Available online at www.sciencedirect.com
ScienceDirect
Journal of Hazardous Materials 157 (2008) 510–517
Journal of Hazardous Materials
www.elsevier.com/locate/jhazmat

# Prediction of auto-ignition temperatures of hydrocarbons by neural network based on atom-type electrotopological-state indices

Yong Pan, Juncheng Jiang *, Rui Wang, Hongyin Cao, Jinbo Zhao

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

Received 27 June 2007; received in revised form 21 September 2007; accepted 7 January 2008
Available online 15 January 2008

**Abstract**

A quantitative structure–property relationship (QSPR) model was constructed to predict the auto-ignition temperature (AIT) of 118 hydrocarbons by means of artificial neural network (ANN). Atom-type electrotopological-state indices were used as molecular structure descriptors which combined together both electronic and topological characteristics of the analyzed molecules. The typical back-propagation (BP) neural network was employed for fitting the possible non-linear relationship existed between the atom-type electrotopological-state indices and AIT. The dataset of 118 hydrocarbons was randomly divided into a training set (60), a validation set (16) and a testing set (42). 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 [16-8-1], the results show that most of the predicted AIT values are in good agreement with the experimental data, with the average absolute error being 21.6 °C, and the root mean square error (RMS) being 31.09 for the testing set, which are superior to those obtained by multiple linear regression analysis and traditional group contribution method. The model proposed can be used not only to reveal the quantitative relation between AIT and molecular structures of hydrocarbons, but also to predict the AIT values of hydrocarbons for chemical engineering.


*Keywords:* Quantitative structure–property relationship (QSPR); Electrotopological-state indices; Artificial neural network; Auto-ignition temperature; Hydrocarbons

## 1. Introduction

The auto-ignition temperature (AIT) is defined as the lowest temperature at which a material in air begins to ignite in the absence of an external ignition source, such as spark or flame. Auto-ignition occurs when the rate of heat produced by exothermic oxidation reactions overbalances the rate at which heat is discharged to the surroundings. Since auto-ignition occurs in air without the presence of an ignition source, it is an important fire performance parameter in process design and operational procedures. In many common situations, such as the manufacture, handling, transport, and storage of combustible materials, the AIT has been widely used to characterize the hazard potential of chemicals.

The measurement of AIT is dependent on many experimental factors, such as the sample concentration, flow condition, the initial pressure, the volume of the sample as well as the geometry of the experimental vessel, all of which can affect the AIT to a certain extent. Based on combinations of these factors, the measured AIT can vary by hundreds of degrees. Besides, the measurement of AIT is laborious, because the number of compounds for which data are needed is very large. Moreover, for toxic, volatile, explosive, and radioactive compounds, the measurement is more difficult and even impossible. Thus the development of theoretical prediction methods, which are desirably convenient and reliable for predicting the AIT is required.

Many previous studies have shown that the AIT of a compound is very dependent on its structure, and several methods for estimating pure components AIT from their molecular structure alone have been reported in the literature [1–5]. Mitchell and Jurs [3] developed mathematical models which related the structures of a heterogeneous group of organic compounds to their AIT values. The molecular structures of the compounds are represented by calculated numerical descriptors which encode their topological, electronic, and geometric features. These

* Corresponding author at: Mail Box 186, No. 5 Xinmofan Road, Nanjing University of Technology, Nanjing 210009, China. Tel.: +86 25 83587305; fax: +86 25 83587411.
*E-mail addresses:* yongpannjut@163.com (Y. Pan), jcjiang@njut.edu.cn (J. Jiang), hades44@126.com (R. Wang), daqiao517@sohu.com (H. Cao), b0101@163.com (J. Zhao).

0304-3894/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhazmat.2008.01.016

descriptors are used to develop several multiple linear regression (MLR) and artificial neural network (ANN) models for predicting the AIT of hydrocarbons, halohydrocarbons, and compounds containing oxygen, sulfur, and nitrogen, respectively. The models developed were reported to have predictive ability within the range of the experimental error of AIT measurements for heterogeneous group of organic compounds, with root mean square errors (RMS) of testing sets from 5.11 to 32.5 °C. Albahri [4] applied the structural group contribution method to develop a theoretical method for predicting the AIT of pure hydrocarbons. The method was used to probe the structural groups that have significant contribution to the overall AIT property and arrive at a set of 20 structural groups which can best represent the AIT for 138 hydrocarbons. The proposed method was reported to predict the AIT of pure components from only the knowledge of the molecular structure, with an average error of 4.2% and a correlation coefficient of 0.92. These successful studies suggested that more information can be gained through a further investigation of this structure–property relationship.

