Prediction of Flammability Characteristics of Pure Hydrocarbons from Molecular Structures

Yong Pan, Juncheng Jiang, Xiaoye Ding, Rui Wang, and Jiajia Jiang
Jiangsu Key Laboratory of Urban and Industrial Safety, College of Urban Construction and Safety Engineering,
Nanjing University of Technology, Nanjing 210009, China

DOI 10.1002/aic.12007
Published online November 18, 2009 in Wiley InterScience (www.interscience.wiley.com).

A quantitative structure-property relationship study is performed to develop mathematical models for predicting the flammability characteristics of pure hydrocarbons. The molecular structures of the compounds are numerically represented by various kinds of molecular descriptors. Genetic algorithm based multiple linear regression is used to select most statistically effective descriptors on the flash point, the autoignition temperature, and the lower and upper flammability limits of hydrocarbons, respectively. The resulted models are four multilinear equations. These models are very simple and can predict the flash point, the autoignition temperature, and the lower and upper flammability limits for the test set with average absolute errors of 5.41 K, 28.00 K, 0.044 vol %, and 0.503 vol %, respectively. The models are further compared with other published method and are shown to be more superior. The proposed method can be used to predict the flammability characteristics of hydrocarbons from the knowledge of only the molecular structures. © 2009 American Institute of Chemical Engineers *AIChE J,* 56: 690–701, 2010

Keywords: flammability characteristics, hydrocarbons, prediction, QSPR, molecular descriptors

Introduction

In chemical and engineering industry, there is a wide application of various physicochemical data. For example, risk assessment calculations often require a wide range of physicochemical parameter inputs. Similarly, in process design, material and energy balances must be based on accurate data to properly size equipment and determine utility consumption and cost. As a result, reliable and accurate data of physicochemical properties are always required and also considered to be absolutely necessary. However, in the practical industry process, the required data are often absent when needed, especially for the properties that are related to the combustion such as the flash point (FP), the autoignition temperature (AIT), and the flammability limits.

The flammability characteristics of chemicals, which include the FP, the AIT, and the lower flammability limit (LFL) and upper flammability limit (UFL), are very important for safety considerations in manufacturing processes.[1] These characteristics are some of the most important safety specifications that must be considered in assessing the overall flammability hazard potential of a chemical substance, defined as the degree of susceptibility to ignition or release of energy under varying environmental conditions.[2]

For example, the FP is the most important parameter used to characterize the fire and explosion hazard of liquids.[3] The FP of a flammable compound is defined as the temperature at which the substance emits sufficient vapor to form a combustible mixture with air.[4] This parameter provides the knowledge necessary for understanding the fundamental physical and chemical processes of combustion and is of importance in practice for safety considerations in storage, processing, and handling of given compounds.

The AIT is defined as the lowest temperature at which the substance will produce hot-flame ignition in air at

Additional Supporting Information may be found in the online version of this article.

Correspondence concerning this article should be addressed to Y. Pan at ypannjut@gmail.com.

atmospheric pressure without the aid of an external ignition source such as spark or flame.[5] It gives an indication of the temperature at which a material will spontaneously burst into flames when exposed to the atmosphere.[2] This parameter 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.[6]

Flammability limits, which are also referred to as the explosive limits, provide the range of fuel concentration, usually in percentage volume (vol %) at 298 K, within which an explosive reaction is possible when an external ignition source is introduced.[7] The LFL is defined as the minimum concentration of a combustible substance that is capable of propagating a flame through a homogeneous mixture of the substance and the air, whereas the UFL is defined as the maximum one. Knowledge of flammability limits values is essential to maximize safety in process design and operational procedures, such as starting up a reactor without passing through a flammable range and storing or shipping the flammable product safely.

All the discussed above showed that reliable and accurate flammability characteristics data are always required and also considered to be absolutely necessary in the practical industry process.

