Journal of Physical and Theoretical Chemistry
of Islamic Azad University of Iran, 18 (2) 75-90: Summer 2021
(J. Phys. Theor. Chem. IAU Iran)
ISSN 1735-2126

Application of Monte Carlo Method and a novel modelling-optimization approach on QSAR Study of Etoposide drugs

Omid Alizadeh[1], Robabeh Sayyadikordabadi[1]*, Ghasem Ghasemi[1], Babak Motahary[2]

[1] Department of Chemistry and Chemical Engineering, Rasht Branch, Islamic Azad University, Rasht, Iran
[2] Department of Computer Engineering, Rasht Branch, Islamic Azad University, Rasht, Iran

Received February 2021; Accepted November 2021

ABSTRACT

Monte Carlo and Multiple Linear Regression (MLR) and Imperialist Competitive Algorithm (ICA) were used to select the most appropriate descriptors. Examining the quality of the model by comparing the mean squared error (MSE) and correlation coefficient (R^2), indicated that 140 is the most appropriate number of empires for the gas phase. In the Monte Carlo method, CORAL software was used and the data were randomly divided into training, calibration, and test subsets in three splits. The correlation coefficient (R^2), cross-validated correlation coefficient (Q^2) and standard error of the model were calculated to be respectively 0.9301, 0.7377, and 0.595 for the test set with an optimum threshold of 4. It was concluded that simultaneous utilization of MLR-ICA and Monte Carlo method can lead to a more comprehensive understanding of the relationship between physico-chemical, structural or theoretical molecular descriptors of drugs to their biological activities and facilitate designing of new drugs.

Keywords: Etoposides; QSAR; ICA Algorithm; Monte Carlo method

1. INTRODUCTION

Etoposide is an anti-cancer drug that belongs to topoisomerase inhibitor family of medication that is used for treatments of testicular, bladder, prostate, lung, stomach, and uterine, cancers [1]. Quantitative structure–activity relationships (QSAR) are mathematical methods that correlate physicochemical and molecular descriptors to the biological activity of chemicals and are widely used for providing insights into the key structural features that affect the biological responses (e.g., half maximal inhibitory concentration (IC50)) of drugs [2,3]. A QSAR approach includes two main parts; modelling and optimization. The former part correlates the descriptors to the response, while the latter's duty is to probe for the most significant descriptors.

*Corresponding author: Sayyadi_04@yahoo.com & sayyai@iaurasht.ac.ir

Many linear and non-linear models coupled with variety of optimization algorithms have been employed in QSAR studies [4, 5]. Imperialist Competitive Algorithm (ICA) is a new population-based optimization algorithm that was proposed by Atashpaz-Gargari and Lucas in 2007 [6] and since then it was employed in solving a variety of optimization problems [7-9]. The algorithm starts with an initial population. The individuals (countries) are two type: imperialists and colonies. The most powerful countries are selected as imperialists and the rest as the colonies of these imperialists. The total power of an empire depends on both power of the imperialist country and power of its colonies [7].

The most powerful empires tend to increase their power while weak empires collapse. All empires try to take possession of colonies of other empires and control them. This is modeled by just picking some of the weakest colonies of the weakest empires and making a competition among all empires to create these colonies.

Recently, CORAL has been proposed as competent software for the QSAR studies. It uses Monte Carlo method to find the most important simplified molecular input-line entry system (SMILES)-based descriptors and calculate their correlation weights to predict an endpoint (e.g., -log(IC_{50})). SMILES are lines of symbols, representing the molecular structure [12-14].

In the present study, CORAL software and a MLR-ICA approach were used to investigate the QSAR in 25 Etoposide anticancer drugs.

2. Theory and Computational Methods

2.1. Selection of Descriptors Using MLR-ICA Approach

Details of geometry optimizations of compounds were described in our previous work [10]. Geometries of 25 Etoposide anti-cancer drugs were optimized using B3lyp/6-311g at Gaussian 03W. Modeling and optimizing calculations for QSAR were performed by MATLAB 2014a software [19, 20]. The 1092 and 977 SPSS screened descriptors [11] were used as the feed to ICA-MLR approach as the population matrix in order to find the best descriptors for the gas and solution phases. The numbers of the most effective descriptors (i.e., 4 for the gas) chosen by a stepwise multiple linear regression procedure in our previous work was used as a basis for the number of descriptors in this work.

The employed ICA of this work is depicted in Figure 1. In this algorithm, the initial countries, which are equivalent to chromosomes in the genetic algorithm are indices of the descriptors matrix. They are set of values of a candidate solution for the optimization problem. Empires are sub-populations of countries. Assimilation, which can be considered as a primitive form of Particle Swarm Optimization moves all non-best countries (called colonies) in an empire toward the best country (called imperialist) in the same empire [15] to find the colonies with lowest error (RMSE of predicted – log(IC50) using MLR versus empirical values).

Different number of decision variables (nDes) and different number of empires (nEmp) were investigated to obtain the least RMSE and highest R^2 using ICA.

2.2. Monte Carlo Method

CORAL [17] software was used for calculation of descriptor correlation weight (DCW) of the 25 Etoposide compounds with a hybrid optimization scheme including hydrogen-suppressed molecular graph (HSG) and SMILES representation of molecular structures. Modelling using CORAL software was carried out for thresholds of 1 up to 5 and 100 epochs (i.e., an overall number of 1500 runs were

performed). Each sequence of computations for finding a new set of modified correlation weights of the model is named an epoch [16]. The SMILES-based and Graph -based optimal descriptors are achieved using the following equations [12, 16]:

