Regular Article

ORGANIC CHEMISTRY RESEARCH
Published by the
Iranian Chemical Society
www.orgchemres.org

Org. Chem. Res., Vol. 7, No. 2, 176-189, July 2021.
DOI: 10.22036/org.chem.2023.325681.1273

Novel Modelling-optimization Approach and Monte Carlo Method on QSAR Study of Bortezomib Drugs

O. Alizade, R. Sayyadi Kord Abadi* and G. Ghasemi
Department of Chemistry and Chemical Engineering, Rasht Branch, Islamic Azad University, Rasht, Iran
(Received 20 January 2022, Accepted 22 June 2023)

Multiple linear regression (MLR) as modeling tool and Imperialist Competitive Algorithm (ICA) as optimization techniques employed to choose the best set of descriptors and The CORAL software has been used as a tool for linear prediction of -log(IC_{50}) (empirical negative logarithm of half of maximal inhibitory concentration) for Bortezomib derivatives. A high predictive ability was observed for the MLR-ICA model with the best number of empires/imperialists (nEmp) 90 with root-mean-sum-square errors (RMSE) of 0.0121 and correlation coefficient ($R^2_{predict}$) of 0.9896 in gas phase. The 25 data sets were randomly splitted into the training set, the calibration set, the test set in the Monte Carlo method and the number of compounds in the each set (n), correlation coefficient (R^2), cross-validated correlation coefficient (Q^2), standard error(s) were calculated 13, 0.9826, 0.9780, 0.161 in training set; and n = 6, R^2 = 0.8463, Q^2 = 0.7377, s = 0.715 in test set in the Threshold (T) of 2 and probe of 3, respectively. From the MLR-ICA method, it was revealed that Espm15u, R5p+, B06 [O-O], F03[N-N], F07[C-O], MATs3m, RDF125v are the most important descriptors. From Monte Carlo simulations, it was found that the presence of double bond and ring, absence of halogens are the most important molecular features affecting the biological activity of the drug. It was concluded that simultaneous utilization of MLR-ICA and Monte Carlo method can lead to a more comprehensive understanding of the relation between physico-chemical, structural or theoretical molecular descriptors of drugs to their biological activities and facilitate designing of new drugs.

Keywords: Bortezomib, QSAR, ICA Algorithm, Monte Carlo method

INTRODUCTION

Bortezomib is important in the treatment of multiple myeloma that It is also being investigated for the treatment of other hematological malignancies and solid tumors as a single agent or as part of a combined therapy [1-5].

One of the most efficacious approaches for designing new chemical identities and understanding the action mechanisms of drugs is quantitative structure activity relationship (QSAR) [6-9] with several variable selection models including multiple linear regression (MLR), genetic algorithm (GA), simulated annealing algorithm (SA) *etc.* [10].

Imperialist Competitive Algorithm (ICA) is a new population-based optimization algorithm that has recently been introduced for dealing with different kinds of optimization problem [11-14]. The total power of an empire depends on both the power of its colonies and power of the imperialist country because the ICA starts with an initial population called countries and most powerful countries are selected as imperialists and the rest form the colonies of these imperialists and 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 [13,15-16].

The CORAL (Correlation And Logic) software with the Monte Carlo method [23-24] was utilized to find simplified molecular input-line entry system (SMILES)-based

*Corresponding author. E-mail: Sayyadi@iaurasht.ac.ir

descriptors and calculated the correlation weights of the related SMILES [25-27] attributes.

In the present study, multiple linear regressions as linear modeling tool and Imperialist Competitive Algorithm as optimization method and Monte Carlo method were applied to investigate the QSAR in some Bortezomib anticancer drugs.

COMPUTATIONAL METHODS

Selection of Descriptors Using MLR-ICA Method

Details of geometry optimizations of compounds were given in our previous work [22]. As it was described, B3lyp/6-311g by Gaussian 09W [17,34] was utilized to optimize the geometries of 25 Bortezomib anti-cancer drugs and Dragon program [35] was used for calculation of 3226 molecular descriptors for each of the 25 compounds [22]. Modeling and optimizing calculations were carried out using Matlab. 2014a [37].

The 776 SPSS [18] screened descriptors [22] were used as the feed to ICA-MLR approach as the population matrix in order to find the best descriptors for the gas phase. The numbers of the most effective descriptors (7 for the gas phase) 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 Fig. 1. An efficient ICA algorithm using random points (matrix indices of descriptors) called the initial Countries that are the counterpart of Chromosomes in GA and it is a set of values of a candidate solution for optimization problem.

Empires are sub-populations of countries. Assimilation, which can be considered as a primitive form of Particle Swarm Optimization [19-21] moves all non-best countries (called colonies) in an empire toward the best country (called imperialist) in the same empire 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.

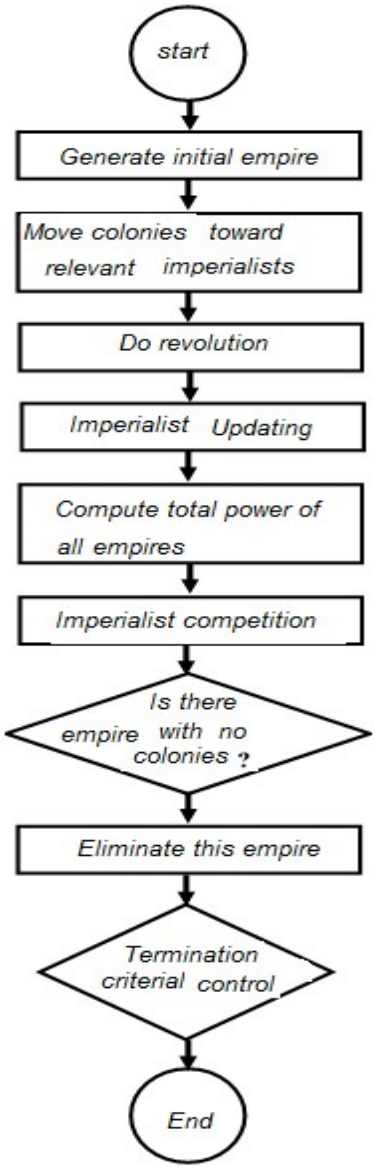

Fig. 1. Flowchart of the employed MLR-ICA algorithm.

