Indian Journal of Chemistry
Vol. 59A, October 2020, pp. 1484-1493

In silico de novo design of NNRTIs of HIV-1: Functional group based computational molecular modelling approach

U Raghuvanshi & N S Sapre*

Department of Applied Chemistry, SGSITS, Indore, India

Email: sukusap@yahoo.com

Received 08 May 2019; revised and accepted 10 July 2020

Seven novel lead compounds, acting as NNRTIs of HIV-1, are extracted from a database of, *in silico de novo* designed, 500 compounds. Functional group based computational molecular modelling techniques are used for such design of Acylthiocarbamate derivatives. Effect of structural characteristics on the antiviral activity of these derivatives has also been studied. Statistical regression techniques namely, Non-linear (Back Propagation Neural Network, Support Vector Machine) and linear (Multiple Linear) chemometric regression methods are used in developing the relationships of Kier-Hall Electrotopological State Indices ($E_{RingA,}$ $E_{O8,}$ $E_{N9,}$ $E_{O14,}$ $E_{S16,}$ $E_{N17,}$ $E_{O19,}$ $E_{R,}$ and E_{R1}) with the HIV-1 antiviral activity. The relative potentials of these methods are also assessed and the results suggest that BPNN (r^2 = 0.845, MSE = 0.142, q^2 = 0.818) describes the relationship between the descriptors and antiviral activity in a relatively better manner than SVM-ε-radial (r^2 = 0.844, MSE = 0.144, q^2 = 0.807) and MLR (r^2 = 0.836, MSE = 0.150, q^2 = 0.805).

Keywords: Back Propagation Neural Networks (BPNN), *De Novo* Design, Molecular Modeling, Multiple Linear Regression (MLR), NNRTIs, Support Vector Machine (SVM)

In the voyage of clinical management of AIDS, NNRTIs play key role which help eradicate the infection caused by HIV-1[1-5]. As NNRTIs are impeded the conversion of single stranded viral RNA into double stranded pro-viral DNA in the HIV-1 life cycle at very initial stage. Despite the efficiency of NNRTIs, the genetic mutation in virus, toxicity, difficult treatment regimens, inadequate pharmacology (bioavailability and tissue distribution) and side effects of present medications, still confronts the journey of AIDS treatment[6-8]. To conquer these challenges there is an urgent need to develop innovative potent drug(s) with broad spectrum of pharmacokinetic profile that are able to provide higher genetic barrier to resistance and reduced safety problems.

From past few decades computational modelling techniques have been established as valuable tools in assisting new drug discovery process[9-11]. These are relatively less expensive techniques, which speed up the drug discovery process and help in producing novel and potent molecules with desired biological activity. 2D/3D-QSAR/QSPR, molecular docking, virtual screening etc. are some of the common computational modelling techniques used in drug development and discovery[12-14].

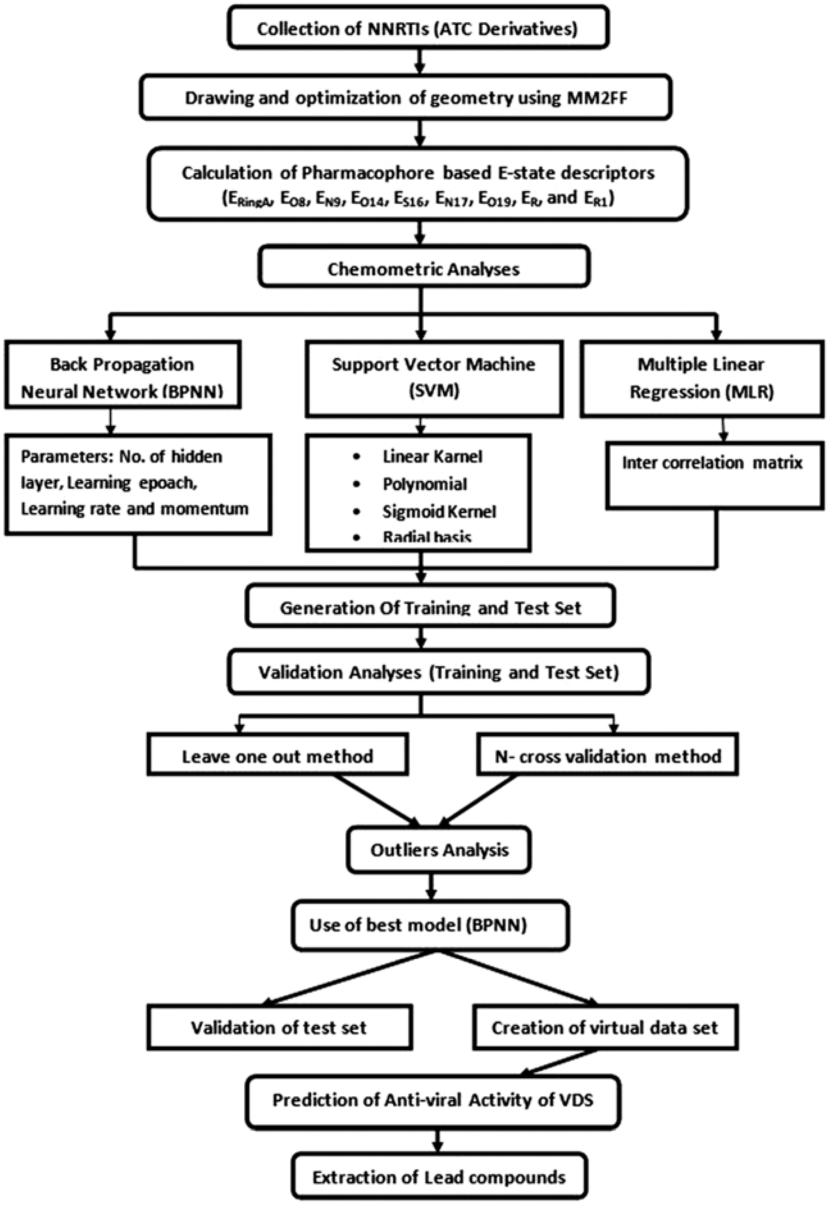
Materials and Methods

Molecular Dataset and Computational methods

In the present work *de novo* design of novel NNRIs of HIV-1 is carried out using functional group based computational molecular modelling techniques using 3D-Kier-Hall Electrotopological state (E-state) indices. Also, work is performed with an additional goal to get insight into the effect of structural characteristics on the antiviral activity of a dataset of 78 Acylthiocarbamate (ATC) derivatives, a diverse class of compounds acting as a NNRTIs of HIV-1[15-19]. The structures are drawn and optimized using ChemDraw Ultra version 7.0.0 and Chem3D Ultra version 7.0.0 respectively[20]. Kier-Hall E-state indices for various functional groups are calculated using Toxicity Estimation Software Tool[21]. Molegro Data Modeller tool of the Molegro Virtual Docker software 2.6.0 is used for regression analyses and deriving correlation of E-state indices with the antiviral activity (pEC_{50}, in μM terms) [22]. The flow of work is presented as Scheme 1.

