Prediction of Melting Point for Drug-like Compounds Using Principal Component-Genetic Algorithm-Artificial Neural Network

Aziz Habibi-Yangjeh,* Eslam Pourbasheer, and Mohammad Danandeh-Jenagharad

Department of Chemistry, Faculty of Science, University of Mohaghegh Ardabili, P. O. Box 179, Ardabil, Iran
**E-mail: ahabibi@uma.ac.ir*
Received December 9, 2007

Principal component-genetic algorithm-multiparameter linear regression (PC-GA-MLR) and principal component-genetic algorithm-artificial neural network (PC-GA-ANN) models were applied for prediction of melting point for 323 drug-like compounds. A large number of theoretical descriptors were calculated for each compound. The first 234 principal components (PC's) were found to explain more than 99.9% of variances in the original data matrix. From the pool of these PC's, the genetic algorithm was employed for selection of the best set of extracted PC's for PC-MLR and PC-ANN models. The models were generated using fifteen PC's as variables. For evaluation of the predictive power of the models, melting points of 64 compounds in the prediction set were calculated. Root-mean square errors (RMSE) for PC-GA-MLR and PC-GA-ANN models are 48.18 and 12.77 ºC, respectively. Comparison of the results obtained by the models reveals superiority of the PC-GA-ANN relative to the PC-GA-MLR and the recently proposed models (RMSE = 40.7 ºC). The improvements are due to the fact that the melting point of the compounds demonstrates non-linear correlations with the principal components.

Key Words : Quantitative structure-property relationship, Melting point, Drug-like compounds, Genetic algorithm, Artificial neural network

Introduction

Melting point is a fundamental physical property of organic compounds, which has found wide use in chemical identification, as a criterion of purity and for the calculation of other important physicochemical properties such as vapor pressure and aqueous solubility.[1,2] The solubility of a compound in water is strongly correlated with its melting point. An estimate of the water-solubility of a compound before it is synthesized, or available in sufficient purity for analytical measurements, would be most useful.[3] Adequate aqueous solubility is necessary for a compound to be transported to the active site within an organism. As noted above, melting point affects solubility, and solubility controls toxicity in that, if a compound is only poorly soluble, its concentration in the aqueous environment may be too low for it to exert a toxic effect.[4,5] Thus, it would be helpful to be able to estimate the melting point of a compound from its chemical structure.[6,7] Prediction methods for melting point, mainly can be categorized as property-property relationship (PPR), group contribution, and quantitative structure-property relationship (QSPR).[8,9] Comprehensive reviews of the subject reveal that many studies involved hydrocarbons and homologous compounds.[10-12] This is because of the difficulty of melting point prediction for various organic compounds, since the numerous factors that control it are not easy to quantify.

The prediction of physicochemical and biological properties/ activities of organic molecules are the main objective of quantitative structure-property/activity relationships (QSPRs/ QSARs). The QSPR/QSAR models now correlate chemical structure to a wide variety of physical, chemical, biological (including biomedical, toxicological, ecotoxicological) and technological properties.[13-17] QSPR/QSAR models are obtained on the basis of the correlation between the experimental values of the property/activity and descriptors reflecting the molecular structure of the compounds. To obtain a significant correlation, it is crucial that appropriate descriptors be employed. A wide variety of molecular descriptors has been reported for using in QSPR/QSAR models.[18] However, as the number of descriptors (variables) increases, the model becomes complicated, and its interpretation is difficult if many variables are used in modeling. Therefore, the application of these techniques usually requires variable selection for building well-fitted models. A better predictive model can be obtained by ortogonalization of the variables by means of principal component analysis (PCA).[19,20] The principal component analysis was used to compress the descriptor groups into principal components (PC's). In order to reduce the dimensionality of the independent variable space, a limited number of PC's are used.[21] Hence, selecting the significant and informative PC's is the main problem in all of the PCA-based calibration methods.[22-25] Different methods have been addressed to select the significant PC's for calibration purposes. The simplest and most common one is a top-down variable selection where the PC's are ranked in the order of decreasing eigenvalues and the PC's with highest eigenvalue is considered as the most significant one and, subsequently, the PC's are introduced into the calibration model. However, the magnitude of an eigenvalue is not necessarily a measure of its significance for the calibration.[25] In the other method,

which is called correlation ranking, the PC's are ranked by their correlation coefficient with the property and selected by the procedure discussed for eigenvalue ranking.[22,23] Better results are often achieved by this method. Recently, genetic algorithm (GA) has been applied for the selection of the most relevant PC's instead of the older methods. Comparison of the results obtained using GA principal component selection with the two above-mentioned methods shows that GA gives a better result and close to the correlation ranking.[26-28] GA is a stochastic method to solve optimization problems applying evolution hypothesis of Darwin and different genetic functions, *i.e*., cross-over and mutation.[29,30] Genetic algorithm is robust, global and generally more straightforward to apply in situations where there is little or no a *priori* knowledge about the process to be controlled.[29]

Artificial neural networks (ANNs) have become popular in QSPR/QSAR models due to their success where complex non-linear relationships exist amongst data.[31,32] An ANN is formed from artificial neuron, connected with coefficients (weights), which constitute the neural structure and are organized in layers. The layers of neurons between the input and output layers are called hidden layers. Neural networks do not need explicit formulation of the mathematical or physical relationships of the handled problem. These give ANNs an advantage over traditional fitting methods for some chemical applications. For these reasons in recent years, ANNs have been applied to a wide variety of chemical problems.[33-42]

Very recently, QSPR models have been applied for prediction of the melting point of 323 set of drug-like compounds.[43] Ability of these models for prediction of the melting point is poor (for example, root-mean square error of the models is approximately 40.7 ºC). In order to predict accurately melting point of the same compounds, in the present work, principal component-genetic algorithm-multiparameter linear regression (PC-GA-MLR) and principal component-genetic algorithm-artificial neural network (PC-GA-ANN) models were employed to generate QSPR models between the principal components and melting point of the compounds and the results were compared with each other, the previous work and the experimental values.

