Journal of Chemical Health Risks

Journal of Chemical Health Risks (2016) 6(1), 49–67

ORIGINAL ARTICLE

Novel QSPR Study on the Melting Points of a Broad Set of Drug-Like Compounds Using the Genetic Algorithm Feature Selection Approach Combined With Multiple Linear Regression and Support Vector Machine

Alireza Jalali, Mehdi Nekoei*, Majid Mohammadhosseini

Department of Chemistry, College of Basic Sciences, Shahrood Branch, Islamic Azad University, Shahrood, Iran

(Received: 2 October 2015 Accepted: 8 December 2015)

KEYWORDS

Melting point;
QSPR;
Genetic algorithm (GA);
Support vector machine;
Multiple linear regression;
Molecular descriptors

ABSTRACT: A robust and reliable quantitative structure-property relationship (QSPR) study was established to forecast the melting points (MPs) of a diverse and long set including 250 drug-like compounds. Based on the calculated descriptors by Dragon software package, to detect homogeneities and to split the whole dataset into training and test sets, a principal component analysis (PCA) approach was used. Accordingly, there was no outlier in the constructed cluster. Afterwards, the genetic algorithm (GA) feature selection strategy was used to select the most impressive descriptors resulting in the best-fitted models. In addition, multiple linear regression (MLR) and support vector machine (SVM) were used to develop linear and non-linear models correlating the molecular descriptors and the melting points. The validation of the obtained models was confirmed applying cross validation, chance correlation along with statistical features associated with external test set. Our computational study exactly showed a determination coefficient and of 0.853 and a root mean square error (RMSE) of 11.082, which are better than those MLR model (R^2=0.712, RMSE 15.042%) accounting for higher capability of SVM-based model in prediction of the theoretical values related to melting points. In fact, using the GA approach resulted in selection of powerful descriptors having useful information concerning effective variables on MPs, which can be utilized in further designing of drug-like compounds with desired melting points.

* Corresponding author: m_nekoei1356@yahoo.com (M. Nekoei).

INTRODUCTION

The term melting point (M.P.) for a distinct substance implies its conversion from its study the status into liquid. In a variety of investigations, this term serves a key role specifically for

i) Fast measurement of the purity of the materials

ii) The identity of matters

iii) Predicting other related characteristics such as solubility in water, boiling points (BP), etc.[1, 2]. In turn, solubility is a very important character in drug design and in an assessment of the effective toxicity of chemicals and materials [3].

In fact, there is a close and strong relationship between the melting point of the compound and its solubility. Taking into account this point, proper modeling the solubility of a chemical substance prior to its synthesis is of prime importance [2-4].

Furthermore, sufficient solubility for transportation of a compound to active sites present in an organism is unavoidable. Due to a mutual relationship between the melting point and the solubility, a comprehensive attention should be paid in such sorts of studies. It is evident that for compounds having low solubility in water, one cannot expect appreciable toxicity in an aqueous medium. It also seems logical that chemical structure of a compound has a crucial impact in its corresponding melting point. To predict the MPs of chemical compounds, common strategies are quantitative structure-property relationship (QSPR), group contribution as well as property-property relationship (PPR). Comprehensive reviews concerning the subject explicitly show that in the majority of these studies hydrocarbons and their homologous compounds are the main subject of respective attempts.

This is due to the difficulty of the melting point prediction for various organic compounds, since the numerous factors affecting and controlling M.P. are not easy to quantify [5].

Quantitative structure-property relationship (QSPR) models are capable of relating the property of interest, with a set of molecular descriptors. These descriptors encode the chemical information and are related to certain physicochemical properties of the molecule [6]. In such studies, numerous physical properties of molecular systems have been successfully modeled, including enthalpy of vaporization, aqueous solubility, melting points and electrical conductivity of ionic liquids and half-wave potentials [7-14].

In linear QSPR modeling approaches, some methodologies like multiple linear regression (MLR), partial least squares (PLS) are frequently used. However, for nonlinear models diverse types of artificial neural networks (ANN) are being employed [9, 10]. In the case of complex and nonlinear systems, linear models face a big challenge. It should be noted that the main drawbacks of ANN-based models are overtraining, the way of training, optimization of the network, overfitting and insufficient reproducibility in the obtained results. Due to these reasons, a more accurate and informative modeling technique is desirably needed, which can be effectively used in QSPR-based analyses.

The support vector machine (SVM) is fairly a new and a very promising classification and regression method developed by Vapnik [18]. The SVM approach automatically controls the flexibility of the resulting classifier on the training data. Accordingly, by the design of the algorithm, the deteriorating effect of the input dimensionality on the generalization ability is largely suppressed. Regarding the remarkable generalization performance of the SVM approach, it has gained much attention and extensive application in a variety of QSPR simulations [19-23].

The main purpose of this study was to search for efficient methods to build accurate quantitative relationship between the molecular structures and the melting points by using GA-MLR and GA-SVM techniques.

MATERIAL AND METHODS

Hardware and software characteristics of the computer

To perform our computations, an advanced Pentium IV computer (CPU at 3.06 GHz) having Windows 7.0 as operating system was used. Optimization of these structures was carried out using HyperChem software and processing of the descriptors was performed using Dragon 2.1 software. The linear modeling was done with SPSS software while the advanced calculations were conducted by MATLAB (Version 12, Math Works, Inc.).

Data set, structure optimization and molecular descriptors generation

The data set of the melting points of 250 drug-like compounds was taken from the values reported by Eddington et al. (Table1) [24]. At first, after drawing of all the chemical structures (250 drug -like compounds), the optimization of their geometry was performed using the AM1 algorithm. Regarding this fact that the numerical values of the molecule or descriptors depend on some general characteristics like bond length, angles, bound energy and …, the optimization step is of prime importance. To start the modeling process, we used the Dragon software to calculate the molecular descriptors. Consequently, 1481 molecular descriptors, from 18 different types of theoretical descriptor, were calculated for each molecule. Molecular descriptors include general characteristics of a chemical compound. These variables could be considered as results of rational and mathematical-based processing being performed on each molecule. The success of each modeling is strongly depends on the most effective descriptors correlating with the activity or property of the studied compounds.

Table1. The data set, - experimental and predicted values of melting points of a diverse set of drug-like compounds using GA-MLR and GA-SVM strategies for the training and test sets

No	Compound name	Exp.[a]	Pred. MLR [b]	Pred. SVM [c]
1	6,6-Dimethyl-2-oxo-4-phenethylamino-cyclohex-3-enecarboxylic acid methyl ester	130	131.45	136.98
2	4-Benzylamino-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid tert-butyl ester	152	139.93	151.43
3	4-(4-Chloro-benzylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	173	165.31	173.01
4	4-(4-Chloro-benzylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	169	151.11	163.53
5	4-(4-Chloro-benzylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid tert-butyl ester	182	155.26	164.93
6	6-Methyl-4-(4-methyl-benzylamino)-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	160	158.25	160.01
7	6-Methyl-4-(4-methyl-benzylamino)-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	134	142.47	145.27
8[d]	6-Methyl-2-oxo-4-phenethylamino-cyclohex-3-enecarboxylic acid phenethyl-amide	171	159.86	155.55
9[d]	4-(4-Cyano-benzylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	204	186.18	197.51
10[d]	4-(4-Cyano-benzylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	184	173.09	183.99
11	3-Benzylamino-cyclohex-2-enone	125	141.99	129.42
12	3-Benzylamino-5-methyl-cyclohex-2-enone	137.5	143.33	137.51
13	3-Benzylamino-5,5-dimethyl-cyclohex-2-enone	124	157.73	138.25
14[d]	3-(4-Chloro-benzylamino)-cyclohex-2-enone	170	166.47	169.99
15	3-(4-Chloro-benzylamino)-5-methyl-cyclohex-2-enone	186	167.13	169.73
16	3-(4-Chloro-benzylamino)-5,5-dimethyl-cyclohex-2-enone	159	177.78	163.72
17	3-(4-Methyl-benzylamino)-cyclohex-2-enone	153	142.83	139.43