Kier-Hall electrotopological state (E-State) indices

The Kier-Hall Electrotopological state indices (E-State) are atom level descriptors encoding both the electronic character and topological environment of each skeletal atom. They are formulated using intrinsic value I_i


Collection of NNRTIs (ATC Derivatives)
Drawing and optimization of geometry using MM2FF
Calculation of Pharmacophore based E-state descriptors
(E_{RingA}, E_{O8}, E_{N9}, E_{O14}, E_{S16}, E_{N17}, E_{O19}, E_R, and E_{R1})
Chemometric Analyses
Back Propagation Neural Network (BPNN)
Support Vector Machine (SVM)
Multiple Linear Regression (MLR)
Parameters: No. of hidden layer, Learning epoach, Learning rate and momentum
Linear Karnel
Polynomial
Sigmoid Kernel
Radial basis
Inter correlation matrix
Generation Of Training and Test Set
Validation Analyses (Training and Test Set)
Leave one out method
N- cross validation method
Outliers Analysis
Use of best model (BPNN)
Validation of test set
Creation of virtual data set
Prediction of Anti-viral Activity of VDS
Extraction of Lead compounds

A scheme presenting the flow of work

Scheme 1

and a perturbation term ΔI_i, arising from the electronic interactions within the molecular topological environment of each atom in molecule[23-31]. Using E-State descriptors one can demonstrate structural specificity of a molecule at an atomic or fragmental level.

Molecular modeling and chemometric analyses

A functional group based 3D-quantitative structure activity relationship (3D-QSAR) is developed using the Kier-Hall Electrotopological indices. E-state indices for various functional groups are calculated. Non-linear (BPNN and SVM) and linear (LR and MLR) regression methods are used in deriving the relationships and understanding the correlation potential of the methods. The potential of the Kier-Hall E-state indices and structural attributes responsible for affecting biological activity of the molecules are studied. A variety of chemometric methods are used for handling multivariate data and are responsible for reliable QSAR interpretations[32-34]. A brief account of chemometric methods used in the study is presented herewith.

Back propagation neural network (BPNN)

Neural networks resemble human brain neuron network and can handle complex and non-linear data and thus extract the hidden relationships between the dependent and independent variables[35]. Rumelhart *et al*[36]., developed the Back-Propagation Neural Network (BPNN) as a solution to the problem of training multi-layer perceptrons[37-40].

Support vector machine (SVM)

SVM is based on the structural risk minimization (SRM) principle which is least sensitive to data over fitting[41]. SVM techniques are introduced by Boser, Guyon and Vapnik[42]. This method can be applied to linear as well as nonlinear classification and are trained faster[43]. SVM has been successful in correlating various quantitative structure activity/property relationships in the areas of computer-aided drug design methods[44-48]. It is a supervised learning method and support vectors are used with suitable kernel functions. For the present study ν- and ε-support vector regressions based on LIBSVM are considered and in each case linear, polynomial, sigmoid, and radial basis functions are used.

Multiple linear regressions (MLR)

Multiple linear regression (MLR) is a method where the values of the regression coefficients (bn's) are evaluated using least squares curve fitting method[49,50].

$$y = b_1x_1 + b_2x_2 + b_3x_3 + \wedge\wedge\, b_n\, x_n + c \qquad \ldots(1)$$

Where, 'y' is the dependent variable, 'x_1, x_2 .... x_n' are the independent variables, 'b_1, b_2 .... b_n' are the regression coefficients and 'c' is the intercept on Y axis and is constant.

This is the most widely used method owing to its fast and easy interpretability. However, for complex systems, such as a biological system, the linear combination of descriptor information can often lead to a model with limited accuracy, simply due to the assumption of linearity in the data.

Results and Discussion

Supplementary Data, Table S1 records the structure of 78 ATC derivatives along with the position of substituents, antiviral activity ($pEC_{50,}$ in μM terms) and the E-state indices (E_{RingA}, E_{O8}, E_{N9}, E_{O14}, E_{S16}, E_{N17}, E_{O19}, E_R, and E_{R1}). The dataset is split into a training set (n=53) and test set (n=19). BPNN, SVM and MLR techniques are used for generating regression models to establish correlation between the E-state indices (descriptors) and antiviral activity (pEC_{50}), thereby establish the effect of substitution on the activity. The results thus obtained are then used for generating a virtual dataset (VDS) of ATC analogues. Using the best model thus generated, the pEC_{50} of VDS is estimated and compounds exhibiting high anti-viral activity are extracted.

Correlation analyses

To assess the effect of substituents, E-state descriptors are correlated with the antiviral activity, generating various nonlinear and linear-regression models. Uni-variate, bi-variate and multi-variate models are generated for assessing the referred potential.

Univariate correlation

A univariate structure activity relationship is developed between the E-state values and antiviral activity for the training set. The impact of individual substituent on the antiviral activity is determined using linear equation expressed as pEC_{50}= bX+c, where X is the independent variable (descriptor), 'b' is the coefficient and 'c' is the constant. The results of correlations and impact of descriptors (substituent) are presented in Table 1. From this table it is observed that the relationship of descriptors with the activity shows following order of correlation (assessed in terms of correlation coefficient, r^2): E_{R1}(0.135) > E_{RingA} (0.075) > E_{N9}(0.069) > E_{O14}(0.055) > E_{S16}(0.054) > E_{N17}(0.044)> E_R (0.008) > E_{O19}(0.001) = E_{O8}(0.00). While the effect (impact) of each descriptor on the antiviral activity is expressed in terms of the coefficient of respective descriptors and is referred as Impact Coefficient (IC). It follows the following order: E_{N9}(−6.114) > E_{O14}(−3.224) > E_{S16}(−3.184) > E_{N17}(−1.415) > E_{RingA} (−0.968) > E_{O8}(−0.319) > E_{O19}(0.146) > E_{R1}(0.040) > E_R (−0.023).

