

Modelling of Normal Boiling Points of Hydroxyl Compounds by Radial Basis Networks

Liangjie Jin, Peng Bai

School of Chemical Engineering and Technology, Tianjin University, Tianjin, PR China

Email address:

jinliangjie868@sina.com (Liangjie Jin), chemeng114_tju@163.com (Peng Bai)

To cite this article:

Liangjie Jin, Peng Bai. Modelling of Normal Boiling Points of Hydroxyl Compounds by Radial Basis Networks. *Modern Chemistry*. Vol. 4, No. 2, 2016, pp. 24-29. doi: 10.11648/j.mc.20160402.12

Received: February 18, 2016; **Accepted:** March 11, 2016; **Published:** May 4, 2016

Abstract: Radial basis networks (RBN) were applied to link molecular descriptor and boiling points of 168 hydroxyl compounds. The total database was randomly divided into a training set(134), a validation set(17) and a testing set(17). Each compound in the lowest energy conformation was numerically characterized with E-dragon software. Then 8 molecular descriptors were selected to develop the RBN model. Simulated with the final optimum RBN model [8-35₍₆₄₎-1], the root mean square errors (RMSE) for the training, the validation and the testing set were 5.55, 4.28, and 5.33, and the correlation coefficients $R=0.994$ (training), 0.994(validation), 0.993(testing). The final RBN model was compared with the multiple linear regression approach and showed more satisfactory results.

Keywords: Radial Basis Networks, Normal Boiling Point, Hydroxyl Compounds, QSPR Model

1. Introduction

The normal boiling point (NBP) can be defined as the temperature at which a pure saturated liquid has a vapor pressure of 760 mm Hg. NBP can be used to estimate many key physical and physicochemical properties such as critical temperature, enthalpy of vaporization and vapor pressure, etc [1-2]. So having an accurate knowledge of the NBP is very important for the chemical industry. The direct measurement of normal boiling point of organic compounds may be costly, laborious, and even dangerous to the researcher or the environment if the compound has some hazardous properties. Therefore, it is essential to develop reliable methods for estimating the NBP of the compounds.

The NBP of compounds is an indicator of the strength of the intermolecular forces which bind them together. The stronger the intermolecular forces, the more tightly packed the atoms and, therefore, the higher NBP. NBP is directly correlated to the chemical structure of a molecule. The classical approach based on chemical structures to predict NBP is the group contribution methods [3-4], where each molecule is considered as made of fundamental groups, each one giving a constant increment to the value of the NBP for a compound [5-6]. The method is applicable only to the compounds for which all group contributions have been

established. Another well-known solution is quantitative structure-property relationships (QSPR) approach [7-9]. In this approach, a QSPR model is introduced by developing a correlation between the NBP and a variety of molecular features.

In recent years, neural networks have become an important modeling technique in the field of QSPR. The advantage of them is in their inherent ability to incorporate nonlinear and cross-product terms into the model. Besides, they do not require knowledge of the mathematical function to be known in advance. Q. F. Li et al. [10] used radial basis function neural networks to link the molecular structures with the boiling points of 106 compounds. The final 10-parameters model showed satisfactory prediction results. Simulated with the final model, the predictive correlation for the training, the validation, and the testing set were 0.998, 0.998 and 0.991, respectively. Gharageizi et al. [11] optimized a three-layer feed forward artificial neural network (ANN) with 44 molecular descriptors to predict the NBP of a very large database. The final model gave $R^2=0.943$ with an RMS error of 22°C for a training set of 14216 compounds, an RMS error of 21°C for a validation set of 1776 compounds and an RMS error of 21°C for a test set of 1776 compounds. The results indicated the ANN would be a promising strategy to predict the NBP of pure chemicals.

In this paper, the QSPR method was applied to predict the boiling points of 168 hydroxyl compounds at standard pressure using radial basis networks. A large number of molecular descriptors were calculated from the chemical structure and were used to describe the structure of hydroxyl compounds. Some of these descriptors were selected and quantitatively related to the boiling points of 168 hydroxyl compounds by using radial basis networks. The results obtained were validated and tested.