In this work, we used the quantitative structure–property relationship (QSPR) method to investigate the quantitative mathematical relationships between AIT and molecular structures of pure hydrocarbons. Moreover, the electrotopological-state (E-state) indices were employed as descriptors to encode structural characteristics of the studied compounds. The atom-type E-state indices were recently introduced by Hall and Kier [6] for the description of molecules at the atomic level. These indices not only combine together both electronic and topological characteristics of the analyzed atom, but also take into account the binding environment of the atom in the analyzed molecule, and have proven to be effective descriptors in QSPR studies for predicting many physical and chemical properties of pure compounds, such as the critical temperature [7], boiling point [6,7], aqueous solubility [8,9], partition coefficient [10], log $P$ [11], the toxicity [12], and so on. These properties covered almost all the primary aspects of pure component properties except that related to the flammability characteristics of compounds such as flash point, upper and lower flammability limits and AIT.

The main goal of the present work is to further verify the potential of E-state indices for application to AIT prediction. Through an extension of these indices by introducing more detailed indices for $-CH_2-$, >CH–, >C<, =CH–, and =C< groups, as well as the employment of the widely used non-linear modeling technique of artificial neural network(ANN), we wish that this study could be a new attempt for predicting the AIT values and improve the prediction results.

## 2. Methods

The QSPR models in this study were developed using artificial neural network program and multiple linear regression routine based on the atom-type E-state indices. All computations were executed on a Pentium PC with 512M RAM and CPU speed of 2.4 G. The flowchart in Fig. 1 outlines the procedure used in this study.

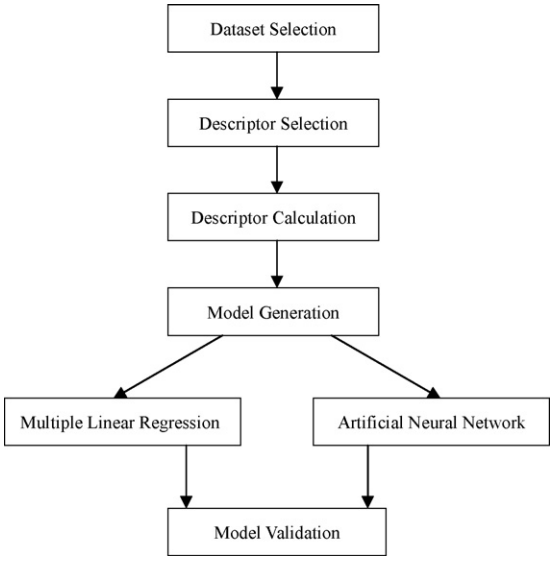


Fig. 1. Flowchart of the method used to develop AIT prediction model.

### 2.1. Dataset

The experimental AIT values of hydrocarbons utilized in the present study were collected from the following sources:

(1) International Chemical Safety Cards (ICSCs) on the Internet [13].
(2) Chemical and other safety information database of Physical and Theoretical Chemistry Laboratory at Oxford University (UK) [14].
(3) Chemical database of the department of chemistry at the University of Akron (USA) [15].
(4) Chemical manufacturer's MSDSs [16].
(5) Lange's Handbook of Chemistry [17].

Dealing with such a large amount of data raises issue of reliability. As we know, the experimental AIT values are dependent on many experimental factors, so the overlapped compounds in different sources frequently possess different experimental AIT values, especially for compounds with higher molecular weights. The experimental values sometimes can differ by as much as 71 °C. For example, Ref. [17] supplies experimental AIT value of 258 °C for methylcyclopentane, while Chemical manufacturer's MSDSs (Sigma–Aldrich Inc.) gives 329 °C. Such tremendous discrepancies would disturb pure structure investigations and influence the establishment of reliable QSPR models.

As we know, most organizations assess the reliability of their reported experimental values and also updated the data when necessary. Such as the International Chemical Safety Cards (ICSCs) on the Internet, most data of which were reviewed after 2000, which are judged to be more reliable than those published long before. Thus the ICSCs are considered as a major source of reliable data by organizations such as EU, UNO, etc. Consequently, for QSPR study of the AIT here, we chose the most

recently reported experimental data as the final data for each compound among different reported values.

The dataset finally selected consisted of 118 structurally diverse hydrocarbons which comprised of alkane, olefin, alkyne, and aromatic hydrocarbons. The number of carbon atoms in the compounds varied from 2 to 16. The AIT values for these compounds were in the range from 202 to 640 °C.