Data and Methodology

Data set and theoretical descriptors. Melting points were taken from the recently published paper.[43] The data are mostly for the compounds that are solid at room temperature but also include some liquids and gaseous compounds. The melting points are spread between −118 and 345 ºC. The z-matrices (molecular models) were constructed with HyperChem 7.0 and molecular structures were optimized using AM1 algorithm.[44] In order to calculate the theoretical descriptors, *Dragon* package version 2.1 was used.[45] For this propose the output of the HyperChem software for each compound fed into the *Dragon* program and the descriptors were calculated. As a result, a total of 1481 theoretical descriptors were calculated for each compound in data sets (323 compounds).

Data pretreatment. The theoretical descriptors were reduced by the following procedure: 1) descriptors that are constant have been eliminated (292 descriptors). 2) in addition, to decrease the redundancy existing in the descriptors data matrix, the correlation of descriptors with each other and with melting point of the molecules are examined, and collinear descriptors ($R > 0.9$) are detected. Those of the descriptors which have the pair wise correlation coefficient above 0.9 and having the lower correlation with melting point values are removed from the data matrix (758 descriptors). 3) before statistical analysis, the descriptors are scaled to zero mean and unit variance (autoscaling procedure). The data matrix (431 descriptors) is subjected to principal component analysis using Matlab software package.[46] Multiparameter linear regression was obtained using spss software.[47]

Genetic algorithm (GA). To select the most relevant principal components, evolution of population was simulated.[48-52] Each individual of the population defined by a chromosome of binary values represented a subset of principal components. The number of genes at each chromosome was equal to the number of principal components. The population of the first generation was selected randomly. A gene took a value of 1 if its corresponding principal component was included in the subset; otherwise, it took a value of zero. The number of genes with a value of 1 was kept relatively low to have a small subset of principal components,[52] that is, the probability of generating 0 for a gene was set greater (at least 60%) than the value of 1. The operators used here were crossover and mutation. The probability of the application of these operators was varied linearly with generation renewal (0-0.1% for mutation and 60-90% for crossover). The population size was varied between 50 and 250 for different GA runs. For a typical run, the evolution of the generation was stopped when 90% of the generations took the same fitness. The GA program was written in Matlab 6.5.[53]

Artificial neural network (ANN). A feed forward artificial neural network with a back-propagation of error algorithm was used to process the non-linear relationship between the selected principal components and the melting point. The number of input nodes in the ANN was equal to the number of PC's. The ANN models confined to a single hidden layer, because the network with more than one hidden layer would be harder to train. A three-layer network with a sigmoid transfer function was designed. The initial weights were randomly selected between 0 and 1. Optimization of the weights and biases was carried out according to the resilient back-propagation algorithm. The data set was randomly divided into three groups: a training set, a validation set and a prediction set consisting of 195, 64 and 64 molecules, respectively. The training and validation sets were used for the model generation and the prediction set was used for evaluation of the generated model. The performances of training, validation and prediction of models are

evaluated by the mean percentage deviation (MPD) and root mean square error (RMSE), which are defined as follows:

$$\text{MPD} = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{P_i^{\exp} - P_i^{cal}}{P_i^{\exp}}\right| \quad (1)$$

$$\text{RMSE} = \sqrt{\sum_{i=1}^{N}\frac{(P_i^{\exp} - P_i^{cal})^2}{N}} \quad (2)$$

where $P_i^{\exp}$ and P_i^{cal} are experimental and calculated values of melting point with the models and N denote the number of data points. Individual percent deviation (IPD) is defined as follows:

$$\text{IPD} = 100 \times \left(\frac{P_i^{cal} - P_i^{\exp}}{P_i^{\exp}}\right) \quad (3)$$

The processing of the data was carried using Matlab 6.5.[46] The neural networks were implemented using Neural Network Toolbox Ver. 4.0 for Matlab.[54]

Results and Discussion

Principal component analysis. After the elimination of the constant and one of the collinear ones, 431 descriptors remained from 1481 theoretical descriptors calculated for the compounds. The results of application of PCA on the descriptors data matrix were shown that 99.9% of the variances in the descriptors data matrix are explained by 234 first PC's. Therefore, we focused our analysis on these PC's, and the reminders, which are noisy factors, were not considered.

Principal component-genetic algorithm-multiparameter linear regression. Obtaining the number of significant principal components is the main problem in the PCA-based methods. The first 234 principal components (PC's) were found to explain more than 99.9% of variances in the original data matrix. As noted previously, not all of the PC's is informative for QSAR/QSPR modeling.[25-27] Then, we used GA for the selection of the most relevant PC's instead of the older methods. The selected PC's are PC1, PC2, PC3, PC4, PC5, PC6, PC7, PC9, PC15, PC32, PC33, PC36, PC37, PC39 and PC86. As can be seen, the selected principal components are not based on their eigenvalue. For example, PC9 and PC15 are selected and PC8 is not considered in the model. This is due to the fact the information contents of some extracted PC's may not be in the same direction of the activity data. Multiparameter linear correlation of melting point values for 195 compounds in training set was obtained using the fifteen principal components. The calculated values of melting point for the compounds in training, validation and prediction sets using the PC-GA-MLR model have been plotted *versus* the experimental values of it (Figure 1).