2. Database and Mathematical Methods

2.1. Database

Experimental data set of the normal boiling points of 168 compounds containing the group “-OH” are taken from literature [12], which is a handbook of boiling points drawn from the primary chemical literature. From these 168 hydroxyl compounds, 134 compounds were randomly chosen as the training set to generate the structure of the neural networks, 17 for validation of the generated neural networks set and 17 for testing.

2.2. Determination of Molecular Descriptors

Molecular descriptors are numerical characteristics associated to the chemical structures of compounds [13]. The optimized molecular structures are a requisite for the calculation of molecular descriptors. In this paper, the molecular structure of each compound was sketched via the drawing capabilities of Materials studio. Then these chemical structures were initially energy-minimized with compass molecular mechanics method and subsequently subjected to AM1 semi-empirical quantum chemical method for final geometry optimization. The optimized molecular structures were loaded into E-dragon software, which can calculate molecular descriptors free of charge [14-15]. E-dragon software is capable of calculating 1666 descriptors from 22 diverse blocks: Constitutional descriptors, Topological descriptors, Walk and path counts, Connectivity indices, Information indices, 2D-autocorrelation indices, Edge adjacency indices, Burden eigenvalue descriptors, Topological charge indices, Eigenvalue-based indices, Randic molecular profiles, Geometrical descriptors, RDF descriptors, 3D-MoRSE descriptors, WHIM descriptors, GETAWAY descriptors, Functional groups, Atom-centered fragments, Charge descriptors, and Molecular properties.

Among the huge number of the calculated molecular descriptors, a pre-selection was performed to remove some information-poor descriptors by a series of objective methods. Descriptors matching any of the following criteria were eliminated: (1) descriptors were not available for all structures; (2) the values of descriptors were constant for all structures; (3) descriptors with the R^2 value of the one-parameter correlations were lower than 0.1. Besides, among the collinear descriptors, whose pair-correlation coefficient value was greater than 0.98, the one having the highest R^2 value with the boiling points was retained while the rest were discarded.

The next step was to generate an optimal subset of descriptors for the QSPR model. Sequential forward selection was used for descriptors selection. All the remaining descriptors were listed in decreasing order according to the one-parameter R^2 . Starting from the top descriptor, other descriptors were added sequentially. At each step, the probability of the F -value was evaluated to determine the descriptor entry or removal. If the probability of the F -value was below 0.05, the variable was entered, and if the probability of F -value was above 0.10, the variable was removed. The process was repeated until the average absolute relative deviation (AARD) was less than a threshold value 1%. The mathematical definition of AARD is presented as follows:

$$\%AARD = \frac{100}{N} \sum_i^N \frac{|pred(i) - exp(i)|}{exp(i)} \quad (1)$$

where pred and exp stand for the predicted value by model and its corresponding experimental value, respectively.

2.3. RBN Model Development

The selected molecular descriptors were introduced to the radial basis network (RBN) for the final model development. The RBN model is a three layered network where the connections (the hidden layer) are feed-forward between the input and the output layers. The input layer consists of the selected descriptors and the values of the output layer are the target values. The hidden layer neurons have a Gaussian activation function that determines the excitation level of the neurons depending on how close the input data is located with respect to the neuron's center of the activation functions. In this study, the RBN was designed with an artificial neural network toolbox in MATLAB. In MATLAB subroutine, the generation of the RBN model involved the determination of the optimum number of neurons n in the hidden layer and the appropriate Gaussian function parameter *spread* capable of predicting the target with minimum error. The number of neurons n of radial basis functions greatly influences the performance of the radial basis neural networks. If the number is too low, the networks may not calculate a proper estimation of the data. On the other hand, if too many hidden layer units are used, the networks tend to overfit the training data. The larger that *spread* is the smoother the function approximation will be. Too large a *spread* means a lot of neurons will be required to fit a fast changing function. Too small a *spread* means many neurons will be required to fit a smooth function, and the networks may not generalize well. The generalized network configuration could be represented by $[P-n_{spread}-a2]$, where P was the input and $a2$ was the target. The optimal values of n and *spread* were obtained by minimization of the objective function. The root mean square error (RMSE) between the outputs of RBN and the experimental data was set as the objective function. The mathematical definition of RMSE is presented as follows:

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

where *pred* and *exp* stand for the predicted value by model and its corresponding experimental value in literature, respectively. The problem of “over-fitting” may occur during neural networks training. Over-fitting means that the networks have memorized the training examples, but have not learned to generalize to new situations. In this situation, the error of the training set is driven to a very small value, but when new data is presented to the networks the error is large. To prevent over-fitting, the RMSE value of the training set and the validation set were monitored simultaneously during the training phase. If the RMSE value was still decreasing on the training set but began to increase on the validation set, the RBN model began to over-fit. The optimum values of *n* and *spread* were selected at the minimum RMSE in the validation set. Such a simple method was found sufficient for the considered problem. The test set was not used in the model development, but was applied to assess the predictive capability of the model.