Table 1. Continued

18	4-(4-Chloro-benzylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	173	158.22	166.99
19	5-Methyl-3-(4-methyl-benzylamino)-cyclohex-2-enone	146	147.69	145.99
20	5,5-Dimethyl-3-(4-methyl-benzylamino)-cyclohex-2-enone	139	155.44	150.54
21	3-(4-Methoxy-benzylamino)-5,5-dimethyl-cyclohex-2-enone	159	149.90	158.99
22	4-[(5,5-Dimethyl-3-oxo-cyclohex-1-enylamino)-methyl]-benzonitrile	215	208.57	214.99
23	4-Benzoylamino-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	178	175.13	179.41
24	6-Methyl-4-(4-methyl-benzylamino)-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	160	146.40	149.68
25[d]	4-(4-Chloro-benzoylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	200	188.53	196.34
26[d]	4-(4-Chloro-benzoylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	160	174.46	173.56
27	6-Methyl-4-(4-methyl-benzoylamino)-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	157	169.78	163.94
28	6-Methyl-4-(4-methyl-benzoylamino)-2-oxo-cyclohex-3-enecarboxylic acid tert-butyl ester	136	158.00	147.58
29	4-(4-Methoxy-benzoylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	166	166.30	165.99
30[d]	4-(4-Methoxy-benzoylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid tert-butyl ester	142	158.41	145.85
31	4-[(4-Methoxycarbonyl-5-methyl-3-oxo-cyclohex-1-enylamino)-methyl]-benzoic acid	231	206.45	214.96
32[d]	4-(4-Cyano-benzoylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	233	210.62	217.11
33	4-(4-Cyano-benzoylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	187	196.77	198.30
34	4-(4-Cyano-benzoylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid tert-butyl ester	212	201.36	204.69
35[d]	N-(3-Oxo-cyclohex-1-enyl)-benzamide	172	167.86	172.01
36	4-Chloro-N-(3-oxo-cyclohex-1-enyl)-benzamide	196	184.89	188.82
37	4-Chloro-N-(5-methyl-3-oxo-cyclohex-1-enyl)-benzamide	170	186.08	184.58
38	4-Chloro-N-(5,5-dimethyl-3-oxo-cyclohex-1-enyl)-benzamide	196	198.79	196.01
39	3-Benzylamino-5,5-dimethyl-cyclohex-2-enone	124	158.55	139.26
40	4-Methyl-N-(5-methyl-3-oxo-cyclohex-1-enyl)-benzamide	174	167.48	172.92
41	N-(5,5-Dimethyl-3-oxo-cyclohex-1-enyl)-4-methyl-benzamide	183	183.21	182.99
42	4-Methoxy-N-(3-oxo-cyclohex-1-enyl)-benzamide	171	156.87	167.16
43	4-Methoxy-N-(5-methyl-3-oxo-cyclohex-1-enyl)-benzamide	166	162.15	166.01
44	4-Cyano-N-(3-oxo-cyclohex-1-enyl)-benzamide	207	212.82	206.99
45	4-Cyano-N-(5-methyl-3-oxo-cyclohex-1-enyl)-benzamide	219	211.98	218.99
46	4-Cyano-N-(5,5-dimethyl-3-oxo-cyclohex-1-enyl)-benzamide	238	219.86	226.52
47	4-(4-Bromo-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid tert-butyl ester	192	166.56	160.81
48	6-Methyl-2-oxo-4-(4-trifluoromethoxy-phenylamino)-cyclohex-3-enecarboxylic acid methyl ester	158	178.56	167.52
49	6-Methyl-2-oxo-4-(4-trifluoromethoxy-phenylamino)-cyclohex-3-enecarboxylic acid ethyl ester	162	167.05	162.01
50	6-Methyl-2-oxo-4-(4-trifluoromethoxy-phenylamino)-cyclohex-3-enecarboxylic acid tert-butyl ester	168	179.44	168.01
51	4-(4-Iodo-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	193	174.17	175.98
52	4-(4-Iodo-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	160.5	159.84	159.46
53[d]	4-(4-Ethoxycarbonyl-5-methyl-3-oxo-cyclohex-1-enylamino)-benzoic acid	230	212.51	206.76
54	4-(4-Hydroxy-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	160	179.74	177.66
55	4-(4-Amino-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	164	192.77	179.56
56	4-(4-Amino-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	187	177.19	175.86
57	4-(4-Fluoro-benzylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	174	161.33	173.99
58	4-(4-Methoxycarbonyl-5-methyl-3-oxo-cyclohex-1-enylamino)-benzoic acid ethyl ester	180	168.07	159.26
59[d]	4-(4-Cyano-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	210	188.45	198.37
60	4-(4-Cyano-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	190	182.38	189.99
61	4-(4-Methoxy-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	177	148.10	154.78

Table 1. Continued

62	4-(3-Chloro-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	137	144.39	139.67
63	4-(3-Bromo-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	159	172.05	167.86
64[d]	4-(3-Bromo-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	135	154.49	148.36
65	6-Methyl-2-oxo-4-(3-trifluoromethoxy-phenylamino)-cyclohex-3-enecarboxylic acid ethyl ester	151	143.15	150.99
66	4-(3-Iodo-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	163	169.29	163.01
67[d]	4-(3-Iodo-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	120	150.38	144.79
68	3-(4-Ethoxycarbonyl-5-methyl-3-oxo-cyclohex-1-enylamino)-benzoic acid	201	198.56	187.86
69	4-(3-Hydroxy-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	188	186.49	187.43
70	4-(3-Hydroxy-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	165	176.41	165.01
71[d]	4-(3-Cyano-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	165	168.33	164.99
72	4-(3-Methoxy-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	111	127.84	124.35
73	6-Methyl-4-(3-nitro-phenylamino)-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	166.5	173.56	166.49
74	6-Methyl-4-(3-nitro-phenylamino)-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	155	166.28	155.01
75	4-(3-Ethyl-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	146	138.65	141.89
76[d]	6,6-Dimethyl-4-morpholin-4-yl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	132	145.53	132.01
77	6-Methyl-2-oxo-4-phenylamino-cyclohex-3-enecarboxylic acid methyl ester	141	140.80	141.01
78	4-(3-Ethyl-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	116	121.03	117.04
79	4-(3-Fluoro-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	151	148.26	150.99
80[d]	4-(2-Chloro-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	152	142.82	142.36
81	4-(2-Amino-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	151	152.88	150.99
82	4-(2-Bromo-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	157	153.28	155.15
83	4-(2-Iodo-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	148	154.77	152.06
84	4-(2-Hydroxy-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	164	179.22	172.23
85	4-(2-Carbamoyl-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	165	180.89	165.01
86	6-Methyl-2-oxo-4-(N'-phenyl-hydrazino)-cyclohex-3-enecarboxylic acid methyl ester	167	156.48	163.57
87[d]	4-(2-Cyano-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	155	167.70	161.10
88	4-(2-Methoxy-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	159	134.78	147.22
89	6-Methyl-4-(2-nitro-phenylamino)-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	154	150.31	154.01
90	6-Methyl-2-oxo-4-o-tolylamino-cyclohex-3-enecarboxylic acid methyl ester	138	127.85	137.00
91	6-Methyl-2-oxo-4-(2-trifluoromethyl-phenylamino)-cyclohex-3-enecarboxylic acid methyl ester	166	136.30	149.21
92[d]	4-(2-Fluoro-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	132	155.40	152.26
93[d]	4-(2-Methoxy-5-methyl-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	110	125.45	118.53
94[d]	4-(4-Ethyl-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	153.5	150.94	150.45
95[d]	4-(3-Chloro-4-methoxy-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	133	152.80	144.92
96	4-(2-Chloro-5-methoxy-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	138	155.00	147.47
97	4-(2,4-Dichloro-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	153	178.69	179.62
98	4-(2,4-Dichloro-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	153	158.15	154.15
99	4-(2,5-Dichloro-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	160	155.36	152.88
100	4-(3,4-Dichloro-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	171	171.51	170.99
101[d]	4-(2,6-Dichloro-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	205	163.78	164.81
102[d]	4-(3,5-Dichloro-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	212	184.63	184.17
103	4-(3,5-Dichloro-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	191	159.88	161.50
104	6-Methyl-2-oxo-4-(2,3,4-trichloro-phenylamino)-cyclohex-3-enecarboxylic acid methyl ester	205	196.68	191.23
105	6-Methyl-2-oxo-4-(2,3,4-trichloro-phenylamino)-cyclohex-3-enecarboxylic acid ethyl ester	165	176.40	165.01