Table 1 — The univariate correlation (r^2) and impact (IC) coefficients of e-state descriptors with pEC_{50} (μM) and linear equation for ATC analogues (Training set)

Descriptor	r^2	Impact Coefficient (IC)	Equation
E_{RingA}	0.075	-0.9689	pEC_{50} = -0.9689 * E_{RingA} + 14.54
E_{O8}	0.000	-0.3197	pEC_{50} = -0.3197 * E_{O8} + 11.28
E_{N9}	0.069	-6.1149	pEC_{50} = -6.1149 * E_{N9} + 13.86
E_{O14}	0.055	-3.2248	pEC_{50} = -3.2248 * E_{O14} + 25.41
E_{S16}	0.054	-3.1841	pEC_{50} = -3.1841 * E_{S16} + 24.40
E_{N17}	0.044	-1.4155	pEC_{50} = -1.4155 * E_{N17} + 8.90
E_{O19}	0.001	0.1465	pEC_{50} = 0.1465 * E_{O19} + 5.33
E_R	0.008	-0.0232	pEC_{50} = -0.0232 * E_R + 7.58
E_{R1}	0.135	0.0409	pEC_{50} = 0.0409 * E_{R1} + 6.58

It is observed that there is no direct association between impact coefficient(IC) and linear correlation coefficient (r^2). From the values of impact coefficient it is observed that E_{N9} has highest, though in a highly retarding manner, while E_{O19} has a little enhancing and other substitutents impart moderate to low, impacts on anti viral activity. To assess interrelationship between individual descriptors and antiviral activity numerous nonlinear analyses are performed. The relative potential of each descriptor on the antiviral activity (in terms of impact coefficient of each descriptor) using univariate regression technique is presented in Fig. 1.

Back propagation neural network (BPNN) analyses

To assess the interdependence and relative level (relevance score) of effect of the E-state descriptors on antiviral activity of ATC derivatives, Back Propagation Neural Network (BPNN) analyses are performed. The following parameters are set to train the network: maximum training epoch = 10000, learning rate = 0.30, output layer learning rate = 0.30, momentum = 0.20, data range normalization = 0.1-0.9, number of neurons = 1, and initial weight = ±0.50. Leave-one-out (LOO) and N-cross validation (N-CV) methods are used in validating the results. The results obtained suggest that LOO method performed better with the higher r^2 and lower MSE than other validation methods for BPNN.

The order of relevance score of correlation for E-state descriptors with antiviral activity is as follows:

$E_{S16}(100) > E_{O14}(81) > E_{R1}(69) > E_{N9}(67) > E_{O8}(61) > E_{RingA}(58) > E_{N17}(54) > E_R(39) > E_{O19}(13)$.

Assessment of above order indicates that E_{S16} has highest impact while the E_{O19} has the lowest impact on antiviral activity. The other E-state descriptors namely E_{O14}, E_{R1}, E_{N9}, E_{O8}, $E_{RingA,}$ E_{N17}, E_R have moderate to low impact on antiviral activity. The relevance score for each descriptor, as estimated using BPNN technique, is presented in the Fig. 2.

Multiple linear regression (MLR) analyses

MLR is performed to evaluate relative potential of E-state descriptors on the antiviral activity of ATC derivatives in multivariate linear terms. The best model for relating the descriptor values with antiviral activity (pEC_{50}) derived using MLR method is presented in Eqn (2):

$$\begin{aligned} pEC_{50} = {} & 75.4858(\pm 21.3577)\mathrm{E}_{\mathrm{RingA}} \\ & + 157.523(\pm 5.96248)\mathrm{E}_{O8} \\ & - 298.35(\pm 12.8421)\mathrm{E}_{\mathrm{N9}} \\ & - 964.786(\pm 70.4247)\mathrm{E}_{O14} \\ & + 681.255(\pm 49.6534)\mathrm{E}_{\mathrm{S16}} \\ & + 75.9518(\pm 11.3036)\mathrm{E}_{\mathrm{N17}} \\ & + 3.62796(\pm 0.692889)\mathrm{E}_{\mathrm{O19}} \\ & + 0.23869(\pm 0.947825)\mathrm{E}_{\mathrm{R}} \\ & + 0.235618(\pm 2.11682)\mathrm{E}_{\mathrm{R1}} \\ & - 590.445 \qquad \ldots(2) \end{aligned}$$

(n=53 r^2=0.8368 r^2adj=0.8027 spearman rho=0.7652 MSE=0.1501).

Eqn (2) is exhibiting the coefficients for all the functional groups (presented in the form of E-State descriptors). This suggests that these can be divided in to three categories: (a) activity enhancing, (b) activity retarding and (c) moderately affecting.

The high positive coefficients of "E_{S16}", "E_{O8}", "E_{N17}" and " E_{RingA} " suggest that these functional groups impart an activity enhancing effect on the antiviral activity while high negative coefficients for "E_{O14}" and "E_{N9}" indicate that presence of bridging oxygen atom at '14' position and nitrogen atom in the five member ring will adversely affect the antiviral activity and thus will retard the activity. Though, the coefficients of E_{O19}, E_R and E_{R1} are positive but are

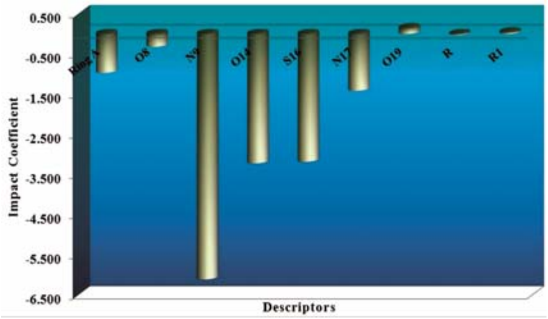


Fig. 1 — Bar chart presenting the impact of each E-state index on anti-HIV-1 activity of ATC derivatives as obtained from univariate linear analysis.