The experimental values are the main source of the flammability characteristics data used in production. However, as we know, the measurement of flammability characteristics is very much dependent on the apparatus and the test methods used, and the measured flammability characteristics values reported by different literatures are often inconsistent, sometimes quite different. Besides, the measurement of flammability characteristics is expensive and time consuming, and for toxic, volatile, explosive, and radioactive compounds, the measurement is more difficult and even impossible. Therefore, to support and expand the flammability characteristics dataset used for industry, the development of theoretical prediction methods that are desirably convenient and reliable for predicting the flammability characteristics is required.

There have been already several methods reported in the literatures for predicting various flammability characteristics of pure compounds, which can be classified into several categories containing empirical correlations, group contribution models, and the quantitative structure-properties relationship (QSPR) models. These methods have been extensively reviewed by Albahri,[2] Taskinen and Yliruusi,[8] and Vidal et al.[9]

However, the empirical correlations suffer from some important disadvantages. First, the use or application of these models requires unconventional physicochemical properties, and the availability or the lack of which may result in some limitations on their applicability range. Moreover, the prediction accuracy of these models is quite dependent on the accuracy of the needed physicochemical properties.

The most important disadvantage of group contribution models is their limitations in use. For example, the applicability range of these models are too related to the studied dataset, and the new chemicals with functional groups not included in those used for the model development will be out of the model applicability range and, thus, will not be predicted. Moreover, group contribution models also provide a weak ability in distinguishing the isomeric compounds.

A current trend in predicting various physicochemical properties is the use of QSPR method. QSPR is a mathematical method that relates the properties of interest to the molecular structures of compounds, which are represented by a variety of molecular descriptors. Molecular descriptors are various molecular-based theoretical parameters, which can be calculated using known mathematical algorithms solely from molecular structures. Several molecular descriptors are always selected as the QSPR input to correlate the desired property of compounds with special principles. The QSPR method possesses some obvious advantages. First, the number of descriptors selected in the QSPR method is almost always lower than those selected in the group contribution method for the same studied dataset. This fact may bring on more robust models. Second, the descriptors used in the QSPR models have definite physical meanings, which would be useful to probe the physicochemical information that has significant contribution to the targeted properties. Third, because only theoretical descriptors derived solely from the molecular structure would be involved and have continuous values, the QSPR models developed should theoretically be applicable to any organic compound. Consequently, the QSPR method has been widely used in predicting various physicochemical properties, such as boiling point, melting point, vapor pressure, critical properties, water solubility, octanol/water coefficients, and so on.[10–18]

In this work, a QSPR study is presented to predict the flammability characteristics of a large number of hydrocarbon compounds, which are of industrial importance. The main purpose is to develop reliable QSPR models for predicting various flammability characteristics of pure hydrocarbons from their molecular structures alone.

Methodology

QSPR method is one of the modern property prediction methods. A basic assumption in the QSPR approach is that various physicochemical properties of a compound are closely related to its molecular structure. By encoding the structures of compounds with numerical values, termed descriptors, an indirect mathematical relationship can be found, which correlates structures to physicochemical properties. Once a reliable relationship has been obtained, it is possible to use it to predict the same property for other compounds not yet measured or even not yet prepared. This approach is invaluable today, because industries continually search libraries for various physicochemical data of hundreds of thousands of chemicals for use in their products and applications.

The main task in QSPR studies is to establish a numerical relationship between certain molecular property and molecular descriptors by means of statistics or some other methods as follows:

Property = f (molecular structure) = f (molecular descriptors).

Generally speaking, the QSPR studies can be divided into the following two main steps:

Step 1—to design and generate molecular descriptors;

Step 2—to construct QSPR models with some proper descriptors.

The success of the QSPR approach can be explained by the insight offered into the structural determination of physicochemical properties and the possibility to estimate the properties of new compounds without the need to synthesize and test them. Following the QSPR, studies will be carried out to develop prediction models for various flammability characteristics of hydrocarbons step by step.