3. Results and Discussions

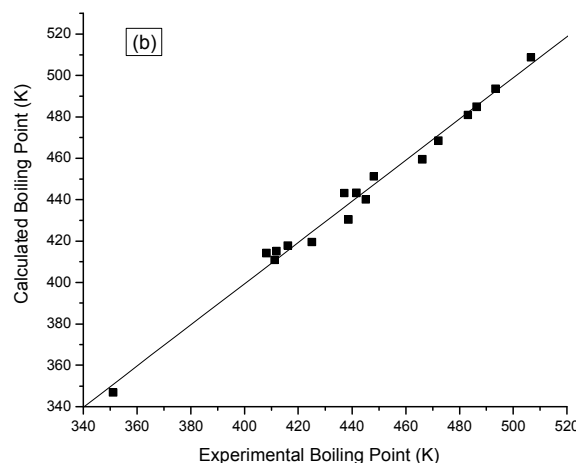
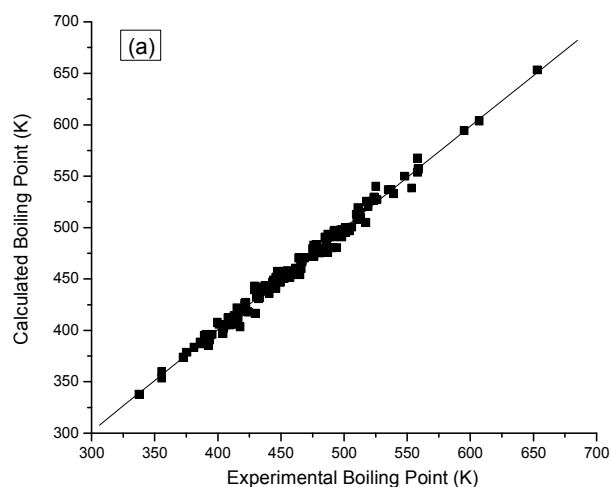
The selected descriptors as well as their definitions and their blocks are shown in Table 1. Eight descriptors belong to 6 descriptor blocks [16]. The Constitutional descriptor here represents information related to the electronic and topological state of the atom in the molecule [17]. The Topological descriptors depict the topological information in the molecule from different aspects [16]. The GETAWAY descriptor (GEometry, Topology, and Atom-Weights Assembly) match 3D-molecular geometry provided by the molecular influence matrix and atom relatedness by molecular topology, with chemical information by using atomic weightings [18-19]. WHIM descriptors are calculated from (x, y, z)-coordinates of a molecule with the atom charge distribution related weighting scheme in a straightforward manner [20]. The Molecular property descriptor here digitizes the hydrophilic properties of the molecules caused by the group “-OH” [16]. The 3D-MoRSE descriptor reflects 3D molecular structure based on electron diffraction [21].

Table 1. Descriptors of the RBN model.

Descriptors	Definitions	Blocks
Ss	Sum of Kier-Hall electrotopological states	Constitutional
GNar	Narumi geometric topological index	Topological
PW4	Path/walk 4 - Randic shape index	Topological
R1p	R autocorrelation of lag 1/weighted by atomic polarizabilities	GETAWAY
E1s	1st component accessibility directional WHIM index / weighted by atomic electrotopological states	WHIM
MAXDN	Maximal electrotopological negative variation	Topological
Hy	Hydrophilic factor	Molecular properties
Mor13u	3D-MoRSE - signal 13/unweighted	3D-MoRSE

The best RBN model generated with the hidden neurons *n*=35 and the Gaussian function parameter *spread*=64. The optimum networks configuration can be represented by [8-35_[64]-1]. The RMSE for the training set, validation set and testing set are 5.55, 4.28, and 5.33, respectively, indicating the accuracy of the RBN model. Table 2 reports the experimental and predicted data for each compound, as well as the relative error (*A%D*). By analyzing the relative errors in Table 2, the large portion of the investigated NBP values were successfully predicted. 90.5% of the investigated data were predicted within a promising range of 0-2%. 7.7% of investigated values were predicted within the range of 2-3%. Only 3 compounds (1.8%) of the predicted values have the relative error greater than 3%. The results indicate that the RBN model has an acceptable predictive capability.

