European Journal of Medicinal Chemistry 77 (2014) 298–305

Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

journal homepage: http://www.elsevier.com/locate/ejmech

Original article

QSAR models for HEPT derivates as NNRTI inhibitors based on Monte Carlo method

Alla P. Toropova [a], Andrey A. Toropov [a], Jovana B. Veselinović [b], Filip N. Miljković [b], Aleksandar M. Veselinović [b,*]

[a] IRCCS – Istituto di Ricerche Farmacologiche Mario Negri, Milano, Italy
[b] University of Niš, Faculty of Medicine, Department of Chemistry, Bulevar Dr Zorana Djindjića 81, 18000 Niš, Serbia

ARTICLE INFO

Article history:
Received 18 December 2013
Received in revised form
31 January 2014
Accepted 5 March 2014
Available online 11 March 2014

Keywords:
QSAR
SMILES
HEPT
CORAL software
Computer-aided drug design
Monte Carlo method

ABSTRACT

A series of 107 1-[(2-hydroxyethoxy)-methyl]-6-(phenylthio) thymine (HEPT) with anti-HIV-1 activity as a non-nucleoside reverse transcriptase inhibitor (NNRTI) has been studied. Monte Carlo method has been used as a tool to build up the quantitative structure–activity relationships (QSAR) for anti-HIV-1 activity. The QSAR models were calculated with the representation of the molecular structure by simplified molecular input-line entry system and by the molecular graph. Three various splits into training and test set were examined. Statistical quality of all build models is very good. Best calculated model had following statistical parameters: for training set $r^2 = 0.8818$, $q^2 = 0.8774$ and $r^2 = 0.9360$, $q^2 = 0.9243$ for test set. Structural indicators (alerts) for increase and decrease of the IC_{50} are defined. Using defined structural alerts computer aided design of new potential anti-HIV-1 HEPT derivates is presented.

1. Introduction

Acquired immunodeficiency syndrome (AIDS) is a fatal disorder resulting from a chronic persistent infection by the human retrovirus, human immunodeficiency virus (HIV) [1]. Up to now successful chemotherapy has not been developed. A key enzyme which is responsible for the process of HIV-1 replication is reverse transcriptase (RT). RT of HIV-1 is required for early provial DNA synthesis making it a prime target for antiviral therapy against AIDS. Inhibition of the RT-catalyzed polymerization of DNA from viral RNA inhibits virus replication [2–4]. Mechanism of inhibition is proposed to be by locking the polymerase active site in an inactive conformation, reminiscent of the conformation observed in the inactive p51 subunit [5]. Several classes of compounds have been synthesized and tested as highly specific inhibitors of HIV-1 for AIDS therapy [6–9].

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), first synthesized by Tanaka et al. as a non-nucleoside reverse transcriptase inhibitor (NNRTI) [6–9]. The experimental studies of crystal structures of two ligand–enzyme complexes revealed an active HEPT conformation that binds the p66 enzyme unit [10]. Studies have shown the importance of so called switching effect, which forces an enzyme Tyr181 to adopt a conformation that enhances the drug–receptor interactions [11]. Also, it was observed that NNRTI inhibitors bind in a common way filling the allosteric pocket in a so-called butterfly mode [12]. One of the butterfly wings binds with p electron-rich moiety with a hydrophobic pocket that includes mainly the side chains of the enzymes Tyr181, Tyr188, Phe227, Trp229, and Tyr318, while the other wing having the hydrogen donor/acceptor interacts with Lys101 and Lys108. A hydrophobic moiety of the butterfly body interacts with an enzyme receptor site formed by the side chains of Lys103, Val106, and Val 179 [13]. HEPT derivatives have potent anti-HIV-1 activity at nanomolar concentration [6]. Further advantages of NNRTIs are lower metabolism and clearance rates [6–9]. Moreover, these compounds are less toxic and more stable than nucleoside RT inhibitors [14].

The importance of quantitative structure–activity relationship (QSAR) methods in modern drug design is well established since QSAR can make the early prediction of activity-related characteristics of drug candidates and can eliminate molecules with undesired properties [15]. A number of QSAR studies have been reported

* Corresponding author.
E-mail address: aveselinovic@medfak.ni.ac.rs (A.M. Veselinović).

http://dx.doi.org/10.1016/j.ejmech.2014.03.013
0223-5234/

for HEPT compounds [16–28]. Thousands of molecular descriptors used in QSAR studies have been defined to encode chemical and structural features of molecules [29,30].

QSAR analysis widely uses descriptors calculated on the basis of molecular graphs [31,32]. The simplified molecular input line entry system (SMILES) is an alternative to molecular graph and it can be used for elucidation of molecular structures [33]. Recent published papers have reported the applicability of solo SMILES based descriptors in QSAR analysis [34–38] as well as SMILES based descriptors in combination with topological descriptors [39–41]. All QSAR models are based on Monte Carlo optimization method where appropriate activity is treated as random event [34–41].

The aim of this study is to build QSAR models based on SMILES and graph optimal descriptors using Monte Carlo method for HEPT derivatives as NNRTIs inhibitors and an attempt to define the molecular structure responsible for stated inhibitory effect. Further, we used build models for computer aided drug design of new potentially promising anti-HIV-1 compounds.

2. Method

2.1. The data set

A series of 107 HEPT derivatives from literature [16] were used to generate canonical SMILES. SMILES used in the present study were generated with the ACD/ChemSketch program (ACD/ChemSketch v. 11.0). General structure of HEPT molecules is presented in Fig. 1. Anti-HIV-1 activity was expressed with RT inhibition data (IC_{50}, μM) and was taken from the literature where IC_{50} (log 1/C) was defined as dependant variable in which C represents the molar concentration of compound needed to achieve 50% protection of MT-4 cells against the cytopathic effect of HIV-1 (HTLV-IIIB strain) [16].

The QSAR models were built up for three random splits (20% of compounds were used in test set). All three splits were selected according to the following principles: a) the range of the endpoint is approximately the same for each sub-set b) the splits are not identical (Table S1). The level of identity of these splits has been checked according to published method [39]. Between split 1 and 2 level of identity was 22.73%, between 1 and 3 13.64 and between 2 and 3 4.55%. Results indicate that splits are different since the level of identity between them is low.