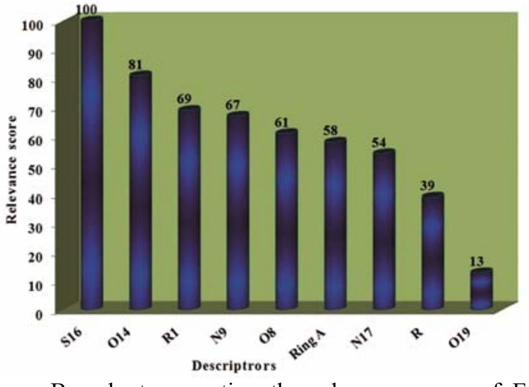


Fig. 2 — Bar chart presenting the relevance score of E-state indices of ATC derivatives as obtained from BPNN analysis.

very low and thus their activity enhancing impact on the antiviral activity will be to a lower extent and thus can be classified as moderately affecting groups.

Support vector machine regression (SVM) analyses

ε-support vector regression and ν-support vector regression with variable kernels [linear (SVM-LK), polynomial (SVM-PK), sigmoid (SVM-SK), and radial basis function (SVM-RBFK)] are considered and eight models are generated using a random seed 3485805689. Optimal parameter settings are fine-tuned and accordingly results are obtained. The following parameters, Cost: 100000, Gamma= 0.0003, Epsilon (ε): 0.001/ Nu (ν): 0.5, Termination criterion tolerance: 0.01 are chosen for performing the regression analyses. It is observed that the radial basis function kernel performs best, followed by polynomial and linear kernels in ε and ν techniques both. In all the cases the correlation coefficients are comparable. In either case the sigmoid kernels perform poorly. The results obtained using SVM method can be attributed to non-linearity among the various parameters and also signifies the robustness of the derived models.

Comparative Analyses

On comparison of the results obtained from the three methods it is observed that BPNN regression method show highest correlation potential followed by SVM regression methods (ε-RBFK and ν-RBFK) while the MLR method shows lowest. Table 2 presents the comparative analyses of all the three methods.

Cross validation with test set

To validate the QSAR models thus obtained from BPNN, SVM and MLR methods a test set of 19 compounds is constructed. The regression models represent a good harmony between calculated and predicted pEC_{50} values. Like the training set, magnitude of squared regression coefficient is higher for BPNN (r^2= 0.805), MLR (r^2= 0.604) as compared to SVM (r^2= 0.575) which indicates BPNN is the best method to be used for further assessment than MLR as well as SVM. Table 3 presents the observed and calculated pEC_{50} values for the training and test sets of ATC derivatives using MLR, BPNN and SVM methods.

Table 2 — Comparative analyses of models build by multiple linear regression (MLR), back propagation neural network (BPNN) and support vector machine (SVM) techniques (Training set)

S.No.	Model	K	r^2	r^2adj	rho (ρ)	PRESS	MSE	q^2
MLR								
1	MS	9	0.836	0.802	0.765	-	0.150	0.805
2	LOO	9	0.731	0.675	0.651	13.968	0.258	0.699
3	NCV (N=10)	9	0.746	0.693	0.661	13.143	0.247	0.722
BPNN								
1	MS	9	0.845	-	0.743	-	0.142	0.818
2	LOO	9	0.774	-	0.666	10.937	0.206	0.729
3	NCV (N=10)	9	0.785	-	0.692	10.517	0.198	0.745
SVM(ε-radial) 53SV: RBFK								
1	MS	9	0.845	-	0.806	-	0.144	0.807
2	LOO	9	0.701	-	0.666	-	0.277	0.620
3	NCV (N=10)	9	0.734	-	0.681	-	0.244	0.649
SVM(ε-polynomial) 53SV:PK								
1	MS	9	0.840	-	0.803	-	0.147	0.803
2	LOO	9	0.730	-	0.707	-	0.248	0.653
3	NCV(N=10)	9	0.750	-	0.716	-	0.230	0.672
SVM(ε-sigmoid) 53SV:SK								
1	MS	9	0.759	-	0.771	-	0.221	0.678
2	LOO	9	0.604	-	0.624	-	0.364	0.330
3	NCV (N=10)	9	0.557	-	0.533	-	0.407	0.236
SVM(ε-linear) 53SV:LK								
1	MS	9	0.826	-	0.745	-	0.161	0.766
2	LOO	9	0.730	-	0.710	-	0.279	0.717
3	NCV (N=10)	9	0.653	-	0.618	-	0.358	0.620

(*Contd*.)

Table 2 — Comparative analyses of models build by multiple linear regression (MLR), back propagation neural network (BPNN) and support vector machine (SVM) techniques (Training set) (*Contd.*)

S.No.	Model	K	r^2	r^2adj	rho (ρ)	PRESS	MSE	q^2
SVM(ν-radial) 29SV:RBFK								
1	MS	9	0.845	-	0.813	-	0.156	0.738
2	LOO	9	0.738	-	0.727	-	0.241	0.638
3	NCV (N=10)	9	0.739	-	0.730	-	0.242	0.590
SVM(ν-plynomial) 29SV:PK								
1	MS	9	0.842	-	0.804	-	0.157	0.737
2	LOO	9	0.743	-	0.717	-	0.236	0.637
3	NCV(N=10)	9	0.748	-	0.754	-	0.236	0.584
SVM(ν-sigmoid) 29SV:SK								
1	MS	9	0.807	-	0.748	-	0.197	0.621
2	LOO	9	0.675	-	0.655	-	0.307	0.373
3	NCV (N=10)	9	0.700	-	0.669	-	0.292	0.371
SVM(ν-linear 29SV:LK								
1	MS	9	0.829	-	0.743	-	0.160	0.812
2	LOO	9	0.741	-	0.671	-	0.249	0.715
3	NCV (N=10)	9	0.754	-	0.675	-	0.241	0.736

MS = Manual Selection, RBFK = Radial Basis Function Kernel, PK = Polynomial Kernel, LK = Linear Kernel. 'k' is the no. of descriptors, 'r^2' is the correlation coefficient, 'q^2' is cross validated 'r^2' from the (LOO) and N-CV procedures, rho (ρ) is the Spearman rank correlation coefficient, MSE is the mean squared error and PRESS is the predictive sum of squares.