Table 1. Continued

106	6-Methyl-2-oxo-4-(2,3,5-trichloro-phenylamino)-cyclohex-3-enecarboxylic acid methyl ester	200	195.55	189.58
107	6-Methyl-2-oxo-4-(3,4,5-trichloro-phenylamino)-cyclohex-3-enecarboxylic acid methyl ester	181	210.26	186.35
108	6-Methyl-2-oxo-4-(N'-phenyl-hydrazino)-cyclohex-3-enecarboxylic acid ethyl ester	148	149.79	148.01
109	4-(5-Chloro-pyridin-2-ylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	174	148.31	152.52
110[d]	4-Cyclohexylamino-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	141	129.52	134.97
111	4-(4-Chloro-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	161	155.21	153.43
112	4-(4-Bromo-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	151	157.30	151.45
113	4-(4-Iodo-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	160.5	160.49	160.49
114	4-(4-Fluoro-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	150	155.00	155.48
115	6-Methyl-2-oxo-4-(4-trifluoromethyl-phenylamino)-cyclohex-3-enecarboxylic acid ethyl ester	184.5	179.20	180.75
116	4-(4-Cyano-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	190	177.16	183.14
117	6-Methyl-4-(4-nitro-phenylamino)-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	173	179.27	173.01
118[d]	6-Methyl-2-oxo-4-p-tolylamino-cyclohex-3-enecarboxylic acid ethyl ester	134.5	145.58	135.17
119	6-Methyl-2-oxo-4-phenylamino-cyclohex-3-enecarboxylic acid ethyl ester	155	129.11	135.37
120	6-Methyl-4-(4-nitro-phenylamino)-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	186	189.71	186.01
121	4-(3-Chloro-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	137	144.42	137.01
122	4-(3-Bromo-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	135	148.02	138.02
123	4-(3-Fluoro-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	138	151.07	148.19
124[d]	6-Methyl-2-oxo-4-(3-trifluoromethyl-phenylamino)-cyclohex-3-enecarboxylic acid ethyl ester	164	171.20	164.01
125	4-(3-Cyano-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	165	173.51	165.01
126[d]	6-Methyl-4-(3-nitro-phenylamino)-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	155	177.36	156.96
127	6-Methyl-2-oxo-4-m-tolylamino-cyclohex-3-enecarboxylic acid ethyl ester	125	141.64	136.24
128	4-(3-Methoxy-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	110	130.43	128.79
129[d]	4-(2-Methoxy-5-methyl-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	160.5	139.66	145.19
130	3-(4-Chloro-phenylamino)-5-methyl-cyclohex-2-enone	198	192.06	197.99
131	3-(4-Bromo-phenylamino)-5-methyl-cyclohex-2-enone	213	196.98	199.71
132	3-(4-Fluoro-phenylamino)-5-methyl-cyclohex-2-enone	182	187.23	189.25
133	5-Methyl-3-(4-trifluoromethyl-phenylamino)-cyclohex-2-enone	205	202.69	197.20
134	5-Methyl-3-(4-trifluoromethoxy-phenylamino)-cyclohex-2-enone	174	186.64	174.01
135[d]	4-(5-Methyl-3-oxo-cyclohex-1-enylamino)-benzonitrile	234	213.27	223.01
136	5-Methyl-3-(4-nitro-phenylamino)-cyclohex-2-enone	208	205.60	207.99
137	5-Methyl-3-p-tolylamino-cyclohex-2-enone	194	167.18	173.82
138	5-Methyl-3-phenylamino-cyclohex-2-enone	158	160.42	158.01
139	3-(3-Chloro-phenylamino)-5-methyl-cyclohex-2-enone	177	179.31	176.99
140[d]	3-(3-Bromo-phenylamino)-5-methyl-cyclohex-2-enone	172	185.13	178.05
141	3-(3-Fluoro-phenylamino)-5-methyl-cyclohex-2-enone	163	173.72	170.67
142	5-Methyl-3-(3-trifluoromethyl-phenylamino)-cyclohex-2-enone	179	184.72	179.01
143	5-Methyl-3-(3-trifluoromethoxy-phenylamino)-cyclohex-2-enone	156	157.43	157.26
144	3-(5-Methyl-3-oxo-cyclohex-1-enylamino)-benzonitrile	175.5	201.91	193.29
145	4-(4-Methoxycarbonyl-5-methyl-3-oxo-cyclohex-1-enylamino)-benzoic acid	226	213.44	219.76
146[d]	6-Methyl-4-morpholin-4-yl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	125	138.18	127.31
147	4-(4-Hydroxy-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	199.5	185.10	199.49
148	4-(4-Chloro-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	178	182.77	183.23
149	4-Benzylamino-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	154	147.48	154.01

Table 1. Continued

150	4-(4-Ethyl-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	153.5	156.17	153.51
151	6-Methyl-2-oxo-4-phenethylamino-cyclohex-3-enecarboxylic acid methyl ester	116	141.25	145.82
152	6-Methyl-2-oxo-4-(3-phenyl-propylamino)-cyclohex-3-enecarboxylic acid methyl ester	163	140.73	157.12
153	4-(4-Fluoro-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	161	179.91	182.77
154	4-(4-Bromo-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	188	186.96	187.99
155	6-Methyl-2-oxo-4-(4-trifluoromethyl-phenylamino)-cyclohex-3-enecarboxylic acid methyl ester	169	195.81	170.29
156[d]	6,6-Dimethyl-2-oxo-4-phenylamino-cyclohex-3-enecarboxylic acid methyl ester	168	150.10	159.67
157	4-(4-Chloro-phenylamino)-6,6-dimethyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	141	177.80	177.40
158	4-(4-Chloro-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	161	159.78	161.01
159	4-(4-Bromo-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	151	164.25	162.05
160	6-Methyl-2-oxo-4-pyrrolidin-1-yl-cyclohex-3-enecarboxylic acid methyl ester	138	126.70	132.77
161	6-Methyl-2-oxo-4-p-tolylamino-cyclohex-3-enecarboxylic acid ethyl ester	134.5	146.64	142.70
162	4-(4-Ethyl-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	148.5	138.94	143.71
163	6-Methyl-4-(4-nitro-phenylamino)-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	173	177.69	173.01
164	4-(2,5-Dimethoxy-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	137	137.71	136.99
165	4-(4-tert-Butyl-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	165.5	150.22	148.32
166	6-Methyl-2-oxo-4-(4-trifluoromethyl-phenylamino)-cyclohex-3-enecarboxylic acid ethyl ester	184.5	181.30	167.81
167	4-(2-Benzoyl-4-chloro-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	154	170.85	154.01
168	4-(4-Fluoro-benzylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid 4-fluoro-benzylamide	183	200.32	183.01
169	3-Benzylamino-5,5-dimethyl-cyclohex-2-enone	124	157.44	137.94
170	5,5-Dimethyl-3-phenylamino-cyclohex-2-enone	181	182.59	180.99
171[d]	3-(4-Chloro-phenylamino)-5,5-dimethyl-cyclohex-2-enone	208	215.29	207.99
172	5,5-Dimethyl-3-(4-nitro-phenylamino)-cyclohex-2-enone	242	224.61	209.84
173	5,5-Dimethyl-3-p-tolylamino-cyclohex-2-enone	203	192.86	194.06
174[d]	3-(4-Ethyl-phenylamino)-5,5-dimethyl-cyclohex-2-enone	201	171.81	183.81
175	3-(4-Methoxy-phenylamino)-5,5-dimethyl-cyclohex-2-enone	189	182.41	188.99
176	3-(4-Amino-phenylamino)-5,5-dimethyl-cyclohex-2-enone	211.5	225.09	211.49
177	3-(4-tert-Butyl-phenylamino)-5,5-dimethyl-cyclohex-2-enone	206	180.59	180.13
178	3-(4-Chloro-phenylamino)-5-methyl-cyclohex-2-enone	198	189.42	190.17
179[d]	3-(4-Chloro-phenylamino)-cyclohex-2-enone	190	183.11	189.99
180[d]	3-Benzylamino-cyclohex-2-enone	125	141.86	129.34
181[d]	3-(4-tert-Butyl-phenylamino)-cyclohex-2-enone	185	170.44	174.86
182	3-Benzylamino-cyclopent-2-enone	139	135.30	138.99
183	3-Phenethylamino-cyclohex-2-enone	136	134.22	136.01
184	4-Benzylamino-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	154	138.44	146.56
185[d]	6-Methyl-2-oxo-4-phenethylamino-cyclohex-3-enecarboxylic acid phenethyl-amide	171	158.91	153.24
186	4-(4-Chloro-benzylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	173	165.12	172.99
187	4-(4-Fluoro-benzylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	174	165.37	174.01
188	6-Methyl-2-oxo-4-(N'-phenyl-hydrazino)-cyclohex-3-enecarboxylic acid methyl ester	167	159.84	166.99
189	6-Methyl-2-oxo-4-p-tolylamino-cyclohex-3-enecarboxylic acid methyl ester	144	164.45	159.43
190	6-Methyl-4-(4-nitro-phenylamino)-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	186	191.20	185.99
191	4-(2-Methoxy-5-methyl-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	160.5	148.91	160.49
192	4-(2,5-Dimethoxy-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	134	145.40	146.06
193	4-Benzylamino-6,6-dimethyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	138	144.06	146.38

