Available online at www.sciencedirect.com
ScienceDirect
Journal of Hazardous Materials 147 (2007) 424–430
Journal of Hazardous Materials
www.elsevier.com/locate/jhazmat

Quantitative structure–property relationship studies for predicting flash points of alkanes using group bond contribution method with back-propagation neural network

Yong Pan *, Juncheng Jiang, Zhirong Wang

Institute of Safety Engineering, Nanjing University of Technology, Nanjing 210009, China

Received 20 November 2006; received in revised form 6 January 2007; accepted 8 January 2007
Available online 12 January 2007

Abstract

Models of relationships between structure and flash point of 92 alkanes were constructed by means of artificial neural network (ANN) using group bond contribution method. Group bonds were used as molecular structure descriptors which contained information of both group property and group connectivity in molecules, and the back-propagation (BP) neural network was employed for fitting the possible nonlinear relationship existed between the structure and property. The dataset of 92 alkanes was randomly divided into a training set (62), a validation set (15) and a testing set (15). The optimal condition of the neural network was obtained by adjusting various parameters by trial-and-error. Simulated with the final optimum BP neural network [9-5-1], the results showed that the predicted flash points were in good agreement with the experimental data, with the average absolute deviation being 4.8 K, and the root mean square error (RMS) being 6.86, which were shown to be more accurate than those of the multilinear regression method. The model proposed can be used not only to reveal the quantitative relation between flash points and molecular structures of alkanes, but also to predict the flash points of alkanes for chemical engineering.

Keywords: Quantitative structure–property relationship (QSPR); Back-propagation (BP) neural network; Flash point; Group bond contribution method; Alkane

1. Introduction

The flash point of a flammable (or combustible) compound is the temperature at which the vapor pressure of the substance is such as to produce a concentration of vapor in the air that corresponds to the lower flammable limit [1]. This parameter provides the knowledge necessary for understanding the fundamental physical and chemical processes of combustion. Moreover, it is of importance in practice for safety considerations in storage, processing, and handling of a given compound, and is one of the major flammability characteristics used to assess the fire and explosion hazards of chemical compounds.

Reliable values of flash points are always desirable, and some of them can be measured by two currently-accepted experimental methods, which are the closed cup test and the open cup test [2]. However, for many other compounds, the experimental flash point values are scarce or too expensive to obtain. What is more, for toxic, volatile, explosive, and radioactive compounds, the experimental determination of flash point values is more difficult. Hence the development of estimation methods which are desirably convenient for predicting the flash points in short is required.

Quantitative structure–property relationships (QSPR) method which relates descriptors of the molecular structure to the properties of chemical compounds, has been reported quite extensively in the literature for the prediction of flash point [3–10]. For example, the first method for estimating the flash point of organic compounds from their molecular structure was developed by Suzuki et al. The 25 atomic and group contributions were employed for predicting the flash points of 33 aliphatic and 26 aromatic hydrocarbons with an average absolute deviation of 12.2 and 6.1 °C, respectively. The average deviation for the 59 hydrocarbon compounds tested was 9.5 °C. In another work, Tetteh et al. used a radial basis function neural network for the estimation of flash points for a large set of 400 compounds from different classes. The structures were described simply with a molecular connectivity index and counts of the 25 functional groups present in the molecules.

* Corresponding author. Tel.: +86 25 83587305; fax: +86 25 83587411.
E-mail address: yongpannjut@163.com (Y. Pan).

0304-3894/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhazmat.2007.01.025

The average absolute error for the test set in flash point prediction was 11.9 °C using a RBFNN with a 26-36-2 configuration. Katritzky et al. studied quantitative structure–flash point relationships for a diverse set of 271 compounds. The general three-parameter QSPR model provided $R^2 = 0.9020$ and $s = 16.1$ K. When the boiling point was used as a descriptor in the model, the correlation was improved to $R^2 = 0.9529$. Meanwhile, the study on mixture flash points [9,10] which can display non-ideal behavior with important safety consequences has also been developed. Liaw et al. proposed a mathematical model, which could be used for predicting the flash point of aqueous-organic solutions, and the results revealed that this model was able to precisely predict the flash point over the entire composition range of binary aqueous-organic solutions by way of utilizing the flash point data pertaining to the flammable component.