Dataset preparation

The accuracy of the QSPR model can be directly affected by the reliability of experimental dataset. Hence, it is very important to select a reliable database for choosing the studied dataset. Currently, one of the best evaluated databases presented for various physical properties of organic compounds is the DIPPR 801 project,[19] which is recommended by the American Institute of Chemical Engineers and considered to be the world's best source of critically evaluated thermophysical, environmental, safety, and health property data. In this study, we selected this database for reliable and accurate flammability characteristics values. Moreover, to build a QSPR model with a wide applicability range, applying an exhaustive dataset to develop the model is preferred, which could extend the applicability of the developed model. An exhaustive dataset is a dataset containing maximum possible number of involved compounds, as well as maximum possible diverse chemical families. In this work, a total of 457 hydrocarbons were selected from the DIPPR 801, and their corresponding FP, AIT, LFL, and UFL values were extracted. Finally, the data of 314 hydrocarbons on the FP, 153 on the AIT, 354 on the LFL, and 278 on the UFL were achieved and used in this study. A complete list of the compounds and their corresponding flammability characteristics values were presented as the Supporting Information.

Descriptor calculation and reduction

To obtain a QSPR model, compounds must be represented using molecular descriptors. A wide variety of descriptors have been reported for QSPR analysis,[20,21] such as topological, geometrical, electrostatic, and quantum chemical descriptors. Each type of these descriptors is related to the special types of interaction between chemical groups in a molecule. In this work, the molecular descriptors used to search for the best models of the flammability characteristics prediction are calculated by the Dragon program (version 5.4, DRAGON is copyrighted by TALETE srl),[22] which is a sophisticated program for the calculation of molecular descriptors. Because the values of many descriptors are related to the bonds length and bonds angles, the chemical structure of every molecule must be optimized before calculating its molecular descriptors. For this reason, chemical structures of all 457 hydrocarbons were drawn using the Hyperchem software (version 7.5, Hyperchem is copyrighted by Hypercube) and optimized using the MM+ molecular mechanics force field and AM1 semiempirical method to obtain the minimized energy molecular models. After optimizing the chemical structures, the molecular descriptors were calculated using Dragon software. The detailed description on the types of the molecular descriptors that Dragon can calculate and the procedure of calculation of the descriptors can be referred from Dragon user manual.[22] In all, a total of 1664 descriptors were calculated for each compound in the dataset.

After the calculation of molecular descriptors, those stayed constant and near constant for all studied compounds were removed from the descriptor pool, because those descriptors were not encoding the structural differences between compounds that account for their different flammability characteristics values. Further reduction of the descriptor pool was attained by examining pairwise correlations between descriptors, so that only one descriptor was retained from a pair contributing similar information (correlation coefficient >0.96 in this study). These reductions resulted in reduced pools of 567, 591, 565, and 603 descriptors for further studies of FP, AIT, LFL, and UFL, respectively.

Descriptor selection and model development

The basic strategy of QSPR analysis is to find optimum quantitative relationships between the molecular descriptors and desired property, which can be then used for the prediction of the property from only molecular structures. One of the most important problems involved in QSPR studies is to select optimal subset of descriptors that have significant contribution to the desired property. The well-known genetic algorithm (GA) is just a well-accepted method for solving this problem.

GA is a powerful optimization method to search for the global optima of solutions. This algorithm is developed to mimic some of the processes observed in natural evolution. The detailed description of which can be found in Ref. [23]. In recent years, GA has been successfully applied to feature selection in QSPR studies. In this study, the GA, along with multiple linear regressions (MLR) method (GA-MLR), was used to find the optimal subset of descriptors that accurately represented the relationships between molecular structures and FP, AIT, LFL, and UFL, respectively. GA-MLR is a sophisticated hybrid approach that combines GA as a powerful optimization method with MLR as a popular statistical method for variable selection. This algorithm was presented by Leardi et al.[24] for the first time. In this study, the program required to perform GA-MLR was written in MATLAB M-file in our laboratory. The chromosome and its fitness function in the species correspond to a set of descriptors and the root mean square error of cross-validation (RMSECV), respectively.