Table 1. Continued

194[d]	4-(4-Bromo-phenylamino)-pent-3-en-2-one	125	124.52	125.01
195	4-(4-Nitro-phenylamino)-pent-3-en-2-one	144	132.71	143.30
196	4-(4-Methoxy-phenylamino)-pent-3-en-2-one	118	112.32	118.01
197[d]	4-Phenethylamino-pent-3-en-2-one	138	120.38	136.04
198	4-(4-Ethyl-phenylamino)-pent-3-en-2-one	109	107.23	110.52
199	4-p-Tolylamino-pent-3-en-2-one	100	122.70	117.83
200	4-Benzylamino-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	134	138.38	144.10
201[d]	6-Methyl-4-(4-methyl-benzylamino)-2-oxo-cyclohex-3-enecarboxylic acid tert-butyl ester	123	141.03	145.31
202	4-(4-Methoxy-benzylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	168.5	150.87	152.05
203	4-(4-Methoxy-benzylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	154	137.64	141.97
204	4-(4-Cyano-benzylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid tert-butyl ester	172	167.32	173.52
205	3-(4-Methoxy-benzylamino)-cyclohex-2-enone	159	133.33	137.87
206[d]	3-(4-Methoxy-benzylamino)-5-methyl-cyclohex-2-enone	160	145.16	149.68
207	4-Benzoylamino-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	161	159.92	160.20
208	4-Benzoylamino-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid tert-butyl ester	193	163.12	190.77
209[d]	4-(4-Chloro-benzoylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid tert-butyl ester	166	172.27	171.28
210	N-(5,5-Dimethyl-3-oxo-cyclohex-1-enyl)-4-methoxy-benzamide	153	171.88	178.53
211	6-Methyl-4-(4-nitro-benzylamino)-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	174	181.84	175.99
212	4-(4-Iodo-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid tert-butyl ester	186	165.17	158.12
213	4-(4-Carbamoyl-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	228	204.99	220.83
214	6-Methyl-2-oxo-4-(4-sulfamoyl-phenylamino)-cyclohex-3-enecarboxylic acid methyl ester	210	223.94	196.11
215	4-(3-Chloro-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	137	167.44	138.99
216	6-Methyl-2-oxo-4-(3-trifluoromethoxy-phenylamino)-cyclohex-3-enecarboxylic acid methyl ester	173	155.84	159.79
217	4-(3-Carbamoyl-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	200	203.81	189.13
218	6-Methyl-2-oxo-4-m-tolylamino-cyclohex-3-enecarboxylic acid methyl ester	165	141.52	142.81
219	6-Methyl-2-oxo-4-m-tolylamino-cyclohex-3-enecarboxylic acid ethyl ester	125	130.37	128.15
220	6-Methyl-2-oxo-4-(3-trifluoromethyl-phenylamino)-cyclohex-3-enecarboxylic acid methyl ester	167	169.33	170.62
221[d]	6-Methyl-2-oxo-4-(3-trifluoromethyl-phenylamino)-cyclohex-3-enecarboxylic acid ethyl ester	164	159.43	162.62
222	4-(3-Fluoro-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	138	143.16	142.59
223	4-(2-Chloro-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	157	148.91	149.14
224	6-Methyl-2-oxo-4-(2-sulfamoyl-phenylamino)-cyclohex-3-enecarboxylic acid methyl ester	197	181.90	177.74
225	4-(3-Chloro-4-methoxy-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	175	168.52	165.74
226[d]	4-(2,5-Dichloro-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	190	172.54	175.11
227	4-(3,4-Dichloro-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	160	190.22	187.45
228	4-(4-Chloro-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	178	166.04	168.44
229	6-Methyl-2-oxo-4-(2,3,5-trichloro-phenylamino)-cyclohex-3-enecarboxylic acid ethyl ester	183	175.75	164.17
230[d]	6-Methyl-2-oxo-4-p-tolylamino-cyclohex-3-enecarboxylic acid methyl ester	144	152.18	150.61
231[d]	6-Methyl-2-oxo-4-(4-trifluoromethoxy-phenylamino)-cyclohex-3-enecarboxylic acid ethyl ester	162	165.97	166.60
232	6-Methyl-2-oxo-4-(3-trifluoromethoxy-phenylamino)-cyclohex-3-enecarboxylic acid ethyl ester	151	139.39	146.81
233	4-(2,5-Dimethoxy-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	134	135.94	138.11
234	3-(3-Iodo-phenylamino)-5-methyl-cyclohex-2-enone	185	187.57	179.10
235	5-Methyl-3-(3-nitro-phenylamino)-cyclohex-2-enone	187	192.08	176.10
236	5-Methyl-3-m-tolylamino-cyclohex-2-enone	147	151.37	155.85
237	3-(3-Methoxy-phenylamino)-5-methyl-cyclohex-2-enone	140	143.61	150.42

Table 1. Continued

238	5,5-Dimethyl-3-phenylamino-cyclohex-2-enone	181	183.08	181.83
239	4-(5-Chloro-pyridin-2-ylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	206	193.74	205.17
240	4-(4-tert-Butyl-phenylamino)-6,6-dimethyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	170	168.67	161.96
241[d]	4-(4-Fluoro-phenylamino)-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	150	160.46	169.83
242	6-Methyl-2-oxo-4-phenylamino-cyclohex-3-enecarboxylic acid ethyl ester	155	131.33	157.96
243	4-Benzylamino-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid ethyl ester	134.5	170.85	154.01
244	5,5-Dimethyl-3-(4-trifluoromethyl-phenylamino)-cyclohex-2-enone	240.5	225.01	230.75
245[d]	3-(4-Trifluoromethyl-phenylamino)-cyclohex-2-enone	203	200.87	181.84
246	4-Benzylamino-6,6-dimethyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	138	153.19	156.97
247	6,6-Dimethyl-2-oxo-4-phenethylamino-cyclohex-3-enecarboxylic acid methyl ester	130	135.63	146.60
248	4-(4-Chloro-phenylamino)-pent-3-en-2-one	113	126.53	130.34
249[d]	4-Benzylamino-pent-3-en-2-one	107	113.16	119.85
250	4-Benzylamino-6-methyl-2-oxo-cyclohex-3-enecarboxylic acid methyl ester	154	147.90	154.49

[a] Experimental melting point of each drug-like compound
[b] Predicted melting point of each drug-like compound using GA-MLR approach
[c] Predicted melting point of each drug-like compound using GA-SVM approach
[d] Test set

Data pretreatment

The calculated descriptors were first analyzed for the existence of constant or near-constant variables in the preliminary step, and those detected were removed. In addition, to reduce redundancy in the descriptor data matrix, correlation of the descriptors with each other and with the melting points of the molecules was examined and the collinear descriptors (i.e. r> 0.9) were detected. Among the collinear descriptors, that with the highest correlation with melting point was retained while the others were removed from the data matrix. Then, the remaining descriptors were collected in an n×m data matrix (D), where n=250 and m=348 are the numbers of the compounds and the descriptors, respectively.