The group bond contribution method was recently proposed by Wang [11] for the description of molecules structure. This method combined together information of both group property and connectivity in the analyzed molecules, and has been successfully used in the estimation of physical and chemical properties, such as density [11] and boiling point [12].

In recent years, the modeling technique of artificial neural network (ANN) has been widely used in the field of QSPR [4,8,13,14]. ANN is a powerful tool for correlating and estimating chemical properties and one of a group of intelligence technologies for data analysis that differ from other classical analysis techniques. The advantage of ANN is in its inherent ability to incorporate nonlinearity and cross-product terms into the model. Besides, it is also able to acquire an estimate function from studied samples while the form of the mathematical function is unknown.

In this paper, we developed a method to estimate the flash points of 92 alkanes based on the back-propagation (BP)-ANN using group bond contribution method. Group bonds which obtained from chemical structure are used as molecular structure descriptors, and these descriptors are quantitatively related to flash points of 92 alkanes by BP neural network.

2. Group bond contribution method

One of the most widely used methods proposed for prediction of properties from molecular structure is group contribution method. It is based on the assumption that the contribution of a certain group is completely the same in different molecules, and the properties of compounds are considered as the contribution addition of groups which constituted the compounds. The group contribution method works well for a large number of compounds, however, difficulties may arise in decomposing some structures into appropriate groups whose constants are available. Several correction factors are also needed for some molecular interactions, for group contribution method takes into account only the contribution of groups in the molecule but the interaction between groups and chemical bonds. Besides, group contribution method has a weak ability for distinguishing the isomeric compounds. For instance, the structure difference of 2-methylhexane, 3-methyhexane and 3-ethylpentane cannot be distinguished from only molecular groups, because there are 3 "–CH_3"group, 3 "–CH_2–" group and 1 "≺CH–" group in each of the three compounds above, while the flash point values of them were 269, 258 and 255 K, respectively [15]. Because of such causations, group contribution method has some limitations in QSPR studies for property calculation.

Table 1
Group bonds presented in the alkane molecules

No.	Group bond	No.	Group bond	No.	Group bond
1	CH_3–CH_2–	4	–CH_2–CH_2–	7	≻CH–CH≺
2	CH_3–CH≺	5	–CH_2–CH≺	8	≻CH—C—
3	CH_3—C—	6	—CH_2—C—	9	—C—C—

However, the group bond contribution method can cover these shortages above. Group bond which is defined as an integration of two molecular groups and the chemical bond between them, contains information of both group and chemical bond. Strict quantitative relationships among the number of group bond, number of group and number of chemical bond are existed, and the number of group and chemical bond can be confirmed from the number of group bond in molecule, but the species and number of group bond cannot be ascertained from group and chemical bond. Thus the group bond contribution method contains both group contribution and bond contribution, and could have better and more comprehensive prediction abilities than the group contribution method to a certain extent. Besides, the group bond contribution method takes into account both group property and connectivity in the analyzed molecules, so it may have a great ability for the identification of isomeric compounds.

In this work, the flash point values of alkanes were regarded as the concerted contribution of numerous group bonds constituting the alkane molecules. There were four species of groups (–CH_3, ≻CH–, –CH_2–, ≻C≺) existed in the molecular structure of alkanes, which constituted a set of nine species of group bonds presented in all the alkane molecules except ethane. All the nine group bonds were listed in Table 1.

Furthermore, because of the possibility that highly nonlinear interaction may exist among the group bonds in the chemical molecules, the contribution value of each group bond obtained by mathematical regression method could not satisfactorily show the difference of contributions to property for one certain group bond in different molecules. However, ANN could describe such nonlinear interaction between the group bonds satisfactorily for its inherent ability of nonlinear fitting. So in this work we combined together both group bond contribution method and ANN for the flash point prediction of alkanes.