The scatter plots of the predicted boiling point versus the experimental data, which could provide a prompt indication of the accuracy of the RBN model, are reported in Fig. 1. With the RBN model, the performance of prediction for the training set is $NBP_{Exp} = (5.19 \pm 4.25) + (0.990 \pm 0.009) \times NBP_{Pre}$ ($R=0.994$, $n=134$), for the validation set is $NBP_{Exp} = (1.49 \pm 2.79) + (0.995 \pm 0.029) \times NBP_{Pre}$ ($R=0.994$, $n=17$), and for the testing set is $NBP_{Exp} = (-5.15 \pm 13.48) + (1.008 \pm 0.030) \times NBP_{Pre}$ ($R=0.993$, $n=17$).



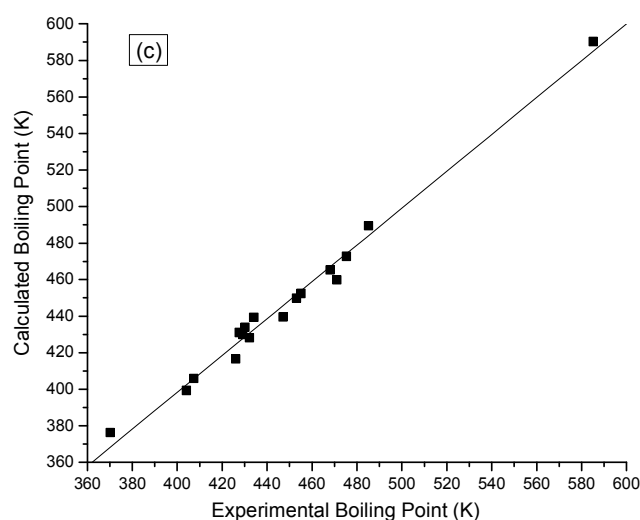


Figure 1. Plots of experimental vs. calculated boiling points for the training (a), validation (b), and test set (c).

A multiple linear regression approach was also employed to describe the relation between NBP and their molecular descriptors. By using these descriptors selected above, the best multiple linear regression function can be obtained as follows:

$$\begin{aligned} \text{NBP}_{\text{Pred}} = & -132.688(\pm 12.687) + 8.958(\pm 0.327) \times (\text{Ss}) \\ & + 103.693(\pm 9.461) \times (\text{GNar}) + 84.688(\pm 29.490) \times (\text{PW4}) \\ & - 62.163(\pm 12.337) \times (\text{R1p}) + 49.552(\pm 7.333) \times (\text{E1s}) \\ & + 12.374(\pm 1.116) \times (\text{Hy}) - 11.064(\pm 3.762) \times (\text{MAXDN}) \\ & - 6.308(\pm 2.544) \times (\text{Mor13u}) \end{aligned} \quad (3)$$

With the multiple linear regression model, the performance of prediction for the training set is $\text{NBP}_{\text{Exp}} = (7.24 \pm 5.63) + (0.984 \pm 0.012) \times \text{NBP}_{\text{Pre}}$ ($R=0.990$, $n=134$), and for the test set (combined with the validation set) is $\text{NBP}_{\text{Exp}} = (43.31 \pm 8.05) + (0.993 \pm 0.018) \times \text{NBP}_{\text{Pre}}$ ($R=0.993$, $n=34$). Compared with the multiple linear regression method, the calculated slope and intercept of the RBN model are more close to the ideal values of 1 and 0, respectively. It is clear that RBN model can give more satisfactory predicted results.

Table 2. Experimental and predicted boiling points for the 168 compounds.