Genetic algorithm

The GA feature selection approach as a stochastic method is capable of solving a variety of optimization problems, which are defined through fitness criteria. The basis of this strategy goes back to the evolution hypothesis given by Darwin. Furthermore, genetic functions encompassing crossover and mutation are very important in this algorithm. As usual, simulation of the population evolution is among the most significant preliminary steps [25-34]. We have recently reported all the detailed information concerning the general performance of the GA approach [35].

Support vector machine (SVM)

Support vector machine is a novel type of machine learning method, and is gaining popularity due to many attractive features as well as promising empirical performance. The main advantage of SVM is that it adopts the structure risk minimization (SRM) principle, which has been shown to be superior to the traditional empirical risk minimization (ERM) principle, employed by conventional neural networks. SRM minimizes an upper bound of the generalization error on Vapnik-Chernoverkis (VC) dimension, as opposed to ERM that minimizes the training error [36, 37]. For the case of regression approximation, suppose there are a given set of data points are you (x_i is the input vector, d_i the desired value, and n is the total number of data patterns) drawn independently and identically from an unknown function, SVMs approximate the function with three distinct characteristics: (i) SVMs estimate the regression in a set of linear functions, (ii) SVMs define the regression estimation as the problem of risk minimization with respect to the ε-insensitive loss function, and (iii) SVMs minimize the risk based on the SRM principle whereby elements of the structure are defined by the inequality $\frac{1}{2}\|\omega\|^2 \leq$ constant. The linear function is formulated in the high dimensional feature space, with the form of function (eq. 1).

$$y = f(x) = w\phi(x) + b \qquad \text{(eq. 1)}$$

Where $\phi(x)$ is the high dimensional feature space, which is non-linearly mapped from the input space *x*. The aforementioned characteristics (ii and iii) are reflected in the minimization of the regularized risk function (eq. 2) of SVMs, by which the coefficients *w* and *b* are estimated. The goal of this risk function is to find a function that has at most ε deviation from the actual values in all the training data points and at the same time is as flat as possible.

$$R_{SVMs}(C) = C\frac{1}{n}\sum_{i=1}^{n} L_{\varepsilon}(d_i, y_i) + \frac{1}{2}\|\omega\|^2 \qquad \text{(eq. 2)}$$

$$L_{\varepsilon}(d,y) = \begin{Bmatrix} |d-y| - \varepsilon, & |d-y| \geq \varepsilon, \\ 0 & otherwise \end{Bmatrix} \qquad \text{(eq. 3)}$$

The first term $C\frac{1}{n}\sum_{i=1}^{n} L_{\varepsilon}\left(d_i, y_i\right)$ is the so-called empirical error (risk), which is measured by the ε-insensitive loss function (eq. 3). This loss function provides the advantage of using sparse data points to represent the designed function (1). The second term $\frac{1}{2}\|\omega\|^2$, on the other hand, is called the regularized term. ε implies for called the tube size of SVMs, and C is the regularization constant determining the trade-off between the empirical error and the regularized term. Introduction of the positive slack variables ξ, ξ^* leads to eq. (4) with the following constrained function:

Minimize

$$R_{SVMs}\left(\omega, \xi^*\right) = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}\left(\xi_i + \xi_i^*\right) \text{(eq. 4)}$$

Where *i* represents the data sequence, with *i*=1 being the most recent observation and *i*=1 being the earliest observation. Finally, by introducing Lagrange multipliers and exploiting the optimality constraints, decision function (eq. 5) takes the following form:

$$f\left(x, a_i^*\right) = \sum_{i=1}^{n}\left(a_i - a_i^*\right)K\left(x, x_i\right) + b \qquad \text{(eq. 5)}$$

Where a_i, a_i^* are the introduced Lagrange multipliers? So far, by exploiting the Karush–Kuhn–Tucker (KKT) conditions, only a number of coefficients among a_i and a_i^* will be non-zero and the data points associated with them could be referred to support vectors. In this equation, *K* refers to kernel function, including linear, polynomial, splines, and radial basis function. In support vector regression, the Gaussian radial basis function (RBF) (eq. 6) is commonly used, which has the following form:

$$k\left(\bar{x}_i, \bar{x}_j\right) = \exp(-\gamma\left\|\bar{x}_i - \bar{x}_j\right\|^2) \qquad \text{(eq. 6)}$$

RESULTS AND DISCUSSION

Regression analysis

Principal components analysis (PCA) was performed with the calculated structure descriptors for the whole data set to detect the homogeneities in the data set and to show spatial location of samples to assist separation of the data into training and test sets. The PCA results show that two principal components (PC1 and PC2) describe 52.04% of the overall variables, as follows: PC1 = 34.61%, PC2 = 17.43%. Because almost all variables can be accounted for by the first two PCs, their score plot is a reliable representation of the spatial distribution of the points for the data set. The plot of PC1 against PC2 (Figure 1) displays the distribution of compounds over the first two principal components space.

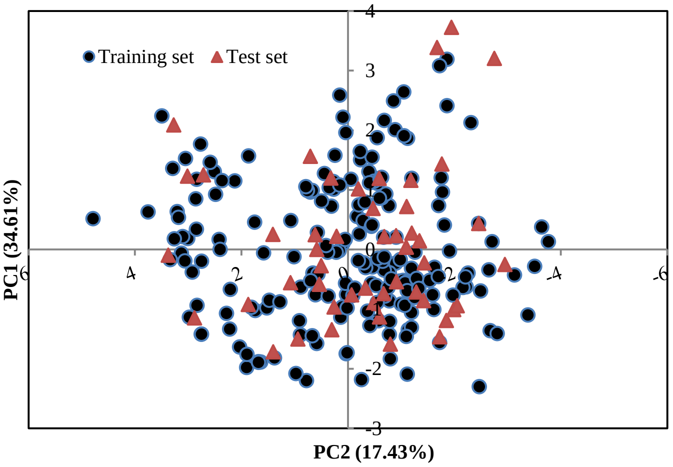

Figure 1. Principal components analysis of the training and test sets

According the results of PCA, all the data were divided into a training set of 200 compounds to develop the models and a test set of 50 compounds to evaluate the models based on two rules:

1. The range of the activity values of both the training set and the test set should be covered from the lowest to the highest;

2. The points corresponding to the training set in the PCA plot should not be out of the main clusters.

The two training and tests sets are listed in Table 1. For the selection of the most important descriptors, genetic algorithm variable subset selection method was used combined with MLR approach.

The GA-MLR analysis led to the derivation of one model possessing eight descriptors. The descriptors obtained are shown in Table 2. Since co-linearity between the variables degrades the performance of the MLR-based QSAR models, before a multi-parametric analysis was undertaken, the correlation between each pair of the variables used in this study was examined. The correlation matrix itself exhibits how the used descriptors were mutually correlated (Table 3). From Table 3, it could be seen that the correlation coefficient value of each pair of molecular descriptors was at the most 0.679, confirming that the selected descriptors behave independently.

In addition, in order to check the inter-correlation of descriptors, variation inflation factor (VIF) analysis was performed. The VIF value is calculated from 1/1- r^2, where r^2 is the multiple correlation coefficient of one descriptor's effect regressed on the remaining molecular descriptors. If VIF equals to 1.0, no inter-correlation exists for each variable; if VIF falls into the range of 1.0- 5.0, the related model is acceptable; and if VIF is larger than 10.0, the related model is unstable and an exhaustive re-check is necessary [38]. The VIF values of the selected descriptors are shown in Table 2. As can be seen from this table, the majority of the variables have VIF values less than 5, indicating that the obtained model has obvious statistical significance.