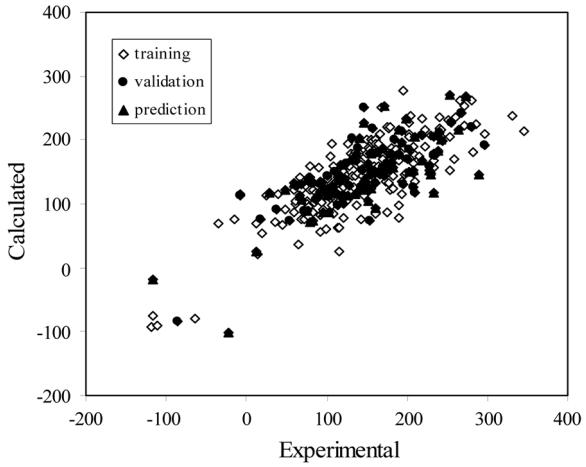

Figure 1. Plot of calculated values of the melting point using the PC-GA-MLR model *versus* the experimental values of it for training, validation and prediction sets.

Principal component-genetic algorithm-artificial neural network. To process the non-linear relationships exists between the melting point and the PC's, the ANN modeling method combined with PCA for dimension reduction and GA for feature selection was employed. A principal component-genetic algorithm-artificial neural network (PC-GA-ANN) model, which combines the PC's with ANN, is another PC-based calibration technique for non-linear modeling between the PC's and dependent variables.[25-28] The input vectors were the set of PC's, which were selected by GA, and therefore, the number of nodes in the input layer was dependent on the number of selected PC's. In the PC-GA-MLR model it is assumed that the PC's are independent of each other and truly additive relevant to the property under study. ANNs are particularly well-suited for QSAR/QSPR models because of their ability to extract non-linear information present in the data matrix. For this reason the next step in this work was generation of the ANN model. There are no rigorous theoretical principles for choosing the proper network topology; so different structures were tested in order to obtain the optimal hidden neurons and training cycles.[34-42] Before training the network, the number of nodes in the hidden layer was optimized. In order to optimize the number of nodes in the hidden layer, several training sessions were conducted with different numbers of hidden nodes (from one to thirty two). The root mean square error of training (RMSET) and validation (RMSEV) sets were obtained at various iterations for different number of neurons at the hidden layer and the minimum value of RMSEV was recorded as the optimum value. Plot of RMSET and RMSEV *versus* the number of nodes in the hidden layer has been shown in Figure 2. It is clear that the twenty nine nodes in hidden layer is the optimum value.

This network consists of fifteen inputs (including PC1, PC2, PC3, PC4, PC5, PC6, PC7, PC9, PC15, PC32, PC33, PC36, PC37, PC39 and PC86), the same PC's in the PC-GA-MLR model, and one output for melting point. Then an ANN with architecture 15-29-1 was generated. It is noteworthy that training of the network was stopped when the RMSEV started to increases *i.e.* when overtraining begins.

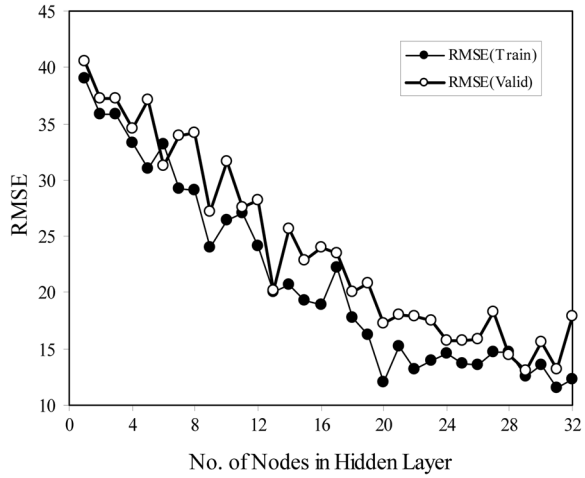

Figure 2. Plot of RMSE for training and validation sets *versus* the number of nodes in hidden layer.

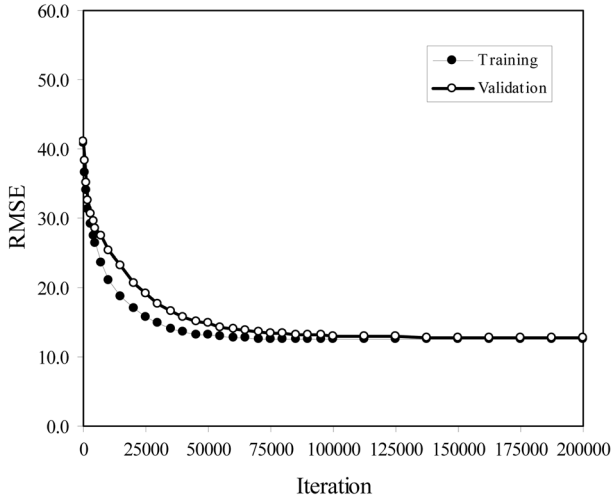

Figure 3. Plot of RMSE for training and validation sets *versus* the number of iterations.

The overtraining causes the ANN to loose its prediction power.[31] Therefore, during training of the network, it is desirable that iterations are stopped when overtraining begins. To control the overtraining of the network during the training procedure, the values of RMSET and RMSEV were calculated and recorded to monitor the extent of the learning in various iterations. Results showed that overfitting did not see in the optimum architecture (Figure 3).

The generated ANN was then trained using the training and validation sets for the optimization of the weights and biases. For the evaluation of the predictive power of the generated ANN, an optimized network was applied for prediction of the melting point values in the prediction set, which were not used in the modeling procedure (Table 1). The calculated values of melting point for the compounds in training, validation and prediction sets using the ANN model have been plotted *versus* the experimental values of it in Figure 4.