In ICA-MLR method the number of decision variables (nDes)) and number of empires/imperialists (nEmp) were considered 1 up to 8 in the gas phase and 10, 20, 30, respectively.

MONTE CARLO METHOD

CORAL [33] software was used for calculation of

descriptor correlation weight (DCW) of the 25 Bortezomib compounds with a hybrid optimization scheme including hydrogen-suppressed molecular graph (HSG), hydrogen-filled graphs (HFG) and SMILES representation of molecular structures. Modelling using CORAL software was carried out for thresholds of 1 up to 7 and 100 epochs (*i.e.*, an overall number of 2100 runs were performed). The threshold (T) is a coefficient used to classify SMILES attributes into two groups (1) noise; and (2) active. The threshold that gives desirable statistical quality is denoted by T*. The noise attributes are blocked (their correlation weights are set to zero). Each sequence of computations for finding a new set of modified correlation weights of the model is named an epoch (29). The SMILES-based and Graph -based optimal descriptors are achieved using the following equations [28]:

$$DCW(T,Nepoch)^{SMILES} = \alpha\sum CW(Sk) + \beta\sum CW(SSk) + \gamma\sum CW(SSSk) + x.CW(NOSP) + y.CW(HALO) + z.CW(BOND) \quad (1)$$

$$DCW(T,Nepoch)^{Graph} = \sum CWAk + \alpha\sum CW({}^{0}ECk) + \beta\sum CW({}^{1}ECk) + \gamma\sum CW({}^{2}ECk) + \delta\sum CW({}^{3}ECk) \quad (2)$$

Where, S_k, SS_k, and SSS_k denote one, two, and three component SMILES attributes. The presence or absence of chemical elements are demonstrated by NOSP (nitrogen, oxygen, sulfur, and phosphorus) and HALO (fluorine, chlorine, and bromine). Also "BOND" denotes double (=), triple (#), or stereo chemical bonds (@ or @@). A_k in Eq. (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:

$$DCW(T,Nepoch)^{Hybrid} = DCW(T,Nepoch)^{SMILES} + DCW(T,Nepoch)^{Graph} \quad (3)$$

RESULTS AND DISCUSSION

Molecular Descriptors Generation with MLR-ICA Method

All studied Bortezomib compounds (22) have been presented in Fig. 2.

As a first trial, 1000 number of iterations were done 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. 3. It implies that there is no variation in the best cost (MSE) after about 300 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) was investigated and the results are plotted in Fig. 4. 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 check the robustness of the selected descriptors by varying the model parameters, the effects of number of empires on them were studied by changing the number of empires from 10 to 20 and 30.

For choosing the most suitable number of empires, the model was run using different number of empires and 7 number of descriptors (according to our previous work [22]. The results are revealed in Fig. 5. According to these results, the optimum number of empires for the gas phase is 90.

A plot of the predicted versus empirical values of -logIC50 is depicted Fig. 6. The figure implies that the developed model possesses a high correlation coefficient, indicating that the experimental and predicted values are well correlated. The observed and predicted values of -logIC50 using Matlab program are shown in Table 1.

The last chosen descriptors with nEmp = 90 and nDes = 7 using MLR-ICA Method have been presented in Table 2 in the gas phase.