2.2. Optimal descriptors

The molecular structure can be represented both with SMILES and molecular graph. In some cases "hybrid" representation of the molecular structure, i.e., by SMILES together with molecular graph, can give a model characterized by higher statistical quality than model which is based on the representation of the molecular structure by solely SMILES or solely molecular graph. The hybrid optimal descriptors used to build up model for the IC_{50} were calculated according to Eq. (1):

$$^{\text{Hybrid}}\text{DCW}\left(T, N_{\text{epoch}}\right) = {}^{\text{Graph}}\text{DCW}\left(T, N_{\text{epoch}}\right) + {}^{\text{SMILES}}\text{DCW}\left(T, N_{\text{epoch}}\right) \quad (1)$$

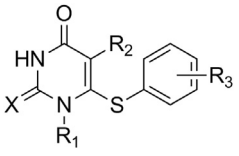

Fig. 1. General molecular structure of used molecules.

T and N_{epoch} are parameters used in Monte Carlo optimization method. T is threshold used for definition of rare molecular features. If a molecular feature, x, that is calculated from SMILES or molecular graph in the training set takes place less than T times, then the x should be removed from the building up of the model, therefore the correlation weight of the x, $\text{CW}(x) = 0$. That molecular feature is defined as rare. E_{poch} is the number of the epochs of the Monte Carlo optimization.

For SMILES based descriptors DCW is calculated according to Eq. (2).

$$\begin{aligned} ^{\text{SMILES}}\text{DCW}\left(T, N_{\text{epoch}}\right) &= \text{CW}(\text{ATOMPAIR}) + \text{CW}(\text{NOSP}) \\ &+ \text{CW}(\text{BOND}) + \text{CW}(\text{HALO}) \\ &+ \alpha \sum \text{CW}(\text{Sk}) + \beta \sum \text{CW}(\text{SSk}) \\ &+ \gamma \sum \text{SCW}(\text{SSSk}) \end{aligned} \quad (2)$$

Sk is SMILES attribute, mainly one symbol ("C", "O", "=", etc.) or in some cases two symbols, which cannot be examined separately (e.g., "Cl", "Br", etc.). For calculating all SMILES based descriptors methodology from literature was used [30]. Building QSAR models took into consideration all SMILES descriptors both indices and fragments (ATOMPAIR, BOND, NOSP, HALO, Sk, SSk and SSSk).

Graph based optimal descriptors are calculated according to published methodology [34] and according to Eq. (3).

$$\begin{aligned} ^{\text{Graph}}\text{DCW}\left(T, N_{\text{epoch}}\right) &= \sum \text{CW}(A_k) + \alpha \sum \text{CW}(^0\text{EC}_k) \\ &+ \beta \sum \text{CW}(^1\text{EC}_k) \end{aligned} \quad (3)$$

where A_k is chemical element, such as, C, N, O, etc., for HSG and HFG; or atomic orbitals, such as, 1s1, 2p3, 3d10, etc. for GAO; $^0\text{EC}_k$, $^1\text{EC}_k$,… is the hierarchy of the Morgan's extended connectivity. In presented study three types of the graph were studied: (1) hydrogen-suppressed graph (HSG), (2) hydrogen-filled graph (HFG) and (3) graph of atomic orbitals (GAO).

For building QSAR models CORAL software was used (http://www.insilico.eu/coral). Search for the best model was performed according to published methodology [36]. The search for most predictive combination of T and N_{epoch} for all splits was performed from values 0–10 for T and 0–60 for N_{epoch}.

Having numerical data on these CW (CW for SMILES and graph descriptors in hybrid modeling) one can calculate $\text{DCW}(T, N_{\text{epoch}})$ for compounds of training and test set. These data can be used for calculation by Least Squares method model of view according to Eq (4):

$$\text{IC}_{50} = \text{C}_0 + \text{C}_1 \times \text{DCW}\left(T, N_{\text{epoch}}\right) \quad (4)$$

2.3. Validation of QSAR model

Developing a robust model capable of predicting the property of new molecules in objective, reliable and precise manner is the main goal of QSPR modeling [42]. Three strategies for QSAR models validation are suggested in literature [43]. Stated methodology was applied to presented research for assessment of robustness and reliability of developed model. Three used methods were:

1. Internal validation or cross-validation using the training set compounds.
2. External validation using the test set compounds.
3. Data randomization or Y-scrambling.

Leave-one-out (LOO) cross validation technique was used to developed models as an internal validation. LOO is based on principle that one molecule is randomly omitted from data set in each cycle and than the rest of molecules is used to model development. The process is repeated until all the compounds are eliminated once. Cross-validated Q^2 determinates the predictive ability of the model [44]. Higher value of Q^2 means better model prediction. The cross-validated Q^2 is expressed as:

$$Q^2 = 1 - \sum \left(Y_{obs} - Y_{pred}\right)^2 / \sum \left(Y_{obs} - \overline{Y}_{train}\right)^2 \quad (5)$$

In Eq. (5), Y_{obs} is observed property of the training set compounds, Y_{pred} is LOO-predicted property of the training set compounds and $\overline{Y}_{train}$ is mean observed property of the training set compounds. The predictive ability of model is considered as acceptable when Q^2 is greater than 0.5.

Same principles and statistical methodology can be applied for external validation. The predictive ability of a model is determined by calculating Q^2_{ext} which is defined as:

$$Q^2_{ext} = 1 - \sum \left(Y_{obs(test)} - Y_{pred(test)}\right)^2 / \sum \left(Y_{obs(test)} - \overline{Y}_{train}\right)^2 \quad (6)$$

In Eq. (6), $Y_{obs(test)}$ is the observed property of the test set compounds, $Y_{pred(test)}$ is the predicted property of the test set compounds and $\overline{Y}_{train}$ is mean observed property of the training set compounds. The value Q^2_{ext} for an acceptable model should be greater than 0.5.

Novel statistical metric (R^2_m) was used to validate a true predictive potential of developed QSPR models [45,46]. For calculating R^2_m metric an open access web application "R^2m calculator" is available at http://203.200.173.43:8080/rmsquare/.

Y-randomization test was used for checking the robustness of the model. For a suitable QSPR model, the average correlation coefficient (R_r) of randomized models should be less than the correlation coefficient (R) of non-randomized model. A parameter ${}^cR^2_p$ penalizes the model R^2 for a small difference between squared mean correlation coefficient (R^2_r) of randomized models and squared correlation coefficient (R^2) of the non-randomized model [47]. The parameter ${}^cR^2_p$ is defined as:

$${}^cR^2_p = R\sqrt{\left(R^2 - R^2_r\right)} \quad (7)$$

For an acceptable QSPR model, the value of ${}^cR^2_p$ should be greater than 0.5.