It is clear that the calculated values of melting point are in good agreement with those of the experimental values. The

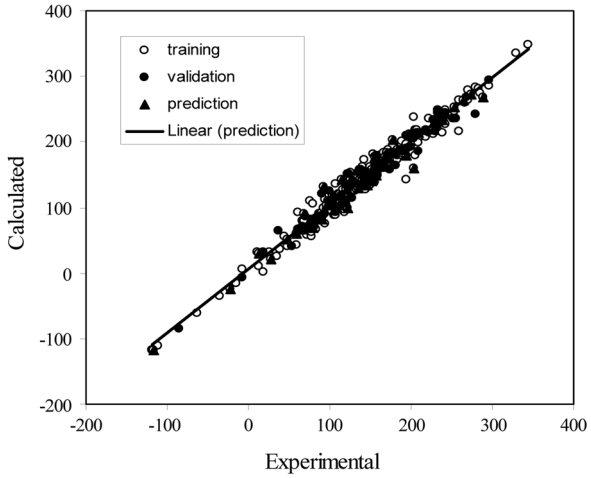

Figure 4. Plot of calculated values of the melting point using the PC-GA-ANN model *versus* the experimental values of it for training, validation and prediction sets.

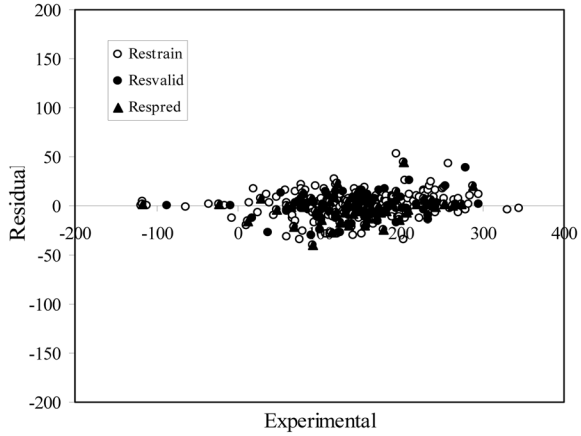

Figure 5. Plot of the residual for calculated values of the melting point using the PC-GA-ANN model *versus* the experimental values of it.

correlation equation for all of the calculated values of melting point (Mp) from the ANN model and the experimental values is as follows:

$$Mp(cal) = 0.969\ Mp(exp) + 4.381 \quad (4)$$

(R = 0.9850; MPD = 9.326; RMSE = 12.623; F = 10445.99)

Similarly, correlation of Mp(cal) *versus* Mp(exp) values in the prediction set gives equation (5):

$$Mp(cal) = 0.972\ Mp(exp) + 5.623 \quad (5)$$

(R = 0.9843; MPD = 9.119; RMSE = 12.767; F =1930.99)

Plot of the residual for melting point values in the training, validation and prediction sets *versus* the experimental values of it has been illustrated in Figure 5. It is clear that the propagation of errors in both sides of zero is random. Then there is not systematic error in the model.

As a result, it was found that properly selected and trained neural network could fairly represent dependence of melting point for the drug-like compounds on the PC's. Then the

Table 1. Experimental and calculated values of melting point for the drug-like compounds in training, validation and prediction sets using PC-GA-MLR and PC-GA-ANN models along with the residual for the calculated values by PC-GA-ANN model

No.	Compound	Experimental	Cal (PC-GA-ANN)	Res.
	Training			
1	Halothane	−118	−118.3	0.3
2	Diethyl ether	−116.3	−120.2	3.9
3	Ethylene oxide	−111.3	−112.0	0.7
4	Chloroform	−63.7	−62.3	−1.4
5	Methoxyflurane	−35	−35.9	0.9
6	Benzyl alcohol	−15.3	−14.9	−0.4
7	Nicotinyl alcohol	−7.7	5.1	−12.8
8	Amphetamine	11.3	31.8	−20.5
9	Glyceryl trinitrate	13.5	10.5	3.0
10	Propofol	19	2.2	16.8
11	Nikethamide	25	32.1	−7.1
12	Ephedrine	36	24.7	11.3
13	Methyl nicotinate	39	36.7	2.3
14	Trimipramine	45	55.0	−10.0
15	Phencarbamide	48	39.7	8.3
16	Hyoscine	59	43.4	15.6
17	Prometazine	60	57.5	2.5
18	Gemfibrozil	61	92.0	−31.0
19	Procaine	61	65.5	−4.5
20	Dichloralphenazone	65.5	67.1	−1.6
21	Etomidate	67	80.3	−13.3
22	Lignocaine	67.5	79.9	−12.4
23	Penbutolol	68	78.5	−10.5
24	Betaxolol	71	86.4	−15.4
25	Mephenesin	71.5	57.3	14.2
26	Phenadoxone	75	71.1	3.9
27	Ibuprofen	76	110.7	−34.7
28	Mebutamate	77	71.4	5.6
29	Oxprenolol	77.5	56.5	21.0
30	Methadone	78	61.1	16.9
31	Allylestrenol	80	80.0	0.0
32	Bamifylline	80	106.1	−26.1
33	Nabumetone	80	67.0	13.0
34	Anileridine	83	83.3	−0.3
35	Fentanyl	83	67.2	15.8
36	Amphetaminil	85	84.2	0.8
37	Methdilazine	87	91.1	−4.1
38	Noxythiolin	88	90.1	−2.1
39	Vinylbital	90	83.4	6.6
40	Phenindamine	91	92.9	−1.9
41	Carisoprodol	92	87.6	4.4
42	Beclamide	92.5	99.1	−6.6
43	Perphenazine	94	110.8	−16.8
44	Thenalidine	95	75.2	19.8
45	Tropicamide	96.5	96.6	−0.1
46	Aldicarb	99	97.6	1.4
47	Acetylpheneturide	100	96.2	3.8
48	Phenocoll	100.5	117.7	−17.2
49	Piperidione	102	106.1	−4.1
50	Isoxsuprine	102.5	94.8	7.7