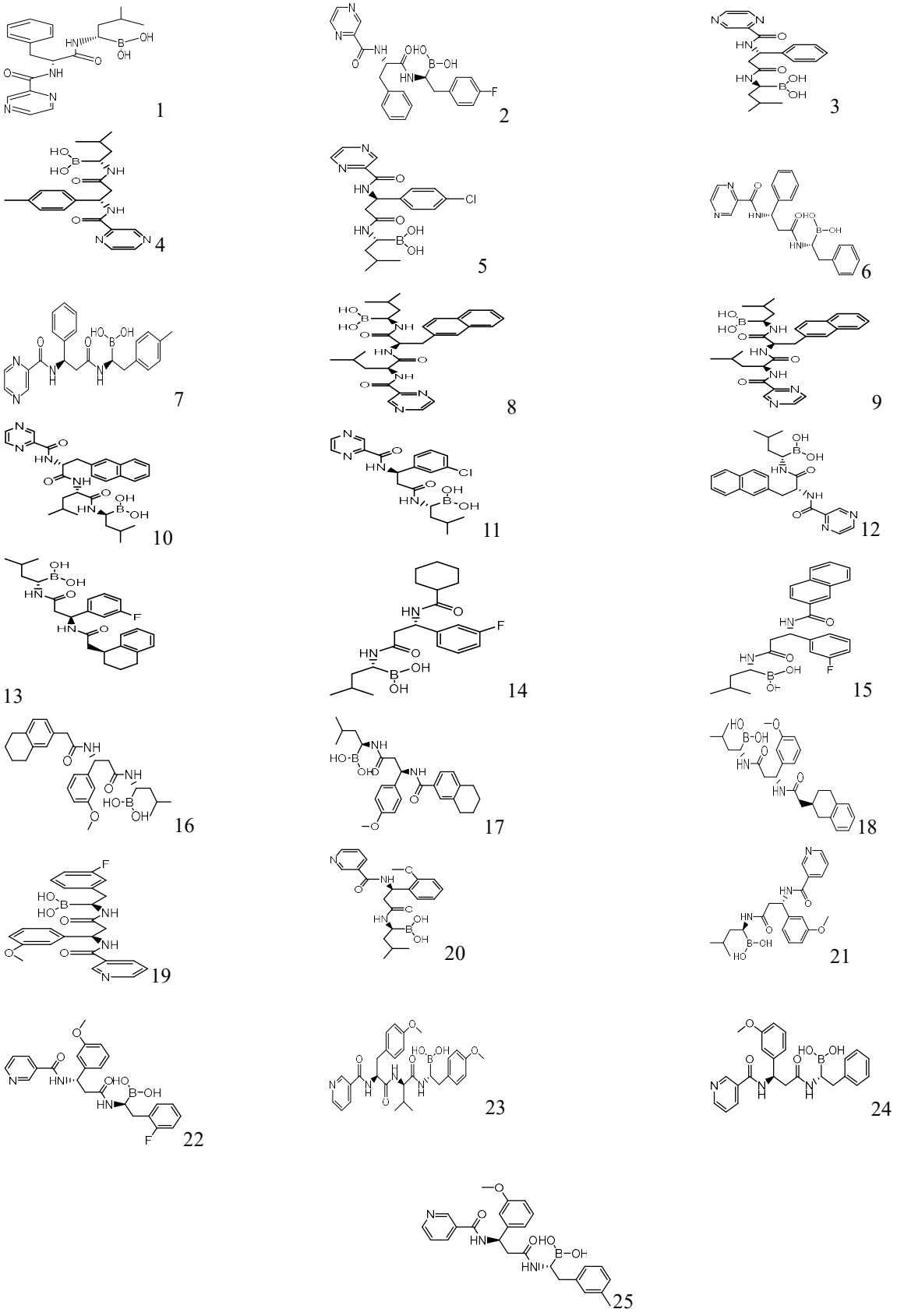

Fig. 2. Optimized structure of the compounds used to build QSAR models by B3lyp/6-31g in gas phase [22].

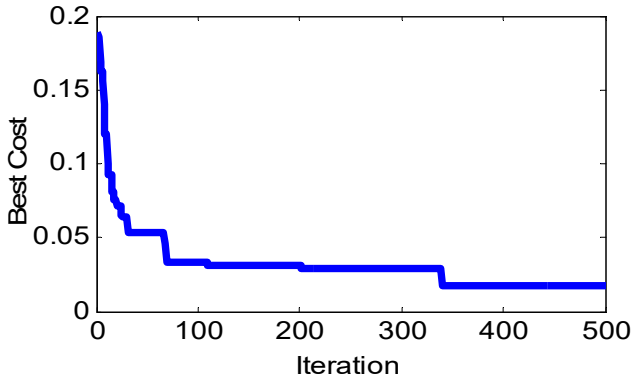

Fig. 3. Plot between Best Cost values compared to the variation of Iteration.

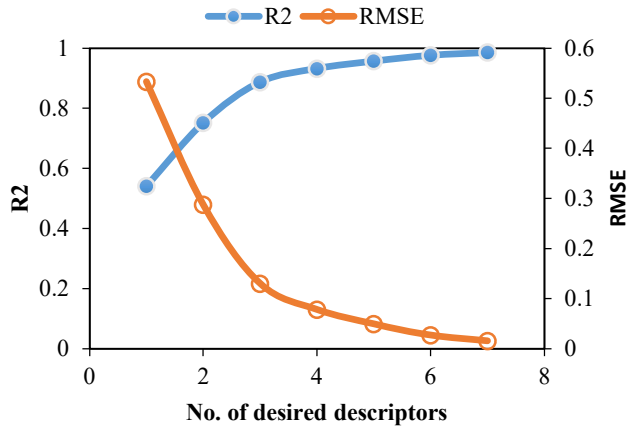

Fig. 4. Effects of number of descriptors on R^2 and RMS of the model.

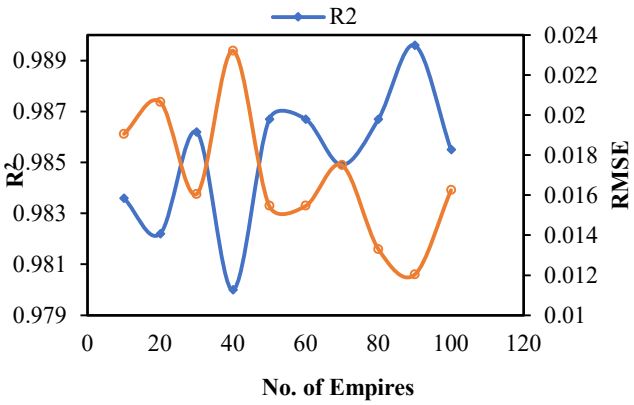

Fig. 5. Variation of R^2 and MSE by varying the number of empires (nDes = 7).

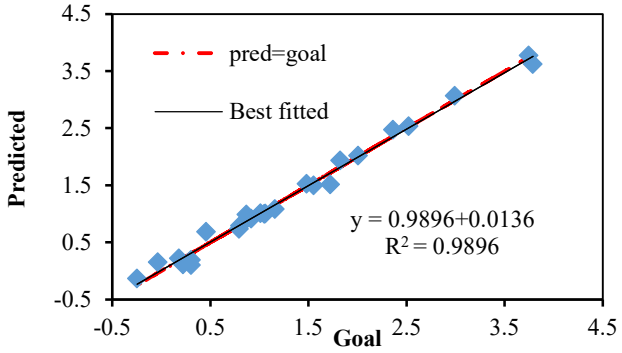

Fig. 6. Plot between predicted values compared to Goal with nDes = 7 and nEmp = 90.

Table 1. Observed and Predicted Values of -logIC50 by Using ICA-MLR

Predicted	-logIC50	Predicted	-logIC50
3.624822	3.788	0.100947	0.304
1.010875	1.01	2.021953	2.004
3.773883	3.745	1.933037	1.824
2.534196	2.521	1.514501	1.721
0.789456	0.801	0.191983	0.301
0.737076	0.793	0.999218	1.05
1.497436	1.551	0.109389	0.22
0.149343	-0.037	1.005477	1.06
-0.13414	-0.248	2.476321	2.36
0.985788	0.866	0.683296	0.456
1.085169	1.156	1.5252	1.48
0.914262	0.917	3.066555	2.991
0.215956	0.178		

It shows that polarizabilities (R5p+) and Presence/absence of O-O (B06[O-O]), Frequency of N-N(F03[N-N]), Frequency of C-O(F03[C-O]), atomic van der Waals volumes (RDF125v), and atomic masses (MATs3m) in the gas phase are important for designing this class of drugs.

The graphs of Espm15u, R5p+, B06 [O-O], F03[N-N], F07[C-O], MATs3m, RDF125v descriptors in the gas phase *versus* the half maximal inhibitory concentration (IC50) were plotted (Fig. 7).