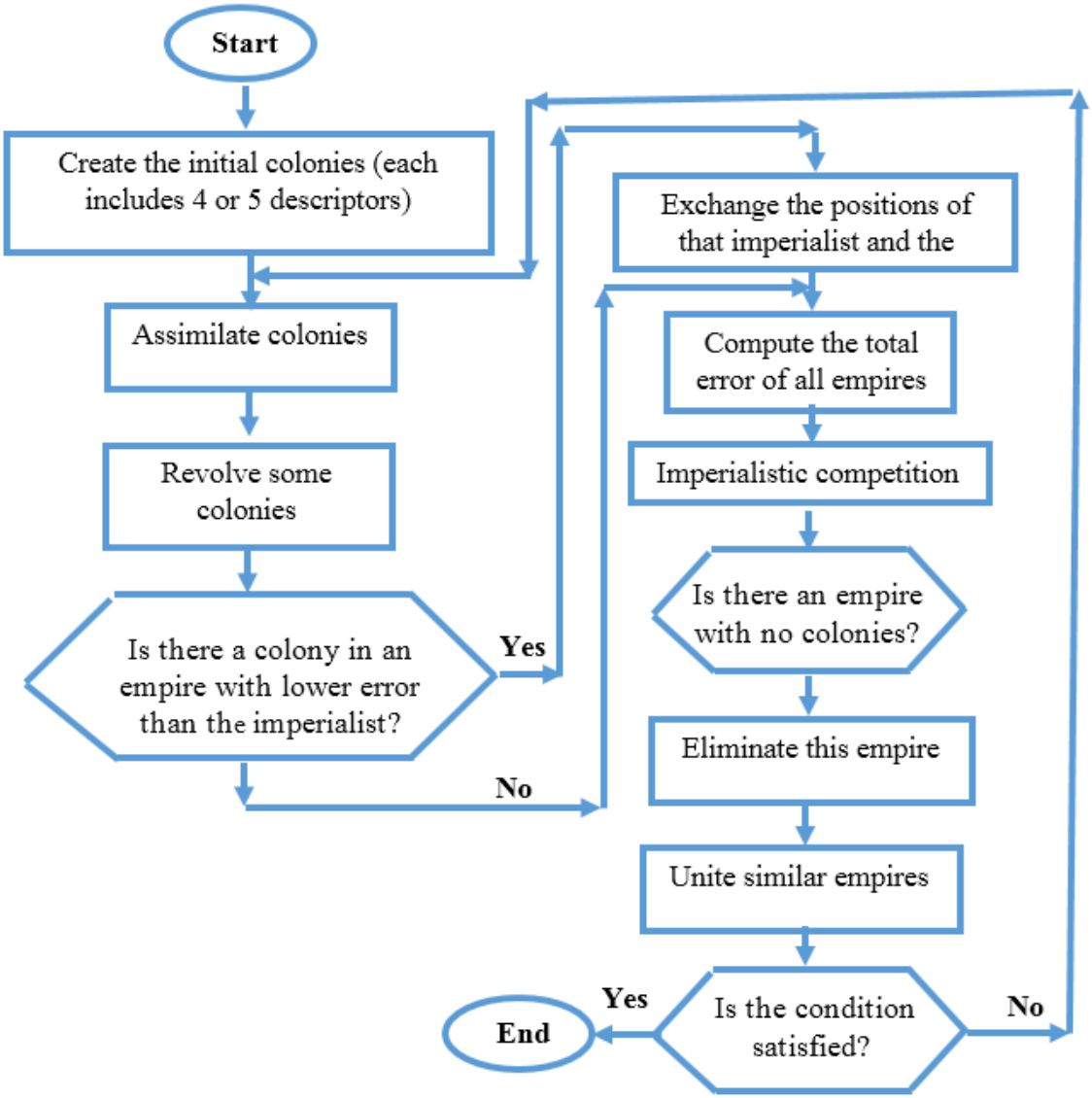


Fig. 1. Flowchart of the Imperialist Competitive Algorithm.

$$\mathrm{DCW}(\mathrm{T},\mathrm{N}_{\mathrm{epoch}})^{SMLLES} = \alpha\Sigma CW(Sk) + \beta\Sigma CW(SSk) + \gamma\Sigma CW(SSSk) + x \cdot CW(NOSP) + y.CW(HALO) + z \cdot CW(BOND) \quad (1)$$

$$\mathrm{DCW}\left(\mathrm{T},\mathrm{N}_{\mathrm{epoch}}\right)Graph = \sum \mathrm{CWA}_k + \alpha\sum \mathrm{CW}(0\mathrm{ECk}) + \beta\sum \mathrm{CW}(1\mathrm{ECk}) + \gamma\sum \mathrm{CW}(2\mathrm{ECk}) + \delta\sum \mathrm{CW}(3\mathrm{ECk} \quad (2)$$

Where, S_k, SS_k, and SSS_k denote SMILES attributes. The NOSP (nitrogen, oxygen, sulfur, and phosphorus) and HALO (fluorine, chlorine, and bromine) are demonstrated the presence or absence of chemical elements. Also "BOND" are shown double (=), triple (#), or stereo chemical bonds (@ or @@). A_k in equation (2) indicates the occurrence of the C, N, O atoms in the HSG and HFG molecular graphs. The α, β, γ, and δ coefficients and combinations of their values are used to define various versions of the graph-based optimal descriptor and can be 1 or 0. The hybrid objective function for finding the optimal descriptors is defined as [14]:

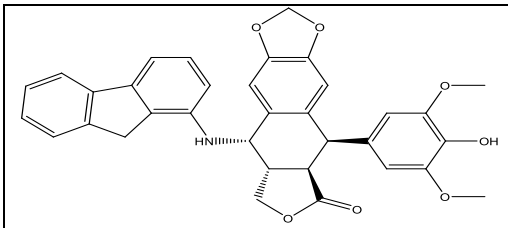
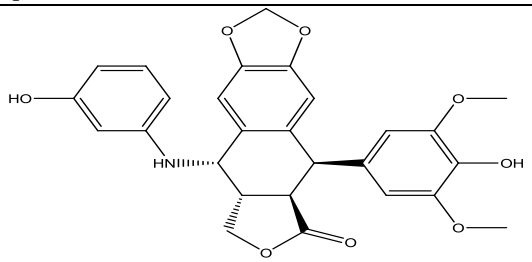
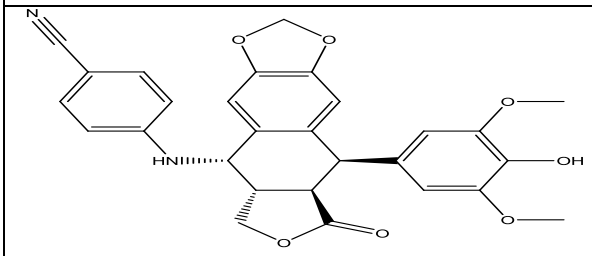
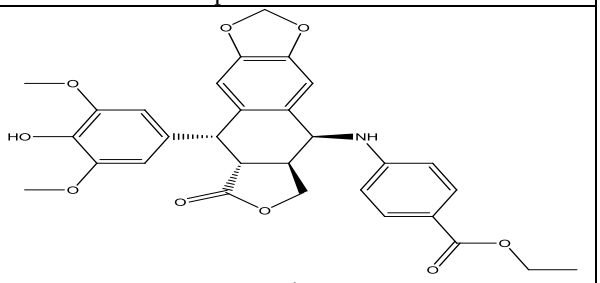
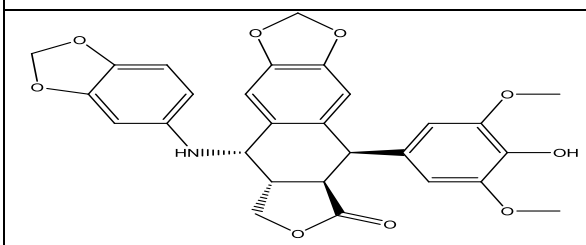
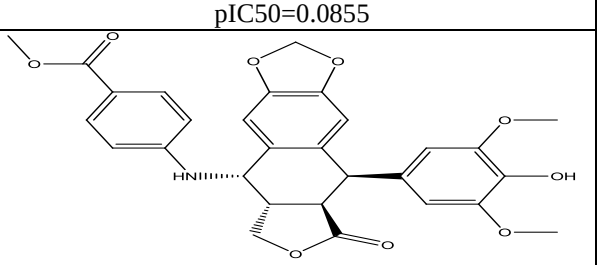
$$\mathrm{DCW}(\mathrm{T},\mathrm{N}_{\mathrm{epoch}})^{\mathrm{Hybrid}} = \mathrm{DCW}(\mathrm{T},\mathrm{N}_{\mathrm{epoch}})^{\mathrm{SMILES}} + \mathrm{DCW}(\mathrm{T},\mathrm{N}_{\mathrm{epoch}})^{\mathrm{Graph}} \quad (3)$$