### 2.2. Partition of the atom-type E-state indices

The key to the QSPR study is the selection of molecular descriptor. An efficient descriptor must reflect all of the structural information as accurately as possible. In this paper, the problem was expected to be tackled by employing the widely used atom-type E-state indices [6–12], which combined together both electronic and topological characteristics of the analyzed molecules. A detailed description of the calculation of atom-type E-state indices can be referred in the original work of Hall and Kier [6]. For each atom type in a molecule, the E-state indices were summed and can be used in a group contribution manner. For the whole 118 hydrocarbon compounds studied in the datasets, a set of 11 atom-type E-state indices would have been selected for the analysis according to Ref. [6]. However, the atom types provided by Hall and Kier [6] were only based on a general partition for some atom groups, such as the –CH2–, >CH–, >C<, =CH–, and =C< groups, of which the binding environment (straight chain or cyclic ring) has not been distinguished. So in the present study, the atom type of each group mentioned above was further extended to be a straight chain one and a cyclic ring one. The scheme of all the atom types and their E-state indices symbols was shown in Table 1. A extended set of 16 atom-type E-state indices were obtained, which were expected to make the partition of the atom types more reasonable and distinguish the isomeric compounds more effectively.

Table 1
Sixteen atom types and their coding number and E-state indices symbols

| No. | Atom-type[a] | E-state indices symbol[b] |
|---|---|---|
| 1 | $-CH_3$ | SsCH3 |
| 2 | $=CH_2$ | SdCH2 |
| 3 | ≡CH | StCH |
| 4 | ≡C– | StsC |
| 5 | $-CH_2-$ | SssCH2 |
| 6 | $(-CH_2-)_R$ | |
| 7 | >CH– | SsssCH |
| 8 | $(>CH-)_R$ | |
| 9 | >C< | SssssC |
| 10 | $(>C<)_R$ | |
| 11 | =CH– | SdsCH |
| 12 | $(=CH-)_R$ | |
| 13 | aCHa | SaaCH |
| 14 | =C< | SdssC |
| 15 | $(=C<)_R$ | |
| 16 | saCa | SsaaC |

[a] *R* referred to the atom-type in cyclic ring compound.
[b] According to Ref. [6].

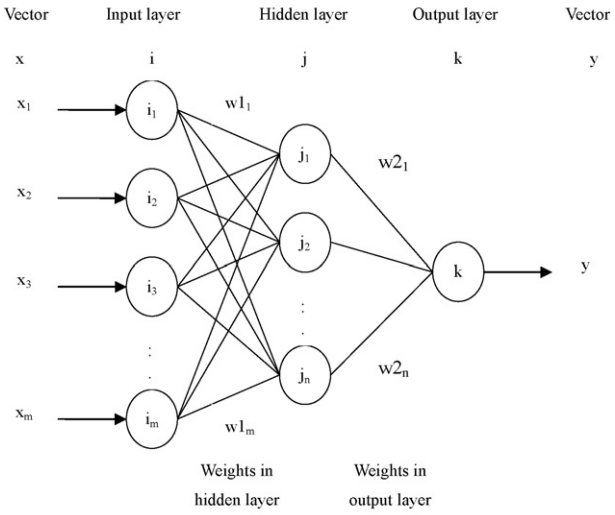


Fig. 2. The neural network architecture used in this work.

The E-state indices of all the atom types for each compound were calculated. The results showed that the structure of various hydrocarbon compounds can be characterized by various atom-type E-state indices, that is to say, atom-type E-state indices can be used to distinguish the molecular structure of these 118 hydrocarbon compounds successfully, with the structure distinguishability of 100%. In such case, the weak ability in distinguishing the isomeric compounds by traditional group contribution method can be overcome.

### 2.3. Modeling methods

Both the ANN method and the MLR method were employed to model the relationship between the AIT values of hydrocarbons and the atom-type E-state indices selected above for the purposes of comparison. The dataset was randomly divided into a training set with 76 compounds and a testing set with 42 compounds for modeling. The training set was used for model development, while the testing set used for model validation. In addition, the training set and testing set used for both MLR analysis and ANN modeling consisted of the same compounds.

For the MLR analysis, it was performed with the SPSS software (Version 11.0; SPSS Inc., Chicago, IL) running on a Pentium PC. The quality of the calculated model was judged by statistical characteristics such as correlation coefficient, *R*, the root mean square error, RMS, and Fischer significance value, *F*. Moreover, for the atom-type E-state indices which were found to be non-significant in the regression model ($P > 0.05$), they would be omitted from the final equation.

For the ANN modeling, it was carried out using the STATISTICA Neural Networks (SNN) software. Several ANN architectures were tried and the one that best simulated the AIT was retained. The final network structure used in this work is shown in Fig. 2, which was a fully connected, feed-forward, three-layer neural network. As can be seen from Fig. 2, the net-

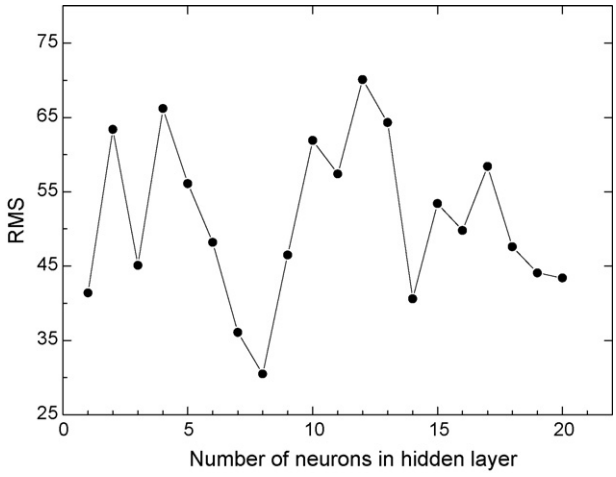


Fig. 3. RMS as a function of the number of neurons in the hidden layer.