To obtain the best QSPR model, the descriptor space is then thoroughly explored by GA-MLR analysis. Models with varying numbers of descriptors are examined. Because the minimum number of possible descriptors must be tested at the starting point, the program is started with one descriptor. After running the program, the best regression model should be obtained. Then a stepwise addition of further descriptor scales is performed to find the best multiparameter regression models with the desired number of descriptors. The first rule to determine if one descriptor is useful is the fitness function (RMSECV). When adding another descriptor did not improve significantly the obtained RMSECV, it was determined that the optimum subset of descriptors that yield to the best MLR model had been achieved.

Table 1. Descriptors Selected for the Presented Model for Prediction of FP

Descriptor	Type	Definition
SCBO	Constitutional descriptors	Sum of conventional bond orders (H-depleted)
VEp2	Eigen value-based indices	Average Eigen vector coefficient sum from polarizability weighted distance matrix
C-002	Atom-centered fragments	CH_2R_2

Model validation

For QSPR studies, model validation is of crucial importance for the developed models. The calibration and predictive capability of developed QSPR models should be tested through model validation. The most widely used squared correlation coefficient for fitting (R^2) can provide a reliable indication of the fitness of the model; thus, in this study, it was used to validate the calibration capability of QSPR models. As for the validation of predictive capability of QSPR models, both the widely used cross-validation (CV) and external validation are used.

The CV is one of the most often used methods for internal validation. A good CV result (Q^2) often indicates a good robustness and high internal predictive ability of a QSPR model. In this work, the leave-many-out (LMO, 20% out) CV is used.

The external validation is a significant and necessary validation method used to determine both the generalization performance and the true predictive capability of the QSPR models for new chemicals, by splitting the available dataset into a training set and an external test set. The training set is used for descriptor selection and model development, while the external test set is used for model validation. Moreover, as we know, a QSPR model cannot be verified for its predictivity by checking only a few compounds, because in such cases, the results could be obtained by chance, and it is impossible to obtain general conclusions.[25] Consequently, the model must be tested on a sufficiently large number of compounds that are not used in the model development (at least 20% of the complete dataset is recommended).[25] Hence, in this work, for each flammability property, the whole dataset is randomly divided into a training set with 80% compounds and a test set with 20% compounds of the dataset.

Both useful parameters RMSE and the average absolute error (AAE) calculated on the dataset were also used to evaluate the performance of developed models.

Results and Discussion

Results of FP prediction

For the QSPR study of FP, the dataset of 314 hydrocarbons on the FP is randomly divided into a training set with 251 compounds and a test set with 63 compounds. By performing the GA-MLR procedure on the training set, an optimum subset of three descriptors is achieved. The types and definitions of these descriptors are presented in Table 1. The corresponding best MLR model is presented as following:

$$\mathrm{FP} = 367.069\ (\pm 6.939) + 6.300\ (\pm 0.154)\ \mathrm{SCBO} - 397.659\ (\pm 15.906)\ \mathrm{VEp2} + 1.616\ (\pm 0.109)\ \mathrm{C} - 002 \quad (1)$$

$$\text{range}:\ 169.15\ \mathrm{K} \leq \mathrm{FP} \leq 517\ \mathrm{K}$$

$$R^2 = 0.989, Q^2_{\mathrm{LMO}} = 0.988, s = 6.99, n = 251$$

where n is the number of compounds used in the model and s is the standard error of the model.

Equation 1 shows that the FP can be predicted using three molecular descriptors. The physical meanings of these descriptors are presented as following:

SCBO is a constitutional descriptor, which corresponds to a sum of conventional bond orders of the incident bonds to each atom of the molecule (without considering hydrogen's). It is a measure of the degree of insaturation in each compound. VEp2 is an Eigen value-based index, which is related to the molecular geometry and molecular size. C-002 is an atom-centred fragment, which represents the number of methylene groups in a molecule. It is a measure of the degree of molecular branching in each compound.[20]

In conclusion, all the three selected descriptors are significantly related to the simple features of the compounds like molecular size and shape and represent the degree of insaturation in each compound. Therefore, the overall FP property of hydrocarbons can be reasonably explained by their steric effects.