Table 2. The Best Selected Descriptors Using MLR-ICA Method with nDes = 7 and nEmp = 90 in Gas Phase

Descriptor	Definition	Type
Espm15u	Spectral moment 15 from edge adjacent matrix	Edge adjacency indices
R5p+	R maximal autocorrelation of lag 5/weighted by atomic polarizabilities	GETAWAY descriptors
B06[O-O]	Presence/absence of O-O at topological distance 06	2D Binary fingerprints
F03[N-N]	Frequency of N-N at topological distance 03	2D Frequency fingerprints
F03[C-O]	Frequency of C-O at topological distance 07	2D Frequency fingerprints
MATs3m	Moran autocorrelation-lag3/weighted by atomic masses	2D Autocorrelations
RDF125v	Radial Distribution Function-12.5/Weighted by atomic van der Waals volumes	RDF Descriptors

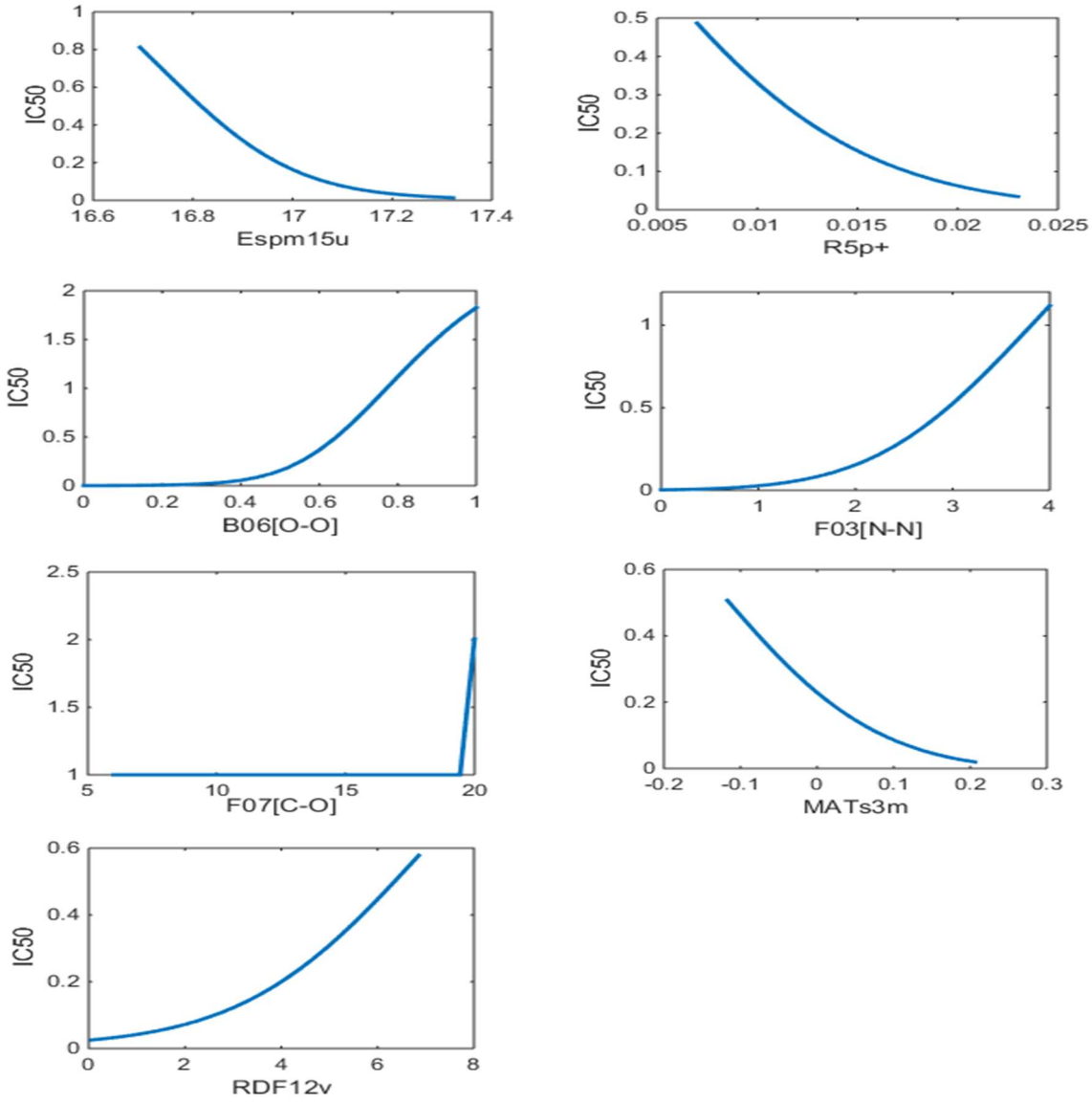

Fig. 7. Plot between IC50 experimental *versus* the Espm15u, R5p+, B06[O-O], F03[N-N], F07[C-O], MATs3m, RDF125v descriptors.

The charts in the gas phase show that IC50 value decreases *via* increasing Espm15u, R5p+ and, MATs3m descriptors and decreasing B06[O-O], F03[N-N], and RDF125v descriptors. As the F07[C-O] descriptor increased from 15 to near 20, no changes in IC50 value was observed. Thus during this period, a bar was seen in the response.

Statistical parameter and QSAR model of the compounds from the previous literatures are presented on the Table 2 [22,6,7,36]. It shows that the results of

ICA-MLR method in this work (Table 3) is better than the other QSAR models in previous studies.

Result of the Monte Carlo Method

The statistical parameters of the models obtained with molecular graphs (HSG) and SMILES with Eq. (3) are shown in Tables 4-6. The performance of the employed splits were compared with each other by the criterion of the predictability (Rm^2) in test set which should be larger than 0.5 [30] and correlation coefficient (R^2) in each set, cross-validated correlation coefficient (Q^2), standard error of estimation (s), Fischer F-ratio (F). The criterion of the predictability (Rm^2) is mean of R^2m and R'^2m values in test set.