work consists of an input layer, an output layer and one hidden layer. The input layer has a number of neurons which is equal to the number of atom-type E-state indices being investigated. The hidden layer is a single layer with a certain number of neurons, which will be discussed as following, and the output layer consists of one neuron representing the predicted AIT property. Signals from input layer are transferred to output layer through hidden layer. In this paper, a multilayer feed-forward network with the algorithm of back-propagation was used as the simulator, and a logistic $f(x) = 1/[1 + \exp(-x)]$ transfer function was employed both for hidden and output neurons. The inputs to the network algorithm are E-state indices of each atom type in the given substance. If a certain input atom type did not exist in a molecule, an input value of zero was assigned to that atom type in the network.

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: the learning rate, the momentum constants, the number of neurons in the hidden layer, and how to prevent overtraining.

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 0.01 and 0.1, respectively [18].

For the optimal number of neurons in the hidden layer, as discussed elsewhere [18], it was determined by varying the number of hidden neurons and observing the root mean square error (RMS), which was used as a measure of the prediction error of the trained model. 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. 3), the optimal number of neurons in the hidden layer was 8.

In addition, it was very important to know whether the network has been over-trained. However, the model built on only a training set and a testing set cannot be confidently considered to be an un-overtrained one. Thus the training set was further divided randomly into a smaller training set with 60 compounds and a cross-validation set with 16 compounds. The optimal training endpoint and network architecture was determined on the basis of the cross-validation set. As we known, the training error always decreases with an increase of training epochs, while the validation error has the lowest point. When the epoch reaches the point, the process of training is stopped. At that time, the best training epoch can be achieved. In this work, the optimal training epoch is about 10,000.

## 3. Results and discussion

The MLR analysis was performed with the SPSS software running on a Pentium PC. Two atom-type E-state indices $(>C<)_R$ and $(=C<)_R$ were found to be non-significant in the regression analysis and were thus omitted from the final equations. As a result, the following regression equation was obtained for the training set ($x_1$–$x_{16}$ referred to the atom-type 1–16 in Table 1):

$$\begin{aligned} \mathrm{AIT} = {} & 498.858 - 25.162x_1 - 10.264x_2 - 24.232x_3 \\ & -6.075x_4 - 13.065x_5 - 22.332x_6 + 58.852x_7 \\ & +24.950x_8 + 335.108x_9 - 21.480x_{11} + 8.581x_{12} \\ & -6.408x_{13} + 20.274x_{14} + 51.376x_{16} \end{aligned} \quad (1)$$

$$R = 0.870, \quad \mathrm{RMS} = 39.07, \quad F = 12.95, \quad n = 76$$

In this equation, $n$ is the number of compounds used in the model. As can be seen from the equation, the quality of the calculated model was not as good as expected. An analysis of residuals showed that there were four outliers (1,7-octadiene, 1-heptene, 1,4-diethylbenzene and cyclododecene) whose estimation errors were higher than two times RMS error. The source of error for this type of outlier can be attributed either to the observed AIT data or to structural features of the molecule that are not properly encoded in the model but that have a large influence on the observed AIT. After excluding these four compounds, the RMS error and average absolute error were 37.86 and 28.6 °C for the remaining 72 compounds.

The model was also used to predict the AIT values for 42 compounds in the testing set, with RMS error of 40.76 and average absolute error of 32.4 °C, which was a little higher than the reported experimental error of +30 °C. The predicted results of testing set were shown in Table 2, in which the predicted AIT of each testing compound can be compared with the experimental one.

Since the attempt to develop a linear model by MLR analysis was not as successful as expected, it is of reasonable confidence to believe that some strong non-linear dependencies may exist between the atom-type E-state indices and AIT of hydrocarbon compounds. In such case, an application of non-linear method of data analysis might provide improvement in the modeling.