No.	Compounds	Exp. NBP	Pred. NBP	A%D
1	Methanol	337.75	337.73	0.01
2	Ethanol*	351.15	346.87	1.22
3	Propanol**	370.25	376.24	1.62
4	2-propanol	355.55	353.60	0.55
5	Butanol	390.75	396.03	1.35
6	2-methyl-1-propanol	381.25	383.49	0.59
7	2-butanol	372.65	373.94	0.35
8	2-methyl-2-propanol	355.55	359.87	1.22
9	Pentanol*	411.15	410.91	0.06
10	2-pentanol	392.45	385.69	1.72
11	3-pentanol	389.35	395.01	1.45
12	3-methyl-2-butanol	386.05	388.58	0.66

No.	Compounds	Exp. NBP	Pred. NBP	A%D
13	2-methyl-2-butanol	375.45	378.60	0.84
14	3-methyl-1-butanol**	404.15	399.31	1.20
15	2-methyl-1-butanol	401.15	405.73	1.14
16	Hexanol	430.75	432.98	0.52
17	3-isopropylphenol	501.15	498.95	0.44
18	4-pentylphenol	523.65	529.63	1.14
19	4-tert-amylphenol	539.15	532.93	1.15
20	Thymol*	506.65	508.71	0.41
21	phenol	455.15	457.82	0.59
22	2-phenylphenol	548.15	549.96	0.33
23	2-naphthol	558.65	556.82	0.33
24	4-methyl-1,2-dihydroxybenzene	525.15	540.15	2.86
25	1,2-benzenediol	519.05	520.35	0.25
26	1,3-benzenediol	553.65	538.25	2.78
27	4-hydroxydiphenylmethane	595.15	594.28	0.15
28	2-tert-butylhydroquinone	558.15	553.63	0.81
29	3-methyl-1-pentano**	426.15	416.67	2.22
30	2-ethyl-1-butanol	420.15	420.38	0.05
31	2-methyl-1-pentanol	422.15	427.25	1.21
32	4-methyl-1-pentanol*	425.05	419.49	1.31
33	3,3-dimethyl-1-butanol	417.65	403.26	3.45
34	2,3-dimethyl-1-butanol	416.15	412.13	0.97
35	2-hexanol	413.15	405.73	1.80
36	2,2-dimethyl-1-butanol	409.65	405.15	1.10
37	3-hexanol*	408.15	414.18	1.48
38	4-methyl-2-pentanol	404.75	402.05	0.67
39	2-methyl-3-pentanol	399.65	407.72	2.02
40	3-methyl-3-pentanol	395.55	396.21	0.17
41	2-methyl-2-pentanol	394.25	395.75	0.38
42	3-methyl-2-pentanol**	407.45	405.84	0.40
43	3,3-dimethyl-2-butanol	393.55	390.69	0.73
44	2,3-dimethyl-2-butanol	391.55	389.92	0.42
45	Heptanol	449.55	450.56	0.22
46	4-methyl-1-hexanol	446.15	440.64	1.24
47	5-methyl-1-hexanol	443.15	444.62	0.33
48	3-methyl-1-hexanol	442.15	441.19	0.22
49	2-methyl-1-hexanol*	437.15	443.16	1.37
50	2-heptanol	432.15	430.66	0.34
51	2-methyl-2-hexanol	416.15	417.20	0.25
52	5-methyl-2-hexanol	424.15	417.58	1.55
53	2,4-dimethyl-1-pentanol**	432.15	428.10	0.94
54	2,3-dimethyl-2-pentanol	412.85	413.94	0.26
55	2,4-dimethyl-2-pentanol	406.25	404.97	0.32
56	2,3,3-trimethyl-2-butanol	404.15	396.67	1.85
57	3-heptanol**	430.15	433.91	0.87
58	3-methyl-3-hexanol	416.15	418.72	0.62
59	5-methyl-3-hexanol	421.15	426.31	1.23
60	2-methyl-3-hexanol*	416.15	417.85	0.41
61	3-ethyl-3-pentanol	415.15	421.89	1.62
62	2,2-dimethyl-3-pentanol	408.15	412.11	0.97
63	2,3-dimethyl-3-pentanol	412.85	409.44	0.83
64	2,4-dimethyl-3-pentanol*	411.85	415.13	0.80
65	Octanol	468.45	470.31	0.40
66	4-methyl-1-heptanol	461.15	455.69	1.18
67	6-methyl-1-heptanol	461.75	460.45	0.28
68	2,5-dimethyl-1-hexanol	452.65	450.38	0.50
69	4-heptanol**	429.15	430.04	0.21
70	2-methyl-2-heptanol	429.15	439.33	2.37