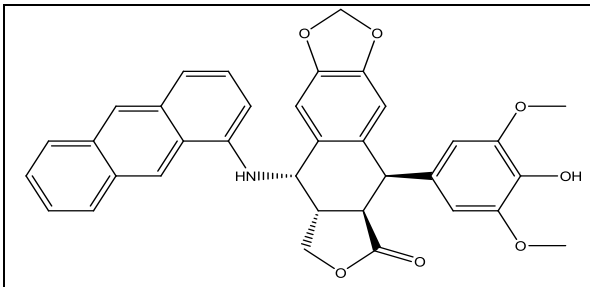
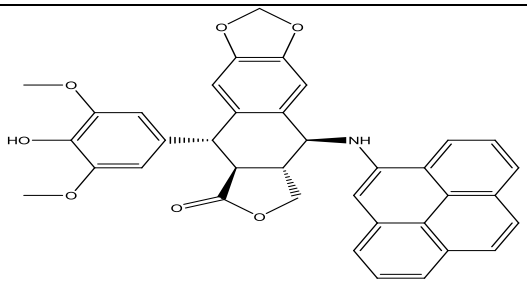
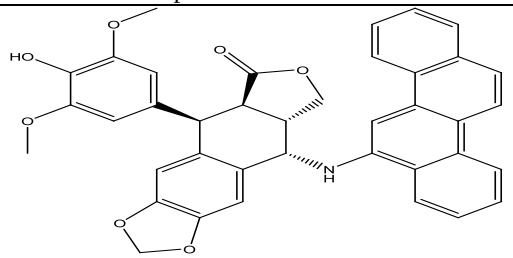
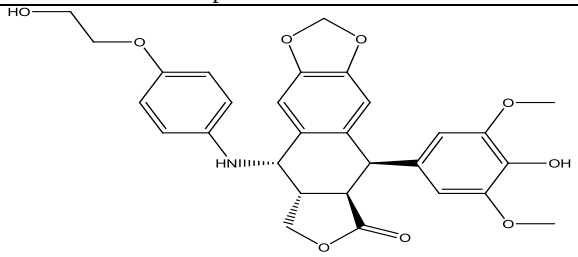
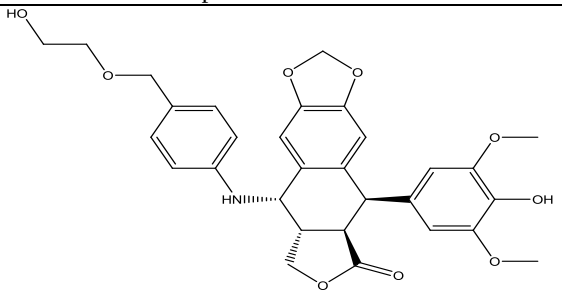
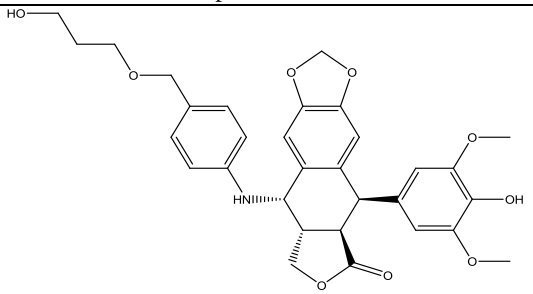
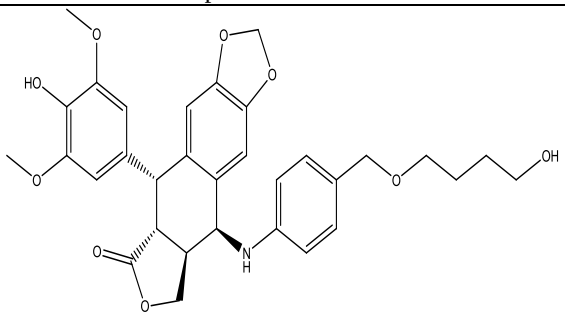
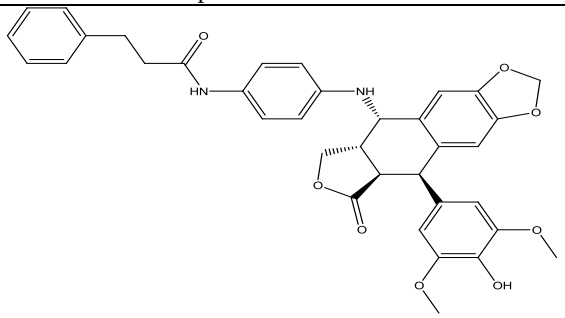
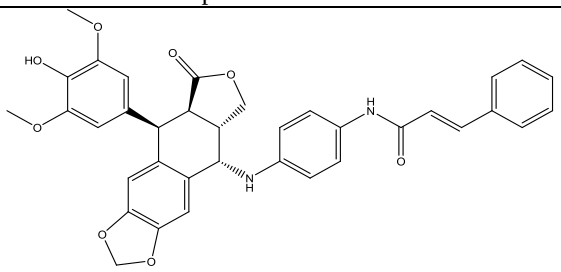
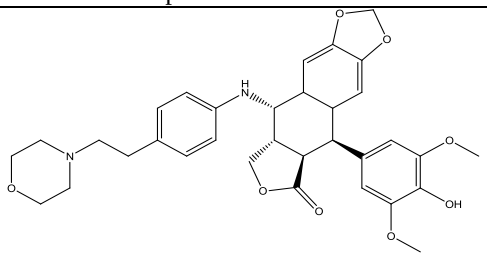
3. RESULTS and DISCUSSION

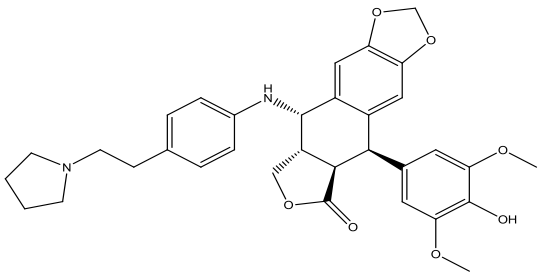
3.3. Molecular Descriptors Generation with MLR-ICA Approach

All studied Etoposide compounds have been presented in table 1 [10].

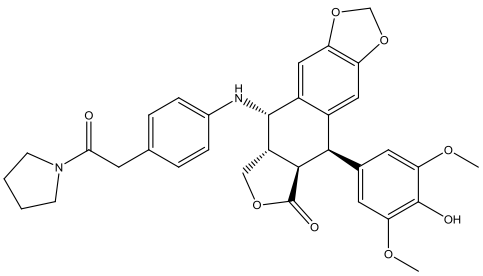
Table 1. Optimized structure of the Etoposides derivatives used to build QSAR models with B3lyp/6-31g in gas phase [10]

 1 pIC50=2.7064	 2 pIC50=0.3573
 3 pIC50=0.2025	 4 pIC50=0.0855
 5 pIC50=0.1771	 6 pIC50=0.0829

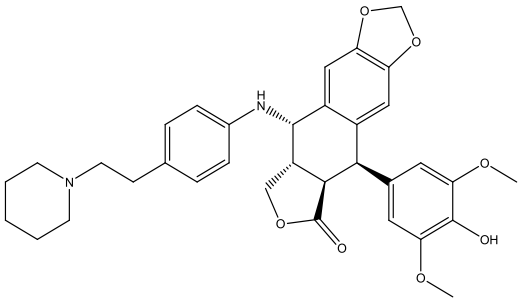
 7 pIC50=-0.3808	 8 pIC50=3.9873
 9 pIC50=3.3039	 10 pIC50=3.2292
 11 pIC50=3.2405	 12 pIC50=2.29381
 13 pIC50=2.1841	 14 pIC50=3.4392
 15 pIC50=3.3008	 16 pIC50=3.2098