2.4. Applicability domain

Applicability domain (AD) of a QSPR model is defined as the biological, structural or physico-chemical space, knowledge or information on which the model of the training set is developed, and for which it is applicable to make predictions for new compounds. If predicted compounds are within applicability then the predictions of a QSPR model are more reliable. When a compound is very much dissimilar to all compounds of the modeling set, reliable prediction of its property is uncertain. For defining AD method from literature was applied [39].

3. Results and discussion

From data set presented at Table S1 three random splits were generated. Generated training and test set had 85 and 22 molecules respectively. The molecules in all sets were carefully selected so the ranges of endpoint in training and test set are set to be approximately equivalent. The search for the best model included models with SMILES descriptors and molecular graph based descriptors, with different combination of connectivity indices. Model 1 had only SMILES descriptors, model 2 had SMILES descriptors and HSG descriptors with ${}^0EC_k$ connectivity, model 3 had SMILES descriptors and HSG descriptors with ${}^1EC_k$ connectivity, model 4 had SMILES descriptors and HFG descriptors with ${}^0EC_k$ connectivity, model 5 had SMILES descriptors and HFG descriptors with ${}^1EC_k$ connectivity, model 6 had SMILES descriptors and GAO descriptors with ${}^0EC_k$ connectivity and model 7 had SMILES descriptors and GAO descriptors with ${}^1EC_k$ connectivity.

Applied methodology for defining applicability domain revealed that all compounds have typical ('average') behavior and all were included in developing QSAR models.

Table 1 shows the statistical quality of the built models. From presented data it can be observed that there is the reproduction of the statistical quality for calculated models in three runs of the Monte Carlo optimization for all splits. Results suggest that prediction with CORAL is robust since there is similar statistical quality for all splits. The best calculated model was model 7 for split 3. The search for preferable T and N_{epoch} revealed that preferable T is 3 and preferable N_{epoch} 6. Fig. 2 present best models for all splits graphically where Monte Carlo optimization run with highest value for r^2 for the best model is represented.

Statistical criteria of the predictability of the best models for all splits are represented in Table 2 [44–46]. Results from Table 2 show that the predictability for best models for all splits is very good. In addition, all presented models for IC_{50} are satisfactory from the point of view of new criteria suggested by Roy et al. [45]. Table S2 (supplementary material) shows Y-randomization [47] which also confirms the robustness of suggested models.

The application of above mentioned T and the N_{epoch} gives the following models for the IC_{50} calculated according to Eq. (4):

Split 1

$$IC_{50} = -6.1361(\pm 0.0522) + 0.1202(\pm 0.0005) \times DCW(4,7) \quad (8)$$

(training set: $r^2 = 0.8474$, $q^2 = 0.8410$, $s = 0.565$, MAE = 0.444, $F = 461$, test set: $r^2 = 0.8947$, $q^2 = 0.8739$, $s = 0.783$, MAE = 0.641, $F = 170$).

Split 2

$$IC_{50} = -2.7180(\pm 0.03781) + 0.1451(\pm 0.0006) \times DCW(10,11) \quad (9)$$

(training set: $r^2 = 0.8430$, $q^2 = 0.8365$, $s = 0.578$, MAE = 0.459, $F = 446$, test set: $r^2 = 0.9299$, $q^2 = 0.9164$, $s = 0.514$, MAE = 0.406, $F = 265$).

Split 3

$$IC_{50} = -0.4962(\pm 0.0213) + 0.0594(\pm 0.0002) \times DCW(3,6) \quad (10)$$

(training set: $r^2 = 0.8818$, $q^2 = 0.8774$, $s = 0.506$, MAE = 0.382, $F = 619$, test set: $r^2 = 0.9360$, $q^2 = 0.9243$, $s = 0.466$, MAE = 0.359, $F = 293$).

In the Eqs. (8)–(10), r^2 is the correlation coefficient, q^2 is the leave-one-out cross-validated correlation coefficient; s is the standard error of estimation; MAE is mean absolute error and F is Fischer ratio. Proposed models are for the best model for each split where best Monte Carlo optimization run was taken into consideration. For splits 1, 2 and 3 best models were model 6, model 6 and model 7, respectively.

Table 1
Statistical quality of QSAR models of IC_{50} for three random splits.