No.	Compounds	Exp. NBP	Pred. NBP	A%D
71	3-methyl-2-heptanol	439.25	442.49	0.74
72	5-methyl-2-heptanol	443.15	446.88	0.84
73	2,4,4-trimethyl-1-pentanol*	441.65	443.21	0.35
74	2,3-dimethyl-2-hexanol	433.15	434.60	0.33
75	2,5-dimethyl-2-hexanol**	427.65	431.01	0.79
76	2-methyl-3-ethyl-2-pentanol	429.15	439.40	2.39
77	2,3,4-trimethyl-2-pentanol	420.65	419.58	0.25
78	2-octanol**	453.15	449.87	0.72
79	3-octanol	444.15	448.89	1.07
80	3,4-dimethyl-2-hexanol*	438.65	430.42	1.88
81	2-methyl-3-heptanol	440.65	435.63	1.14
82	3-methyl-3-heptanol	436.15	437.85	0.39
83	4-methyl-3-heptanol	443.15	445.37	0.50
84	5-methyl-3-heptanol*	445.15	439.99	1.16
85	6-methyl-3-heptanol**	447.15	439.60	1.69
86	2-methyl-3-ethyl-3-pentanol	433.15	431.62	0.35
87	2,2,4-trimethyl-3-pentanol	423.65	418.49	1.22
88	2,3,4-trimethyl-3-pentanol	429.65	416.38	3.09
89	4-octanol	449.45	446.52	0.65
90	2-methyl-4-heptanol	437.15	443.20	1.38
91	3-methyl-4-heptanol	435.15	442.10	1.60
92	4-methyl-4-heptanol**	434.15	439.34	1.20
93	2,2-diethyl-1-pentanol	465.15	453.99	2.40
94	3,5,5-trimethyl-1-hexanol	466.15	459.93	1.33
95	Nonanol*	486.45	484.90	0.32
96	2-nonanol	466.65	467.91	0.27
97	2-methyl-2-octanol	451.15	455.46	0.96
98	2,6-dimethyl-2-heptanol	446.15	450.87	1.06
99	3-nonanol**	468.15	465.19	0.63
100	2,2,3-trimethyl-3-hexanol	429.15	443.09	3.25
101	4-nonanol	465.65	462.97	0.58
102	5-nonanol*	466.15	459.47	1.43
103	2,6-dimethyl-4-heptanol	447.65	450.94	0.73
104	3,5-dimethyl-4-heptanol	460.15	456.56	0.78
105	7-methyl-1-octanol	479.15	476.32	0.59
106	4-ethyl-4-heptanol**	455.15	452.36	0.61
107	3,6-dimethyl-3-heptanol	446.15	451.87	1.28
108	2,6-dimethyl-3-heptanol*	448.15	451.22	0.69
109	Decanol	504.25	496.82	1.47
110	3,7-dimethyl-1-octanol	485.65	484.18	0.30
111	2-decanol	484.15	478.95	1.07
112	3-ethyl-3-octanol*	472.15	468.50	0.77
113	3-ethyl-2-methyl-3-heptanol	466.15	469.73	0.77
114	4-decanol	483.65	479.26	0.91
115	2,6-dimethyl-4-octanol	468.15	470.90	0.59
116	2,7-dimethyl-3-octanol	466.65	466.36	0.06
117	3,6-dimethyl-3-octanol**	475.35	472.75	0.55
118	Undecanol	518.15	525.37	1.39
119	2-ethyl-1-hexanol	457.15	451.49	1.24
120	2,2-dimethyl-1-propanol	386.65	386.75	0.03
121	2,6,8-trimethyl-4-nonanol	498.35	498.23	0.02
122	triphenylmethanol	653.15	653.19	0.01
123	benzyl alcohol	478.45	483.06	0.96
124	1-phenyl-1-propanol	492.15	496.16	0.81
125	2-phenyl-2-propanol	475.15	479.72	0.96
126	2,5-dimethylcyclohexanol	452.15	457.23	1.12
127	2,6-dimethylcyclohexanol	445.15	446.02	0.20
128	3,4-dimethylcyclohexanol	462.15	457.72	0.96