Table 3. Statical Parameter and QSAR Model from the Previous Literatures

Compounds	QSAR model	Statistical parameters
Sulfonamide derivatives	GA-ANN	$R^2pred = 0.9894$
BORTEZOMIB derivatives	GA-ANN	$R^2pred = 0.9728$
Taxol derivatives	GA-ANN	$R^2pred = 0.82$
Several etoposides	GA-ANN	$R^2pred = 0.966$

Table 4. The Split Models in Monte Carlo Method

Split 1: (T = 2)
-logIC50 = -11.9140264 (± 0.1194294) + 0.1899727 (± 0.0015612) * DCW(2,100)
n = 13, $R^2 = 0.9826$, $Q^2 = 0.9780$, s = 0.161 (training set)
n = 6, $R^2 = 0.9998$, $Q^2 = 0.9995$, s = 0.987 (calibration set)
n = 6, $R^2 = 0.8463$, $Q^2 = 0.7377$, s = 0.715 (test set), R^2m TEST = 0.6699
Spit 2: (T = 2)
-logIC50 = -26.7778528 (± 0.1746202) + 0.3154204 (± 0.0019979) * DCW(2,100)
n = 11, $R^2 = 0.9983$, $Q^2 = 0.9973$, s = 0.045 (training set)
n = 8, $R^2 = 0.9983$, $Q^2 = 0.9956$, s = 0.254 (calibration set)
n = 6, $R^2 = 0.8334$, $Q^2 = 0.7186$, s = 1.21, R^2m TEST = 0.8163
Spit 3: (T = 2)
-logIC50 = -22.9169943 (± 1.3511914) + 0.4943667 (± 0.0283896) * DCW(7,100)
n = 9, $R^2 = 0.9180$, $Q^2 = 0.8501$, s = 0.360 (training set)
n = 7, $R^2 = 0.8404$, $Q^2 = 0.0.7432$, s = 1.19 (calibration set)
n = 9, $R^2 = 0.694$, $Q^2 = 0.0994$, s = 1.82 (test set), R^2m TEST = 0.6383

Table 5. Statistical Data Calculated with Both HSG, HFG and SMILES 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	0.7366	0.7267	0.7153	0.7262	0.0087
2	0.8160	0.8201	0.8471	0.8277	0.0138
3	0.7507	0.7407	0.7493	0.7469	0.0044
4	0.7502	0.7545	0.7419	0.7488	0.0052
5	0.6698	0.6874	0.6993	0.6855	0.0121
6	0.6260	0.6341	0.5799	0.6133	0.0239
7	0.6183	0.6185	0.6064	0.6144	0.0057
SPLIT2					
1	0.8015	0.7756	0.7981	0.7917	0.0115
2	0.8223	0.8292	0.8339	0.8285	0.0048
3	0.6503	0.6799	0.6732	0.6678	0.0126
4	0.7769	0.8002	0.7882	0.7884	0.0095
5	0.8150	0.8060	0.8199	0.8136	0.0057
6	0.5296	0.5164	0.4975	0.5145	0.0132
7	0.5303	0.4826	0.5048	0.5059	0.0195
SPLIT3					
Continue Table 4					
1	0.6166	0.6160	0.6176	0.6168	0.0007
2	0.7764	0.7760	0.7760	0.7762	0.0002
3	0.6654	0.6636	0.6939	0.6743	0.0139
4	0.6557	0.6636	0.6683	0.6625	0.0052
5	0.6394	0.6417	0.6573	0.6462	0.0079
6	0.6443	0.6633	0.6478	0.6518	0.0083
7	0.6102	0.6091	0.6097	0.6097	0.0004

The results of the splits proved that split 1 with threshold and probe equal 2, 3 (Tables 3, 4, 5), were better than other splits because this model have good predictability according to criterion R^2m [30,32] which should be larger than 0.5.

The given experimental and predicted in Table 7 are plotted against each other in Fig. 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.

The variation of correlation coefficient (test set) with respect to threshold and the number of epochs are plotted in Figs. 9A, 9B. Figure 9A show that with increase in threshold, the correlation coefficient between experimental and calculated values of endpoint for training and test set were reduced and in threshold equai 2 value correlation coefficient for test set has a maximum. In addition, as the number of epochs of the Monte Carlo method optimization increased, the correlation coefficient for the training and calibration and test sets, were increased as well (Fig. 9B).

Table 6. Statistical Quality of Models Calculated with Both HSG, HFG and SMILES for Training, Calibration and Test Sets in Threshold and Probe Equal 2 and 1-3, Respectively. Best Model are Indicated by Bold

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
2-1	0.9855	0.9818	0.147			
Training (n = 13)						
Calib (n = 6)	0.9997	0.9991	0.978			
Test (n = 6)	0.8154	0.6739	0.814	0.6613	0.7982	0.1369
2-2	0.9844	0.9802	0.153			
Training (n = 13)						
Calib (n = 6)	0.9998	0.9996	0.965			
Test (n = 6)	0.8193	0.7071	0.7071	0.6782	0.8107	0.1325
2-3	0.9826	0.9780	0.161			
Training (n = 13)						
Calib (n = 6)	0.9998	0.9995	0.987			
Test (n = 6)	0.8463	0.7377	0.715	0.6699	0.7943	0.1244

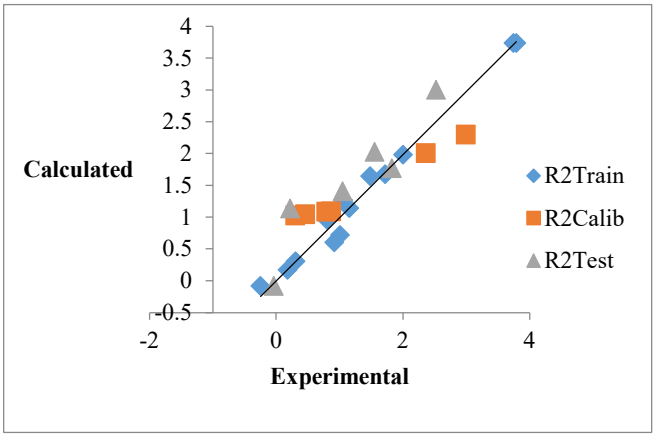

Fig. 8. Correlation between experimental and predicted -logIC50 calculated using Eq. (3).