		1 SPLIT									2 SPLIT									3 SPLIT								
		T	Nac	rs2	ss	fs	r_v^2	sv	fv	R_m^2	*T*	Nac	r_s^2	ss	fs	r_v^2	sv	fv	R_m^2	*T*	Nac	r_s^2	ss	fs	r_v^2	sv	fv	R_m^2
Model 1	1	6	117	0.8267	0.602	396	0.8487	0.832	112	0.7672	5	126	0.8497	0.566	469	0.8910	0.599	164	0.8598	8	105	0.7488	0.738	247	0.893	0.606	167	0.8856
	2	6	117	0.8281	0.600	400	0.8498	0.844	113	0.7663	5	126	0.8579	0.550	501	0.8509	0.692	114	0.8251	8	105	0.7666	0.711	273	0.9048	0.565	190	0.8984
	3	6	117	0.8223	0.610	384	0.8375	0.879	103	0.7367	5	126	0.8528	0.560	481	0.9054	0.566	191	0.8875	8	105	0.7656	0.713	271	0.8906	0.588	163	0.8822
	Av	**6**	**117**	**0.8257**	**0.604**	**393**	**0.8453**	**0.852**	**109**	**0.7568**	**5**	**126**	**0.8535**	**0.559**	**484**	**0.8824**	**0.619**	**156**	**0.8574**	**8**	**105**	**0.7603**	**0.721**	**264**	**0.8961**	**0.586**	**173**	**0.8887**
Model 2	1	8	120	0.8103	0.630	355	0.8598	0.779	123	0.8127	10	108	0.8430	0.578	446	0.9299	0.515	265	0.8814	7	116	0.8151	0.633	366	0.8851	0.589	154	0.8433
	2	8	120	0.8011	0.645	334	0.8452	0.772	109	0.7969	10	108	0.8451	0.575	453	0.9010	0.564	182	0.8935	7	116	0.8037	0.652	340	0.8973	0.577	175	0.8520
	3	8	119	0.8206	0.613	380	0.8796	0.736	146	0.8233	10	108	0.8393	0.585	433	0.9210	0.532	233	0.8743	7	116	0.7979	0.662	328	0.8979	0.596	176	0.8697
	Av	**8**	**120**	**0.8107**	**0.630**	**356**	**0.8615**	**0.762**	**126**	**0.8110**	**10**	**108**	**0.8423**	**0.580**	**444**	**0.9173**	**0.537**	**227**	**0.8831**	**7**	**116**	**0.8056**	**0.649**	**345**	**0.8934**	**0.587**	**168**	**0.855**
Model 3	1	9	117	0.7871	0.668	307	0.8672	0.743	131	0.8159	9	125	0.8701	0.526	556	0.8861	0.621	156	0.8845	7	122	0.8331	0.602	414	0.9194	0.541	228	0.9092
	2	9	117	0.8449	0.570	452	0.8629	0.785	126	0.8112	9	125	0.8673	0.532	542	0.9054	0.547	191	0.8717	7	120	0.8359	0.597	423	0.9285	0.480	260	0.8959
	3	9	118	0.7780	0.682	291	0.8636	0.754	127	0.7933	9	125	0.8633	0.54	524	0.8864	0.601	156	0.8697	7	119	0.8234	0.619	387	0.9224	0.526	238	0.9086
	Av	**9**	**117**	**0.8033**	**0.640**	**350**	**0.8646**	**0.761**	**128**	**0.8068**	**9**	**125**	**0.8669**	**0.533**	**541**	**0.8926**	**0.590**	**168**	**0.8753**	**7**	**120**	**0.8308**	**0.606**	**408**	**0.9234**	**0.516**	**242**	**0.9046**
Model 4	1	4	154	0.8456	0.569	455	0.8736	0.787	138	0.8390	5	137	0.8624	0.541	520	0.9188	0.543	226	0.8925	7	117	0.8105	0.641	355	0.8962	0.594	173	0.8385
	2	4	153	0.8313	0.594	409	0.8819	0.775	149	0.8212	5	137	0.8554	0.555	491	0.9149	0.552	215	0.8818	7	117	0.7758	0.697	287	0.8998	0.602	180	0.8834
	3	4	154	0.8310	0.595	408	0.8821	0.779	150	0.8479	5	137	0.8583	0.550	503	0.9106	0.540	204	0.8968	7	117	0.7934	0.669	319	0.8883	0.605	159	0.8459
	Av	**4**	**154**	**0.8360**	**0.586**	**424**	**0.8792**	**0.780**	**146**	**0.8360**	**5**	**137**	**0.8587**	**0.549**	**505**	**0.9148**	**0.545**	**215**	**0.8904**	**7**	**117**	**0.7932**	**0.669**	**320**	**0.8948**	**0.600**	**171**	**0.8559**
Model 5	1	3	171	0.8601	0.541	510	0.8245	0.904	94	0.8203	7	129	0.8563	0.553	494	0.9127	0.54	209	0.8964	8	122	0.8843	0.501	634	0.9273	0.538	255	0.8635
	2	3	172	0.8479	0.564	463	0.8345	0.887	101	0.8341	7	129	0.8763	0.513	588	0.9063	0.554	193	0.8912	8	122	0.8842	0.501	634	0.9203	0.559	231	0.8478
	3	3	172	0.8525	0.556	480	0.8104	0.918	86	0.7831	7	129	0.8779	0.51	597	0.9125	0.546	208	0.8849	8	122	0.8876	0.494	656	0.9189	0.578	227	0.8701
	Av	**3**	**172**	**0.8535**	**0.554**	**484**	**0.8231**	**0.903**	**94**	**0.8125**	**7**	**129**	**0.8702**	**0.525**	**560**	**0.9105**	**0.547**	**203**	**0.8908**	**8**	**122**	**0.8854**	**0.499**	**641**	**0.9222**	**0.558**	**238**	**0.8605**
Model 6	1	4	173	0.8469	0.566	459	0.8678	0.793	131	0.8596	5	153	0.8560	0.554	493	0.9214	0.519	234	0.9110	7	131	0.7868	0.680	306	0.8951	0.591	171	0.8251
	2	4	173	0.8474	0.565	461	0.8947	0.782	170	0.8601	5	153	0.8541	0.558	486	0.9193	0.553	228	0.8934	7	132	0.7832	0.686	300	0.8747	0.635	140	0.8155
	3	4	173	0.8416	0.576	441	0.8779	0.807	144	0.8353	5	153	0.8568	0.552	497	0.9192	0.537	228	0.9083	7	133	0.7975	0.663	327	0.8623	0.659	125	0.7635
	Av	**4**	**173**	**0.8453**	**0.569**	**454**	**0.8801**	**0.794**	**148**	**0.8517**	**5**	**153**	**0.8556**	**0.555**	**492**	**0.9200**	**0.536**	**230**	**0.9042**	**7**	**132**	**0.7892**	**0.676**	**311**	**0.8774**	**0.628**	**145**	**0.8014**
Model 7	1	9	163	0.8981	0.462	732	0.7795	1.039	71	0.7376	2	270	0.9017	0.458	761	0.9025	0.570	185	0.8343	3	223	0.8831	0.503	627	0.9271	0.501	254	0.9213
	2	9	162	0.8938	0.472	698	0.8226	0.958	93	0.7978	2	270	0.9006	0.460	752	0.9162	0.520	219	0.8444	3	223	0.8818	0.506	619	0.9360	0.468	293	0.9347
	3	9	163	0.8903	0.479	674	0.7886	1.013	75	0.7503	2	269	0.9003	0.461	750	0.9145	0.531	214	0.8415	3	222	0.8810	0.508	615	0.9294	0.486	263	0.9082
	Av	**9**	**163**	**0.8941**	**0.471**	**701**	**0.7969**	**1.003**	**80**	**0.7619**	**2**	**270**	**0.9009**	**0.456**	**754**	**0.9111**	**0.540**	**206**	**0.8401**	**3**	**222**	**0.8820**	**0.506**	**620**	**0.9308**	**0.485**	**270**	**0.9214**

T − Threshold.
Nac − Number of attributes which are not blocked for a given *T*.
r_s^2, r_v^2 − Correlation coefficient for training and test set.
ss, sv − Standard error of estimation for training and test set.
fs, fv − Fischer *F*-ratio for training and test set.
R_m^2 − Metric suggested by Roy et al. [45].