The developed model (Eq. 1) is then used to predict the FP values of 63 hydrocarbons in the test set for external validation. The predicted FP values are obtained and presented as the Supporting Information. The main statistical parameters of the model are shown in Table 2. A plot of the predicted FP values vs. the observed ones for both the training and test sets is shown in Figure 1.

The obtained results presented in Table 2 showed that the resulting AAE values for both training and test sets are within the experimental error of FP determination, which is around ±10 K.[26] Also, it is noteworthy that both the AAE and RMSE values were not only low but also as similar as possible for the training and external test sets, which suggests that the proposed model has both predictive ability (low values) and generalization performance (similar values).[25]

Table 2. The Main Statistical Parameters of the Developed Model (Eq. 1)

Statistical Parameters	Training Set	Test Set	The Whole Dataset
Squared correlation coefficient for fitting (R^2)	0.989	0.985	0.988
Squared correlation coefficient for LMO CVs (Q^2_{LMO})	0.988	–	–
AAE	5.41	5.41	5.41
RMSE	6.97	7.47	7.08
No. compounds (n)	251	63	314

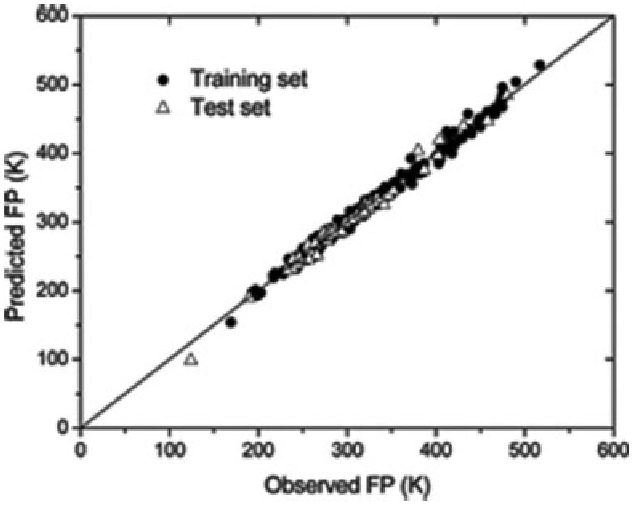

Figure 1. Correlation between the predicted and observed FP values for both the training and test sets.

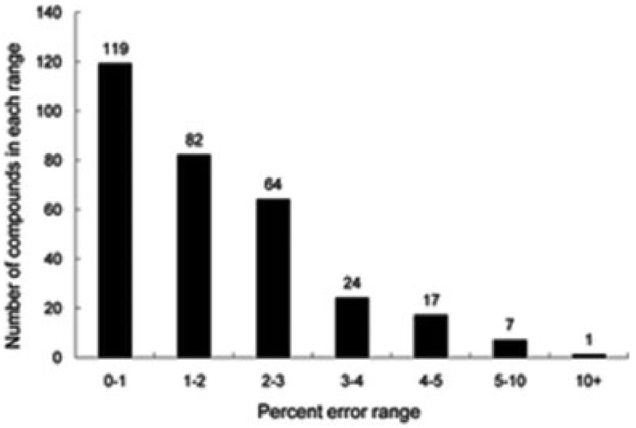

Figure 2. Percent errors of predicted FP by Eq. 1, and the number of compounds in each range.

Moreover, the predicted percentage error of all the 314 compounds was calculated. The obtained average percentage error for these compounds was 1.76%, while the maximum percentage error was 20.55%. The results were shown in detail in Figure 2. As can be seen from Figure 2, the model predictions for the majority compounds are very accurate, with the percentage errors lower than 2%, and very few compounds are above the 5% error range.