Table 2
Comparison of predicted and experimental AIT for the 42 compounds in the testing set

| No. | Compound | Experimental AIT (°C) | Predicted AIT (°C) | | |
|---|---|---|---|---|---|
| | | | MLR | ANN | SGCM [4] |
| 1 | Propane | 432 | 375.6 | 413.6 | 407.3 |
| 2 | Butane | 405 | 354.6 | 319.9 | 358.8 |
| 3 | 2-Methyl-2-butene | 240 | 326.0 | 212.6 | 291.1 |
| 4 | 2,2-Dimethylpropane | 450 | 446.2 | 468.5 | 475.6 |
| 5 | *cis*-2-Pentene | 288 | 289.3 | 279.7 | 283.9 |
| 6 | 1-Hexyne | 263 | 265.7 | 244.0 | 231.6 |
| 7 | Methylcyclopentane | 329 | 333.3 | 316.6 | 329.1 |
| 8 | 4-Methylpentene | 300 | 341.2 | 322.2 | 358.0 |
| 9 | Cyclohexane | 260 | 297.9 | 282.5 | 284.8 |
| 10 | Ethylcyclopentane | 260 | 317.9 | 300.7 | 287.1 |
| 11 | 2,4-Dimethylpentane | 337 | 356.1 | 363.8 | 351.0 |
| 12 | Methylcyclohexane | 283 | 299.2 | 291.5 | 294.4 |
| 13 | Toluene | 422 | 448.6 | 466.7 | 539.6 |
| 14 | *n*-Heptane | 223 | 294.3 | 213.7 | 241.8 |
| 15 | Isoheptane | 280 | 327.0 | 271.4 | 289.4 |
| 16 | 1-Octene | 250 | 275.9 | 235.0 | 230.7 |
| 17 | 2,4,4-Trimethyl-1-pentene | 391 | 391.8 | 387.5 | 367.7 |
| 18 | 2,2,3-Trimethylpentane | 430 | 441.1 | 439.0 | 400.2 |
| 19 | 2,2,4-Trimethylpentane | 410 | 419.3 | 429.8 | 400.2 |
| 20 | Styrene | 490 | 418.3 | 489.6 | 527.5 |
| 21 | Propylcyclopentane | 269 | 296.1 | 262.4 | 252.3 |
| 22 | Octane | 210 | 274.4 | 206.2 | 218.8 |
| 23 | 3,4,4-Trimethyl-2-pentene | 325 | 330.9 | 321.6 | 308.9 |
| 24 | 1,3,5-Trimethylcyclohexane | 314 | 295.1 | 328.0 | 314.6 |
| 25 | Isopropylcyclohexane | 283 | 294.9 | 274.7 | 272.3 |
| 26 | 2,2,3,3-Tetramethylpentane | 430 | 444.8 | 448.4 | 449.6 |
| 27 | 2,3,3,4-Tetramethylpentane | 437 | 411.0 | 442.5 | 420.7 |
| 28 | 1-Methyl-2-ethylbenzene | 448 | 467.3 | 467.6 | 436.3 |
| 29 | Alpha-methyl styrene | 445 | 455.6 | 454.8 | 498.2 |
| 30 | 2-Methylnonane | 214 | 267.3 | 229.1 | 209.6 |
| 31 | *tert*-Butylbenzene | 450 | 433.3 | 502.9 | 548.6 |
| 32 | Cyclodecane | 235 | 163.9 | 141.6 | 205.3 |
| 33 | 4-Ethyloctane | 235 | 275.5 | 233.8 | 209.6 |
| 34 | 2,3-Dimethyloctane | 231 | 298.1 | 275.3 | 237.4 |
| 35 | 1,3-Diethylbenzene | 450 | 451.0 | 445.0 | 440.7 |
| 36 | 4-Isopropyl-1-methylcyclohexane | 306 | 293.1 | 294.7 | 281.2 |
| 37 | 1,2-Diethylbenzene | 395 | 455.7 | 446.2 | 387.4 |
| 38 | *p-tert*-Butyltoluene | 510 | 457.1 | 516.1 | 529.9 |
| 39 | 1-Methyl-3,5-diethylbenzene | 461 | 481.4 | 458.9 | 478.5 |
| 40 | Undecane | 240 | 220.2 | 320.6 | 202.8 |
| 41 | Tridecane | 202 | 175.8 | 230.2 | 222.8 |
| 42 | Tetradecane | 235 | 156.1 | 239.2 | 234.0 |
| | The average absolute error (°C) | | 32.4 | 21.6 | 24.8 |
| | RMS | | 40.76 | 31.09 | 34.01 |
| | *R* | | 0.902 | 0.952 | 0.949 |

Then artificial neural network with back-propagation learning algorithm was used to explore the presence of non-linear dependencies and develop improved models.

The same set of 16 atom-type E-state indices as used in MLR analysis was employed as the inputs in neural network simulation. With the optimum network architecture represented by [16-8-1], the simulation process was repeated 10 times with different random starting weights between neurons. Then the averaged AIT value of each compound was calculated, which was regarded as the final predicted AIT value. As a result, the average absolute error of training, validation and testing sets were 12.8, 26.9 and 21.6 °C, respectively. The RMS were 17.49, 34.78 and 31.09, and the correlation *R* were 0.987, 0.955 and 0.952, respectively. The predicted results of training set and validation set were shown in Table 3, while the predicted results of testing set were shown in Table 2. Obviously, compared with the aforementioned MLR model, the ANN model can provide better results here.