Table 3 — Observed and calculated values of pEC_{50} (in μM terms) for the training and test sets of ATC derivatives using MLR, BPNN and SVM Techniques

S.No.	Comp no.	pEC_{50} (μM)	MLR	BPNN	SVM ε:RBFK	SVM ε:PK	SVM ε:SK	SVM ε:LK	SVM ν:RBFK	SVM ν:PK	SVM ν:SK	SVM ν:LK
1.	1[a]	6.40	7.41	7.61	7.42	7.43	8.03	7.17	7.36	7.39	7.78	7.29
2.	2	4.96	5.44	5.26	5.56	5.61	5.71	5.62	5.45	5.45	5.87	5.34
3.	3	4.96	4.49	4.49	4.88	4.92	4.29	4.87	4.66	4.66	4.83	4.42
4.	4	4.96	5.64	5.60	5.75	5.76	6.42	5.56	5.87	5.89	6.27	5.45
5.	5	4.96	4.71	4.62	4.95	4.96	4.96	4.88	5.02	5.04	5.21	4.57
6.	6[a]	5.00	6.49	6.70	6.57	6.57	6.81	6.50	6.41	6.41	6.79	6.38
7.	7[a]	4.96	7.33	6.39	7.37	7.16	7.26	7.80	7.05	7.23	6.79	7.17
8.	8[a]	5.22	6.56	6.22	6.57	6.53	6.19	6.93	6.36	6.35	6.30	6.47
9.	9[a]	4.96	7.37	7.10	7.23	7.17	8.10	7.27	7.21	7.22	7.55	7.14
10.	10[a]	5.22	6.81	6.86	6.75	6.74	7.27	6.77	6.68	6.67	7.06	6.64
11.	11[a]	5.92	6.62	6.76	6.63	6.63	6.97	6.62	6.51	6.50	6.89	6.49
12.	12	6.42	6.97	6.46	6.92	6.85	6.95	7.30	6.75	6.78	6.71	6.84
13.	13[a]	5.46	6.57	6.50	6.58	6.55	6.52	6.74	6.44	6.44	6.57	6.45
14.	14	7.52	7.47	7.62	7.25	7.15	7.36	7.49	7.20	7.20	7.28	7.30
15.	15	7.00	7.64	7.53	7.59	7.57	8.30	7.57	7.40	7.37	7.78	7.47
16.	16	7.00	7.44	7.46	7.40	7.39	7.64	7.44	7.31	7.31	7.49	7.35
17.	17	7.60	7.74	7.76	7.70	7.69	7.79	7.78	7.64	7.64	7.70	7.74
18.	19	8.10	7.36	7.44	7.30	7.29	7.77	7.31	7.13	7.11	7.51	7.22
19.	21	8.22	7.82	7.80	7.71	7.69	7.86	7.85	7.68	7.68	7.74	7.81
20.	22	7.46	7.31	7.43	7.33	7.32	7.46	7.27	7.22	7.23	7.40	7.22
21.	23	8.00	7.20	7.40	7.19	7.18	7.51	7.13	7.01	7.00	7.37	7.06
22.	24	8.10	7.78	7.86	8.06	7.96	8.10	7.92	7.76	7.76	7.61	7.68
23.	25	8.00	7.76	7.76	8.00	7.97	7.95	7.98	7.61	7.61	7.58	7.71
24.	28	7.40	6.88	6.95	7.09	7.12	7.12	6.81	7.06	7.10	7.15	6.84
25.	29	7.15	7.39	7.46	7.16	7.16	7.30	7.39	7.21	7.24	7.37	7.45

(*Contd.*)

Table 3 — Observed and calculated values of pEC$_{50}$ (in μM terms) for the training and test sets of ATC derivatives using MLR, BPNN and SVM Techniques (*Contd.*)