3. Experiment

3.1. Data sets

The applicability and accuracy of a flash point estimation model are directly affected by the size and quality of the training

set. For the fact that flash points are experimentally determined data, the experimental flash point values reported by different authors as well as organizations, can differ by as much as 30 K. For example, Ref. [16] supplies experimental values 355 K for 1-decanol, while International Chemical Safety Cards (ICSCs) [17] gives 381 K. Most organizations assess the reliability of the reported experimental and predicted values and also verify if such values are obtained in the most similar conditions using as much as possible the same methods. Thus, in order to have homogeneous training and test sets, all the flash points of alkanes used in this work were taken from the chemical database of the department of chemistry at the University of Akron (USA) [15], except 2,7-dimethyloctane. The flash point value of 2,7-dimethyloctane acquired from the chemical database referred above differs much from most of other sources, which demonstrates that the value of 354 K is more than probably erroneous. Thus here we adopted the mostly used value of 314 K in Ref. [16] instead of 354 K in Ref. [15] for 2, 7-dimethyloctane. From the set of 92 alkanes with the number of carbon atoms from 3 to 16, 62 alkanes were randomly chosen for the training set, 15 alkanes used for validation set, and the rest 15 for testing. Flash point values of these compounds were in the range from −104 to 135 °C.

3.2. Back-propagation neural networks

A three-layer feed-forward neural network utilizing the back-propagation algorithm was used to model the flash points. The typical back-propagation network consists of an input layer, an output layer and at least one hidden layer. Each layer contains neurons and each neuron is a simple micro-processing unit which receives and combines signals from many neurons. The number of neurons presented in the input and output layer depends on the number of variables (in this work molecular group bonds and flash point, respectively). Besides, the number of neurons used for the hidden layer is optimized by trial-and-error training assays.

Each neuron has weighted inputs, summation function, transfer function and output. The behavior of a back-propagation network is mainly determined by the transfer functions of its neurons. At first, summation function is computed from the weighted sum of all input neurons entering each hidden neuron and the weighted sum of the inputs constitutes the activation of the neuron. Then the activation signal is passed on to the transfer function for further processing. The role of the transfer function is to translate the summed information into outputs. In this work, a logistic $f(x) = 1/[1 + \exp(-x)]$ transfer function was applied both for hidden and output neurons.

For a given input and a desired output, the back-propagation neural network system can be trained by the following steps:

(1) The input vector is presented to the input layer of the network, and then propagates through the hidden layer to the output layer, where all of the summed inputs and output states for each processing element in the network are set and an output value is produced.

$$\mathrm{net}_j^h = \sum_i \omega_{ij} O_i \tag{1}$$

$$O_j^h = f(\mathrm{net}_j^h) = \frac{1}{1 + \exp(-\mathrm{net}_j^h)} \tag{2}$$

$$\mathrm{net}_k^o = \sum_j \omega_{jk} O_j^h \tag{3}$$

$$O_k^o = f(\mathrm{net}_k^o) = \frac{1}{1 + \exp(-\mathrm{net}_k^o)} \tag{4}$$

where f is the sigmoid transfer function, and w_{ij}/w_{jk} are the connection weights between hidden units and input/output units.

(2) The actual output value O_k^o is compared with the desired output value T_k, and the error E_p and the global error E are determined, respectively.

$$E_p = \frac{1}{2}\sum_k (T_{pk} - O_{pk}^o)^2 \tag{5}$$

$$E = \frac{1}{2p}\sum_p \sum_k (T_{pk} - O_{pk}^o)^2 \tag{6}$$

(3) The weights are modified to reduce the error associated with the overall error function. In this work, gradient descent method is carried out for the reduction of E. The gradient descent method is an iterative least squares procedure which tries to adjust the connection weight for reducing the error most rapidly, by moving the state of the system downward towards the direction of maximum gradient.

$$\omega_{jk}(n+1) = \omega_{jk}(n) + \eta\delta_{pk} O_{pk} \tag{7}$$

$$\omega_{ij}(n+1) = \omega_{ij}(n) + \eta\delta_{pj} O_{pj} \tag{8}$$

$$\delta_{pk} = \frac{\partial E_p}{\partial \mathrm{net}_k^o} = O_{pk}^o(1 - O_{pk}^o)(T_{pk} - O_{pk}^o) \tag{9}$$

$$\delta_{pj} = \frac{\partial E_p}{\partial \mathrm{net}_j^h} = O_{pj}^h(1 - O_{pj}^h)\sum \delta_{pk}\omega_{jk} \tag{10}$$

(4) For each hidden layer, the training process starts at the layer below the output layer, and ends with the layer above the input layer. And for each processing element in the hidden layer, the global error E is calculated and propagated back through the networks. Furthermore, the delta weights are calculated again.