Table 1. Continued

No.	Compound	Experimental	Cal (PC-GA-ANN)	Res.
51	Meprobamate	104	104.1	−0.1
52	Gentamicin	105	104.9	0.1
53	Physotigmine	105.5	88.2	17.3
54	Bupivacaine	107	89.2	17.8
55	Amidopyrine	108	136.4	−28.4
56	Acecarbromal	109	105.4	3.6
57	Celiprolol	110	107.8	2.2
58	Tolnaftate	111	121.1	−10.1
59	Amphotalide	113	119.7	−6.7
60	Valnoctamide	113.5	111.1	2.4
61	Ifenprodil	114	115.5	−1.5
62	Bamipine	115	104.8	10.2
63	Alverine	116	128.1	−12.1
64	Pericyazine	116	116.1	−0.1
65	Atropine	118	114.8	3.2
66	Morphazinamide	118.5	91.8	26.7
67	Chlophedianol	120	125.4	−5.4
68	Pridinol	120	99.3	20.7
69	Terbutaline	120.5	130.8	−10.3
70	Capobenic acid	121	124.6	−3.6
71	Propizepine	122	150.0	−28.0
72	Nadolol	124	117.9	6.1
73	Bamethan	125	114.0	11.0
74	Nimodipine	125	126.3	−1.3
75	Mecloqualone	126	153.3	−27.3
76	Febantel	129	128.3	0.7
77	Clonidine	130	136.1	−6.1
78	Xylometazoline	131	124.8	6.2
79	Diazepam	133	127.3	5.7
80	Thozalinone	133	133.5	−0.5
81	Aminorex	136	145.7	−9.7
82	Praziquantel	136	128.2	7.8
83	Simvastatin	136.5	142.4	−5.9
84	Butalbital	138	138.8	−0.8
85	Phenazopyridine	139	147.9	−8.9
86	Erythrocentaurin	140	161.0	−21.0
87	Carbaryl	142	144.1	−2.1
88	Fexofenadine	142	141.0	1.0
89	Letosteine	142	149.8	−7.8
90	Acetylsalicylic acid	142.4	172.9	−30.5
91	Tetrazepam	144	126.4	17.6
92	Felodipin	145	140.9	4.1
93	Metoclopramide	146.5	153.7	−7.2
94	Atenolol	147	152.7	−5.7
95	clotrimazole	147	144.6	2.4
96	Salacetamide	148	157.1	−9.1
97	Morazone	149	146.7	2.3
98	Astemizole	149.1	162.3	−13.2
99	Acemetacin	150	134.2	15.8
100	Mafenide	151	140.9	10.1
101	Haloperidol	151.5	148.2	3.3
102	Glymidine	152	152.4	−0.4
103	Azatadine	153	148.1	4.9
104	Testosterone	153	180.9	−27.9

Table 1. Continued

No.	Compound	Experimental	Cal (PC-GA-ANN)	Res.
105	Taurolidine	154	152.8	1.2
106	Colchicane	156	160.7	−4.7
107	moricizine	156	157.3	−1.3
108	Omeprazole	156	150.3	5.7
109	Urapidil	156	137.3	18.7
110	Salicylic acid	157	163.0	−6.0
111	Succisulfone	157	152.0	5.0
112	Lidoflazine	159	153.8	5.2
113	Azacyclonol	160	158.1	1.9
114	Benzydamine	160	164.0	−4.0
115	Didanosine	160	156.1	3.9
116	Ketorolac	160.5	178.2	−17.7
117	Oxaprozin	160.5	161.7	−1.2
118	Aldosterone	164	176.3	−12.3
119	Pizotifen	164	169.8	−5.8
120	Tolrestat	164	175.0	−11.0
121	Lorazepam	166	184.1	−18.1
122	Sulfamethoxazole	167	161.7	5.3
123	Chlortetracycline	168.5	168.3	0.2
124	Glyburide	169	170.0	−1.0
125	Benperidol	170	161.2	8.8
126	Metopimazine	170	160.5	9.5
127	Tolazamide	170	183.7	−13.7
128	Isoniazid	172	188.2	−16.2
129	Hydralazine	172.5	166.3	6.2
130	Nifedipine	173	179.0	−6.0
131	Lovastatin	174.5	159.5	15.0
132	Amisometradine	175	167.6	7.4
133	Acifran	176	179.3	−3.3
134	Melphalan	177	170.2	6.8
135	Propallylonal	177	179.7	−2.7
136	Sulpiride	178	184.0	−6.0
137	Zomepirac	178	177.0	1.0
138	Nomifensine	179	165.8	13.2
139	Sulthiame	180	174.8	5.2
140	Acepromazine	182.5	174.7	7.8
141	Amphenidone	182.5	173.2	9.3
142	Sulfacetamide	183	179.4	3.6
143	Bezafibrate	186	186.8	−0.8
144	Acetohexamide	189	179.6	9.4
145	Pyrazinamide	189	200.8	−11.8
146	Clomipramine	189.5	184.5	5.0
147	Carbamazepine	190	181.0	9.0
148	Embutramide	190.5	181.4	9.1
149	Apronal	194	197.3	−3.3
150	Clebopride	194	141.2	52.8
151	Methotrexate	195	196.4	−1.4
152	Aceglutamide	197	189.8	7.2
153	Aceneocoumarol	197	208.6	−11.6
154	Furonazide	199	209.9	−10.9
155	Polythiazide	202.5	205.2	−2.7
156	Ampicillin	203	237.0	−34.0
157	Picrotoxin	203	199.1	3.9