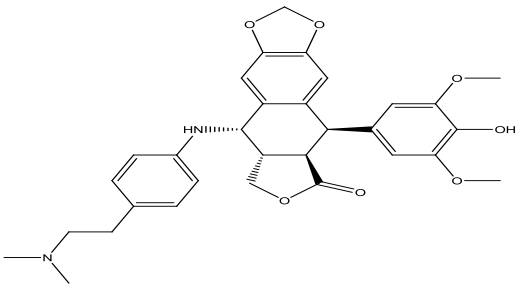
17
pIC50=3.2081



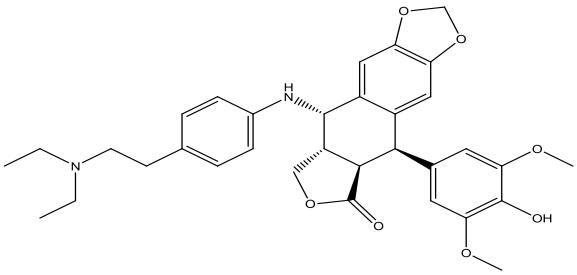
18
pIC50=3.2060



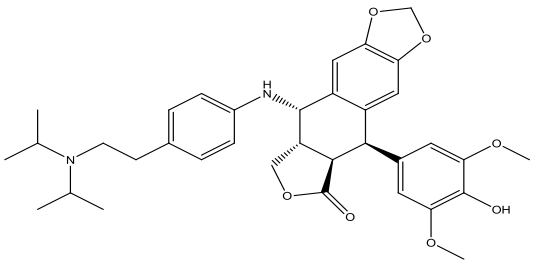
19
pIC50=3.2559



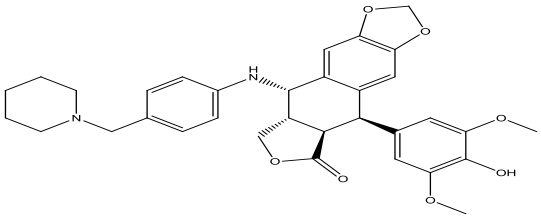
20
pIC50=2.6637



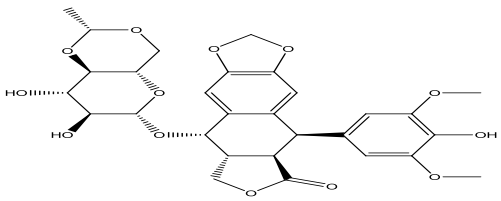
21
pIC50=3.2063



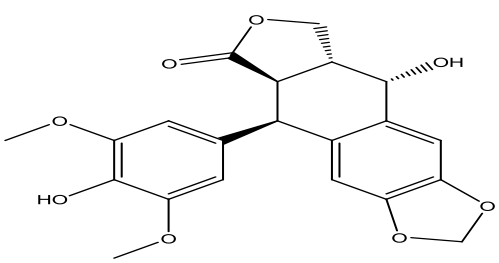
22
pIC50=3.2125



23
pIC50=3.2068



24
pIC50=2.5969



25
pIC50=3.3063

As a first trial, 5000 iterations were performed to find the most powerful empires and, subsequently, the best descriptors. A plot of the Best Cost values versus the number of iterations is represented in Fig. 2. The figure implies that there is no variation in the best cost after about 100 iterations. However, in order to ensure that the best descriptors are captured, the number of iterations for the rest of computations was set to 500.

The effects of number of selected descriptors on the chosen descriptors and the prediction quality (according to R^2 and RMSE) were investigated and the results are plotted in Fig. 3. As it is expected, the model's accuracy regarding to R^2 and RMS increases by increasing the number of model parameters (descriptors in this case).

In order to choose the most suitable number of empires, the model was run using different number of empires and the results are demonstrated in Fig. 4. According to this figure, the optimum number of empires was chosen as 140.

A plot of the predicted versus empirical values of –logIC50 is depicted. The figure implies that the developed model possesses a high correlation coefficient, indicating that the experimental and predicted values are well correlated (Fig. 5).

The chosen descriptors using MLR-ICA approach are presented in Table 2.

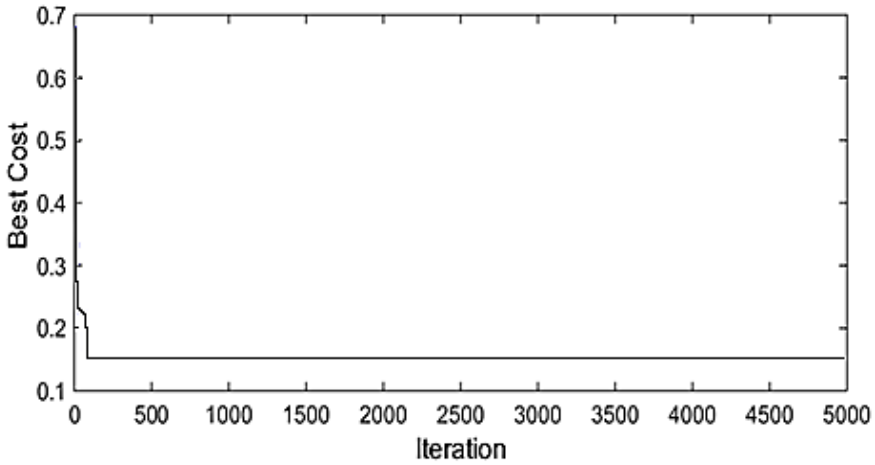


Fig. 2. Plot between Best Cost values compared to the variation of Iteration.

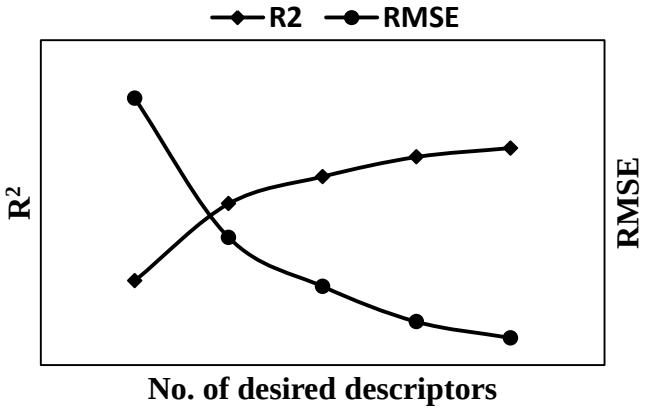


Fig. 3. variation of R^2 and MSE by varying the number of empires for the gas phase

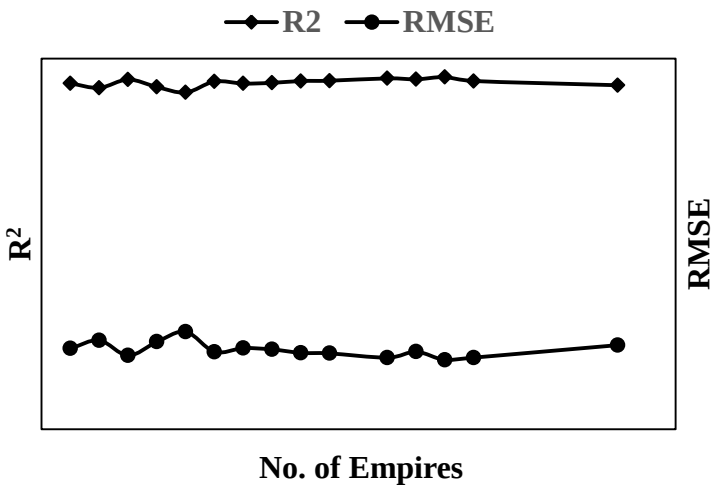


Fig. 4. Variation of R^2 and MSE by varying the number of empires for the gas phase

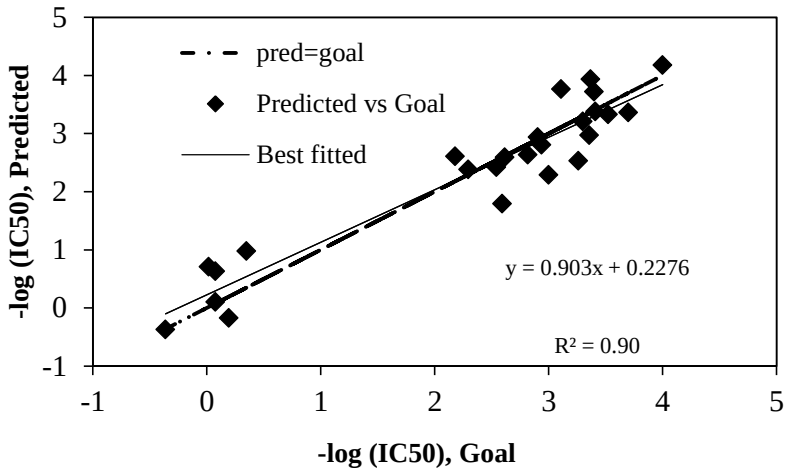


Fig. 5. Plots of predicted versus Goal values of log(ICP50)".

Table 2. Selected descriptors using MLR-ICA Method with nDes=4 and nEmp=140 in the gas phase

Descriptor	Definition	Type
H5p	H autocorrelation of lag 5/weighted by atomic polarizabilities	GETAWAY descriptors
nBnz	Number of benzene like rings	Constitutional descriptors
JGI5	Mean topological charge index of order 5	Topological charge indices
VEA2	Average eigenvector coefficient sum from adjacency matrix	Eigenvalue based indices

The presented information in table 2 show that polarizabilities, number of Benzene like rings, JGI5 (Topological charge indices) and VEA2 (Eigenvalue-based indices) in gas phase are the most important descriptors for designing this class of drugs. Eigenvalue-based indices descriptors are computed from weighted distance matrices of a hydrogen-depleted molecular graph. The following weighted distance matrices are required for the computation of eigenvalue-based indices descriptors. Topology charge index has been proposed to evaluate the charge

transfer between pairs of atoms and the global charge transfer in the molecule [11].