To examine the relative importance as well as the contribution of each descriptor in the model, the value of the mean effect (*MFj*) was calculated for each descriptor. This calculation was performed with the equation below:

$$MF_j = \frac{\beta_j \sum_{i=1}^{i=n} d_{ij}}{\sum_j^m \beta_j \sum_i^n d_{ij}} \qquad \text{(eq. 7)}$$

Table 2: The list of the selected descriptors by the GA-MLR technique

Descriptor	Chemical meaning	MF [a]	VIF [b]
Constant	Intercept	-	-
Mv	Mean atomic van der Waals volume (scaled on Carbon atom)	-0.165	1.301
E1s	1st component accessibility directional WHIM index/weighted by atomic electrotopological states.	-0.011	1.168
HGM	Geometric mean on the leverage magnitude	0.019	3.401
HATS4u	Leverage-weighted autocorrelation of lag 4/unweighted	-0.040	2.528
RTe+	R maximal index/weighted by atomic Sanderson electronegativities	-0.016	1.605
WPSA-3	WPSA-3 Weighted PPSA (PPSA3*TMSA/1000) [Zefirovs PC]	-0.020	5.078
HDCA-2	HA dependent HDCA-2 [Quantum-Chemical PC]	-0.014	1.822
MREC	Max resonance energy for a C-H bond	1.247	1.063

Table 3: Correlation matrix for the eight selected descriptors using bivariate correlation approach

	Mv	**E1s**	**HGM**	**HATS4u**	**RTe+**	**WPSA-3**	**HDCA-2**	**MREC**
Mv	1							
E1s	0.182	1						
HGM	0.355	0.039	1					
HATS4u	0.115	0.011	0.679	1				
RTe+	0.146	0.274	0.299	0.261	1			
WPSA-3	-0.245	-0.027	-0.614	-0.625	-0.316	1		
HDCA-2	0.146	-0.020	-0.081	0.015	0.289	0.312	1	
MREC	0.053	-0.048	0.129	-0.005	0.052	-0.069	0.122	1

MF_j represents the mean effect for the considered descriptor *j*, β_j is the coefficient of the descriptor *j*, d_{ij} stands for the value of the target descriptors for each molecule and, eventually, *m* is the descriptors number in the model [39]. The MF_j value indicates the relative importance of a descriptor, compared with the other descriptors in the model. The mean effect values for selected descriptors as well as their chemical meaning are shown in Table 2. As can be seen the MREC descriptor has a highest mean effect value, and subsequently it exerts the most impact on the constructed model.

The selected variables are Mv, E1s, HGM, HATS4u, RTe, WPSA-3, HDCA-2 and MREC. With the selected eight molecular descriptors, we have built a reliable linear model using the training set data that it is described by the following equation:

$$\mathbf{M_p} = 2063.77\ (\pm 552.08) + 502.45\ (\pm 62.82)\ \mathbf{Mv} - 42.86\ (\pm 7.68)\ \mathbf{E1s} - 5.70\ (\pm 2.27)\ \mathbf{HGM} + 147.75\ (\pm 14.32)\ \mathbf{HATS4u} + 242.61\ (\pm 53.28)\ \mathbf{RTe} + 7.55\ (\pm 2.20)\ \mathbf{WPSA\text{-}3} + 17.25\ (\pm 3.29)\ \mathbf{HDCA\text{-}2} - 213.38\ (\pm 49.39)\ \mathbf{MREC} \qquad (8)$$

The built model was used to predict the test set data. The prediction results are given in Table 1 and shown in Figure 2. The square correlation coefficient R^2 was obtained to be 0.712 for the training set and 0.713 for the test set with root mean square error (RMSE) of 15.042 and 14.919, respectively.

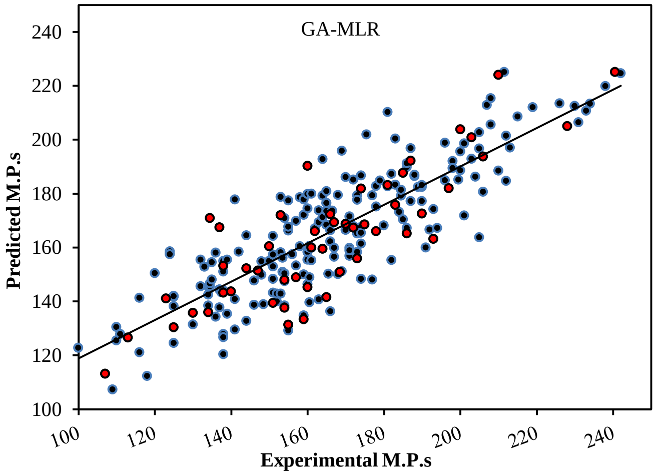

Figure 2. The predicted M.P. values by the GA-MLR modeling vs. the experimental M.P.s

Figure 3 shows the residual plot when using the GA-MLR approach. As shown in this Figure, the normal scattering of the points on two sides of the X-axis confirms that there is no systematic error in this modeling strategy.

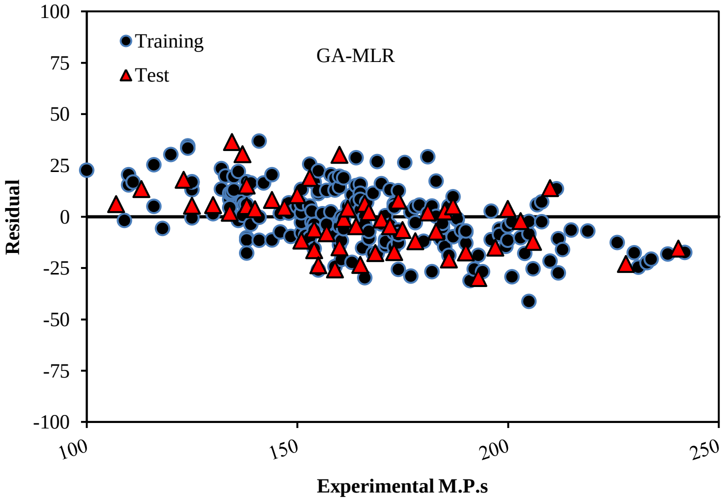

Figure 3. Plot of the residuals against the experimental values of the melting points by GA-MLR

SVM model development

From the obtained results by the multiple linear regressions (MLR), it can be seen that the linear model was not sufficiently accurate. Therefore, a non-linear model was built by SVM-based genetic algorithm approach (GA-SVM) on the same subset of descriptors. LOO cross-validation method implied in SVM was used to build the model by the training set compounds. Performance of SVM for regression depends on the combination of several factors. They are kernel function type, capacity parameter C, ε of ε-insensitive loss function and its corresponding parameters.

Firstly, the kernel function should be decided, which determines the sample distribution in the mapping space. The radial basis function (RBF) is commonly used in many studies because of its good general performance and few parameters to be adjusted. The corresponding parameters, i.e. γ of the kernel function greatly affect the number of support vectors, which has a close relation with the performance of the SVM and training time. Too many support vectors could produce overfitting which increase the time of the training step. In addition, γ controls the amplitude of the RBF function and, therefore, controls the generalization ability of SVM. The plot of RMSE versus γ on the LOO cross-validation is shown in Fig. 4. As can be seen from the figure, the optimal γ was 0.7.

Parameter ε-insensitive prevents the entire training set meeting boundary conditions and so allows for the possibility of sparsity in the dual formulation's solution. The optimal value for ε depends on the type of noise present in the data, which is usually unknown. The RMSE of LOO cross-validation on different epsilon is recorded in Fig. 5 and the optimal value was found to be 0.01. The last parameter C was a regularization parameter that controlled the tradeoff between maximizing the margin and minimizing the training error. The plot of RMSE versus C value is shown in Fig.6 with values γ = 0.7 and ε = 0.01. Accordingly, the optimal value of C was 11.