The experimental and predicted AIT of the training, validation and testing sets were plotted in Fig. 4. Regression lines were used for comparing the values obtained by this model with experimental values. As can be seen from Fig. 4, the calculated slope and intercept did not differ greatly from the "ideal" values of 1 and 0, respectively, and most of the predicted AIT values agreed

Table 3
Experimental and predicted AIT by ANN for 76 compounds in the training set and validation set

| No. | Compound[a] | Experimental AIT (°C) | Predicted AIT (°C) | Deviation (°C) |
|---|---|---|---|---|
| 1 | Ethylene | 450 | 478.4 | 28.4 |
| 2 | Propylene | 460 | 445 | −15 |
| 3 | Cyclopropane | 498 | 504.6 | 6.6 |
| 4 | 2-Butene | 324 | 343.7 | 19.7 |
| 5 | 1,3-Butadiene | 420 | 391 | −29 |
| 6 | Isobutane | 460 | 467.1 | 7.1 |
| 7 | Isobutene | 465 | 463.8 | −1.2 |
| 8 | Cyclopentane | 361 | 358.6 | −2.4 |
| 9 | Cyclopentene | 395 | 381.7 | −13.3 |
| 10 | Isopentane | 420 | 379.1 | −40.9 |
| 11 | Isoprene | 220 | 231.8 | 11.8 |
| 12 | 3-Methyl-1-butene | 365 | 374.7 | 9.7 |
| 13 | *n*-Pentane | 260 | 264.7 | 4.7 |
| 14 | 1-Pentene | 272 | 311.3 | 39.3 |
| 15 | Cyclohexene | 310 | 342.7 | 32.7 |
| 16 | 2,3-Dimethylbutane | 420 | 445.7 | 25.7 |
| 17 | 2,3-Dimethyl-1-butene | 370 | 377.7 | 7.7 |
| 18 | 2,3-Dimethyl-2-butene | 401 | 412.9 | 11.9 |
| 19 | 2-Ethyl-1-butene | 315 | 317.1 | 2.1 |
| 20 | 1-Hexene | 253 | 273 | 20 |
| 21 | *trans*-2-Hexene | 245 | 235.8 | −9.2 |
| 22 | 2-Methylpentane | 306 | 313.7 | 7.7 |
| 23 | 3-Methylpentane | 278 | 317.8 | 39.8 |
| 24 | 1-Heptene | 260 | 249 | −11 |
| 25 | 3-Methylhexane | 280 | 275.2 | −4.8 |
| 26 | 2,2,3-Trimethylbutane | 450 | 457.8 | 7.8 |
| 27 | 1,2-Dimethylcyclohexane | 304 | 307.3 | 3.3 |
| 28 | Ethylcyclohexane | 262 | 267.7 | 5.7 |
| 29 | Ethylbenzene | 432 | 449.4 | 17.4 |
| 30 | 1,7-Octadiene | 230 | 242 | 12 |
| 31 | 1-Octyne | 225 | 229.1 | 4.1 |
| 32 | 2,3,3-Trimethylpentane | 430 | 433.3 | 3.3 |
| 33 | *m*-Xylene | 527 | 524.4 | −2.6 |
| 34 | *o*-Xylene | 463 | 501.5 | 38.5 |
| 35 | *p*-Xylene | 529 | 499.5 | −29.5 |
| 36 | 1,3,5-Trimethylbenzene | 550 | 509 | −41 |
| 37 | 3-Methyloctane | 220 | 236.5 | 16.5 |
| 38 | 4-Methyloctane | 225 | 237.2 | 12.2 |
| 39 | Nonane | 205 | 205.2 | 0.2 |
| 40 | Propylcyclohexane | 248 | 255.8 | 7.8 |
| 41 | 2,4-Dimethyl-3-ethylpentane | 390 | 381.9 | −8.1 |
| 42 | 1,2,3-Trimethylbenzene | 470 | 511.4 | 41.4 |
| 43 | 1,2,4-Trimethylbenzene | 515 | 509.8 | −5.2 |
| 44 | *n*-Butylcyclohexane | 245 | 252 | 7 |
| 45 | Butylbenzene | 412 | 406.7 | −5.3 |
| 46 | *sec*-Butylbenzene | 418 | 405.3 | −12.7 |
| 47 | *p*-Cymene | 436 | 453.3 | 17.3 |
| 48 | Decane | 210 | 208.5 | −1.5 |
| 49 | 1-Decene | 235 | 225 | −10 |
| 50 | 1,4-Diethylbenzene | 430 | 444.6 | 14.6 |
| 51 | Divinylbenzene | 470 | 472.5 | 2.5 |
| 52 | 1,2,3,4-Tetrahydronaphthalene | 385 | 388.8 | 3.8 |
| 53 | Cyclododecene | 258 | 257 | −1 |
| 54 | Dicyclohexyl | 245 | 247 | 2 |
| 55 | Dodecane | 205 | 221.8 | 16.8 |
| 56 | Methyl biphenyl | 482 | 479.8 | −2.2 |
| 57 | Diphenylmethane | 485 | 483.7 | −1.3 |
| 58 | Bibenzyl | 480 | 482.6 | 2.6 |
| 59 | 1-Tetradecene | 235 | 236.7 | 1.7 |
| 60 | Hexadecene | 240 | 249 | 9 |
| 61 | Acetylene | 305 | 400.3 | 95.3 |
| 62 | 1-Butene | 384 | 368.4 | −15.6 |
| 63 | Cyclopentadiene | 640 | 607.9 | −32.1 |
| 64 | 2-Methyl-1-butene | 365 | 379.4 | 14.4 |