The graphs of H5p,nBnz, JGI5 and VEA2 descriptors in the gas phase versus the empirical negative logarithm half maximal inhibitory concentration (–logIC50) were plotted by using the Matlab program (Fig. 6).

The charts show that the empirical negative logarithm half maximal inhibitory concentration (-logIC50) value increases with increasing H5p and nBnz descriptors, and thus the half maximal inhibitory concentration (IC50) value is reduced.

As the JGI5 and VEA2 descriptors increased the empirical negative logarithm half maximal inhibitory concentration (-log IC50) value decreased, and then the increase in these descriptors increased the half maximal inhibitory concentration value.

3.2. Result of the Monte Carlo Method

The statistical parameters of the models obtained using molecular graphs (HSG) and SMILES are shown in Table 3. Performance of the models were compared with each other by the criterion of the predictability in test set (Rm^2) which should be larger than 0.5 [18], correlation coefficient (R^2) in each set, cross-validated correlation coefficient (Q^2) and standard error of estimation (s). The difference between R^2m and R'^2m values ($\Delta RmTEST$) was used as another criterion in this issue. The depicted results in table 3 disclose that for all of the three splits, threshold of 4 gives the best results. The results with threshold of 4 for the three Monte Carlo probes are presented in table 3, 4.

The SMILES-based models are the following:

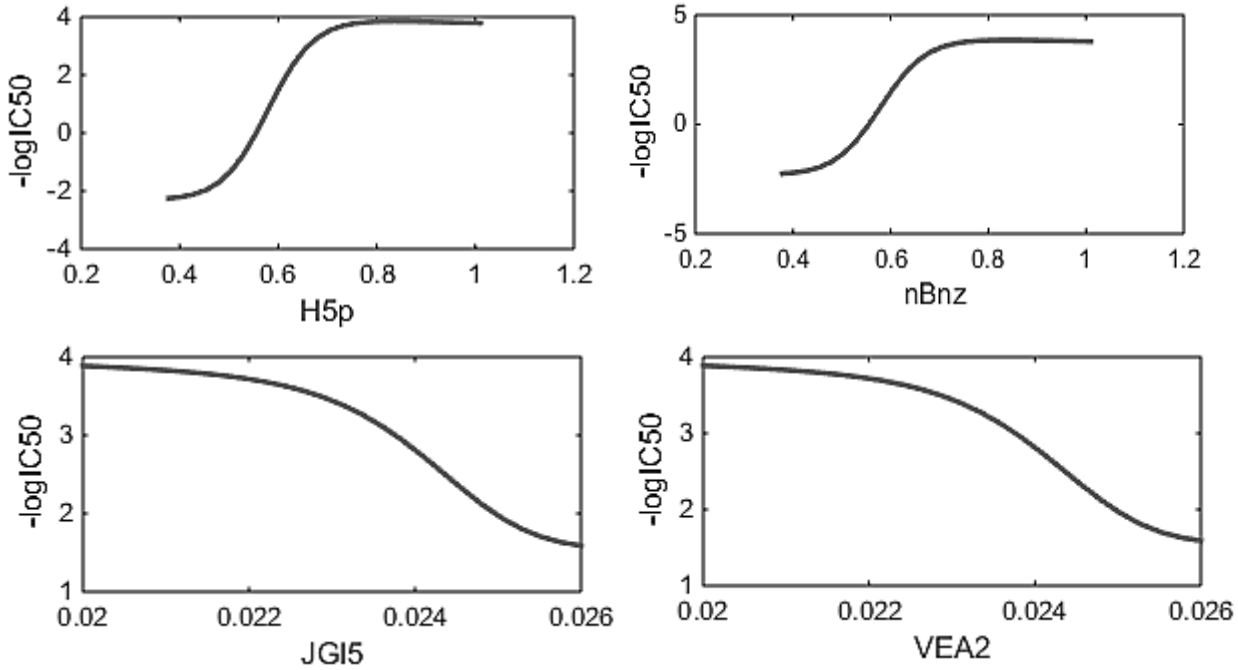


Fig. 6. Variations of -logIC50 in terrms of the MLR-ICA chosen descriptors.

Split 1: (T=4)

-logIC50 = -15.7728463 (± 0.1282099) + 0.1338664 (± 0.0010600) * DCW(4,100)

n=12, R^2= 0.9526, Q^2=0.9423, s = 0.294 (training set)

n=8, R^2 = 0.9772, Q^2 = 0.9673, s = 0.54 (calibration set)

n=5, R^2 = 0.9301, Q^2=0.7377, s = 0.595 (test set), R^2m TEST= 0.6033

Spit 2: (T=4)

-logIC50 = -7.6317696 (± 0.4240583) + 0.0907350 (± 0.0035242) * DCW(4,100)

n=13, R^2 = 0.7324, Q^2=0.6585, s = 0.709 (training set)

n=6, R^2 = 0.9995, Q^2 = 0.9984, s = 0.318 (calibration set)

n=6, R^2 = 0.8030, Q^2 = 0.6188, s = 0.842 (test set), R^2m TEST= 0.5034

Spit 3: (T=4)

-logIC50 = -8.3928675 (± 0.1542551) + 0.0706452 (± 0.0012585) * DCW(4,100)

n=10, R^2 = 0.9629, Q^2 = 0.9483, s = 0.290 (training set)

n=8, R^2 = 0.9997, Q^2 = 0.9995, s = **0.**441 (calibration set)

n=7, R^2 = 0.5187, Q^2 = 0.4634, s = 1.55 (test set), R^2m TEST = 0.4843

The prediction for Split 1 and probe 1 is better than the others.

Table 3. Statistical data calculated with both HSG for three random splits into test set . Best model are indicated by bold

Threshold	R^2 test Probe 1	R^2 test Probe 2	R^2 test Probe 3	R^2 test Average	Dispersion
SPLIT 1					
1	0.7567	0.7666	0.7771	0.7668	0.0084
2	0.8054	0.7927	0.8065	0.8015	0.0063
3	0.9259	0.9234	0.9129	0.9207	0.0056
4	**0.9301**	0.9171	0.9190	0.9221	0.0058
5	0.7528	0.7933	0.8158	0.7873	0.0261
SPLIT2					
1	0.7908	0.8066	0.8067	0.8014	0.0075
2	0.7882	0.8043	0.7897	0.7941	0.0073
3	0.8538	0.8353	0.8680	0.8524	0.0134
4	0.9026	0.8601	0.8739	0.8789	0.0177
5	0.8553	0.8849	0.8535	0.8646	0.0144
SPLIT3					
1	0.6572	0.6506	0.6387	0.6488	0.0076
2	0.6621	0.6659	0.6565	0.6615	0.0039
3	0.6529	0.6725	0.6564	0.6606	0.0085
4	0.7546	0.7308	0.7560	0.7471	0.0115
5	0.7210	0.7241	0.6995	0.7149	0.0109

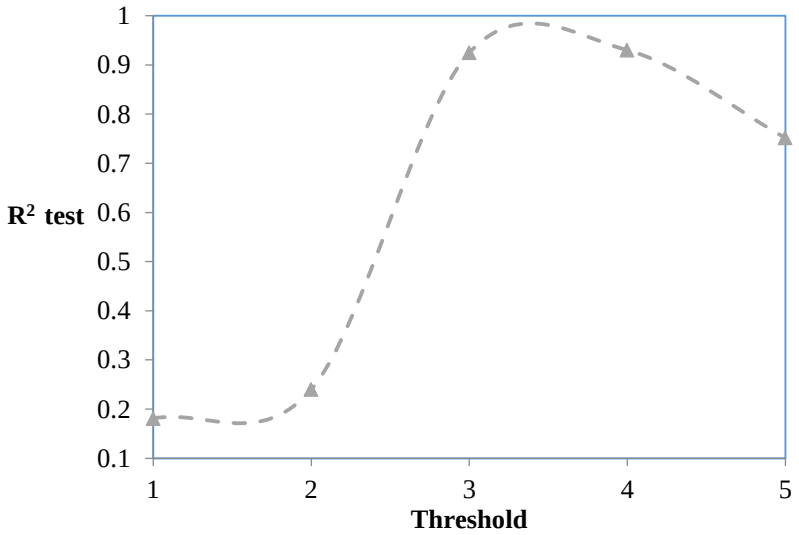
The variation of correlation coefficient (test set) with respect to threshold and the number of epochs are plotted in figure 7. This figure confirms that 4 and 70 are the most appropriate values for threshold and number of epochs, respectively.