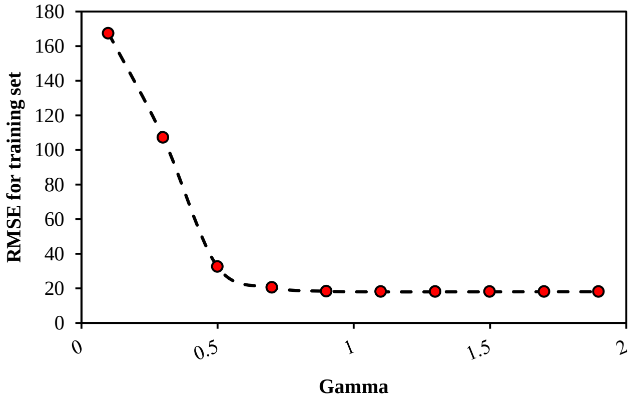

Figure 4: The trends of RMSE vs. the term gamma for the training set

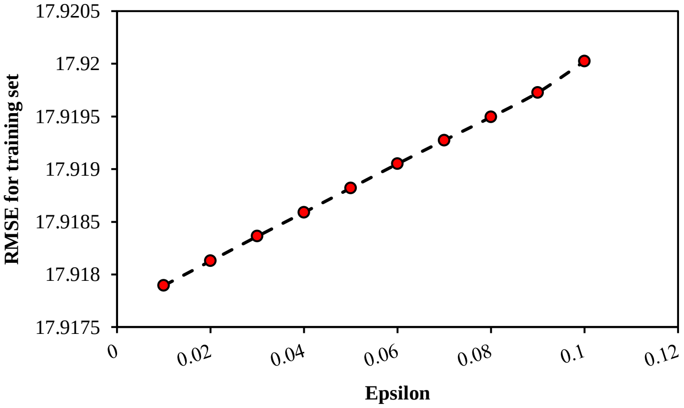

Figure 5. The plot of RMSE as a function of epsilon for the training set

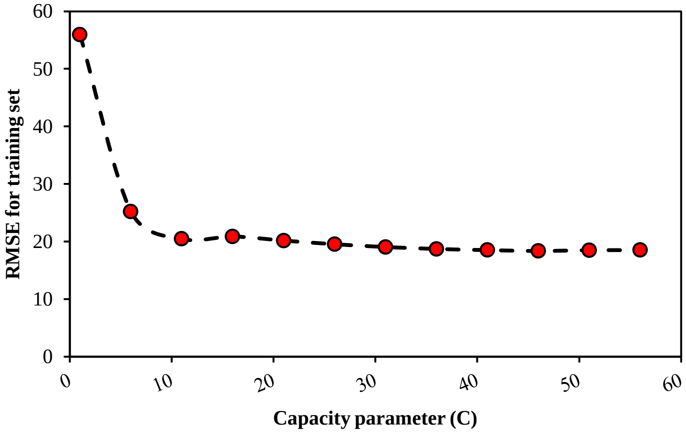

Figure 6. Variation of RMSE vs a capacity parameter C. for the training set

Therefore, the best choices forγ, ε and C were 0.7, 0.01 and 11. For the optimal model, the cross-validated coefficients were 0.577 and 0.578 for Q^2_{LOO} and Q^2_{LGO}, respectively. It gave RMSE of 11.082 for the training set, 13.332 for the test set, and the corresponding correlation coefficients (R^2) were 0.853 and 0.737, respectively.The calculated M.P. values obtained from SVM predictive model are listed in Table 1. Figure 7 shows the predicted versus experimental values of M.P. for the training and test sets using the GA-SVM method.

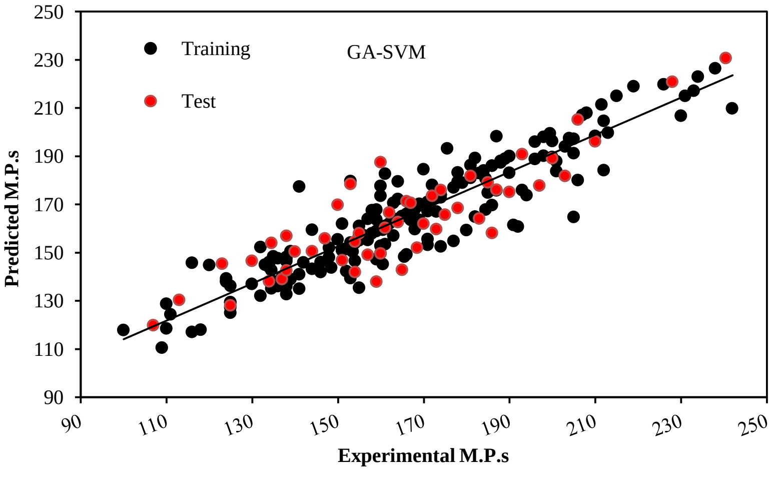

Figure7. The predicted M.P. values by the GA-SVM modeling vs. the experimental M.P. values

Finally, we have plotted the trends of variation in the residuals as a function of experimental melting points (Figure 8). Similarly, to the GA-MLR approach, in this model, it was not observed any systematic error within the modeling process.

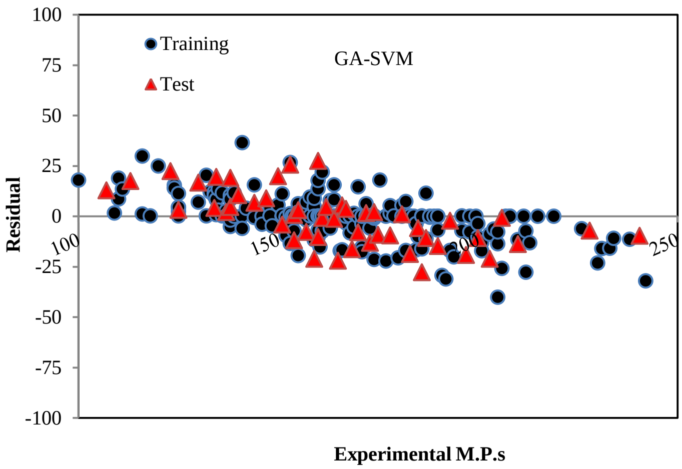

Figure 8. Plot of the residuals against the experimental values of the melting points by GA-SVM

Comparison of the MLR and SVM results

Table 4 presents the statistical parameters of the results obtained from the two studied models for the same set of compounds. The RMSE of SVM model for the training and test data sets were lower than those of models proposed by the GA-MLR method. Moreover, the correlation coefficient (R^2) given by the GA-SVM was higher than that of GA-MLR method. In addition, the results of F-test were obtained as shown in Table 4. From the Table, it can be seen that SVM model gives higher F values, so this model gives the most satisfactory results, compared with the results obtained from MLR method. Consequently, this SVM approach currently constitutes the most accurate method for prediction of the melting points of a variety of organic and/or drug-like compounds.

Table 4. Statistical parameters of the results obtained from the GA-MLR and GA-SVM models

The technique used	Training set			Test set		
	R^2	RMSE	F	R^2	RMSE	F
GA-MLR	0.712	15.042	59.061	0.713	14.919	13.770
GA-SVM	0.853	11.082	106.691	0.737	13.332	14.031

CONCLUSIONS

In this study, a new QSPR model was developed for predicting the melting point compounds, from a wide pool of the molecular structure. The GA approach was used to select the main relevant descriptors and to build a linear model, namely the GA-MLR method. The GA-SVM method was used to construct a non-linear QSPR model based on the same selected parameters. Regarding the statistical parameters of the constructed models, we can conclude that the GA-SVM model produced more satisfactory results than the MLR model confirming its good predictive ability. It was easy to notice that there was a good prospect for the SVM application in the QSPR modeling. This model could accurately predict the M.P. of those components that did not exist in the modeling procedure**.**

ACKNOWLEDGEMENTS

The authors express their special thanks to the Office of Research Affairs of the Islamic Azad University, Shahrood Branch for financial support and the time dedicated to this work. The authors declare that there is no conflict of interests.

REFERENCES

1. Abramowitz R., Yalkowsky S. H., 1990. Melting-point, boiling-point, and symmetry. Pharm Res. 7 (9), 942-947.

2. Katritzky A. R., Jain R., Lomaka A., Petrukhin R., Maran U., Karelson M., 2001. Perspective on the relationship between melting points and chemical structure. Cryst Growth Des. 1 (4), 261-265.

3. Karthikeyan M., Glen R. C., Bender A., 2005. General melting point prediction based on a diverse compound data set and artificial neural networks. J Chem Inf Model. 45 (3), 581-590.

4. Matheson L. E., Chen Y. S., 1995. A quantitative structure-transportability relationship for the release of a series of substituted benzenes and pyridines from a planar polydimethylsiloxane matrix. Int J Pharm. 125 (2), 297-307.

5. Habibi-Yangjeh A., Pourbasheer E., Danandeh-Jenagharad M., 2008. Prediction of melting point for drug-like compounds using principal component-genetic algorithm-artificial neural network. Bull Korean Chem Soc. 29 (4), 833-841.