No.	Compounds	Exp. NBP	Pred. NBP	A%D
129	4,4-dimethylcyclohexanol	459.15	453.98	1.13
130	cyclohexanemethanol	455.15	457.59	0.54
131	2-ethyl-1,3-hexanediol	517.15	504.82	2.38
132	2,5-dimethyl-2,5-hexanediol	487.15	493.51	1.31
133	1,3-butanediol	480.65	475.11	1.15
134	1,4-butanediol	501.15	495.33	1.16
135	2,3-butanediol	455.65	457.60	0.43
136	1,6-hexanediol	523.15	526.38	0.62
137	2-methyl-1,3-pentanediol*	493.45	493.60	0.03
138	2-methyl-2,4-pentanediol**	471.15	459.94	2.38
139	1,3-propanediol	487.15	475.51	2.39
140	1,2-pentanediol	484.95	490.40	1.12
141	1,5-pentanediol	512.55	514.49	0.38
142	pinacol	447.55	457.19	2.15
143	1,7-heptanediol	535.15	536.83	0.31
144	2-ethylphenol	477.65	476.40	0.26
145	3-ethylphenol	487.15	490.15	0.62
146	4-ethylphenol	492.15	490.86	0.26
147	2,3-dimethylphenol	491.15	480.32	2.21
148	2,4-dimethylphenol*	483.15	480.89	0.47
149	3,4-dimethylphenol	498.15	490.62	1.51
150	2,5-dimethylphenol**	485.15	489.46	0.89
151	2,6-dimethylphenol	476.15	471.99	0.87
152	3,5-dimethylphenol	492.65	497.04	0.89
153	2,6-dibutylphenol	526.15	527.26	0.21
154	2-butyl-4-methylphenol	510.15	508.00	0.42
155	2-tert-butylphenol	497.15	497.48	0.07
156	3-tert-butylphenol	513.15	509.70	0.67
157	4-tert-butylphenol	509.65	512.85	0.63
158	2,3,5-trimethylphenol	506.15	500.29	1.16
159	2,4,6-trimethylphenol	494.15	480.49	2.76
160	2-methylphenol	463.95	470.33	1.38
161	3-methylphenol	475.95	482.86	1.45
162	4-methylphenol	475.15	478.45	0.69
163	5-isopropyl-2-methylphenol	511.15	519.42	1.62
164	3-propylphenol	501.15	499.61	0.31
165	4-hexyl-1,3-benzenediol	607.15	603.64	0.58
166	2-methyl-1,3-dihydroxybenzene	537.15	536.52	0.12
167	4-tert-butylcatechol	558.15	567.29	1.64
168	2-hydroxydiphenylmethane**	585.15	590.33	0.89

$$A\%D = \frac{|pred(i) - exp(i)|}{exp(i)} \times 100, *: \text{compounds in the validation set,}$$

***: compounds in the testing set, the other compounds belong to the training set.

4. Conclusion

In this paper, a radial basis network was developed to predict the normal boiling point of hydroxyl compound. The structure features of all compounds in the dataset were numerically characterized by a huge number of molecular descriptors. These descriptors obtained were analyzed carefully and finally eight important descriptors remained by a series of methods. The results revealed that 8 molecular descriptors can be used to construct the RBN model for the prediction of the boiling points. The three-layer radial basis network model can be represented by $[8-35_{[64]}-I]$. With the

same 8 descriptors, a multiple linear regression approach was also applied and was compared with the RBN model. The results showed that the RBN model could provide more accurate values of the predicted boiling point. In summary, the results of the current study indicated that radial basis networks would be a promising strategy for QSPR modeling.