The distribution of SMILES notations in the train, calibration and test sets are reported in table 5.

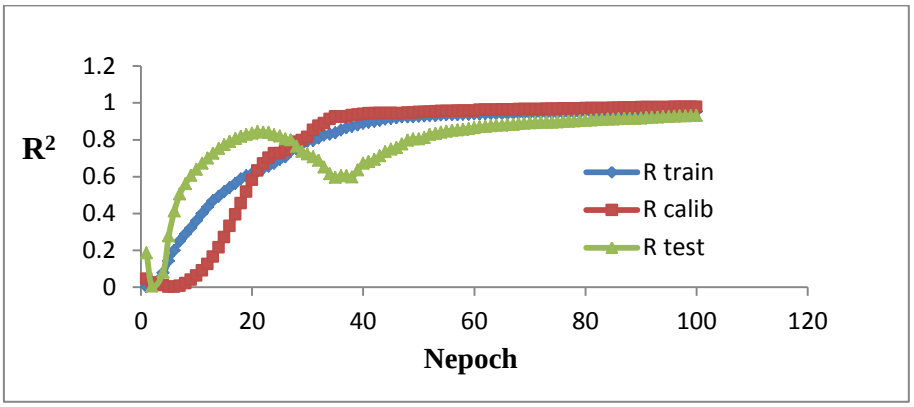
The corresponding values of DCW, experimental and calculated activities (-log IC50) for the sequence of compounds of table 5 are given in table 6. The given experimental and predicted in table 6 are plotted against each other in figure 8. A good correlation between the calculated and empirical values of –log IC50 can be observed in this figure that approves the appropriateness of the developed model.

Table 4. Statistical quality of models calculated with both HSG and SMILES for Training, calibration and test

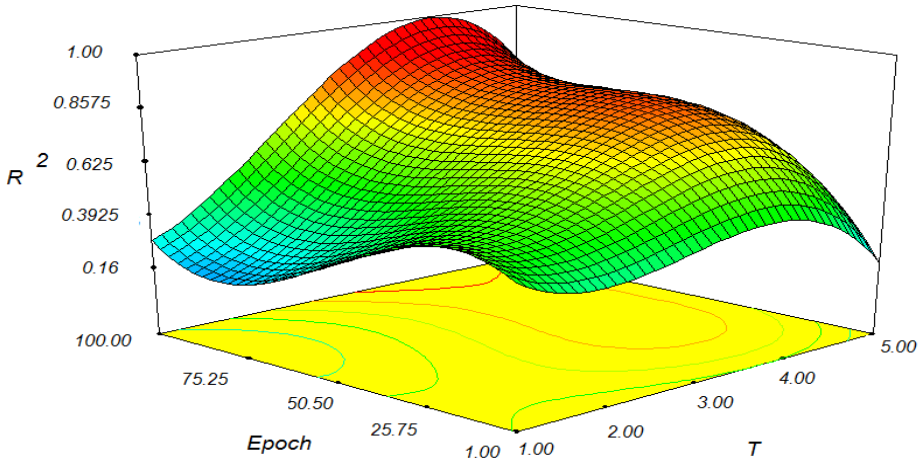
Threshold-probe	**R^2**	**Q^2**	**S**	**R^2m TEST[31] Should be>0.5**	**R^{*2}m TEST[30] Should be>0.5**	**ΔRmTEST[32] Should be<0.2**
4-1						
Training(n=12)	0.9526	0.9423	0.294			
Calib(n=8)	0.9772	0.9673	0.54			
Test(n=5)	0.9301	0.8371		0.5332	0.6033	0.026
4-2						
Training(n=12)	0.9517	0.9412	0.296			
Calib(n=8)	0.9772	0.9675	0.52			
Test(n=5)	0.9171	0.8110	0.64	0.461	0.5677	0.0321
4-3						
Training(n=12)	0.9524	0.9418	0.294			
Calib(n=8)	0.9767	0.9666	0.53			
Test(n=5)	0.9190	0.87706	0.63	0.670	0.5767	0.0305



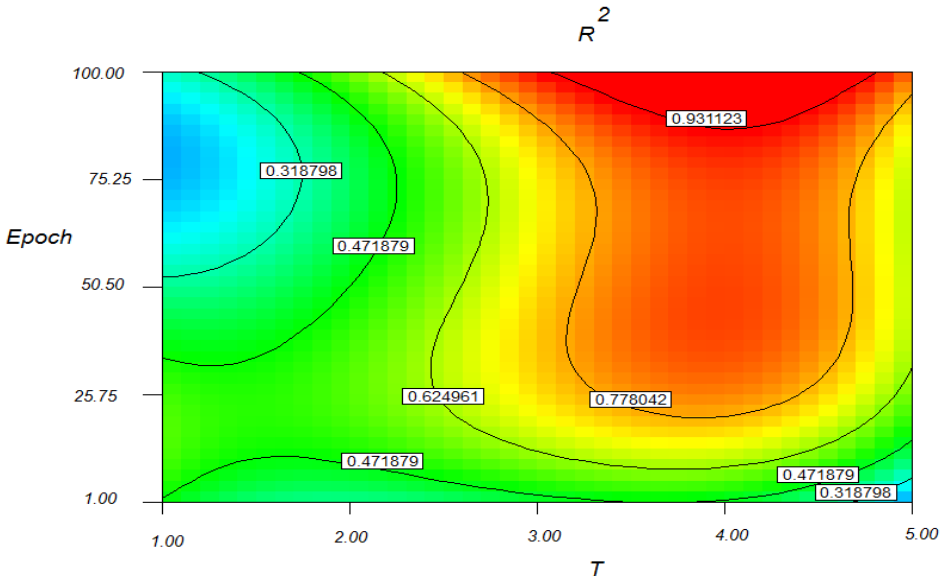
A)



B)



C)



D)

Fig. 7. The variation of correlation coefficient for test set by threshold and number of epochs. (A): effects of threshold. (B) Effects of the number of epochs. (C) 3-D surface plot of R^2 according to the threshold and the number of epochs. (D) Contour plots of R^2 according to the threshold and the number of epochs.