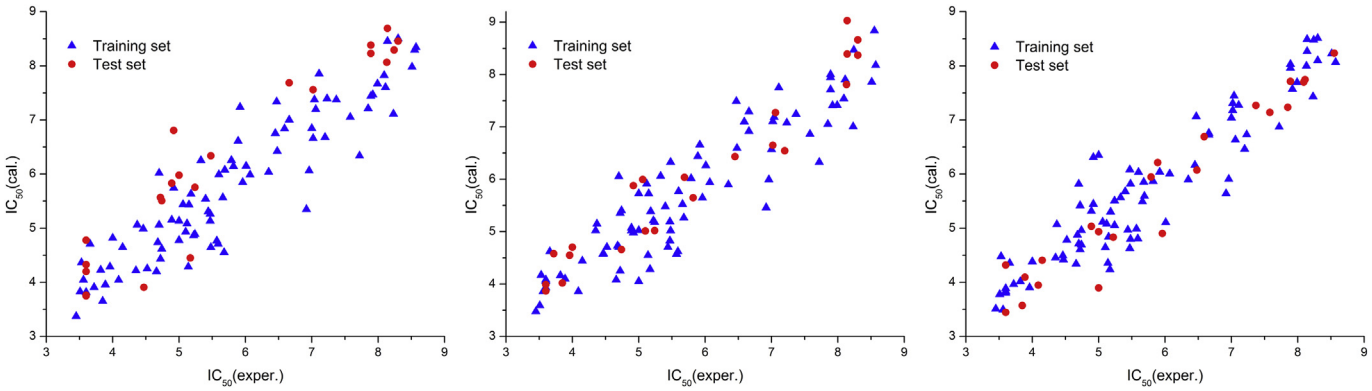

Fig. 2. Graphical representation of QSAR models for used splits for IC_{50}.

Table S3 contains an example of calculation DCW(3,6) for the best Monte Carlo run for best model represented in Table 1. As addition, Table S4 contains the Adjacency Matrix of the molecular graph used for calculation (Supplementary material).

In order to demonstrate the predictive power and accuracy of the Monte Carlo method, 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. Models constructed using Monte Carlo method were described by similar or better statistics and predictive power as compared with the other QSAR models. Comparing coefficients from literature [16] 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 to coefficients from best Monte Carlo model (training set: $r^2 = 0.8818$, $q^2 = 0.8774$ and set: $r^2 = 0.9360$, $q^2 = 0.9243$) reveals similar and better predictivity

Table 2
Criteria of predictability for best QSAR models for three splits.

Split 1	Split 2	Split 3
$r^2_v(x,y)$:		
y-experimental values;		
x-calculated values		
$r^2_v = 0.8947$	$r^2_v = 0.9299$	$r^2_v = 0.9360$
$r^2_0 = 0.8718$	$r^2_0 = 0.9133$	$r^2_0 = 0.9322$
$rr^2_0 = 0.8932$	$rr^2_0 = 0.9271$	$rr^2_0 = 0.9360$
$r_A = 0.0256$	$r_A = 0.0179$	$r_A = 0.0041$
$r_B = 0.0017$	$r_B = 0.0029$	$r_B = 0.0000$
$k = 1.0782$	$k = 1.0299$	$k = 0.9639$
$kk = 0.9191$	$kk = 0.9655$	$kk = 1.0327$
$R^2_m(t) = 0.7593$	$R^2_m(t) = 0.8100$	$R^2_m(t) = 0.8778$
$r^2_v(y,x)$:		
y-calculated values;		
x-experimental values		
$r^2_v = 0.8947$	$r^2_v = 0.9299$	$r^2_v = 0.9360$
$r^2_0 = 0.8932$	$r^2_0 = 0.9271$	$r^2_0 = 0.9360$
$rr^2_0 = 0.8718$	$rr^2_0 = 0.9133$	$rr^2_0 = 0.9322$
$r_A = 0.0017$	$r_A = 0.0029$	$r_A = 0.0000$
$r_B = 0.0256$	$r_B = 0.0179$	$r_B = 0.0041$
$k = 0.9191$	$k = 0.9655$	$k = 1.0327$
$kk = 1.0782$	$kk = 1.0299$	$kk = 0.9639$
$R'^2_m(t) = 0.8601$	$R'^2_m(t) = 0.8814$	$R'^2_m(t) = 0.9347$
$R^2_m(av) = 0.8097$	$R^2_m(av) = 0.8457$	$R^2_m(av) = 0.9063$
$\Delta R^2_m = 0.1008$	$\Delta R^2_m = 0.0714$	$\Delta R^2_m = 0.0569$

$r_A = (r^2 - r^2_0)/r^2$ should be <0.1 [44].
$r_B = (r^2 - rr^2_0)/r^2$ should be <0.1 [44].
Should be $0.85 < k < 1.15$ [44].
Should be $0.85 < kk < 1.15$ [44].
$R^2_m(t)$ and $R'^2_m(t)$; R^2_m of test set and should be >0.5 [45].
$R^2_m(av)$ is average value of R^2_m and should be >0.5 [46].
ΔR^2_m should be lower 0.2 [46].

since PLS and MLR models were constructed without using test set-predicted values to validate the model. Another research [18] used both ANN and MLR techniques for the same data set. Q^2 values for MLR and ANN were calculated as 0.605 and ranged from 0.525 to 0.954, while r^2 values were found to be 0.811 and 0.919, respectively (but no q^2_{ext1} and q^2_{ext2} values). Results presented in this work are comparable to stated results, since high values for q^2 implied the model's high predictive power. According to the current OECD guidelines [48], high q^2 cannot be a single parameter to imply the predictive ability of a model. For this reason several validation methods were used. Results prove that presented approach is a powerful alternative to more popular QSAR methods.

The correlation weights (CW) of molecular features (SAk) calculated using CORAL both for SMILES and graph based descriptors can be used for classification of these features according to their values from three probes for appropriate Monte Carlo model. They could be divided into three categories: features with stable positive values of correlation weights (promoters of increase of an endpoint); features with stable negative values of correlation weights (promoters of decrease of an endpoint); and unstable features which have positive values of correlation weights together with negative correlation weights values for several models [29–34]. If the, for an example, correlation weight of SAk CW(SAk) is >0 in all three runs of the Monte Carlo optimization process then the SAk is promoter of IC_{50} increase. However, if CW(SAk) is <0 in all three runs of the optimization then the SAk is promoter of IC_{50} decrease. In the end if there are both CW(Sk) >0 and <0, or SAk is blocked in three runs of optimization then SAk has an undefined role. Same rule is applied for all SAk. The list of all SAk together with correlation weights for the three runs of the Monte Carlo optimization process for best model for all splits is given in the Table S5 (Supplementary material).