S.No.	Comp no.	pEC$_{50}$ (μM)	MLR	BPNN	SVM ε:RBFK	SVM ε:PK	SVM ε:SK	SVM ε:LK	SVM ν:RBFK	SVM ν:PK	SVM ν:SK	SVM ν:LK
26.	30	8.00	7.42	7.45	7.51	7.50	7.47	7.39	7.42	7.45	7.46	7.37
27.	31	8.82	8.10	8.09	8.01	8.02	7.95	8.07	7.88	7.87	7.96	8.17
28.	32	7.30	7.86	7.70	7.79	7.77	7.73	7.86	7.67	7.69	7.69	7.77
29.	33	8.30	8.44	8.43	9.29	9.40	8.89	8.21	8.66	8.64	8.73	8.70
30.	34[a]	7.22	8.38	8.31	8.58	8.63	8.48	8.29	8.37	8.34	8.38	8.52
31.	35	7.46	7.67	7.69	7.66	7.64	7.56	7.73	7.56	7.58	7.59	7.67
32.	36	6.70	7.05	7.13	6.88	6.86	7.17	7.00	7.03	7.07	7.18	6.98
33.	37	6.70	6.70	6.87	6.99	7.01	6.82	6.65	6.92	6.97	6.99	6.68
34.	38[a]	6.40	7.22	7.42	6.79	6.77	7.02	7.21	6.96	6.99	7.22	7.29
35.	39	7.40	7.37	7.40	7.41	7.40	7.18	7.35	7.29	7.33	7.31	7.29
36.	40	7.70	7.92	8.07	7.67	7.66	7.67	7.86	7.65	7.64	7.81	8.00
37.	41	7.70	7.69	7.65	7.68	7.64	7.43	7.70	7.53	7.56	7.53	7.62
38.	42	7.52	7.74	8.09	7.75	7.72	7.64	7.51	7.67	7.65	7.86	7.90
39.	43	8.10	7.68	7.60	7.90	7.87	7.69	7.66	7.68	7.75	7.51	7.61
40.	44	8.00	7.68	7.81	7.97	7.96	7.83	7.57	7.75	7.80	7.64	7.62
41.	45	7.70	7.68	7.60	7.90	7.87	7.69	7.66	7.68	7.75	7.51	7.61
42.	46	7.70	7.41	7.25	7.74	7.69	7.50	7.33	7.58	7.69	7.29	7.27
43.	47	8.10	7.87	8.00	8.18	8.20	8.01	7.86	7.89	7.89	7.86	7.95
44.	48	8.15	7.95	7.85	8.09	8.08	7.87	8.00	7.80	7.83	7.73	7.96
45.	49	7.00	7.19	7.20	6.99	6.98	7.20	7.19	7.16	7.18	7.26	7.19
46.	50	7.52	7.40	7.50	7.27	7.26	7.42	7.43	7.38	7.40	7.45	7.44
47.	51[a]	5.25	6.73	6.67	6.56	6.56	6.92	6.64	6.83	6.88	6.98	6.69
48.	52	8.05	7.93	7.82	7.71	7.69	7.72	7.93	7.71	7.72	7.73	7.91
49.	53	8.10	7.66	7.61	7.44	7.42	7.51	7.70	7.50	7.51	7.55	7.67
50.	54	6.40	6.74	6.59	6.47	6.44	6.53	6.70	6.71	6.74	6.90	6.81
51.	55	8.10	8.03	8.23	8.09	8.13	8.11	7.94	7.98	7.96	8.12	8.21
52.	56[a]	7.15	7.77	7.94	7.53	7.53	7.58	7.73	7.48	7.48	7.70	7.87
53.	57	7.70	8.41	8.22	8.25	8.25	8.10	8.41	8.06	8.04	8.17	8.57
54.	58	8.15	7.88	7.71	7.83	7.81	7.58	7.90	7.74	7.76	7.66	7.85
55.	60	7.52	7.47	7.77	7.53	7.50	7.55	7.42	7.47	7.52	7.49	7.49
56.	61	7.40	7.49	7.55	7.43	7.38	7.41	7.43	7.39	7.46	7.36	7.43
57.	62[a]	6.00	7.23	7.18	7.22	7.14	7.20	7.15	7.25	7.36	7.13	7.10
58.	63	7.30	7.74	7.81	7.64	7.62	7.58	7.73	7.52	7.55	7.58	7.75
59.	64	6.70	7.10	7.14	6.78	6.75	6.92	7.04	6.97	7.01	7.11	7.03
60.	65[a]	6.52	7.21	7.45	7.06	7.03	7.13	7.19	7.19	7.22	7.29	7.25
61.	66[a]	5.05	6.58	6.58	6.33	6.32	6.63	6.52	6.64	6.69	6.83	6.56
62.	67	7.30	7.73	7.78	7.51	7.47	7.43	7.72	7.53	7.54	7.58	7.73
63.	69[a]	7.05	7.84	8.22	7.73	7.74	7.82	7.72	7.73	7.72	7.97	8.02
64.	70[a]	7.22	8.22	8.20	7.87	7.84	7.81	8.21	7.80	7.78	8.01	8.39
65.	71	7.30	7.72	7.68	7.74	7.70	7.30	7.73	7.61	7.64	7.52	7.70
66.	72[a]	6.30	7.88	7.92	7.75	7.75	8.11	7.77	7.81	7.84	8.02	7.91
67.	73[a]	6.46	8.34	8.26	8.33	8.35	8.49	8.20	8.21	8.23	8.45	8.55
68.	74	7.52	7.42	7.29	7.53	7.51	6.90	7.44	7.40	7.43	7.30	7.49
69.	75	7.70	7.26	7.27	7.50	7.49	6.87	7.14	7.38	7.41	7.28	7.31
70.	76	5.59	5.60	5.55	5.70	5.63	6.17	5.59	5.93	5.96	6.02	5.38
71.	77	4.77	4.64	4.96	4.78	4.79	4.84	4.75	5.09	5.12	5.19	4.65
72.	78	6.16	6.79	6.83	6.70	6.61	7.02	6.86	6.73	6.73	6.81	6.61

[a]Compounds in Test set

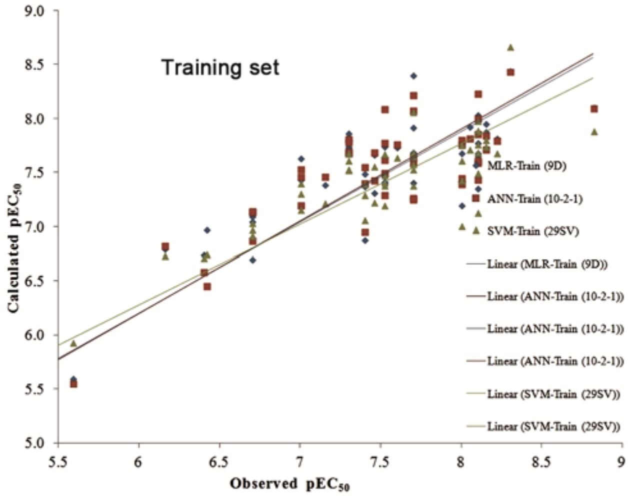


Fig. 3 — A graph of comparative analyses between observed and calculated pEC_{50} using MLR, ANN and SVM techniques for training set of ATC derivatives.

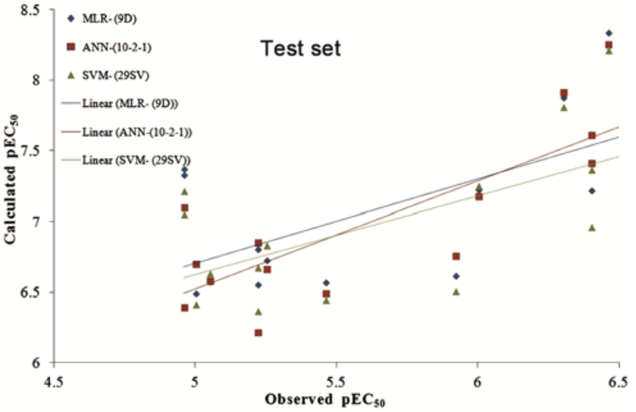


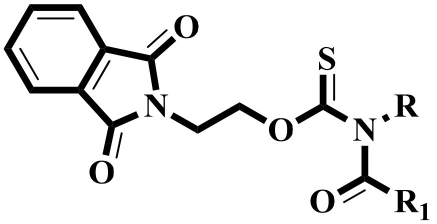
Fig. 4 — A graph of comparative analyses between observed and calculated pEC_{50} using MLR, ANN and SVM techniques for test set of ATC derivatives.