Table 5. SMILES notations 25 compound of Etoposide and train, Calibration and test set

Compound	SMILES	Set
1	COC1=CC(=CC(=C1O)OC)C2C3C(COC3=O)C(NC4=CC=CC5=C4CC6=C5C=CC=C6)C7=CC8=C(OCO8)C=C27	Train
2	COC1=CC(=CC(=C1O)OC)C2C3C(COC3=O)C(NC4=CC(=CC=C4)O)C5=CC6=C(OCO6)C=C25	Train
3	COC1=CC(=CC(=C1O)OC)C2C3C(COC3=O)C(NC4=CC=C(C=C4)C#N)C5=CC6=C(OCO6)C=C25	Train
4	CCOC(=O)C1=CC=C(NC2C3COC(=O)C3C(C4=CC(=C(O)C(=C4)OC)OC)C5=CC6=C(OCO6)C=C25) C=C1	Train
7	+:COC1=CC(=CC(=C1O)OC)C2C3C(COC3=O)C(NC4=CC=CC5=CC6=C(C=CC=C6)C=C45)C7=CC8=C(OCO8)C=C27	Train
10	COC1=CC(=CC(=C1O)OC)C2C3C(COC3=O)C(NC4=CC=C(OCCO)C=C4)C5=CC6=C(OCO6)C=C25	Train
11	COC1=CC(=CC(=C1O)OC)C2C3C(COC3=O)C(NC4=CC=C(COCCO)C=C4)C5=CC6=C(OCO6)C=C25	Train
14	COC1=CC(=CC(=C1O)OC)C2C3C(COC3=O)C(NC4=CC=C(NC(=O)CCC5=CC=CC=C5)C=C4)C6=CC7=C(OCO7)C=C26 :	Train
15	COC1=CC(=CC(=C1O)OC)C2C3C(COC3=O)C(NC4=CC=C(NC(=O)/C=C/C5=CC=CC=C5)C=C4)C6=CC7=C(OCO7)C=C26~:	Train
16	COC1=CC(=CC(=C1O)OC)C2C3C=C4OCOC4=CC3C(NC5=CC=C(CCN6CCOCC6)C=C5)C7COC(=O)C27	Train
19	COC1=CC(=CC(=C1O)OC)C2C3C(COC3=O)C(NC4=CC=C(CCN5CCCCC5)C=C4)C6=CC7=C(OCO7)C=C26	Train
24	COC1=CC(=CC(=C1O)OC)C2C3C(COC3=O)C(OC4OC5COC(C)OC5C(O)C4O)C6=CC7=C(OCO7)C=C26	Train
6	COC(=O)C1=CC=C(NC2C3COC(=O)C3C(C4=CC(=C(O)C(=C4)OC)OC)C5=CC6=C(OCO6)C=C25)C=C1	Calib
9	COC1=CC(=CC(=C1O)OC)C2C3C(COC3=O)C(NC4=C5C=CC=CC5=C6C=CC7=CC=CC=C7C6=C4)C8=CC9=C(OCO9)C=C28~:	Calib
12	COC1=CC(=CC(=C1O)OC)C2C3C(COC3=O)C(NC4=CC=C(COCCCO)C=C4)C5=CC6=C(OCO6)C=C25	Calib
13	COC1=CC(=CC(=C1O)OC)C2C3C(COC3=O)C(NC4=CC=C(COCCCCO)C=C4)C5=CC6=C(OCO6)C=C	Calib
17	COC1=CC(=CC(=C1O)OC)C2C3C(COC3=O)C(NC4=CC=C(CCN5CCCC5)C=C4)C6=CC7=C(OCO7)C=C26	Calib
21	CCN(CC)CCC1=CC=C(NC2C3COC(=O)C3C(C4=CC(=C(O)C(=C4)OC)OC)C5=CC6=C(OCO6)C=C25)C=C1	Calib
22	COC1=CC(=CC(=C1O)OC)C2C3C(COC3=O)C(NC4=CC=C(CCN(C(C)C)C(C)C)C=C4)C5=CC6=C(OCO6)C=C25	Calib
25	COC1=CC(=CC(=C1O)OC)C2C3C(COC3=O)C(O)C4=CC5=C(OCO5)C=C	Calib
5	COC1=CC(=CC(=C1O)OC)C2C3C(COC3=O)C(NC4=CC5=C(OCO5)C=C4)C6=CC7=C(OCO7)C=C26	Test
8	#:COC1=CC(=CC(=C1O)OC)C2C3C(COC3=O)C(NC4=CC5=CC=CC6=CC=C7C=CC=C4C7=C56)C8=C2C=C9OCOC9=C8	Test
18	COC1=CC(=CC(=C1O)OC)C2C3C(COC3=O)C(NC4=CC=C(CC(=O)N5CCCC5)C=C4)C6=CC7=C(OCO7)C=C26	Test
20	COC1=CC(=CC(=C1O)OC)C2C3C(COC3=O)C(NC4=CC=C(CCN(C)C)C=C4)C5=CC6=C(OCO6)C=C25	Test
23	COC1=CC(=CC(=C1O)OC)C2C3C(COC3=O)C(NC4=CC=C(CN5CCCCC5)C=C4)C6=CC7=C(OCO7)C=C26	Test

Table 6. Calculated values for DCW, the experimental activity data (-log IC50) and calculated values for -log IC50 with application of CORAL in split1(T=2)

Compound	Set	DCW	Exp	Calc
1	Train	119.14928	0.079	0.1772
2	Train	147.55382	4	3.9797
3	Train	119.27411	0.194	0.1939
4	Train	137.16075	3.301	2.5884
7	Train	120.43932	0.347	0.3499
10	Train	141.98748	3.26	3.2345
11	Train	138.4508	2.294	2.7615
14	Train	139.15017	2.592	2.8547
15	Train	134.67843	2.541	2.2561
16	Train	142.65475	3.357	3.3238
19	Train	139.19023	2.616	2.86
24	Train	141.05121	3.108	3.1092
6	Calib	136.97304	0.076	0.006
9	Calib	136.67293	-0.362	-0.2
12	Calib	138.64178	3.523	2.7866
13	Calib	138.82949	2.939	2.8118
17	Calib	139.00252	3.699	2.8349
21	Calib	138.81552	3.367	2.8099
22	Calib	138.85869	3.409	2.8157
25	Calib	138.55922	2.903	2.7756
5	Test	136.95114	3	2.5603
8	Test	124.68771	0.017	0.9186
18	Test	136.15678	2.18	2.454
20	Test	139.81184	2.818	2.9433
23	Test	139.00252	3.398	2.8349

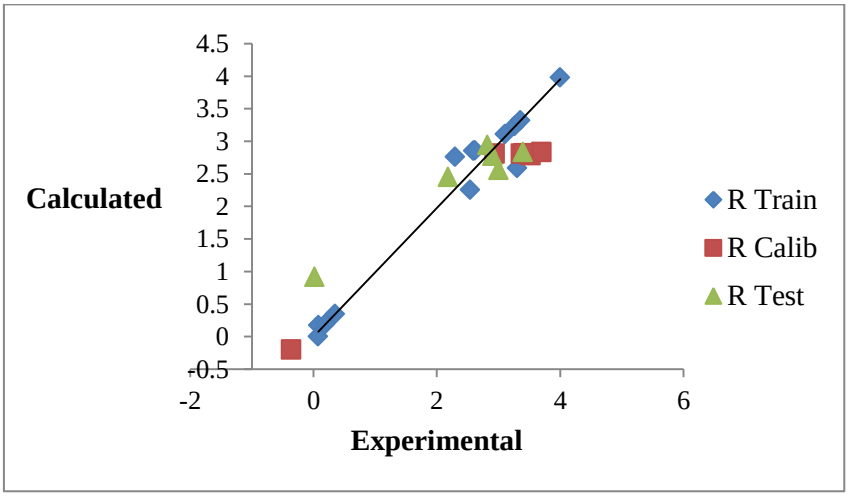


Fig. 8. Correlation between experimental and predicted –logIC50 calculated using Eq.3.

Molecular features are sorted according to their correlation weights and are given in table 7. Molecular feature with negative correlation weights are omitted due to their inverse effect on the -logIC50 value. The higher the correlation weigh of a molecular feature, the lower the value of IC50, therefore, the feature is more significant. Definitions of the molecular features are given in table 8.

According to table 7, presence of Presence of cyclic ring, Absence of

halogens, Presence of double bond, Presence of sp2 carbon connected to ring, Presence oxygen connected to ring, Presence of nitrogen connected to sp3 carbon, Presence of sp2 carbon connected to ring are the most important molecular features that might be considered in designing new drugs.

4. CONCLUSIONS

In this study, MLR-ICA approach was used to study the structure-activity relationships of 25 Etoposide Anticancer Drugs. The best descriptors with nEmp=140 in gas phase was more significant than other descriptors. These results proved that H5p, nBnz, JGI5, VEA2 descriptors in the gas phase were more significant than other descriptors to create QSAR model and predict biological activity of Etoposide substitution patterns.