According to presented results (Table S5) for split 3 (the best among all models) graph based SAk that have stabile positive values and therefore classified as promoters of IC_{50} increase are *EC1-1s2.18..*, *EC1-1s2.42..*, *EC1-1s2.63..*, *EC1-1s2.72..*, *EC1-2p2.18..*, etc. On the other hand SAk that have stabile negative values and therefore are promoters of IC_{50} decrease are *EC1-1s1.21..*, *EC1-1s2.27..*, *EC1-1s2.54..*, *EC1-2p2.36..*, *EC1-2p2.54..*, etc. Some of SMILES based SAk with stabile positive values and therefore promoters of IC_{50} increase are *++++F–S===*, *++++CL–N===*, *BOND10000000*, *N…+……*, *NOSP11100000*, *c…(……* and with stabile negative values and therefore promoters of IC_{50} decrease are *++++B2–B3==*, *++++O–B3==*, *C…#……*, *C…C…C…*, *F……*, *N……*.

Since models calculated with Eqs. (5)–(7) have the mechanistic interpretation the analysis of SMILES attributes given in Table S5

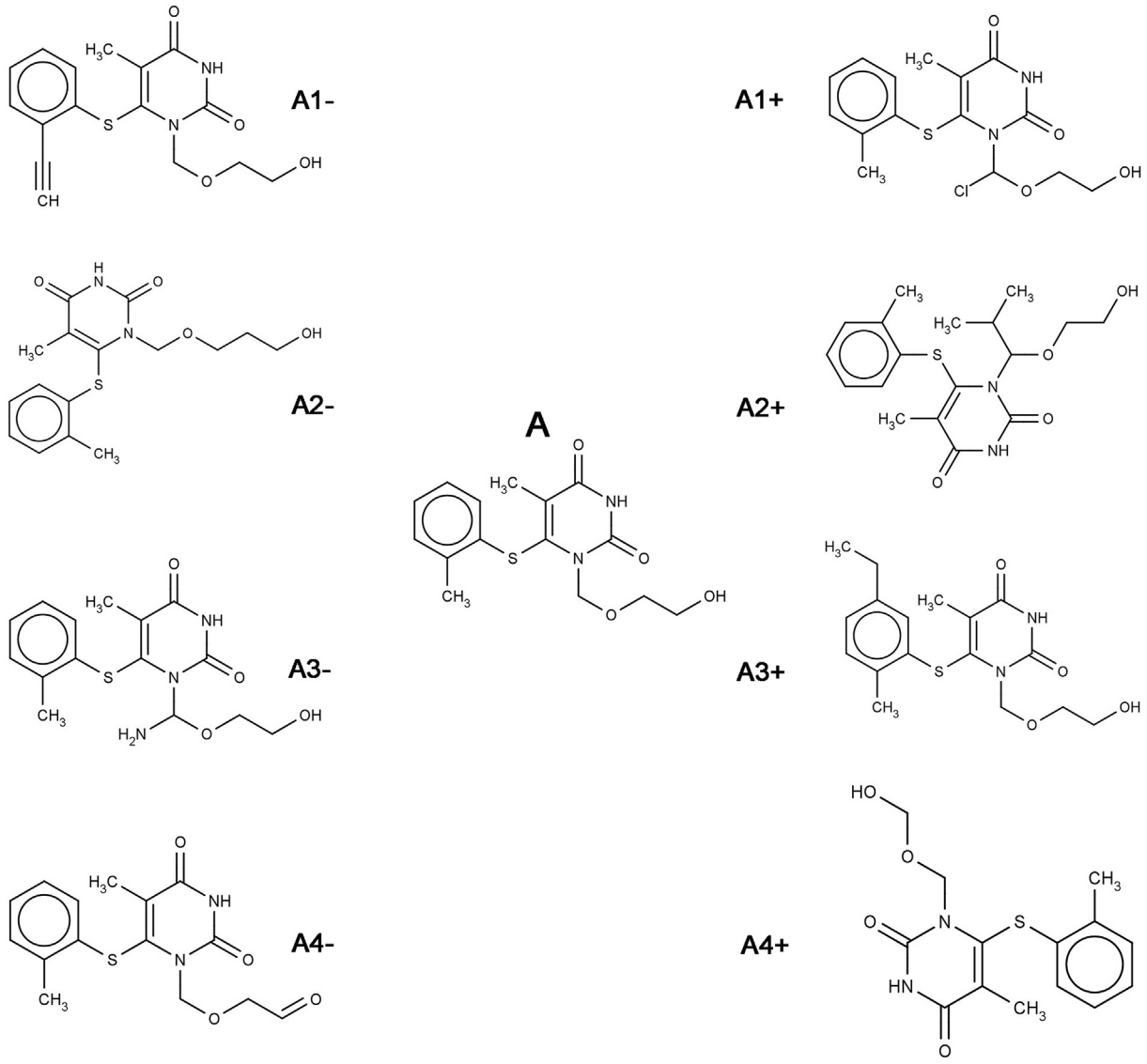

Fig. 3. The design of perspective anti-HIV-1 agents by means of the using of QSAR model calculated with Eq. (7); compound #1 as the basis for the computer aided design was used.

can be useful in searching for novel HEPT derivates with desired IC_{50}. Computer molecular design of new HEPT derivates is presented in Fig. 3. Starting molecule A (molecule 1 form Table S1) for all new molecules is already presented in Table S3 with detail calculation of DCW. When Eq. (7) is applied to calculated DCW, resulting IC_{50} value is 4.407. Results from Table S5 for SAk indicate that triple bond solo #……; triple bond with C atom *C*…#…… and triple bond between two C atoms *C*…#…*C*… have stabile negative value, therefore are promoters of IC_{50} decrease. Designed molecule A1- has addition of all previously stated SAk and the application of Eq (7) give result of 4.213 for IC_{50}. In molecule A2- SAk *C*…*C*… is substituted with *C*…*C*…*C*… which have stabile negative value. That SAk can be considered as three connected sp^3 carbon atoms. Calculated IC_{50} value for molecule A2- is 4.153. Molecule A3- has addition of SAk *N*…… with stabile negative value. That SAk can be considered as aliphatic N atom in amino group. Addition of amino group resulted with 4.207 value for IC_{50}. Molecule A4- has changed hydroxyl group to carbonyl. With this additional SAk = ……, *O*… = …… and *O*… = …*C*… are introduced within molecule A4- all with stabile negative values. Calculated value for IC_{50} is 4.092. When in molecule A SAk with stabile positive values is introduced, value for IC_{50} must be higher in comparison to value for IC_{50} of starting molecule A. Molecule A1+ has addition of chlorine atom SAK are *Cl*…… and HALO01000000, molecule A2+ has group with branching (SAk are *(*…*C*…*(*…, *C*…*(*…… and *C*…*(*…*C*…), molecule A3+ has additional group on benzene ring (SAk is *c*…*(*……). All SAk have stabile positive values. Molecule A4+ has reduced two connected carbon sp^3 atoms to one carbon sp^3 atom (molecule does not have SAk *C*…*C*… which has stabile negative value). Calculated values for IC_{50} for molecules A1+, A2+, A3+ and A4+ are 5.156, 7.519, 6.347 and 4.735 all higher in comparison to IC_{50} of molecule A.