References

- [1] S. Gamba, G. S. Soave, L. A. Pellegrini, "Use of normal boiling point correlations for predicting critical parameters of paraffins for vapour-liquid equilibrium calculations with the SRK equation of state," *Fluid Phase Equil.*, 2009, vol. 276, pp. 133-141.
- [2] E. Panteli, E. Voutsas, K. Magoulas, et al., "Prediction of vapor pressures and enthalpies of vaporization of organic compounds from the normal boiling point temperature," *Fluid Phase Equil.*, 2006, vol. 248, pp. 70-77.
- [3] J. Marrero, and R. Gani, "Group-contribution based estimation of pure component properties," *Fluid Phase Equil.*, 2001, vol. 183/184, pp. 183-208.
- [4] J. Guilherme, J. Maximo, J. A. Antonio, et al., "Boiling point of aqueous d-glucose and d-fructose solutions: Experimental determination and modeling with group-contribution method," *Fluid Phase Equil.*, 2010, vol. 299, pp. 32-41.
- [5] D. Sola, A. Ferri, M. Banchero, et al., "QSPR prediction of N-boiling point and critical properties of organic compounds and comparison with a group-contribution method," *Fluid Phase Equil.*, 2008, vol. 263, pp. 33-42.
- [6] I. Oprisiu, G. Marcou, D. Horvath, S., et al., "Publicly available models to predict normal boiling point of organic compounds," *Thermochim. Acta*, 2013, vol. 553, pp. 60-67.
- [7] Y. M. Dai, Z. P. Zhu, Z. Cao, et al., "Prediction of boiling points of organic compounds by QSPR tools," *J. Mol. Graph. Model.*, 2013, vol. 44, pp. 113-119.
- [8] D. Aboali, M. A. Sobati, "Novel method for prediction of normal boiling point and enthalpy of vaporization at normal boiling point of pure refrigerants: A QSPR approach," *Int. J. Refrig.*, 2014, vol. 40, pp. 282-293.
- [9] V. Zare-Shahabadi, M. Lotfizadeh, A. R. A. Gandomani, "Determination of boiling points of azeotropic mixtures using quantitative structure-property relationship (QSPR) strategy," *J. Mol. Liq.*, 2013, vol. 188, pp. 222-229.
- [10] Q. F. Li, X. G. Chen, Z. D. Hu, "Quantitative structure-property relationship studies for estimating boiling points of alcohols using calculated molecular descriptors with radial basis function neural networks," *Chemometr. Intell. Lab.*, 2004, vol. 72, pp. 93-100.
- [11] F. Gharagheizi, S. A. Mirkhani, P. Ilani-Kashkouli, et al., "Determination of the normal boiling point of chemical compounds using a quantitative structure-property relationship strategy: Application to a very large dataset," *Fluid Phase Equil.*, 2013, vol. 354, pp. 250-258.
- [12] G. Q. Liu, L. X. Ma, S. G. Xiang, "Handbook of chemistry and chemical properties data", 4rd ed, Beijing: Chemical Industry Press, 2013.
- [13] F. Gharagheizi, and M. Sattari, "Prediction of triple point temperature of pure components using their chemical structure," *Ind. Eng. Chem. Res.*, 2010, vol. 49, pp. 929-932.
- [14] VCCLAB, <http://www.vcclab.org>, 2005
- [15] I. V. Tetko, J. Gasteiger, J., R. Todeschini, R. et al. "Virtual computational chemistry laboratory - design and description," *J. Comput. Aid. Mol. Des.*, 2005, vol. 19, pp. 453-63.
- [16] R. Todeschini, and V. Consonni, "Molecular descriptors for chemoinformatics," 2rd ed. Weinheim: Wiley-VCH, 2009.
- [17] N. S. Sapre, N. Pancholi, S. Gupta, "Computational modeling of substitution effect on HIV-1 non-nucleoside reverse transcriptase inhibitors with Kier-Hall electrotopological state (E-state) indice," *Int. Electron. J. Mol. Des.*, 2008, vol. 7, pp. 55-67.
- [18] V. Consonni, R. Todeschini, M. Pavan, "Structure/Response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors," *J. Chem. Inf. Comput. Sci.*, 2002, vol. 42, pp. 682-692.
- [19] V. Consonni, R. Todeschini, M. Pavan, et al., "Structure/Response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies," *J. Chem. Inf. Comput. Sci.*, 2002, vol. 42, pp. 693-705.
- [20] R. Todeschini, P. Gramatica, E. Marengor, et al., "Weighted holistic invariant molecular descriptors. Part 2. Theory development and applications on modeling physicochemical properties of polyaromatic hydrocarbons," *Intell. Lab. Syst.*, 1995, vol. 27, pp. 221-229.
- [21] O. Devinyak, D. Havrylyuk, R. Lesyk, "3D-MoRSE descriptors explained," *J. Mol. Graph. Model.*, 2014, vol. 54, pp. 194-203.