Table 3 (*Continued*)

| No. | Compound[a] | Experimental AIT (°C) | Predicted AIT (°C) | Deviation (°C) |
|---|---|---|---|---|
| 65 | 2,2-Dimethylbutane | 425 | 453.1 | 28.1 |
| 66 | *n*-Hexane | 225 | 231.5 | 6.5 |
| 67 | 2,3-Dimethylpentane | 337 | 376.6 | 39.6 |
| 68 | *trans*-1,3-Dimethylcyclohexane | 306 | 307.8 | 1.8 |
| 69 | 2,4,4-Trimethyl-2-Pentene | 308 | 311.8 | 3.8 |
| 70 | Cumene | 420 | 435.7 | 15.7 |
| 71 | Propylbenzene | 450 | 422.5 | −27.5 |
| 72 | 2-Vinyl toluene | 494 | 512.1 | 18.1 |
| 73 | Decahydronaphthalene | 250 | 274.2 | 24.2 |
| 74 | Isobutylbenzene | 428 | 404.2 | −23.8 |
| 75 | 1-Dodecene | 255 | 227.7 | −27.3 |
| 76 | Hexadecane | 202 | 258.1 | 56.1 |

[a] The compounds from 1 to 60 composed the training sample, those from 61 to 76 were the validation sample.

with the experimental values satisfactorily, with the predicted errors lower than the reported experimental error of +30 °C, for all the training, validation and testing sets. Thus, models have been developed that calculate the AIT values for hydrocarbons with accuracy comparable to experiment.

Moreover, since the predicted results obtained by ANN method were better than those obtained by MLR, it showed a superior prediction ability of the ANN model and supported our aforementioned conjecture that a non-linear relationship may exist between the atom-type E-state indices and AIT of hydrocarbon compounds.

The results of ANN model were also tested for chance effects. A Monte Carlo experiment was adopted for testing, in which the dependent variables were scrambled. As a result, the testing models provided high RMS errors, which were 323.53, 376.46 and 354.15 for the training, validation and testing sets, respectively. Such errors are hundred times the errors obtained when the dependent variables were not scrambled. It indicated that only the proper dependent variables can be used to generate reasonable models, and the results obtained by ANN method here were not due to chance.

After that, a general comparison has been made between the ANN model and the work of Albahri [4], who used the traditional structural group contribution method (SGCM) to build the non-linear model for predicting the AIT values of hydrocarbons. For the same 42 test samples, the predicted results were shown in Table 2. The predicted average absolute error of SGCM was 24.8 °C, and the RMS was 34.01. The comparison in Table 2 implied that: the ANN model was based on the atom-type E-state indices which describe better the structural features of the compounds important for the AIT and thus this model had better statistical characteristics (correlation *R*, the average absolute error and RMS) for prediction than the SGCM model. In addition, the current ANN model possesses a smaller number of input descriptors (16 versus 20), which need less calculations. However, the work of Albahri [4] can provide a direct model and preserve the model provided to be unique. These were just the limitations of the ANN model. Moreover, the SGCM model was based on larger number of compounds for the training sets (131 versus 76).

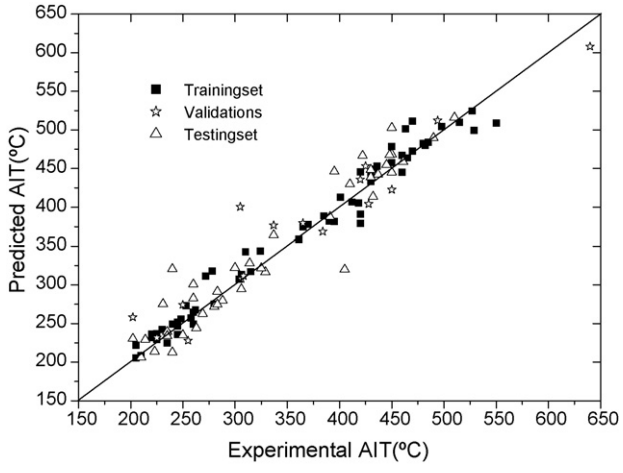


Fig. 4. Correlation between the predicted and experimental AIT for the training, validation and testing sets (95) confidence limits for the intercept and slope.