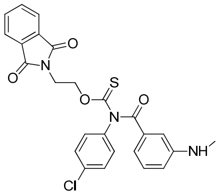
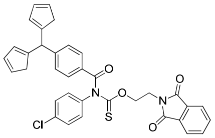
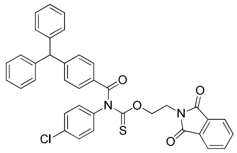
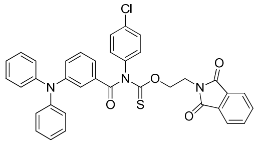
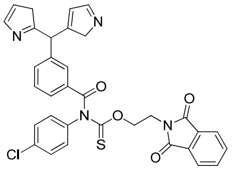
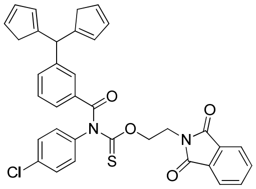
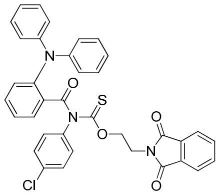
Figs 3 and 4 give a graphical representation of relation of observed and predicted pEC_{50} values for the training and test sets respectively using all the three chemometric, namely MLR, BPNN and SVM, methods.

Virtual dataset

A virtual dataset (VDS) of 500 compounds, by making fragmental changes on the template ATC structure, is created. The antiviral activity of these compounds is predicted using the best derived BPNN model. A set of 26 virtual compounds was found to have predicted anti-viral activity (pEC_{50}) above 8.00μM. Table S2 (Supplementary Data) presents the structural data of all the virtual compounds exhibiting antiviral activity greater than 8.00 μM. Of these 26 compounds 7 compounds have shown predicted antiviral activity greater than 8.5 μM and are given in Table 4.

Table 4 — Structure of lead compounds (Virtual Data Set, pEC_{50} > 8.5 μM) with their predicted pEC_{50}



S.No.	Comp. No.	Structure	Predicted pEC_{50} (μM)
1	VDS01		8.56
2	VDS02		8.55
3	VDS03		8.55
4	VDS04		8.55
5	VDS05		8.53
6	VDS06		8.53
7.	VDS07		8.51

Conclusions

From the present study on following conclusions can be made: The QSAR studies show that structural characteristics of ATC derivatives strongly affect their antiviral activity. Four analyses namely Univariate linear regression (ULR), Multiple linear regression (MLR), support vector machine (SVM) and Back Propagation Neural Network (BPNN) have been performed. The better performance of BPNN model over MLR and SVM models is suggestive of the fact that there exists a non-linear relationship between the independent (descriptors) and dependent (antiviral activity) variables. The univariate correlation was performed solely with an aim to understand potential of descriptors in individual capacity in affecting the antiviral activity. In the case of univariate linear correlation, E_{R1} showed highest while E_{O8} and E_{O9} have lowest linear relationship with an antiviral activity. From the results, higher value of correlation coefficient (r^2) for BPNN and SVM but low value for MLR, it is concluded that there is a nonlinear relationship between E-state descriptors and antiviral activity. In all the regression methods (BPNN, SVM and MLR) manual selection method performed better than leave-one-out (LOO) and N-cross validation method. A test set was used for cross validation of derived model. The results of the test set are encouraging which proved the robustness of models. Six compounds (18, 20, 26, 27, 29 and 68) were observed as outliers due to a vast difference in the observed and calculated antiviral activity due to diverse structural features.

The most significant conclusions of the present study are following: Results suggested that the presence of “S” (sulphur) and bridging “O” (oxygen) atoms present at position ‘16’ and ‘14’ are beneficial for antiviral activity. The model also indicates that presence of “N” (nitrogen) atom and double bond oxygen atom (of pyrrolidine-2,5-dione ring) at positions '9' and '8' respectively in the parent compound is favourable for the antiviral activity. Substitution at position “R1” is more beneficial for antiviral activity than at position “R”. Other descriptors such as “RingA”, tri-substituted amine at position '17', and carbonyl oxygen group at position '19' exhibit a low impact on the antiviral activity. The virtual dataset designed on the basis of aforementioned observations yielded 26 compounds with high biological activity profile which suggest that structural attributes of derivatives play a crucial role in the field of molecular modeling.

For the virtual dataset compounds it is observed that presence of 3-methylaniline phenyl at position ‘R1’ along with 4-chloro benzene group at R position on parent structure produced highest activity enhancing effect (VDS01). Similarly presence of bulky groups such as 4-(dicyclopenta-1,3-dien-1yl)methyl benzene(VDS02), 4-(methylene) dibenzene (VDS03), 3-(diphenylamino) benzene (VDS04), 2-((2H-pyrrol-3-yl)(3H-pyrrol-4-yl)methyl-benzene (VDS05), 3-(dicyclopenta-1,3-diene-1yl) methylbenzene (VDS06), 2-(diphenylamino) benzene (VDS07) groups at position ‘R1’ respectively along with 4-chloro benzene group at R position on parent structure are conducive and impart activity enhancing effect. These 26 compounds can be treated as leads for further refinement to derive compounds with further higher antiviral activity.

Acknowledgment

The authors wish to thank the Director, SGSITS, Indore, for providing necessary facilities to carry out the present work.

Supplementary Data

Supplementary Data associated with this article are available in the electronic form at http://nopr.niscair.res.in/jinfo/ijca/IJCA_59A(10)1484-1493_SupplData.pdf.

References

1 Reynolds C, de Koning C B, Pelly S C, Van Otterlo W A & Bode M L, *Chem Soc Rev*,13 (2012) 465.

2 *Global AIDS update 2016*, http://www.unaids.org/sites/default/files/media_asset/global-AIDS-update-2016_en.pdf.

3 Piot P, Goldman L & Ausiello D, *In Human immunodeficiency virus infection and acquired immunodeficiency syndrome: A global overview* (Goldman-Cecil Medicine, Elsevier Health-US 25th Edition, Philadelphia) 2015, Section XXIV. Chaps. 384.

4 Rai M A, Pannek S & Fichtenbaum C J, *Expert Opin Emerg Drugs*, 23 (2018) 149.

5 Cao Y, Zhanga Y, Wu S, Yanga Q, Suna X, Zhaoa J, Peia F, Guoa Y, Tiana C, Zhanga Z, Wangd H, Mad L, Liua J & Wanga X, *Bioorg Med Chem*, 23 (2015) 149.