Results of AIT prediction

AIT is one of the most difficult properties to estimate or correlate because of its complex dependency on the molecular structure of the compound. In this study, the dataset of 153 hydrocarbons on the AIT is randomly divided into a training set with 122 compounds and a test set with 31 compounds. By performing the GA-MLR procedure on the training set, an optimum subset of five descriptors is achieved. The types and definitions of these descriptors are presented in Table 3. The corresponding best MLR model is presented as following:

$$\begin{aligned}\mathrm{AIT} =\,&725.978\ (\pm 12.760) + 22.820\ (\pm 4.748)\ \mathrm{X2v}\\ &-\ 41.713\ (\pm 11.666)\ \mathrm{BEHe7}\ -\ 122.363\ (\pm 14.775)\ \mathrm{ATS5e}\\ &-12.543\ (\pm 1.906)\ \mathrm{C}-002\ +210.134\ (\pm 13.420)\ \mathrm{ARR}\quad (2)\end{aligned}$$

$$\text{range}:\ 473.15\ \mathrm{K}\ \le\ \mathrm{AIT}\ \le 847.59\ \mathrm{K}$$

$$R^2 = 0.873, Q^2_{\mathrm{LMO}} = 0.860, s = 34.77, n = 122$$

where n is the number of compounds used in the model and s is the standard error of the model.

Equation 2 shows that the AIT can be predicted using five molecular descriptors. The physical meanings of these descriptors are presented as following:

X2v is a connectivity index, which is related to the molecular branching, and contains shape information. BEHe7 is a Burden Eigen value, which gives information about the chemical similarity/diversity of the considered molecules. ATS5e is a 2D autocorrelation descriptor, which is related to the atomic electronegativities of a molecule. C-002 is an atom-centred fragment, which represents the number of methylene groups in a molecule. It is a measure of the degree of molecular branching in each compound. ARR is a constitutional descriptor obtained as a function of the ratio between number of aromatic bonds and the total bonds in the H-depleted molecule. It can roughly be related to the flexibility of the structure.[20]

Among the five selected descriptors, four descriptors (X2v, BEHe7, C-002, and ARR) are more correlated with the simple features of the compounds like molecular branching and give information about the structural similarity/diversity and flexibility of the considered molecules. The other descriptor ATS5e is significantly related to a molecule's electrostatic property. Therefore, the overall AIT property of hydrocarbons can be reasonably explained by their steric and electrostatic effects.

The developed model (Eq. 2) is then used to predict the AIT values of 31 hydrocarbons in the test set for external validation. The predicted AIT values are obtained and presented as the Supporting Information. The main statistical parameters of the model are shown in Table 4. A plot of the

Table 3. Descriptors Selected for the Presented Model for Prediction of AIT

Descriptor	Type	Definition
X2v	Connectivity indices	Valence connectivity index χ^2
BEHe7	Burden Eigen values	Highest Eigen value n = 7 of Burden matrix/weighted by atomic Sanderson electronegativities
ATS5e	2D autocorrelations	Broto–Moreau autocorrelation of a topological structure–lag 5/weighted by atomic Sanderson electronegativities
C-002	Atom-centered fragments	CH_2R_2
ARR	Constitutional descriptors	Aromatic ratio

Table 4. The Main Statistical Parameters of the Developed Model (Eq. 2)

Statistical Parameters	Training Set	Test Set	The Whole Dataset
Squared correlation coefficient for fitting (R^2)	0.873	0.904	0.882
Squared correlation coefficient for LMO CVs (Q^2_{LMO})	0.860	–	–
AAE	25.06	28.00	25.65
RMSE	34.63	33.09	34.32
No. compounds (n)	122	31	153

predicted AIT values vs. the observed ones for both the training and test sets is shown in Figure 3.

The obtained results presented in Table 4 showed that the resulting AAE values for both training and test sets are within the experimental error of AIT determination, which is around ±30 K.[6,27] Also, it is noteworthy that both the AAE and RMSE values were not only low but also as similar as possible for the training and external test sets. As discussed previously, it suggests that the proposed model has both predictive ability (low values) and generalization performance (similar values) to a certain extent.