6. Todeschini R., Consonni V. 2000. Handbook of Molecular Descriptors. Wiley-VCH. Weinheim, Germany.

7. Atabati M., Khandani F., 2012. Ant colony optimization as a descriptor selection in QSPR modeling for prediction of lambda(max) of azo dyes. Chin Chem Lett. 23 (10), 1209-1212.

8. Dai Y.-m., Zhu Z.-p., Cao Z., Zhang Y.-f., Zeng J.-l., Li X., 2013. Prediction of boiling points of organic compounds by QSPR tools. J Mol Graphics Model. 44, 113-119.

9. Gharagheizi F., Sattari M., Ilani-Kashkouli P., Mohammadi A. H., Ramjugernath D., Richon D., 2013. A "non-linear" quantitative structure-property relationship for the prediction of electrical conductivity of ionic liquids. Chem Eng Sci. 101, 478-485.

10. Goudarzi N., Goodarzi M., Mohammadhosseini M. M., Nekooei M., 2009. QSPR models for prediction of half-wave potentials of some chlorinated organic compounds using SR-PLS and GA-PLS methods. Mol Phys. 107 (17), 1739-1744.

11. Liang G., Xu J., Liu L., 2013. QSPR analysis for melting point of fatty acids using genetic algorithm based multiple linear regression (GA-MLR). Fluid Phase Equilibr. 353, 15-21.

12. Sosnowska A., Barycki M., Jagiello K., Haranczyk M., Gajewicz A., Kawai T., Suzuki N., Puzyn T., 2014. Predicting enthalpy of vaporization for persistent organic pollutants with quantitative structure-property relationship (QSPR) incorporating the influence of temperature on volatility. Atmos Environ. 87, 10-18.

13. Toubaei A., Golmohammadi H., Dashtbozorgi Z., Acree W. E., Jr., 2012. QSPR studies for predicting gas to acetone and gas to acetonitrile solvation enthalpies using support vector machine. J Mol Liq. 175, 24-32.

14. Golzar K., Amjad-Iranagh S., Modarress H., 2013. QSPR prediction of the solubility of CO2 and N-2 in common polymers. Measurement. 46 (10), 4206-4225.

15. Maity U., Basu J. K., Sengupta S., 2013. A neural network prediction of conversion of benzothiophene oxidation catalyzed by nano-Ti-beta catalyst. Fuel. 113, 180-186.

16. Qiu P., Ni Y.-N., Kokot S., 2013. Application of artificial neural networks to the determination of pesticides by linear sweep stripping voltammetry. Chin Chem Lett. 24 (3), 246-248.

17. Zheng F., Zhan M., Huang X., Hameed M. D. M. A., Zhan C.-G., 2014. Modeling in vitro inhibition of butyrylcholinesterase using molecular docking, multi-linear regression and artificial neural network approaches. Biorg Med Chem. 22 (1), 538-549.

18. Cortes C., Vapnik V., 1995. Support-Vector Networks. Mach Learn. 20, 273-297.

19. Golmohammadi H., Dashtbozorgi Z., Acree W. E., Jr., 2012. Quantitative structure-activity relationship prediction of blood-to-brain partitioning behavior using support vector machine. Eur J Pharm Sci. 47 (2), 421-429.

20. Hao M., Li Y., Wang Y., Zhang S., 2011. Prediction of P2Y(12) antagonists using a novel genetic algorithm-support vector machine coupled approach. Anal Chim Acta. 690 (1), 53-63.

21. Xuan S., Wu Y., Chen X., Liu J., Yan A., 2013. Prediction of bioactivity of HIV-1 integrase ST inhibitors by multilinear regression analysis and support vector machine. Bioorg Med Chem Lett. 23 (6), 1648-1655.

22. Zhong M., Xuan S., Wang L., Hou X., Wang M., Yan A., Dai B., 2013. Prediction of bioactivity of ACAT2 inhibitors by multilinear regression analysis and support vector machine. Bioorg Med Chem Lett. 23 (13), 3788-3792.

23. Gao T., Sun S.-L., Shi L.-L., Li H., Li H.-Z., Su Z.-M., Lu Y.-H., 2009. An accurate density functional theory calculation for electronic excitation energies: The least-squares support vector machine. J Chem Phys. 130 (18), 184-194.

24. Eddington N. D., Cox D. S., Khurana M., Salama N. N., Stables J. P., Harrison S. J., Negussie A., Taylor R. S., Tran U. Q., Moore J. A., Barrow J. C., Scott K. R., 2003. Synthesis and anticonvulsant activity of enaminones Part 7. Synthesis and anticonvulsant evaluation of ethyl 4- (substituted phenyl)amino -6-methyl-2-oxocyclohex-3-ene-1-carboxylates and their corresponding 5-methylcyclohex-2-enone derivatives. Eur J Med Chem. 38 (1), 49-64.

25. Adimi M., Salimi M., Nekoei M., Pourbasheer E., Beheshti A. S., 2012. A quantitative structure-activity relationship study on histamine receptor antagonists using the genetic algorithm-multi-parameter linear regression method. J Serb Chem Soc. 77 (5), 639-650.

26. Dolatabadi M., Nekoei M., Banaei A., 2010. Prediction of antibacterial activity of pleuromutilin derivatives by genetic algorithm-multiple linear regression (GA-MLR). Monatsh Chem. 141 (5), 577-588.

27. Mohammadhosseini M., Nekoei M., 2013. Quantitative structure-electrochemistry relationship study for prediction of half-wave reduction potentials of some chlorinated organic compounds by genetic algorithm-multiple linear regression. Asian J Chem. 25 (1), 349-352.

28. Nekoei M., Salimi M., Dolatabadi M., Mohammadhosseini M., 2011. Prediction of antileukemia activity of berbamine derivatives by genetic algorithm-multiple linear regression. Monatsh Chem. 142 (9), 943-948.

29. Noorizadeh H., Ardakani S. S., Ahmadi T., Mortazavi S. S., Noorizadeh M., 2013. Application of genetic algorithm-kernel partial least square as a novel non-linear feature selection method: partitioning of drug molecules. Drug Test Anal. 5 (2), 89-95.

30. Noorizadeh H., Farmany A., Narimani H., Noorizadeh M., 2013. QSRR using evolved artificial neural network for 52 common pharmaceuticals and drugs of abuse in hair from UPLCTOF-MS. Drug Test. Anal. 5 (5), 320-324.

31. Noorizadeh H., Farmany A., Noorizadeh M., Kohzadi M., 2013. Prediction of polar surface area of drug molecules: A QSPR approach. Drug Test Anal. 5 (4), 222-227.

32. Pourbasheer E., Riahi S., Ganjali M. R., Norouzi P., 2010. Quantitative structure-retention relationship

(QSRR) models for predicting the GC retention times of essential oil components. Acta Chromatogr. 22 (3), 357-373.

33. Riahi S., Ganjali M. R., Pourbasheer E., Norouzi P., 2008. QSRR study of GC retention indices of essential-oil compounds by multiple linear regression with a genetic algorithm. Chromatographia. 67 (11-12), 917-922.

34. Pourbasheer E., Aalizadeh R., Ganjali M. R., Norouzi P., 2014. QSAR study of IKK beta inhibitors by the genetic algorithm: multiple linear regressions. Med Chem Res. 23 (1), 57-66.

35. Mohammadhosseini M., Deeb O., Alavi- Gharabagh A., Nekoei M., 2012. Exploring novel QSRRs for simulation of gas chromatographic retention indices of diverse sets of terpenoids in *Pistacia lentiscus* L. essential oil using stepwise and genetic algorithm multiple linear regressions. Anal Chem Lett. 2, 80-102.

36. Vapnik N. V. 1998. Statistical Learning Theory. John Wiley & Sons. New York.

37. Vapnik V. N. 1995. The Nature of Statistical Learning Theory. Springer-Verlag

38. Agrawal V. K., Khadikar P. V., 2001. QSAR prediction of toxicity of nitrobenzenes. Biorg Med Chem. 9 (11), 3035-3040.

39. Pourbasheer E., Riahi S., Ganjali M. R., Norouzi P., 2010. Quantitative structure-activity relationship (QSAR) study of interleukin-1 receptor associated kinase 4 (IRAK-4) inhibitor activity by the genetic algorithm and multiple linear regression (GA-MLR) method. J Enzym Inhib Med Chem. 25 (6), 844-853.