(5) Finally, in order to reduce the error, all of the weights in the networks are updated by adding the delta weights to the corresponding previous weights. And the training process of the ANN will be completed when the global error E is minimized.

Before the beginning of the training process, the optimal condition of the neural network was obtained by adjusting various parameters by trial-and-error. These parameters include:

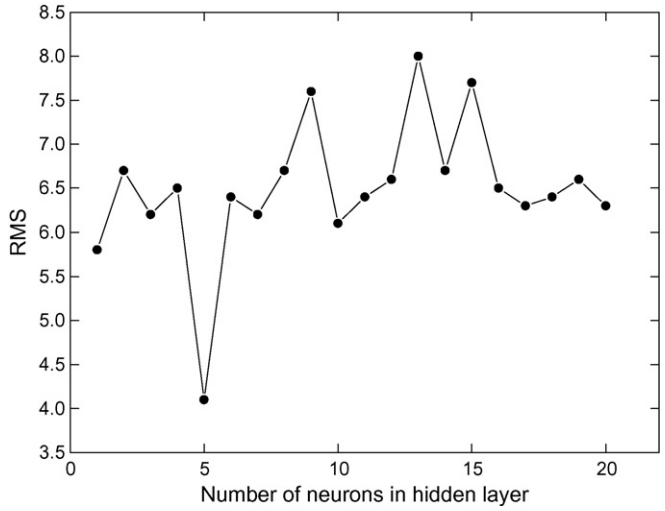

Fig. 1. RMS as a function of the number of neurons in the hidden layer for the testing set.

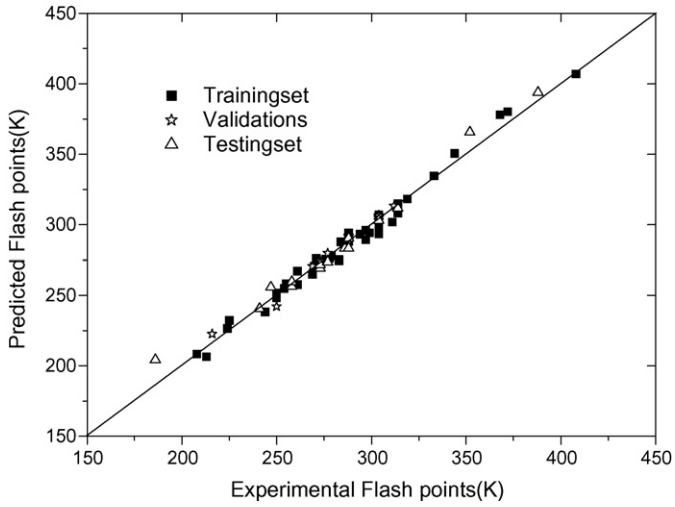

Fig. 2. Correlation between the predicted and experimental flash points for the training, validation and the testing sets.

the learning rate, the momentum constants, the number of neurons in the hidden layer, and the training endpoint. The learning rate determines the speed at which the weights change, and the momentum constant prevents sudden changes in attaining the results. In this work, we empirically set the learning rate and momentum at 0.1 and 0.9, respectively.

The optimal number of neurons in the hidden layer was determined by varying the number of hidden neurons and observing the root mean square error (RMS) [8], which was used as a measure of the prediction error of the trained model and was calculated with the following equation:

$$\mathrm{RMS} = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{n}} \tag{11}$$

where n is the number of compounds in the dataset, and p_i is the predicted output, a_i is the actual output, respectively. Calculations of RMS were performed with leave-one-out cross-validation and the average RMS of 10 runs was adopted. Leave-one-out cross-validation referred to removing one sample in the dataset using for the test set while the rest using as training set. Such process was repeated until all samples of the dataset were used as the test sample. Finally, the number of neurons that gave the lowest RMS was chosen. As can be seen from the plot of RMS versus hidden neurons (Fig. 1), the optimal number of neurons in the hidden layer was 5.

The early stopping technique was used extensively in the current study for avoidance of overfitting [18]. For the determination of optimal training endpoint, a validation set contained 15 compounds was used to monitor the training process as measured by RMS. Thus, the training endpoint giving the lowest RMS for the predictions of the validation set was used.