Besides, from Table 2 we can see that the ANN model and SGCM model can provide better prediction capability here than the MLR model. Such phenomenon strongly suggested a non-linear relationship existing between the atom-type E-state indices and AIT of hydrocarbons once again.

As also can be seen from Table 2, the predicted AIT values for methylcyclopentane by MLR, ANN and SGCM methods were 333.3, 316.6 and 329.1 °C, respectively, all of which were close to the value of 329 °C used in this work, which is just the most recently reported experimental data for methylcyclopentane among different reported values. The fact above demonstrated that the AIT values of hydrocarbons we selected for the experimental dataset in this study were more than probably reasonable and reliable.

## 4. Conclusion

In this study, an ANN-based QSPR model was developed for the prediction of AIT of hydrocarbon compounds using atom-type E-state indices. The extended atom-type E-state indices were used as molecular structure descriptors, and the ANN method was employed for fitting the possible non-linear relationship existed between the structure and property. The results showed that most of the predicted values of AIT agreed with

the experimental values satisfactorily, with the predicted errors within the range of experimental error. Thus ANN-based QSPR model using atom-type E-state indices can be successfully used to correlate the AIT values with the molecular structure of hydrocarbons, and can also enable initial estimation of AIT for new hydrocarbon compounds or for other hydrocarbons for which experimental values are unknown.

## Acknowledgements

This research is supported by National Natural Science Fund of China (No. 29936110) and the Program for New Century Excellent Talents in University (No. NCET-05-0505). Y. Pan acknowledged the support of Jiangsu Graduate Scientific Innovation Projects.

## References

[1] T. Suzuki, Quantitative structure–property relationships for auto-ignition temperatures of organic compounds, Fire Mater. 18 (1994) 81–88.
[2] J. Tetteh, E. Metcalfe, S. Howells, Optimization of radial basis and back-propagation neural networks for modeling auto-ignition temperature by quantitative structure–property relationships, Chemometr. Intell. Lab. Syst. 32 (1996) 177–191.
[3] B.E. Mitchell, P.C. Jurs, Prediction of autoignition temperatures of organic compounds from molecular structure, J. Chem. Inf. Comput. Sci. 37 (1997) 538–547.
[4] T.A. Albahri, Flammability characteristics of pure hydrocarbons, Chem. Eng. Sci. 58 (2003) 3629–3641.
[5] T.A. Albahri, R.S. George, Artificial neural network investigation of the structural group contribution method for predicting pure components auto ignition temperature, Ind. Eng. Chem. Res. 42 (2003) 5708–5714.
[6] L.H. Hall, L.B. Kier, Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information, J. Chem. Inf. Comput. Sci. 35 (1995) 1039–1045.
[7] L.H. Hall, C.T. Story, Boiling point and critical temperature of a heterogeneous data set: QSAR with atom type electrotopological state indices using artificial neural networks, J. Chem. Inf. Comput. Sci. 36 (1996) 1004–1014.
[8] J. Huuskonen, M. Salo, J. Taskinen, Aqueous solubility prediction of drugs based on molecular topology and neural network modeling, J. Chem. Inf. Comput. Sci. 38 (1998) 450–456.
[9] J. Huuskonen, Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology, J. Chem. Inf. Comput. Sci. 40 (2000) 773–777.
[10] J. Huuskonen, D.J. Livingstone, I.V. Tetko, Neural network modeling for estimation of partition coefficient based on atom-type electrotopological state indices, J. Chem. Inf. Comput. Sci. 40 (2000) 947–955.
[11] J. Huuskonen, QSAR modeling with the electrotopological state: TIBO derivatives, J. Chem. Inf. Comput. Sci. 41 (2001) 425–429.
[12] J. Huuskonen, QSAR modeling with the electrotopological state indices: predicting the toxicity of organic chemicals, Chemosphere 50 (2003) 949–953.
[13] http://www.inchem.org/pages/icsc.html.
[14] http://ptcl.chem.ox.ac.uk/MSDS/.
[15] http://ull.chemistry.uakron.edu/erd/index.html.
[16] http://www.msdsxchange.com/english/index.cfm.
[17] J.A. Dean, Lange's Handbook of Chemistry, 15th ed., McGraw-Hill, New York, 1999.
[18] Y. Pan, J.C. Jiang, Z.R. Wang, Quantitative structure–property relationship studies for predicting flash points of alkanes using group bond contribution method with back-propagation neural network, J. Hazard. Mater. 147 (2007) 424–430.