Then, the predicted percentage error of all the 153 compounds was calculated. The obtained average percentage error for these compounds was 4.24 %, while the maximum percentage error was 22.02%. The results were shown in detail in Figure 4. Very few compounds are above the 10% error range, as can be seen in Figure 4.

Results of LFL prediction

For the QSPR study of LFL, the dataset of 354 hydrocarbons on the LFL is randomly divided into a training set with 284 compounds and a test set with 70 compounds. By performing the GA-MLR procedure on the training set, an optimum subset of four descriptors is achieved. The types and definitions of these descriptors are presented in Table 5. The corresponding best MLR model is presented as following:

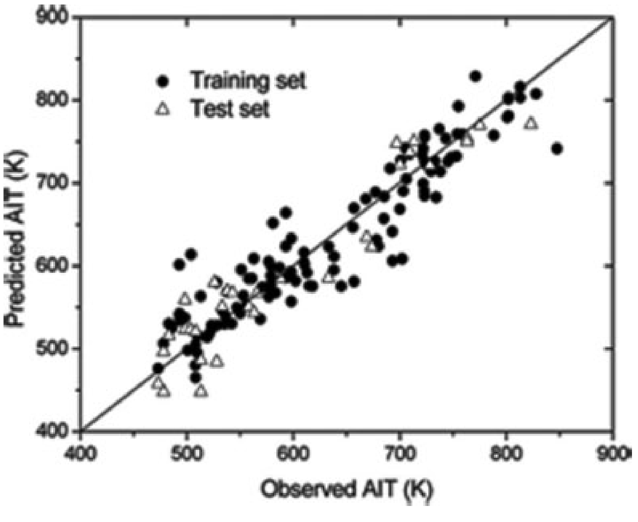

Figure 3. Correlation between the predicted and observed AIT values for both the training and test sets.

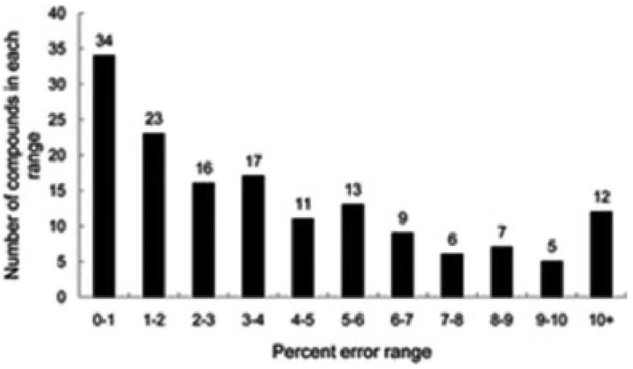

Figure 4. Percent errors of predicted AIT by Eq. 2, and the number of compounds in each range.

$$\mathrm{LFL} = -0.325\,(\pm 0.145) - 3.170\,(\pm 0.106)\,\mathrm{AAC} + 15.545\,(\pm 0.379)\,\mathrm{BIC0} + 0.940\,(\pm 0.083)\,\mathrm{ATS4e} + 0.047(\pm 0.006)\,\mathrm{MLOGP} \quad (3)$$

$$\text{range}: 0.185 \text{ vol }\% \le \mathrm{LFL} \le 2.3 \text{ vol }\%$$

$$R^2 = 0.966, Q^2_{\mathrm{LMO}} = 0.951, s = 0.061, n = 284$$

where n is the number of compounds used in the model and s is the standard error of the model.

