis, A. Koutentis, J. Markopoulos, and O. Igglessi-Markopoulou, *A novel simple QSAR model for the prediction of anti-HIV activity using multiple linear regression analysis*, Mol. Diversity 10 (2006), pp. 405–414.

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

[46] J.D. Walker, J. Jaworska, M.H.I. Comber, T.W. Schultz, and J.C. Dearden, *Guidelines for developing and using quantitative structure–activity relationships*, Environ. Toxicol. Chem. 22 (2003), pp. 1653–1665.

[47] L. Eriksson, J. Jaworska, A.P. Worth, M.T.D. Cronin, R.M. McDowell, and P. Gramatica, *Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs*, Environ. Health Perspect. 111 (2003), pp. 1361–1375.

[48] D.M. Hawkins, S.C. Basak, and D. Mills, *Assessing model fit by cross-validation*, J. Chem. Inf. Comput. Sci. 43 (2003), pp. 579–586.

[49] P.P. Roy and K. Roy, *On some aspects of variable selection for partial least squares regression models*, QSAR Comb. Sci. 27 (2008), pp. 302–313.

[50] P.P. Roy, S. Paul, I. Mitra, and K. Roy, *On two novel parameters for validation of predictive QSAR models*, Molecules 14 (2009), pp. 1660–1701.

[51] P. Willett, *Similarity-based virtual screening using 2D fingerprints*, Drug Discov. Today 11 (2006), pp. 1046–1053.