Table 1. Continued

No.	Compound	Experimental	Cal (PC-GA-ANN)	Res.
158	Glipizide	205	218.5	−13.5
159	Oxazepam	205.5	179.5	26.0
160	Lonidamine	207	206.8	0.2
161	Amodiaquine	208	206.3	1.7
162	Indoramin	208	217.2	−9.2
163	Vigabatrin	209	201.8	7.2
164	Methetion	210	198.8	11.2
165	Pimozide	216	211.9	4.1
166	Oxycodone	219	208.3	10.7
167	Hydroxyprogesterone	222.5	213.5	9.0
168	Hydrocortisone	223	235.9	−12.9
169	Apazone	228	212.5	15.5
170	Acitretin	229	215.0	14.0
171	Nalidixic acid	229.5	219.1	10.4
172	Salinazid	232.5	235.6	−3.1
173	Diaveridine	233	227.8	5.2
174	Phenopyrazone	233	217.5	15.5
175	Pyrimethamine	233.5	230.3	3.2
176	Nicotinic acid	235.5	215.7	19.8
177	Caffiene	238	213.3	24.7
178	Prednisolone	240.5	231.7	8.8
179	Cromolyn	241	238.6	2.4
180	Clometacin	242	225.9	16.1
181	Domperidone	242.5	249.3	−6.8
182	Metolazone	252	235.4	16.6
183	Finasteride	253	245.9	7.1
184	Nifenazone	253	234.4	18.6
185	Pemoline	259	215.5	43.5
186	Dexamethasone	260	262.7	−2.7
187	Ciprofloxacin	266	259.9	6.1
188	Hydroflumethiazide	270.5	262.8	7.7
189	Acefylline	271	278.1	−7.1
190	Dantrolene	279.5	283.8	−4.3
191	Fluorouracil	283	281.2	1.8
192	Prazosin	285	274.5	10.5
193	Enoxolone	296	284.7	11.3
194	Diazoxide	330.5	334.5	−4.0
195	Orotic acid	345	347.6	−2.6
	Validation			
196	Trichlorethylene	−86	−85.7	−0.3
197	Methyl salicylate	−8	−7.5	−0.5
198	Benzyl benzoate	18	31.4	−13.4
199	Prilocaine	37	63.9	−26.9
200	Ethopropazine	53	39.9	13.1
201	Isosorbide	61	66.4	−5.4
202	Fluanisone	67.5	70.6	−3.1
203	Disulfiram	71	67.8	3.2
204	Ethylesterol	77	69.4	7.6
205	Moxaverine	78	81.0	−3.0
206	Pentifylline	82	70.8	11.2
207	Piprozolin	86	83.4	2.6
208	Alclofenac	91	120.5	−29.5
209	Ketoprofen	94	90.4	3.6

Table 1. Continued

No.	Compound	Experimental	Cal (PC-GA-ANN)	Res.
210	Cocaine	98	109.3	−11.3
211	Hycanthone	100.6	124.4	−23.8
212	Benzoyl peroxide	105	103.7	1.3
213	Metaraminol	107.5	92.5	15.0
214	Flurbiprofen	110	99.1	10.9
215	Acetanilide	114	119.3	−5.3
216	Dibenzepin	116	109.3	6.7
217	Antazoline	120	105.2	14.8
218	Acebutolol	121	134.4	−13.4
219	Benzarone	124.3	150.9	−26.6
220	Tolbutamide	128.5	113.8	14.7
221	Benzylmorphine	132	135.8	−3.8
222	Mephenytoin	136	154.0	−18.0
223	Alizapride	139	157.2	−18.2
224	Cimetidine	142	133.5	8.5
225	Carbutamide	144	145.6	−1.6
226	Pyrinoline	146.5	153.9	−7.4
227	Thialbarbital	148	148.2	−0.2
228	Salbutamol	151	143.7	7.3
229	Bufexamac	153	138.0	15.0
230	Ketobemidone	156	167.4	−11.4
231	Dihydromorphine	157	178.3	−21.3
232	Metronidazole	159	148.2	10.8
233	Methallatal	160	158.4	1.6
234	Halazepam	164	160.9	3.1
235	Clobazam	167	159.3	7.7
236	Sumatriptan	169	161.2	7.8
237	Hydroquinine	172	181.5	−9.5
238	Heptabarbital	174	158.1	15.9
239	Mephobarbital	176	181.6	−5.6
240	Ximoprofen	178	183.7	−5.7
241	Androstanolone	181	164.1	16.9
242	Zox azolamine	184	183.8	0.2
243	Verazide	189	186.0	3.0
244	Acediasulfone	194	210.4	−16.4
245	Probenecid	195	190.4	4.6
246	Alphadolone	200	192.0	8.0
247	Ursodiol	203	205.5	−2.5
248	Sotalol	207	209.4	−2.4
249	Acecainide	210	184.6	25.4
250	Propylthiouracil	219	218.1	0.9
251	Azapropazone	228	232.9	−4.9
252	Chlorazanil	233	247.6	−14.6
253	Sulfamerazine	234	243.9	−9.9
254	Amiloride	241	244.4	−3.4
255	Azathioprine	243.5	240.3	3.2
256	Morphine	255	234.7	20.3
257	Fosfosal	268	266.9	1.1
258	Moxestrol	280	241.8	38.2
259	Flucytosine	296	294.2	1.8
	Prediction			
260	Sevoflurane	−116	−116.7	0.7
261	Tetrachloroethylene	−22.3	−24.1	1.8
262	Paraldehyde	12.6	28.9	−16.3
263	Tranylcypromine	28	21.0	7.0
264	Ifosfamide	48	51.7	−3.7
265	Triprolidine	60	59.3	0.7