4. Results and discussion

All input descriptors and output values of all 92 samples were pre-treated to scale the value to between ±1 and yet retain the original proportionality before submission to the network for training or prediction. The programs required to generate the back-propagation networks were written in MATLAB M-file and the programs were executed on a Pentium PC with 512M RAM and CPU speed of 2.4 G. With the optimum network architecture represented by [9-5-1], the predictions were repeated 10 times with different random starting weights between neurons, which were given random values between −0.5 and 0.5, and the averaged flash point values were calculated. The predicted results of training set, validation set and testing set were shown in Table 2. The average absolute deviation of training, validation and testing sets were 3.8, 2.6 and 4.8 K, respectively. The RMS were 4.95, 3.35 and 6.86, and the correlation R were 0.9912, 0.9922 and 0.9902, respectively.

Also, the experimental and predicted flash points of the training, validation and testing sets were plotted in Fig. 2. Regression lines were used for comparing the values obtained by this model with experimental values. As can be seen from the figures, the calculated slope and intercept did not differ greatly from the "ideal" values of 1 and 0, respectively, and the predicted values of flash points agreed with the experimental values satisfactorily for all the training, validation and testing sets.

A multiple linear regression (MLR) method was also employed to describe the relation between flash points (FP) and molecular descriptors. By using these 9 descriptors selected and the same 77 training and validation samples used above, the best MLR model can be obtained as follows (x_1–x_9 referred to the group bond 1–9 in Table 1):

$$\begin{aligned}\mathrm{FP} = {} & 175.234 + 9.180x_1 + 11.232x_2 + 9.796x_3 + 17.194x_4 \\ & + 18.802x_5 + 20.685x_6 + 20.660x_7 + 24.413x_8 \\ & + 29.851x_9, \quad R = 0.986, s = 6.27, n = 77\end{aligned} \tag{12}$$

With this MLR model, the flash points of 15 alkanes in the testing set were calculated, and the results were listed in Table 3. As can be seen from the table, the prediction results obtained by ANN method and MLR method using group bonds