6 Adamson C S & Freed E O, *Antiviral Res J*, 85 (2010) 119.

7 Frankel A D & Young J A T, *Annu Rev Biochem*, 67 (1998) 1.

8 Sironi F, Malnati M, Mongelli N, Cozzi P, Guzzo C, Ghezzi S, Martínez-Romero C, García-Sastre A, Lusso P, Jabes D & Biswas P, *J Transl. Med*, 13 (2015) 1.

9 Cramer C J, In *Essentials of Computational Chemistry: Theories and models*, (John Wiley and Sons Ltd., Chichester, UK), 2004.

10 Yao X J, Panaye A, Doucet J P, Zhang R S, Chen H F, Liu M C, Hu Z D & Fan B T, *J Chem Inf Comput Sci*, 44 (2004) 1257.

11 Roy K & Mandal A S, *J Enzyme Inhib Med Chem*, 23 (2008) 980.

12 Li X, Gao P, Huang B, Zhou Z, Yu Z, Yuan Z, Liu H, Pannecouque C, Daelemans D, De Clercq E, Zhan P & Liu X, *Eur J Med Chem*, 126 (2016) 190.

13 Nizami B, Sydow D, Wolber G & Honarparvar B, *Mol Biosyst*, 12 (2016) 3385.

14 Yee K L, Sanchez R I, Auger P, Liu R, Fan L, Triantafyllou I, Lai M T, Di Spirito M, Iwamoto M & Khalilieh S G, *Antimicrob Agents Chemother*, 61 (2016) pii: e01757.

15 Ranise A, Spallarossa A, Cesarini S, Bondavalli F, Schenone S, Bruno O, Menozzi G, Fossa P, Mosti L, La Colla M, Sanna G, Murreddu M, Collu G, Busonera B, Marongiu M E, Pani A, La Colla P & Loddo R, *J Med Chem*, 48 (2005) 3858.

16 Spallarossa A, Cesarini S, Ranise A, Schenone S, Bruno O, Borassi A, La Colla P, Pezzullo M, Sanna G, Collu G & Secci B, *Eur J Med Chem*, 44 (2009) 2190.

17 Spallarossa A, Cesarini S, Ranise A, Bruno O, Schenone S, La Colla P, Collu G, Sanna G, Secci B & Loddo R, Eur *J Med Chem*, 44 (2009) 1650.

18 Cichero E, Cesarini S, Spallarossa A, Mosti L & Fossa P, *J Mol Model*, 15 (2009) 871.

19 Ranise A, Spallarossa A, Schenone S, Bruno O, Bondavalli F, Vargiu L, Marceddu T, Mura M, La Colla P & Pani A, *J Med Chem*, 46 (2003) 768.

20 Chem Draw Ultra 7.0.0 trial version (www.cambridgesoft.com).

21 Toxicity Estimation Software Tool (T.E.S.T. 4.1), U.S. Environmental Protection Agency, 2012.

22 Molegro Virtual Docker 2.6.0 trial version (http://www.molegro.com).

23 Kier L B & Hall L H, *Quant. Struc-Act Rel*, 12 (1993) 383.

24 Roy K & Mitra I, *Curr Comput Aided Drug Des*, 8 (2012) 135.

25 Hall L H, *Curr Comput Aided Drug Des*, 8 (2012) 93.

26 Kier L B, Hall L H & Frazer J W, *J Math Chem*, 7 (1991) 229.

27 Mandloi M, Sikarwar A, Sapre N S, Karmarkar S & Khadikar P V, *J Chem Inf Comput Sci*, 40 (2000) 57.

28 Hall L H & Kier L B, *J Chem Inf Comput Sci*, 40 (2000) 784.

29 Hall L H & Vaughn A, *Med Chem Res*, 7 (1997) 407.

30 Hall L H & Story C T, *J Chem Inf Comput Sci*, 36 (1996) 1004.

31 Kellogg G E, Kier L B, Gaillard P & Hall L H, *J Comput Aided Mol Des*, 10 (1996) 513.

32 Sabet R, Fassihi A & Moeinifard B, *J Mol Graph Model*, 28 (2009) 146.

33 Roy K & Leonard J T, *Indian J Chem*, 45A (2006) 126.

34 Niazi A, Jameh-Bozorghi S & Nori-Shargh D, *J Hazard Mater*, 151 (2008) 603.

35 Li Y, Qin Y, Chen X & Li W, *PLoS One*, 8 (2013) e73186.

36 Rumelhart D, Hinton G & Williams R, *Nature*, 323 (1986) 533.

37 Hu R, Doucet J, Delamar M & Zhanga R, *Eur J Med Chem*, 44 (2009) 2158.

38 Szaleniec M, Tadeusiewicz R & Witko M, *Neurocomputing*, 72 (2008) 241.

39 Maniezzo V, *IEEE Trans Neural Networks*, 5 (1994) 39.

40 Porto V W & Fogel D B, *IEEE Expert*, 10 (1995) 16.

41 Pourbasheer E, Riahi S, Ganjali M R & Norouzi P, *Eur J Med Chem*, 44 (2009) 5023.

42 Boser B E, Guyon I M & Vapnik V N, *A training algorithm for optimal margin classifiers*, COLT '92 Proceedings of the Fifth Annual Workshop on Comput. Learning Theory (Pittsburgh ACM) 1992, p 144.

43 Chang C C & Lin C J, *Neural Computation*, 14 (2002) 1959.

44 Cortes C & Vapnik V, *Machine Learning*, 20 (1995) 273.

45 Liang G & Li Z, *J Mol Graph Model*, 26 (2007) 269.

46 Darnag R, Mostapha Mazouz E L, Schmitzer A, Villemin D, Jarid A & Cherqaoui D, *Eur J Med Chem*, 45 (2010) 1590.

47 Cong Y, Yang X G, Lv W & Xue Y, *J Mol Graph Model*, 28 (2009) 236.

48 Shi Z, Ma X H, Qin C, Jia J, Jiang Y Y, Tan C Y & Chen Y Z, *J Mol Graph Model*, 32 (2012) 49.

49 Selwood D L, Livingstone D J, Comley J C W, O'Dowd A B, Hudson A T, Jackson P, Jandu K S, Rose V S & Stables J N, *J Med Chem*, 33 (1990) 136.

50 Eriksson L, Johansson E, Muller M & Wold S, *J Chemometrics*, 14 (2000) 599.