4. Conclusion

QSAR models for a non-nucleoside reverse transcriptase inhibitor with HEPT derivates were build. Monte Carlo optimization process incorporated within The CORAL software is able to be an efficient tool to build up robust models with good statistical quality. The predictive potential of the applied approach was tested with three splits into the training and test set. The robustness of model was proven with different methods. The SMILES attributes, which

are promoters of IC_{50} increase/decrease are identified. The suggested modeling process for HEPT derivates as NNRTIs is based on the computational experiments with application of statistically stable structural alerts (promoters of increase or decrease of IC_{50}). This approach can be applied in the search for new potential anti-HIV-1 agents.

Acknowledgment

We would like to thank reviewers whose suggestions have improved our manuscript. This work has been financially supported by Ministry of Education and Science, Republic of Serbia, under Project Number 31060.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.ejmech.2014.03.013.

References

[1] R.A. Koup, V.J. Merluzzi, K.D. Hargrave, J. Adems, K.J. Grozinger, Inhibition of human immunodeficiency virus type 1 replication by the dipyridocliazepinone, The Journal of Infectious Diseases 163 (1991) 966–970.
[2] V.J. Merluzzi, K.D. Hargrave, M. Labadia, K. Grozinger, M. Skoog, J.C. Wu, C.-K. Shih, K. Eckner, S. Hattox, J. Adams, A.S. Rosehthal, R. Faanes, R.J. Eckner, R.A. Koup, J.L. Sullivan, Inhibition of HIV-1 replication by a nonnucleoside reverse transcriptase inhibitor, Science 250 (1990) 1411–1413.
[3] E. De Clercq, Mini-review: the role of non-nucleoside reverse transcriptase inhibitors (NNRTIs) in the therapy of HIV-1 infection, Antiviral Research 38 (1998) 153–179.
[4] E. De Clercq, Non-nucleoside reverse transcriptase inhibitors (NNRTIs) for the treatment of human immunodeficiency virus type 1 (HIV-1) infections: strategies to overcome drug resistance development, Medicinal Research Reviews 16 (1996) 125–157.
[5] R. Esnouf, J. Ren, C. Ross, Y. Jones, D. Stammers, D. Stuart, Mechanism of inhibition of HIV-1 reverse transcriptase by non-nucleoside inhibitors, Nature Structural & Molecular Biology 2 (1995) 303–308.
[6] 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, 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), Journal of Medicinal Chemistry 34 (1991) 349–357.
[7] H. Tanaka, M. Baba, M. Ubasawa, H. Takashima, K. Sekiya, I. Nitta, S. Shigeta, R.T. Walker, T. Miyasaka, Synthesis and anti-HIV activity of 2-, 3-, and 4-substituted analogues of 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio) thymine (HEPT), Journal of Medicinal Chemistry 34 (1991) 1394–1399.
[8] H. Tanaka, H. Takashima, M. Ubasawa, K. Sekiya, I. Nitta, M. Baba, S. Shigeta, T.R. Walker, E. De Clercq, 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, Journal of Medicinal Chemistry 35 (1992) 337–345.
[9] H. Tanaka, H. Takashima, M. Ubasawa, K. Sekiya, I. Nitta, M. Baba, S. Shigeta, T.R. Walker, E. De Clercq, 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, Journal of Medicinal Chemistry 35 (1992) 4713–4719.
[10] D. Kireev, J. Chrétien, D. Grierson, C. Monneret, A 3D QSAR study of a series of HEPT analogues: the influence of conformational mobility on HIV-1 reverse transcriptase inhibition, Journal of Medicinal Chemistry 40 (1997) 4257–4264.
[11] L. Hopkins, J. Ren, M. Esnouf, E. Willcox, Y. Jones, C. Ross, T. Miyasaka, T. Walker, H. Tanaka, K. Stammers, I. Stuart, Complexes of HIV-1 reverse transcriptase with inhibitors of the HEPT series reveal conformational changes relevant to the design of potent non-nucleoside inhibitors, Journal of Medicinal Chemistry 39 (1996) 1589–1600.
[12] W. Schaefer, G. Friebe, H. Leinert, A. Mertens, T. Poll, W. Van der Saal, H. Zilch, B. Nuber, L. Ziegler, Non-nucleoside inhibitors of HIV-1 reverse transcriptase: molecular modeling and X-ray structure investigations, Journal of Medicinal Chemistry 36 (1993) 726–732.
[13] R. Ragno, S. Frasca, F. Mannetti, A. Brizzi, S. Massa, HIV-reverse transcriptase inhibition: inclusion of ligand-induced fit by cross-docking studies, Journal of Medicinal Chemistry 48 (2005) 200–212.
[14] G. Campiani, A. Ramunno, G. Maga, V. Nacci, C. Fattorusso, B. Catalanotti, E. Morelli, E. Novellino, Non-nucleoside HIV-1 reverse transcriptase (RT) inhibitors: past, present, and future perspectives, Current Pharmaceutical Design 8 (2002) 615–657.
[15] C. Hansch, D. Hoekman, H. Gao, Comparative QSAR: toward a deeper understanding of chemicobiological interactions, Chemical Reviews 96 (1996) 1045–1076.
[16] J.M. Luco, F.H. Ferretti, QSAR based on multiple linear regression and PLS methods for the anti-HIV activity of a large group of HEPT derivatives, Journal of Chemical Information and Computer Science 37 (1997) 392–401.
[17] A. Bak, J. Polanski, 4D-QSAR study on anti-HIV HEPT analogues, Bioorganic & Medicinal Chemistry 14 (2006) 273–279.
[18] L. Douali, D. Villemin, D. Cherqaoui, Neural networks: accurate nonlinear QSAR model for HEPT derivatives, Journal of Chemical Information and Computer Science 43 (2003) 1200–1207.
[19] S. Gayen, B. Debnath, S. Samanta, T. Jha, QSAR study on some anti-HIV HEPT analogues using physicochemical and topological parameters, Bioorganic & Medicinal Chemistry 12 (2004) 1493–1503.
[20] M. Arakawa, K. Hasegawa, K. Funatsu, QSAR study of anti-HIV HEPT analogues based on multi-objective genetic programming and counter-propagation neural network, Chemometrics and Intelligent Laboratory Systems 83 (2006) 91–98.
[21] M. Jalali-Heravi, F. Parastar, Use of artificial neural networks in a QSAR study of anti-HIV activity for a large group of HEPT derivatives, Journal of Chemical Information and Computer Science 40 (2000) 147–154.
[22] C. Duda-Seiman, D. Duda-Seiman, M.V. Putz, D. Ciubotariub, QSAR modelling of anti-HIV activity with HEPT derivatives, Digest Journal of Nanomaterials and Biostructures 2 (2007) 207–219.
[23] D. Dana Weekes, G.B. Fogel, Evolutionary optimization, backpropagation, and data preparation issues in QSAR modeling of HIV inhibition by HEPT derivatives, Biosystems 72 (2003) 149–158.
[24] C.N. Alves, J.C. Pinheiroa, A.J. Camargob, M.M.C. Ferreirac, A.B.F. Da Silva, A structure–activity relationship study of HEPT-analog compounds with anti-HIV activity, Journal of Molecular Structure: THEOCHEM 530 (2000) 39–47.
[25] S. Hannongbua, K. Nivesanond, L. Lawtrakul, P. Pungpo, P. Wolschann, 3D-quantitative structure-activity relationships of HEPT derivatives as HIV-1 reverse transcriptase inhibitors, based on ab-initio calculations, Journal of Chemical Information and Computer Science 41 (2001) 848–855.
[26] W. Guo, X. Hu, N. Chu, C. Yin, Quantitative structure-activity relationship studies on HEPTs by supervised stochastic resonance, Bioorganic & Medicinal Chemistry Letters 16 (2006) 2855–2859.
[27] P. Pungpo, S. Hannongbua, P. Wolschann, Hologram quantitative structure-activity relationships investigations of non-nucleoside reverse transcriptase inhibitors, Current Medicinal Chemistry 10 (2003) 1661–1677.
[28] L. Akyüz, E. Sarıpınar, E. Kaya, E. Yanmaz, 4D-QSAR study of HEPT derivatives by electron conformational–genetic algorithm method, SAR and QSAR in Environmental Research 23 (2012) 409–433.
[29] M. Karelson, Molecular Descriptors in QSAR/QSPR, Wiley-Interscience, New York, 2000.
[30] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, Germany, 2000.
[31] P.R. Duchowicz, A. Talevi, L.E. Bruno-Blanch, E.A. Castro, New QSPR study for the prediction of aqueous solubility of drug-like compounds, Bioorganic & Medicinal Chemistry 16 (2008) 7944–7955.
[32] A.R. Katritzky, R. Petrukhin, D. Tatham, S.C. Basak, E. Benfenati, M. Karelson, U. Maran, Interpretation of quantitative structure property and activity relationships, Journal of Chemical Information and Computer Science 41 (2001) 679–685.
[33] Daylight Chemical Information Systems, Inc., 2008. http://www.daylight.com (accessed 10.05.13.).
[34] A.A. Toropov, A.P. Toropova, I. Raska Jr., QSPR modeling of octanol/water partition coefficient for vitamins by optimal descriptors calculated with SMILES, European Journal of Medicinal Chemistry 43 (2008) 714–740.
[35] A.A. Toropov, E. Benfenati, Additive SMILES-based optimal descriptors in QSAR modelling bee toxicity: using rare SMILES attributes to define the applicability domain, Bioorganic & Medicinal Chemistry 16 (2008) 4801–4809.
[36] A.A. Toropov, A.P. Toropova, A. Lombardo, A. Roncaglioni, E. Benfenati, G. Gini, CORAL: building up the model for bioconcentration factor and defining it's applicability domain, European Journal of Medicinal Chemistry 46 (2011) 1400–1403.
[37] A.M. Veselinović, J.B. Milosavljević, A.A. Toropov, G.M. Nikolí, SMILES-based QSAR models for arylpiperazines as high-affinity 5-HT_{1A} receptor ligands using CORAL, European Journal of Pharmaceutical Sciences 48 (2013) 532–541.
[38] A.M. Veselinović, J.B. Milosavljević, A.A. Toropov, G.M. Nikoli&cacute, SMILES-based QSAR models for the calcium channel antagonistic effect of 1,4-dihydropyridines, Archiv der Pharmazie 346 (2013) 134–139.
[39] A.A. Toropov, A.P. Toropova, E. Benfenati, R. Fanelli, The definition of the molecular structure for potential anti-malaria agents by the Monte Carlo method, Structural Chemistry 24 (2013) 1369–1381.
[40] A.P. Toropova, A.A. Toropov, B.F. Rasulev, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, QSAR models for ACE-inhibitor activity of tri-peptides based on representation of the molecular structure by graph of atomic orbitals and SMILES, Structural Chemistry 23 (2012) 1873–1878.
[41] A.A. Toropov, A.P. Toropova, S.E. Martyanov, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, Comparison of SMILES and molecular graphs as the representation of the molecular structure for QSAR analysis for mutagenic potential of polyaromatic amines, Chemometrics and Intelligent Laboratory Systems 109 (2011) 94–100.