Table 2
Experimental and predicted flash point for 92 alkanes

No.	Compound	Experimental FP (K)	Predicted FP (K)	Deviation (K)	No.	Compound	Experimental FP (K)	Predicted FP (K)	Deviation (K)
1	Propane	169	186.2	17.2	47	3-Ethyl-2,2-dimethylhexane	311	301.8	−9.18
2	Butane	213	206.5	−6.5	48	3,3,4,4-Tetramethylhexane	304	304.3	0.3
3	Pentane	224	226.5	2.5	49	2,2,5,5-Tetramethylhexane	304	293.3	−10.7
4	2,2-Dimethylpropane	208	208.3	0.3	50	2,3,3,4-Tetramethylhexane	304	303.6	−0.4
5	Hexane	250	248.4	−1.6	51	2,3,4,5-Tetramethylhexane	304	302.8	−1.2
6	2,2-Dimethylbutane	225	232.2	7.2	52	2,2,4,4-Tetramethylhexane	304	297.9	−6.1
7	2,3-Dimethylbutane	244	238.3	−5.7	53	3,3,5-Trimethylheptane	304	304.7	0.7
8	Heptane	269	264.9	−4.1	54	2,3,5-Trimethylheptane	304	306.8	2.8
9	3,3-Dimethylpentane	254	255.0	1.0	55	3-Ethyl-3-methylheptane	314	310.9	−3.1
10	2,4-Dimethylpentane	261	257.7	−3.3	56	4-Ethyl-3-methylheptane	314	308.0	−6.0
11	3-Ethylpentane	255	258.1	3.1	57	2-Methylnonane	314	314.9	0.9
12	2,2-Dimethylpentane	250	251.7	1.7	58	Undecane	333	334.6	1.6
13	2,2,4-Trimethylpentane	261	267.2	6.2	59	Dodecane	344	350.5	6.5
14	2,2,3-Trimethylpentane	270	270.8	0.8	60	Tetradecane	372	380.1	8.1
15	3-Methyl-3-ethylpentane	276	275.5	−0.5	61	Hexadecane	408	407.0	−1.0
16	3-Ethylhexane	278	276.5	−1.5	62	2,2,4,4,6,8,8-Heptamethylnonane	368	377.9	9.9
17	3-Methylheptane	279	278.1	−0.9	63	2-Methylbutane	216	222.6	6.6
18	3,3-Dimethylhexane	272	273.8	1.8	64	2-Methylpentane	250	242.2	−7.8
19	2,3-Dimethylhexane	283	275.4	−7.6	65	2,3,3-Trimethylpentane	273	274.6	1.6
20	2,4-Dimethylhexane	283	274.5	−8.5	66	2-Methylheptane	277	279.7	2.7
21	2,5-Dimethylhexane	271	276.2	5.2	67	2,2-Dimethylhexane	269	270.7	1.7
22	Nonane	304	301.0	−3.0	68	3-Ethyl-2,2-dimethylpentane	286	285.3	−0.7
23	3,4-Dimethylheptane	288	292.1	4.1	69	2,4,4-Trimethylhexane	288	288.7	0.7
24	2,3-Dimethylheptane	288	293.9	5.9	70	4,4-Dimethylheptane	288	292.1	4.1
25	2,6-Dimethylheptane	299	294.2	−4.8	71	2,3,3-Trimethylhexane	288	291.5	3.5
26	2,3-Dimethyl-3-ethylpentane	288	291.3	3.3	72	2,4,6-Trimethylheptane	304	306.5	2.5
27	2,2,5-Trimethylhexane	286	285.7	−0.3	73	3-Ethyl-2,3,4-trimethylpentane	304	303.0	−1.0
28	3-Methyloctane	297	295.9	−1.1	74	2,3,4,4-Tetramethylhexane	304	306.5	2.5
29	2,2,3,4-Tetramethylpentane	284	287.7	3.7	75	3,4,5-Trimethylheptane	304	304.8	0.8
30	2,2,4,4-Tetramethylpentane	276	274.9	−1.1	76	3-Ethyl-5-methylheptane	304	306.2	2.2
31	4-Ethylheptane	288	294.2	6.2	77	5-Methylnonane	312	313.2	1.2
32	2,2-Dimethylheptane	297	289.2	−7.8	78	2-Methylpropane	186	204.3	18.3
33	2,4-Dimethylheptane	288	292.5	4.5	79	3-Methylpentane	241	240.6	−0.4
34	2,3,4-Trimethylhexane	288	288.7	0.7	80	2,3-Dimethylpentane	258	256.3	−1.7
35	3,3,4-Trimethylhexane	288	289.9	1.9	81	3-Methylhexane	258	259.6	1.6
36	2,3,5-Trimethylhexane	288	290.6	2.6	82	2,2,3-Trimethylbutane	247	255.7	8.7
37	2,2,3-Trimethylhexane	288	287.8	−0.2	83	Octane	286	283.2	−2.8
38	3,5-Dimethylheptane	288	290.7	2.7	84	2,2,3,3-Tetramethylbutane	273	269.2	−13.8
39	Tetraethylmethane	294	293.2	−0.8	85	2,3,4-Trimethylpentane	273	271.8	−1.2
40	Decane	319	318.2	−0.8	86	3,4-Dimethylhexane	277	273.6	−3.4
41	4-Ethyloctane	314	311.4	−2.6	87	3-Ethyl-4-methylhexane	288	290.3	2.3
42	2,4,5-Trimethylheptane	304	306.8	2.8	88	2,2,4-Trimethylhexane	288	283.3	−4.7
43	2,3-Dimethyloctane	314	311.8	−2.2	89	2,7-Dimethyloctane	314	311.6	−2.4
44	3,3-Dimethyloctane	314	309.8	−4.2	90	2,2,3,4-Tetramethylhexane	304	302.9	−1.1
45	3,5-Dimethyloctane	314	308.0	−6.0	91	Tridecane	352	365.6	13.6
46	2,6-Dimethyloctane	314	309.8	−4.2	92	Pentadecane	388	393.9	5.9

The substances from 1 to 62 composed the training sample, those from 63 to 77 were the validation sample, and those from 78 to 92 were the testing sample.

as structure descriptors, were both in good agreement with the experimental values with a average absolute deviation of 4.8 and 6.1 K, respectively, which indicated a success in flash point prediction by using group bond contribution method. Moreover, the predicted results obtained by ANN method were better than those obtained by MLR, which indicated a superior prediction ability of the ANN model and strongly suggested a nonlinear relationship existing between the group bonds and flash points of alkanes.