Table 1. Continued

No.	Compound	Experimental	Cal (PC-GA-ANN)	Res.
266	Chlorambucil	66	67.2	−1.2
267	Ranitidine	69	90.9	−21.9
268	Propoxyphene	75	70.3	4.7
269	Etisazol	78	74.1	3.9
270	Guaiphenesin	80	67.5	12.5
271	Metrifonate	83	84.0	−1.0
272	Benzocaine	90	82.6	7.4
273	Maprotiline	92	132.4	−40.4
274	Tamoxifen	96	102.2	−6.2
275	Metaproterenol	100	106.0	−6.0
276	Difenidol	103.5	118.1	−14.6
277	Pipobroman	106	107.3	−1.3
278	Acetylcysteine	109.5	98.4	11.1
279	Cyproheptad ine	113	108.4	4.6
280	Flupirtine	115	142.8	−27.8
281	Moperone	118	110.0	8.0
282	Temazepam	120	127.7	−7.7
283	Benzoic acid	122.4	99.0	23.4
284	Lofexidine	126	142.9	−16.9
285	Bitoscanate	131	138.2	−7.2
286	Phenacetin	134.5	130.0	4.5
287	Sulfinpyrazone	136.5	156.9	−20.4
288	Aprobarbitone	141	141.2	−0.2
289	Proglumide	142	149.9	−7.9
290	Ketoconazole	146	133.5	12.5
291	Cloricromen	147.5	139.3	8.2
292	Felbamate	151	140.0	11.0
293	Naproxen	152	157.5	−5.5
294	Amobarbital	156	176.3	−20.3
295	Phenallymal	156	158.4	−2.4
296	Warfarin	157	148.4	8.6
297	Bucetin	160	172.4	−12.4
298	Famotidine	163	166.9	−3.9
299	Tyramine	164	169.3	−5.3
300	Acetaminophen	169	176.5	−7.5
301	Risperdone	170	183.1	−13.1
302	Tetracycline	172.5	174.3	−1.8
303	Amoxapine	175.5	182.6	−7.1
304	Oxymetholone	178	202.9	−24.9
305	Dextromoramide	180	182.5	−2.5
306	Clozapine	183	188.3	−5.3
307	Glisoxepid	189	187.6	1.4
308	Spiperone	190	187.4	2.6
309	Hymecromone	194	179.2	14.8
310	Piroxicam	198	212.6	−14.6
311	Caroxazone	203	158.5	44.5
312	Baclofen	207	214.7	−7.7
313	Buprenorphine	209	213.5	−4.5
314	Griseofulvin	219	217.0	2.0
315	Thioacetazone	227.5	220.7	6.8
316	Oxibendazole	230	224.4	5.6
317	Ubenimex	233	231.6	1.4
318	Lotrifen	238	232.8	5.2
319	Zolimidine	242	244.0	−2.0
320	Flumequine	253	252.0	1.0
321	Reserpine	264.5	264.3	0.2
322	Hydrochlorthiazide	274	272.4	1.6
323	Acedapsone	289	268.8	20.2

optimized neural network could simulate the complicated nonlinear relationship between melting point values and the PC's. The RMSE of 48.176 for the prediction set by the PC-GA-MLR model should be compared with the value of 12.77 for the PC-GA-ANN model. As can be seen, ability of the proposed model to predict the melting point is very higher than the QSPR models proposed in recently published paper (RMSE of 12.767 should be compared with 40.7 ºC). It can be seen that although parameters appearing in the PC-GA-MLR model are used as inputs for the generated PC-GA-ANN model, the statistics has shown a large improvement. These improvements are due to the fact that melting point of the compounds shows non-linear correlations with the principal components.

The melting point of a compound is governed by the intermolecular hydrogen-bonding ability of the molecules, the molecular packing in crystals (effects from molecular shape, size, and symmetry), and other intermolecular interactions such as charge transfer and dipole-dipole interactions in the solid phase.[6] The solubility of a compound can be regarded as a partitioning of the compound between its crystal lattice and the solvent. If the forces holding the molecule in the crystal are high, then the solubility will be low. For the same reason the melting point will be high, since melting point is a measure of the energy required to disrupt the crystal lattice. The molar aqueous solubility can be calculated using melting point of compounds by the general solubility equation.[2] Then melting points affect solubility, and solubility controls toxicity in that; if a compound is only poorly soluble, its concentration in the aqueous environment may be too low for it to exert a toxic effect.[5] As a result prediction of melting point of the compounds using the proposed non-linear model is a valuable method in designing new drugs within a specified range of melting point and solubility.

Conclusions

Quantitative-structure property relationships have been applied for prediction of melting point for 323 drug-like compounds by using the principal component-genetic algorithm-multi parameter linear regression (PC-GA-MLR) and principal component-genetic algorithm-artificial neural network (PC-GA-ANN) methods. Comparison of the statistical parameters obtained for training, validation and prediction sets by the PC-GA-MLR and PC-GA-ANN models demonstrate superiority of the PC-GA-ANN model over the PC-GA-MLR model. Root-mean square error of 48.18 for the prediction set by the PC-GA-MLR model should be compared with the value of 12.77 ºC for the PC-GA-ANN model. Since the improvement of the results obtained using non-linear model (PC-GA-ANN) is considerable, it can be concluded that the non-linear characteristics of the principal components on melting point of the compounds is serious.

Acknowledgements. The Authors wish to acknowledge the vice-presidency of research, University of Mohaghegh Ardabili, for financial support of this work.