[42] K. Roy, On some aspects of validation of predictive quantitative structure–activity relationship models, Expert Opinion on Drug Discovery 2 (2007) 1567–1577.
[43] P.P. Roy, J.T. Leonard, K. Roy, Exploring the impact of the size of training sets for the development of predictive QSAR models, Chemometrics and Intelligent Laboratory Systems 90 (2008) 31–42.
[44] A. Golbraikh, A. Tropsha, Beware of q2!, Journal of Molecular Graphics and Modelling 20 (2002) 269–276.
[45] P.P. Roy, K. Roy, QSAR studies of CYP2D6 inhibitor aryloxypropanolamines using 2D and 3D descriptors, Chemical Biology & Drug Design 73 (2009) 442–455.
[46] P.K. Ojha, I. Mitra, R.N. Das, K. Roy, Further exploring rm 2 metrics for validation of QSPR models, Chemometrics and Intelligent Laboratory Systems 107 (2011) 194–205.
[47] P.K. Ojha, K. Roy, Comparative QSARs for antimalarial endochins: importance of descriptor-thinning and noise reduction prior to feature selection, Chemometrics and Intelligent Laboratory Systems 109 (2011) 146–161.
[48] Organisation for Economic Co-operation and Development, Guidance Document on the Validation of (Quantitative) Structure–Activity Relationship [(Q) SAR] Models, in: OECD Series on Testing and Assessment 69, vol. 2, OECD Document ENV/JM/MONO, 2007, p. 2007.