The results obtained by ANN and MLR were also compared with the study of Albahri [6] as well as that of Vazhev et al. [7]. Albahri used structural group contribution method (SGCM) for the prediction of flash points of hydrocarbons including alkanes, and Vazhev et al. applied transformed infrared spectra as descriptors of molecular structure to predict flash points of 85 alkanes. As can be seen from Table 3, for the same 15 test samples, the average absolute deviation of SGCM and infrared spectra method (IRSM) were 5.6 and 3.6 K, and the RMS were 7.90 and 4.15, respectively. Clearly, compared with the three other methods, the SGCM can provide the most satisfactory prediction ability here, followed by the BP-ANN based QSPR model. However, as showed in Table 3, the flash point

Table 3
Comparison of predicted and experimental flash points for the 15 alkanes in the test set

No.	Compound	Experimental flash point (K)	Predicted flash point (K)			
			ANN	MLR	SGCM [6]	IRSM [7]
1	2-Methylpropane	186	204.3	208.9	192.1	166.1
2	3-Methylpentane	241	240.6	242.4	240.0	240.8
3	2,3-Dimethylpentane	258	256.3	257.6	256.7	265.0
4	3-Methylhexane	258	259.6	259.6	260.6	264.5
5	2,2,3-Trimethylbutane	247	255.7	251.5	249.3	249.5
6	Octane	286	283.2	279.6	282.5	282.4
7	2,2,3,3-Tetramethylbutane	273	269.2	263.9	265.8	273.3
8	2,3,4-Trimethylpentane	273	271.8	272.7	270.7	274.2
9	3,4-Dimethylhexane	277	273.6	274.3	274.4	286.4
10	3-Ethyl-4-methylhexane	288	290.3	291.1	292.3	273.0
11	2,2,4-Trimethylhexane	288	283.3	283.3	285.3	287.2
12	2,7-Dimethyloctane	314	311.6	309.3	–	304.3
13	2,2,3,4-Tetramethylhexane	304	302.9	300.1	300.0	302.4
14	Tridecane	352	365.6	365.5	360.2	354.6
15	Pentadecane	388	393.9	399.9	386.0	384.8
The average absolute deviation (K)			4.8	6.1	3.6	5.6
RMS			6.86	8.46	4.15	7.90

value of "2,7-dimethyloctane" cannot be obtained using SGCM proposed by Albahri. In fact, for many other alkanes like "2,7-dimethyloctane", the flash point values cannot be calculated by the aforementioned SGCM, too. The reason is that the group contribution value of each group is based on its location in the molecule, and the groups in the different positions along the HC chain have the different group contribution values. However, in the original literature, the author has only given the group contribution values of group "$>$CH–" in the second, third, fourth, and fifth positions along the HC chain, as well as the values of group "$>$C$<$ " in the second, and third positions, which are not sufficient for the flash point calculation of alkanes with long HC chains, such as "2,7-dimethyloctane" with a "$>$CH–" group in the seventh position, and "2,2,4,4,6,8,8-heptamethylnonane" with a "$>$CH–" group in the sixth position and two "$>$C$<$" groups in the fourth position and the eighth position, respectively. So the proposed group contribution method has limitations in the flash point calculation of alkanes, while the method proposed in this work can be applied to any alkane (only except ethane).

As also can be seen from Table 3, the predicted flash point values for 2,7-dimethyloctane obtained by ANN, MLR and IRSM method were 311.6, 309.3 and 304.3 K, respectively, all of which were close to the value of 314 K used in this work. The fact above demonstrated that the value of 354 K in Ref. [15] was more than probably erroneous for 2,7-dimethyloctane.

A Monte Carlo experiment has also been employed to test the results obtained by BP-ANN method for chance effects. When the dependent variables were scrambled, the testing models provided high RMS errors, which were 143.55, 155.13 and 197.54 for the training, validation and testing sets, respectively. Such errors were hundred times the errors obtained when the dependent variables were not scrambled, which indicated that the results obtained by BP-ANN method here were not due to chance.