Figure 9D show the correlation coefficient for the test set in the Threshold 1-5 and Nepoch (denoted N*) 1-100. It indicate that in Nepoch 80 the correlation coefficient for the test set is maximum.

The distribution of SMILES notations in the train, calibration and test sets are reported in Table 8.

Molecular features are sorted according to their correlation weights and are given in Table 9. 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 9.

Thus according to Table 10, presence of ring, absence of halogens, presence of double bond, present B element, present of F element, presence of nitrogen and oxygen together with absent sulfur and phosphorus, branch in molecular skeleton with B, F elements are increases of the -logIC50 and the half maximal inhibitory concentration (IC50) value decreases. Thus this work predicts in new design for this class of drugs, the presence or absence of these contradictions should be considered.

Table 7. Calculated Values for DCW, the Experimental Activity Data (-log IC50) and Calculated Values for -logIC50 with Application of CORAL in Split1 (T = 2)

Compound	Set	DCW	Exp.	Calc.
1	Train	82.40971	3.788	3.7416
2	Train	66.51002	1.01	0.7211
3	Train	82.40971	3.745	3.7416
5	Train	67.79072	0.801	0.9644
9	Train	62.29971	-0.248	-0.0788
11	Train	68.75969	1.156	1.1484
12	Train	65.90575	0.917	0.6063
13	Train	63.62701	0.178	0.1734
14	Train	64.34652	0.304	0.3101
15	Train	73.17845	2.004	1.9879
17	Train	71.55459	1.721	1.6794
21	Train	69.40098	1.06	1.2703
24	Train	71.40293	1.48	1.6506
6	Calib	68.45861	0.793	1.0912
10	Calib	68.47977	0.866	1.0953
18	Calib	68.13753	0.301	1.0302
22	Calib	69.45434	2.36	2.011
23	Calib	68.23210	0.456	1.0482
25	Calib	69.83551	2.991	2.302
4	Test	80.15323	2.521	3.01
7	Test	76.00969	1.551	2.03
8	Test	62.29971	-0.037	-0.0788
16	Test	72.04633	1.824	1.7728
19	Test	70.14340	1.05	1.4113
20	Test	68.71192	0.22	1.1394

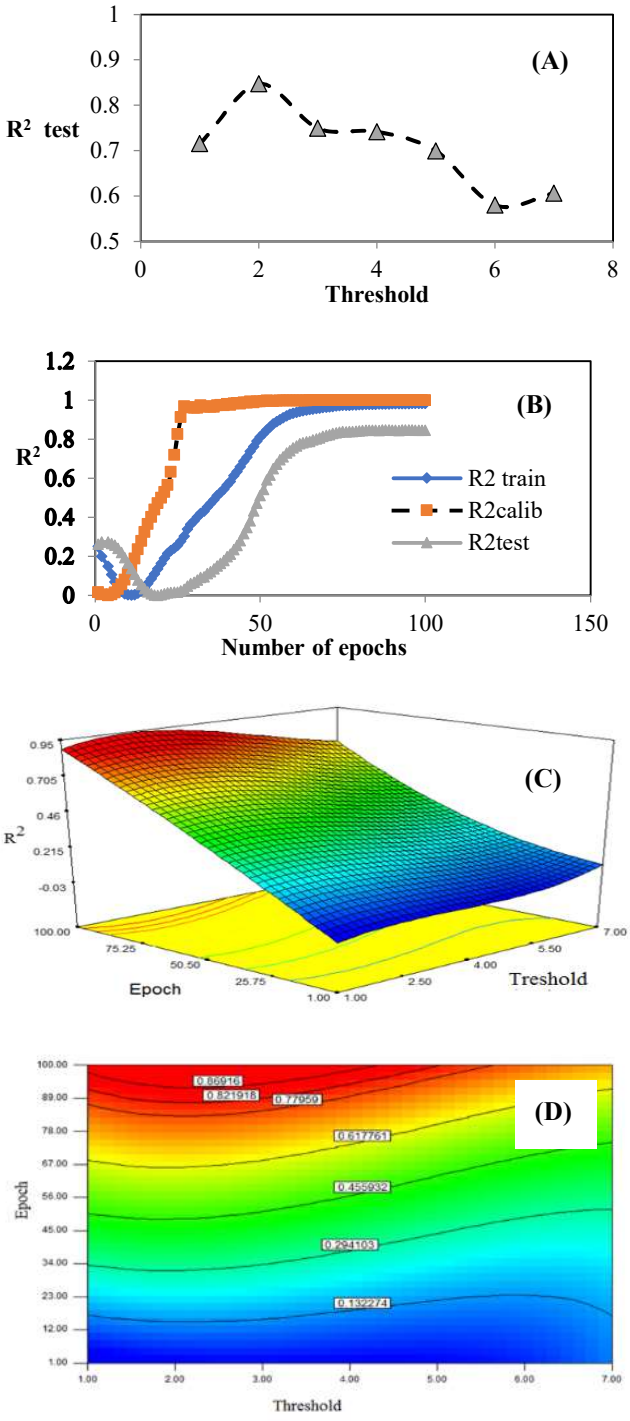

Fig. 9. The variation of correlation coefficient for test set by threshold and number of epochs. (A): effects of threshold. (B) Effects of the number of epochs. (C) 3-D surface plot of R^2 according to the threshold and the number of epochs. (D) Contour plots of R^2 according to the threshold and the number of epochs.