The half maximal inhibitory concentration IC50) value reduced with increasing H5p (weighted by atomic polarizabilities) and nBnz (Number of benzene like rings) descriptors. As the JGI5 (charge transfer between pairs of atoms and the global charge transfer in the molecule) and VEA2 (weighted distance matrices of a hydrogen-depleted molecular graph) descriptors increased the half maximal inhibitory concentration (IC50) value increased.

It was concluded that the simultaneous use of these two methods gives deeper and more comprehensive knowledge of the effect of molecular and structural descriptors on the activity of drugs and provides better insights to design new drugs.

5. ACKNOWLEDGEMENT

The authors would like to acknowledge the Islamic Azad University, Rasht Branch for supporting this study.

Table 7. SMILES attributes with positive correlation weights for split 1

SMILES attributes	CWs	SMILES attributes	CWs
1.....:	5.26003	N...C.....:	6.53305:
2.....	5.34035	O...1.....:	7.76373
3.....:	5.59580	6...(.....:	6.10104
BOND10000000	11.07657	:=...1.....:	5.44523:
:=...4.....:	6.31714	C...1.....:	5.58715:
C...2.....:	5.31651	HALO00000000	5.49792

Table 8. Definition of the promoter of A_k

Attribute Ak	Comment
HALO00000000	Absence of F, Cl, Br
C...C.....	Presence of carbon – carbon bonds (sp3)
C…(…C…	SP3 Carbon atoms with branching
++++O---B2==	Presence of oxygen and double bonds
C…=…….	SP2 Carbon atom
(.....	Branching in molecular skeleton
O.....	Presence of oxygen
1.....	Presence of rings
++++N---B2==	Presence of nitrogen and double bond
=	Double bond
@	Stereo specific bond
#	Triplet bond

References

[1] A. Montecucco, G. Biamonti., Cancer Lett, 252 (2007) 9.

[2] R. SayyadikordAbadi, A. Alizadehdakhel, F. Moosapour, Indian J. Chem. Sect. B, 56 (2017) 677.

[3] R. Sayyadikord Abadi, A. Alizadehdakhel, G. R. Najafi, M. Masomparast, Revue Roumaine de Chimie, 63 (2018) 17.

[4] Khan A. U. Danishuddin, Drug Discov Today, 21 (2016) 1291.

[5] Y. Wang, Q. Zhou, L. Wang, J.-J. Ma., Letters in Organic Chemistry, 15 (2018) 551.

[6] E. Atashpaz-Gargari, C. Lucas., Imperialist competitive algorithm: An algorithm for optimization inspired by imperialistic competition. in *IEEE Congress on Evolutionary Computation*. Singapore. (2007), 4661–4667.

[7] S. Hosseini, A. Al Khaled, Appl. Soft Comput, 24 (2014) 1078.

[8] N. Bigdeli, K. Afshar, A. S. Gazafroudi, M. Y. Ramandi, Renew. Sust. Energ. Rev 27 (2013) 20.

[9] Z. Aliniya, S. A. Mirroshandel, Expert Syst Appl, 117 (2019) 243.

[10] R. Sayyadi kordAbadi, A. Alizadehdakhel, S. Dorani Shiraz, Russ. J. Phys. Chem. B, 11**(**2017) 307.

[11] Todeschini, R., Consonni, V., 2000, *Handbook of Molecular Descriptors*, Wiley-VCH.

[12] A. P. Toropova, A. A. Toropov, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski., J. Comput. Chem, 32 (2011) 2727.

[13] A. A. Toropov, A. P. Toropova, SE. Martyanov, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, Chemom. Intell. Lab., 109 (2011) 94.

[14] J. Veselinović, A. Veselinović, A. Toropov, A. Toropova, I. Damnjanović, G. Nikolić, Scientific Journal of the Faculty of Medicine in Niš, 31(2014) 95.

[15] J.-L. Lin, Y.-H. Tsai, C.-Y. Yu, M.-S. Li., Algorithms, **5** (2012**)** 433.

[16] A. M. Veselinović, JB. Milosavljević, A. A Toropov, GM. Nikolić., Eur. J. Pharm. Sci. 48 (2013) 532.

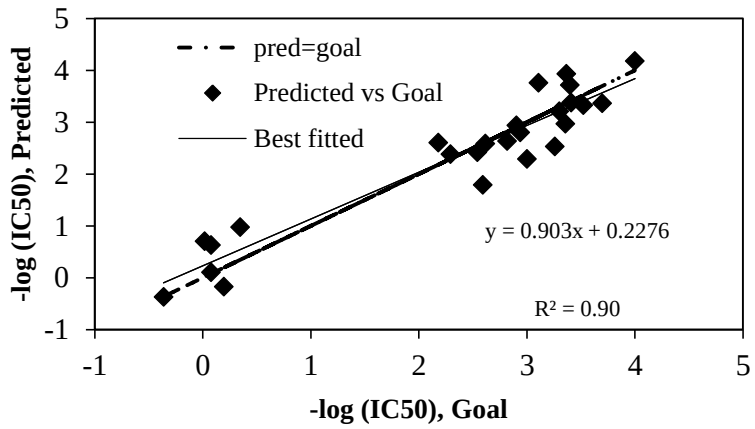
[17] (http://www.insilico.eu/coral)

[18] A. Golbraikh, A. Tropsha,. J .Mol. Graph. Model**.** 20 (2002); 269-276.

[19] S. Hosseini, M. R. Gholami, M. Haghgu., *J. Phys. Theor. Chem.* IAU Iran., 13 **(**2016) 171.

[20] L. Mahdavi, J. Phys. Theor. Chem. IAU Iran., 14 (2017), 103.

Graphical Abstract



مجله شیمی فیزیک و شیمی نظری
دانشگاه آزاد اسلامی واحد علوم و تحقیقات
جلد ۱۸، شماره ۲، تابستان ۱۴۰۰
ISSN ۱۷۳۵-۲۱۲۶

کاربرد روش مونت کارلو و یک روش بهینه‌سازی- مدل سازی جدید در مطالعه QSAR داروهای اتوپوزاید

امید علیزاده [۱]، ربابه صیادی کردآبادی*[۱]، قاسم قاسمی[۱]، بابک مطهری[۲]

[۱] گروه شیمی و مهندسی شیمی، واحد رشت، دانشگاه آزاد اسلامی، رشت، ایران
[۲] گروه مهندسی کامپیوتر، واحد رشت، دانشگاه آزاد اسلامی، رشت، ایران

چکیده

مونت کارلو و رگرسیون خطی چندگانه (MLR) و الگوریتم رقابتی استعماری (ICA) برای انتخاب مناسب‌ترین توصیف کننده‌ها استفاده شد. با بررسی کیفیت مدل با مقایسه میانگین خطای مربع (MSE) و ضریب همبستگی (R^2)، مشخص شد که ۱٤۰ مناسب‌ترین تعداد امپراتوری برای فاز گاز است. در روش مونت کارلو، از نرم‌افزار CORAL استفاده شد و داده‌ها به طور تصادفی به سه زیر مجموعه آموزش، کالیبراسیون و آزمون تقسیم شدند. ضریب همبستگی (R^2)، ضریب همبستگی معتبر متقاطع (Q^2) و خطای استاندارد مدل به ترتیب ۰,۹۳۰۱ ، ۰,۷۳۷۷ و ۰,۵۹۵ برای مجموعه آزمون با آستانه بهینه ٤ محاسبه شد. نتیجه گیری شد که استفاده همزمان از روش MLR-ICA و مونت کارلو می‌تواند به درک جامع‌تری از از رابطه بین توصیف کننده‌های فیزیکی- شیمیایی و ساختاری یا توصیف کننده‌های تئوری داروها با فعالیت‌های بیولوژیکی آن‌ها منجر شود و طراحی داروهای جدید را تسهیل کند.

کلید واژه‌ها: اتوپوزاید؛ QSAR؛ الگوریتم ICA؛ روش مونت کارلو

* مسئول مکاتبات: sayyai@iaurasht.ac.ir & Sayyadi_04@yahoo.com