References

1. Meylan, W. H.; Howard, P. H.; Boethling, R. S. *Environ. Toxicol. Chem.* **1996**, *15*, 100.
2. Ran, Y.; Yalkowsky, S. H. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 354.
3. Nimko, J.; Kukkonen, J.; Riikonen, K. *J. Hazard Mater.* **2002**, *91*, 43.
4. Dearden, J. C. *Sci. Total Environ.* **1991**, *109/110*, 59.
5. Katritzky, A. R.; Jain, R.; Lomaka, A.; Petrukhin, R.; Maran, U.; Karelson, M. *Cryst. Growth Des.* **2001**, *1*, 261.
6. Godavarthy, S. S.; Robinson, R. L.; Gasem, K. A. M. *Ind. Eng. Chem. Res.* **2006**, *45*, 5117.
7. Gao, J.; Wang, X.; Yu, X.; Li, X.; Wang, H. *J. Mol. Model* **2006**, *12*, 521.
8. Krzyzaniak, J. F.; Myrdal, P. B.; Simamora, P.; Yalkowsky, S. H. *Ind. Eng. Chem. Res*. **1995**, *34*, 2530.
9. Karthikeyan, M.; Glen, R. C.; Bender, A. *J. Chem. Inf. Model.* **2005**, *45*, 581.
10. Toropov, A.; Toropova, A.; Ismailov, T.; Bonchev, D. *J. Mol. Struct. (Theochem)* **1998**, *424*, 237.
11. Firpo, M.; Gavernet, L.; Castro, E. A.; Toropov, A. *J. Mol. Struct. (Theochem)* **2000**, *501-502*, 419.
12. Toropov, A.; Toropova, A. *J. Mol. Struct. (Theochem)* **2002**, *581*, 11.
13. Yao, X. J.; Wang, Y. W.; Zhang, X. Y.; Zhang, R. S.; Liu, M. C.; Hu, Z. D.; Fan, B. T. *Chemom. Intell. Lab. Syst.* **2002**, *62*, 217.
14. Consonni, V.; Todeschini, R.; Pavan, M.; Gramatica, P. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 693.
15. Karthikeyan, M.; Glen, R. C.; Bender, A. *J. Chem. Inf. Model* **2005**, *45*, 581.
16. Ajmani, S.; Rogers, S. C.; Barley, M. H.; Livingstone, D. J. *J. Chem. Inf. Model* **2006**, *46*, 2043.
17. Gramatica, P.; Giani, E.; Papa, E. *J. Mol. Graph. Model* **2007**, *25*, 7556.
18. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors* in *Methods and Principles in Medicinal Chemistry*; Mannhold, R.; Kubinyi, H.; Timmerman, H., Eds.; Wiley-VCH: Weinheim, 2000.
19. Sutter, J. M.; Kalivas, J. H.; Lang, P. M. *J. Chemometr.* **1992**, *6*, 217.
20. Malinowski, E. R. *Factor Analysis in Chemistry*; Wiley-Interscience: New York, 2002.
21. Katritzky, A. R.; Tulp, I.; Fara, D. C.; Lauria, A.; Maran, U.; Acree, W. E. *J. Chem. Inf. Model* **2005**, *45*, 913.
22. Hemmateenejad, B.; Akhond, M.; Miri, R.; Shamsipur, M. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1328.
23. Hemmateenejad, B.; Shamsipur, M. *Internet Electron. J. Mol. Des.* **2004**, *3*, 316.
24. Jalali-Heravi, M.; Kyani, A. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1328.
25. Hemmateenejad, B.; Safarpour, M. A.; Miri, R.; Nesari, N. *J. Chem. Inf. Model.* **2005**, *45*, 190.
26. Hemmateenejad, B.; Safarpour, M.; Miri, R.; Taghavi, F. *J. Comput. Chem.* **2004**, *25*, 1495.
27. Depczynski, U.; Frost, V. J.; Molt, K. *Anal. Chim. Acta* **2000**, *420*, 217.
28. Hemmateenejad, B. *Chemom. Intell. Lab. Syst.* **2005**, *75*, 231.
29. Goldberg, D. E. *Genetic Algorithm in Search, Optimization and Machine Learning*; Addison-Wesley-Longman: Reading, MA, USA, 2000.
30. Cho, S. J.; Hermsmeier, M. A. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 927.
31. Despagne, F.; Massart, D. L. *Analyst* **1998**, *123*, 157.
32. Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*; Wiley-VCH: Germany, 1999.
33. Meiler, J.; Meusinger, R.; Will, M. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1169.

34. Habibi-Yangjeh, A.; Nooshyar, M. *Phys. Chem. Liq.* **2005**, *43*, 239.
35. Habibi-Yangjeh, A.; Nooshyar, M. *Bull. Korean Chem. Soc*. **2005**, *26*, 139.
36. Habibi-Yangjeh, A.; Danandeh-Jenagharad, M.; Nooshyar, M. *Bull. Korean Chem. Soc*. **2005**, *26*, 2007.
37. Habibi-Yangjeh, A.; Danandeh-Jenagharad, M.; Nooshyar, M. *J. Mol. Model*. **2006**, *12*, 338.
38. Tabaraki, R.; Khayamian, T.; Ensafi, A. A. *J. Mol. Graph. Model* **2006**, *25*, 46.
39. Habibi-Yangjeh, A. *Phys. Chem. Liq*. **2007**, *45*, 471.
40. Habibi-Yangjeh, A.; Danandeh-Jenagharad, M. *Indian J. Chem.* **2007**, *46B*, 478.
41. Habibi-Yangjeh, A. *Bull. Korean Chem. Soc*. **2007**, *28*, 1472.
42. Habibi-Yangjeh, A.; Esmailian, M. *Bull. Korean Chem. Soc.* **2007**, *28*, 1477.
43. Modarresi, H.; Dearden, J. C.; Modarress, H. *J. Chem. Inf. Model.* **2006**, *46*, 930.
44. *HyperChem Release 7*; HyperCube, Inc.: http://www.hyper.com.
45. Todeschini, R. *Milano Chemometrics and QSPR Group*; http://www.disat.unimib.it/vhm.
46. *Matlab 6.5*. Mathworks; 1984-2002.
47. *SPSS for Windows*, Statistical Package for IBM PC; SPSS Inc.: http://www.spss.com.
48. Cho, S. J.; Hermsmeier, M. A. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 927.
49. Baumann, K.; Albert, H.; Von Korff, M. *J. Chemometr.* **2002**, *16*, 339.
50. Lu, Q.; Shen, G.; Yu, R. *J. Comput. Chem.* **2002**, *23*, 1357.
51. Ahmad, S.; Gromiha, M. M. *J. Comput. Chem.* **2003**, *24*, 1313.
52. Deeb, O.; Hemmateenejad, B.; Jaber, A.; Garduno-Juarez, R.; Miri, R. *Chemosphere* **2007**, *67*, 2122.
53. *Genetic Algorithm and Direct Search Toolbox User's Guide*; The Mathworks Inc.: Massachusetts, 2002.
54. *Neural Network Toolbox User's Guide*; The Mathworks Inc.: Massachusetts, 2002.