5. Conclusion

In this study, a BP-ANN based QSPR model was developed for the prediction of flash points of alkanes using group bond contribution method. The group bonds were used as molecular structure descriptors which required no calculation, and the BP-ANN model was employed for fitting the possible nonlinear relationship existed between the structure and property. The results showed that the predicted values of flash points agreed with the experimental values satisfactorily which can sometimes approach the accuracy of experimental flash point determination. Thus BP-ANN based QSPR model using group bond contribution method can be successfully used to predict the flash points of alkanes and can also enable initial estimation of flash points for new alkane compounds or for other alkanes for which experimental values are unknown. Furthermore, this work is of assistance to the further study on other flammability characteristics, such as auto ignition temperature and flammability limits.

Acknowledgement

This research is supported by National Natural Science Fund of China (No. 29936110).

References

[1] F.P. Lees, Loss Prevention in the Process Industries, vol. 1, 2nd ed., Butterworth-Heinemann, Oxford, 1996.
[2] W.J. Lyman, W.F. Reehl, D.H. Rosenblatt, Handbook of Chemical Property Estimation Methods, McGraw-Hill, New York, 1982.
[3] T. Suzuki, K. Ohtaguchi, K. Koide, A method for estimating flash points of organic compounds from molecular structures, J. Chem. Eng. Jpn. 24 (1991) 258–261 (in Japanese).
[4] J. Tetteh, T. Suzuki, E. Metcalfe, S. Howells, Quantitative structure–property relationships for the estimation of boiling point and flash point

using a radial basis function neural network, J. Chem. Inf. Comput. Sci. 39 (1999) 491–507.
[5] A.R. Katritzky, R. Petrukhin, R. Jain, M. Karelson, QSPR analysis of flash points, J. Chem. Inf. Comput. Sci. 41 (2001) 1521–1530.
[6] T.A. Albahri, Flammability characteristics of pure hydrocarbons, Chem. Eng. Sci. 58 (2003) 3629–3641.
[7] V.V. Vazhev, M.K. Aldabergenov, N.V. Vazheva, Estimation of flash points and molecular masses of alkanes from their IR spectra, Petrol. Chem. 46 (2006) 136–139.
[8] Y. Pan, J.C. Jiang, Quantitative structure–property relationship study for estimating flash points of organic compounds using back-propagation artificial neural networks, in: P. Huang, Y.J. Wang, S.C. Li, C.X. Zheng, Z.H. Mao (Eds.), Progress in Safety Science and Technology, vol. VI, Science Press, Beijing, 2006, pp. 1462–1465.
[9] H.J. Liaw, Y.Y. Chiu, The prediction of the flash point for binary aqueous-organic solutions, J. Hazard. Mater. 101 (2003) 83–106.
[10] H.J. Liaw, Y.Y. Chiu, A general model for predicting the flash point of miscible mixtures, J. Hazard. Mater. 137 (2006) 38–46.
[11] K.Q. Wang, A new method for predicting the densities of alkanes from the information of molecular structure-group bond contribution method, Chin. J. Org. Chem. 19 (1999) 304–308 (in Chinese).
[12] K.Q. Wang, J. Wang, A new group contribution method for calculating the boiling point of alkane, Chin. J. Org. Chem. 21 (2001) 751–754 (in Chinese).
[13] A.P. Bünz, B. Braun, R. Janowsky, Quantitative structure–property relationships and neural networks: correlation and prediction of physical properties of pure components and mixtures from molecular structure, Fluid Phase Equilibr. 158–160 (1999) 367–374.
[14] J. Taskinen, J. Yliruusi, Prediction of physicochemical properties based on neural network modeling, Adv. Drug Deliv. Rev. 55 (2003) 1163–1183.
[15] http://ull.chemistry.uakron.edu/erd/index.html.
[16] J.A. Dean, Lange's Handbook of Chemistry, 15th ed., McGraw-Hill, New York, 1999.
[17] http://www.inchem.org/pages/icsc.html.
[18] I.V. Tetko, D.J. Livingstone, A.I. Luik, Comparison of overfitting and overtraining, J. Chem. Inf. Comput. Sci. 35 (1995) 826–833.