Table 8. SMILES Notations 25 Compound of Bortozomib and Train, Calibration and Test Set

Compound	SMILES	Set
1	CC(C)CC(NC(=O)C(CC1=CC=CC=C1)NC(=O)C2=NC=CN=C2)B(O)O	Train
2	OB(O)C(CC1=CC=C(F)C=C1)NC(=O)C(CC2=CC=CC=C2)NC(=O)C3=NC=CN=C3	Train
3	CC(C)CC(NC(=O)CC(NC(=O)C1=NC=CN=C1)C2=CC=CC=C2)B(O)O	Train
5	CC(C)CC(NC(=O)CC(NC(=O)C1=NC=CN=C1)C2=CC=C(Cl)C=C2)B(O)O	Train
9	CC(C)CC(NC(=O)C(CC(C)C)NC(=O)C(CC1=CC2=CC=CC=C2C=C1)NC(=O)C3=NC=CN=C3)B(O)O	Train
11	COC1=CC=C(C=C1)C(CC(=O)NC(CC(C)C)B(O)O)NC(=O)C2=CC=CC=C2	Train
12	CC(C)CC(NC(=O)C(CC1=CC2=CC=CC=C2C=C1)NC(=O)C3=NC=CN=C3)B(O)O	Train
13	CC(C)CC(NC(=O)CC(NC(=O)CC1CCCC2=C1C=CC=C2)C3=CC(=CC=C3)F)B(O)O	Train
14	CC(C)CC(NC(=O)CC(NC(=O)C1CCCCC1)C2=CC(=CC=C2)F)B(O)O	Train
15	CC(C)CC(NC(=O)CC(NC(=O)C1=CC2=CC=CC=C2C=C1)C3=CC=CC(=C3)F)B(O)O	Train
17	COC1=CC=C(C=C1)C(CC(=O)NC(CC(C)C)B(O)O)NC(=O)C2=CC3=C(CCCC3)C=C2	Train
21	COC1=CC(=CC=C1)C(CC(=O)NC(CC(C)C)B(O)O)NC(=O)C2=CC=CN=C2	Train
24	COC1=CC(=CC=C1)C(CC(=O)NC(CC2=CC=CC=C2)B(O)O)NC(=O)C3=CN=CC=C3	Train
6	OB(O)C(CC1=CC=CC=C1)NC(=O)CC(NC(=O)C2=NC=CN=C2)C3=CC=CC=C3	Calib
10	CC(C)CC(NC(=O)CC(NC(=O)C1=NC=CN=C1)C2=CC=CC(=C2)Cl)B(O)O	Calib
18	COC1=CC(=CC=C1)C(CC(=O)NC(CC(C)C)B(O)O)NC(=O)CC2CCC3=C(C2)C=CC=C3	Calib
22	COC1=CC(=CC=C1)C(CC(=O)NC(CC2=C(F)C=CC=C2)B(O)O)NC(=O)C3=CN=CC=C3	Calib
23	COC1=CC=C(CC(NC(=O)C(NC(=O)C(CC2=CC=C(OC)C=C2)NC(=O)C3=CC=CN=C3)C(C)C)B(O)O)C=C1~	Calib
25	COC1=CC=CC(=C1)C(CC(=O)NC(CC2=CC(=CC=C2)C)B(O)O)NC(=O)C3=CC=CN=C3	Calib
4	CC(C)CC(NC(=O)CC(NC(=O)C1=NC=CN=C1)C2=CC=C(C)C=C2)B(O)O	Test
7	CC1=CC=C(CC(NC(=O)CC(NC(=O)C2=NC=CN=C2)C3=CC=CC=C3)B(O)O)C=C1	Test
8	CC(C)CC(NC(=O)C(CC1=CC=C2C=CC=CC2=C1)NC(=O)C(CC(C)C)NC(=O)C3=NC=CN=C3)B(O)O	Test
16	COC1=CC(=CC=C1)C(CC(=O)NC(CC(C)C)B(O)O)NC(=O)CC2=CC3=C(CCCC3)C=C2	Test
19	COC1=CC=CC(=C1)C(CC(=O)NC(CC2=CC=CC(=C2)F)B(O)O)NC(=O)C3=CC=CN=C3	Test
20	COC1=C(C=CC=C1)C(CC(=O)NC(CC(C)C)B(O)O)NC(=O)C2=CC=CN=C2	Test

Table 9. SMILES Attributes with Positive Correlation Weights for Split 1

SMILES attributes	CWs	SMILES attributes	CWs
=...1.....	12.35152	=...(.....	1.96591
HALO00000000	10.68296	EC0-C...4...	1.91069
2...(.....	10.60929	O...=.....	1.88524
=...2.....	7.17354	EC0-F...1...	1.83472
3...(.....	6.75474	O...C.....	1.78698
EC0-B...3...	6.7398	F.....	1.74197
1.....	6.42055	C...=.....	1.66038
C6.....2...	6.22602	N...(.....	1.17831
B...(.....	6.1294	C...C.....	1.0846
NOSP11000000	5.6663	EC0-N...3...	0.83306
2.....	5.19031	EC0-N...2...	0.71064
C6....H.2...	4.99018	C...(.....	0.70107
B.....	4.93344	N...=.....	0.67825
BOND10000000	4.45183	EC0-O...2...	0.48824
1...(.....	2.52393	N...C.....	0.43716
F...(.....	2.46682	O...(.....	0.23297
EC0-O...1...	2.27044	EC0-C...3...	0.19768

Table 10. Definition of the SMILES Attributes

SMILES attributes	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

CONCLUSIONS

In this study, MLR-ICA and Monte Carlo investigations were used to study the structure-activity relationships of 25 Bortezomib Anticancer Drugs. The best descriptors with ICA and nEmp = 90 proved that Espm15u, R5p+, B06[O-O], F03[N-N], F07[C-O], MATs3m, RDF125v descriptors in the gas phase were more significant than other descriptors to create QSAR model and predict biological activity of Bortezomib substitution patterns.

Noting that the aforementioned descriptors are the most effective descriptors in MLR-ICA methods, atomic polarizabilities and atomic masses should be maximized and absence of O-O, Frequency of N-N, atomic van der Waals volumes should be minimized in designing new drugs

The biological activity of the Bortezomib inhibition was predicted with tree random splits into the sub-training, calibration, and test sets in the Monte Carlo method. The best results were obtained in split 1 with n = 6, $R^2 = 0.8463$, $Q^2 = 0.7377$, s = 0.715 in test set and by both molecular graph (HSG) and SMILES.

Monte Carlo method revealed that presence of B element, presence of nitrogen and oxygen together with absent sulfur and phosphorus are the most important molecular features

The most important physicochemical and structural descriptors were presented and discussed. It was concluded that the simultaneous use of Monte Carlo and linear and non-linear methods gives deeper and more comprehensive knowledge about the effects of molecular and structural descriptors on the activity of drugs and provides better insights to design new drugs.

ACKNOWLEDGEMENT

The support of Rasht Branch Islamic Azad University, is gratefully acknowledged.

REFERENCES

[1] M.A. Gertz, Am. J. Hematol. 93 (2018) 1169.

[2] R.C. Kane, A.T. Farrell, R. Sridhara, R. Pazdur, Clin. Cancer Res. 12 (2006) 2955.

[3] Q.P. Dou, R.H. Goldfarb, Idrugs. 5 (2002) 828.

[4] R.I. Fisher, S.H. Kahl Bernstein, B.S.B. Djulbegovic, M.J. Robertson, S. de Vos, E. Epner, A. Krishnan, J.P. Leonard, S. Lonial, E.A. Stadtmauer, O.A. O'Connor, H. Shi, A.L. Boral, A. Goy, J. Clin. Oncol. 24 (2006) 4867.

[5] C.N. Papandreou, D.D. Daliani, D. Nix, H. Yang, T. Madden, X. Wang, C.S. Pien, R.E. Millikan, L. Pagliaro, J. Kim, J. Adams, P. Elliott, D. Esseltine, A. Petrusich, P. Dieringer, C. Perez, CJ. Logothetis. J. Clin. Oncol. 22 (2004) 2108.

[6] R. SayyadikordAbadi, A. Alizadehdakhel, Rev. Roum. Chim. 63 (2018) 171.

[7] R. SayyadikordAbadi, A. Alizadehdakhel, S. Tajadodi Paskiabei, J. Korean Chem. Soci. 60 (2016) 225.

[8] V.K. Gupta, D. Dina Brauneis, A.C. Shelton, K. Quillen, S. Sarosiek, J.M. Sloan, V. Sanchorawala, Biology of Blood and Marrow Transplant. 25 (2019) e169.

[9] H.K. Srivastava, F.A. Pasha, P.P. Singh, Int. J. Quantum. Chem. 103 (2005) 237.

[10] V.R. Consonni, M. Todeschini, M. Pavan, J. Chem. Inf. Comput. Sci. 42 (2002) 693.

[11] T. Niknam, E. Taherian Fard, N. Pourjafarian, A. Rousta. Eng. Appl. Artif. Intell. 24 (2011) 306.

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

[13] E. Shokrollahpour, M. Zandieh, B. Dorri, Int. J. Prod. Res. 49 (2011) 3087.

[14] M. Bagher, M. Zandieh, H. Farsijani, Int. J. Adv. Manuf. Tech. 54 (2011) 271.

[15] S.J. Mousavi Rad, T.F. Akhlaghian, K. Mollazade, Int. J. Comput. Appl. 40 (2012) 41.

[16] S. Karami, S.B. Shokouhi, Int. J. Comput. Theory Eng. 4 (2012) 137.

[17] D. Horvath, B. Mao, Molecular Informatics 22 (2003) 498.

[18] S. Putta, J. Eksterowicz, C. Lemmen, R. Stanton, J. Chem. Inf. Comput. Sci. 43 (2003) 1623.

[19] E. Atashpaz-Gargari, C. Lucas, Imperialist competitive algorithm: An algorithm for optimization inspired by imperialistic competition. In IEEE Congress on

Evolutionary Computation. Singapore (2007) 661.

[20] E.A. Gargari, F. Hashemzadeh, R. Rajabioun, C. Lucas, Int. J. Intell. Comput. Cybern. 1 (2008) 337.

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

[22] R. SayyadikordAbadi, A. Alizadehdakhel, Rev. Roum. Chim. 63 (2018) 931.

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

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

[25] D. Weininger, J. Chem. Inf. Comput. Sci. 28 (1988**)** 31.

[26] D. Weininger, A. Weininger, J.L. Weininger, J. Chem. Inf. Comput. Sci. 29 (1989) 97.

[27] D. Weininger, J. Chem. Inf. Comput. Sci. 30 (1990**)** 237.

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

[29] A.M. Veselinović, J.B. Milosavljević, A.A. Toropov, G.M. Nikolić, Eur. J. Pharm. Sci. 48 (2013) 3.

[30] A. Golbraikh, A. Tropsha, J. Mol. Graph. Model. 20 (2002) 269.

[31] A.A. Toropov, A.P. Toropova, E. Emilio Benfenati, Chem. Biol. Drug Des. 73 (2009) 442.

[32] P.K. Ojha, I. Mitra, R.N. Das, K. Roy, Chemometr. Intell. Lab. 107 (2011) 194.

[33] http://www.insilico.eu/coral/.

[34] M.J. Frisch, G.W. Trucks, H. B. Schlegel, G. E. Scuseria, M.A. Robb, J.R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G.A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H.P. Hratchian, A.F. Izmaylov, J. Bloino, G. Zheng, J.L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J.A. Montgomery, J.E. Peralta, F. Ogliaro, M. Bearpark, J.J. Heyd, E. Brothers, K.N. Kudin, V.N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S.S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J.E. Knox, J.B. Cross, V.C. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A.J. Austin, R. Cammi, C. Pomelli, J.W. Ochterski, R.L. Martin, K. Morokuma, V.G. Zakrzewski, G.A. Voth, P. Salvador, J.J. Dannenberg, S. Dapprich, A.D. Daniels, Ö. Farkas, J.B. Foresman, J.V. Ortiz, J. Cioslowski, D.J. Fox, *Gaussian 09* (Gaussian, Inc., Wallingford CT, 2009). https://gaussian.com/glossary/g09/

[35] Dragon 3.0 Evaluation Version. Available online: http://www.disat.unimib.it/chm

[36] R. Sayyadi Kord Abadi, A. Alizadehdakhe, S. Dorani Shiraz. Russ. J. Phys. Chem. B 11 (2017) 307.

[37] https.//www.mathworks.com