Equation 3 shows that the LFL can be predicted using four molecular descriptors. The physical meanings of these descriptors are presented as following:

AAC is an information index to describe each atom by its own atom type and the bond types and atom types of its first neighbors. This descriptor is related to molecular complexity in terms of atom types. BIC0 is an information index, which is related to the 2D structure of compounds. It reveals different aspects of the molecular shape in the action of the combustible and, therefore, the role of steric effects. ATS4e is a 2D autocorrelation descriptor, which is related to the atomic electronegativities of a molecule. MLOGP is an octanol–water partition coefficient, calculated from 13 whole-molecular topological parameters, including the summation of hydrophobic atoms and hydrophilic atoms, unsaturated bonds, and other specific functionalities. It represents the extent of hydrophilic/hydrophobic interactions and recommends higher hydrophobicity of a molecule to improve the activity.[20]

Table 5. Descriptors Selected for the Presented Model for Prediction of LFL

Descriptor	Type	Definition
AAC	Information indices	Mean information index on atomic composition
BIC0	Information indices	Bond information content (neighborhood symmetry of 0 order)
ATS4e	2D autocorrelations	Broto–Moreau autocorrelation of a topological structure—lag 4/weighted by atomic Sanderson electronegativities
MLOGP	Molecular properties	Moriguchi octanol–water partition coefficient (logP)

Table 6. The Main Statistical Parameters of the Developed Model (Eq. 3)

Statistical Parameters	Training Set	Test Set	The Whole Dataset
Squared correlation coefficient for fitting (R^2)	0.966	0.971	0.967
Squared correlation coefficient for LMO CVs (Q^2_{LMO})	0.951	–	–
AAE	0.043	0.044	0.043
RMSE	0.061	0.057	0.060
No. compounds (n)	284	70	354

Among the four selected descriptors, two descriptors (AAC and BIC0) are mainly related to the dimensional features of the molecules like molecular shape and molecular complexity. One descriptor ATS4e is significantly related to molecular electrostatic property. The other descriptor MLOGP mainly represents the extent of hydrophilic/hydrophobic interactions. Therefore, the overall LFL property of hydrocarbons can be reasonably explained by their steric and electrostatic effects.

The developed model (Eq. 3) is then used to predict the LFL values of 70 hydrocarbons in the test set for external validation. The predicted LFL values are obtained and presented as the Supporting Information. The main statistical parameters of the model are shown in Table 6. A plot of the predicted LFL values vs. the observed ones for both the training and test sets is shown in Figure 5.

The obtained results presented in Table 6 showed that the resulting AAE values for both training and test sets are within the experimental error of LFL determination, which is around ±0.1 vol %.[9] Also, both the AAE and RMSE values were not only low but also as similar as possible for the training and test sets, which suggests that the proposed model has both predictive ability and generalization performance.

The predicted percentage error of all the 354 compounds was calculated. The obtained average percentage error for these compounds was 6.07 %, while the maximum percentage error was 39.15 %. The results were shown in detail in Figure 6. A few compounds are above the 20% error range, as can be seen in Figure 6. For these compounds, the model prediction errors are higher and should be used with caution.

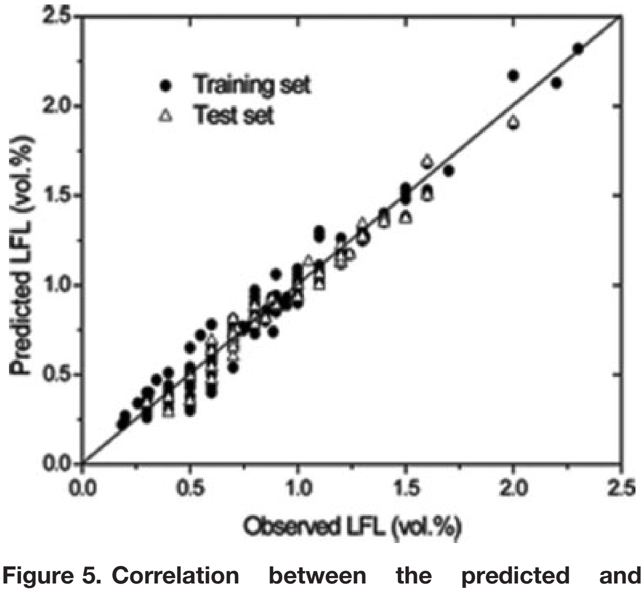

Figure 5. Correlation between the predicted and observed LFL values for both the training and test sets.

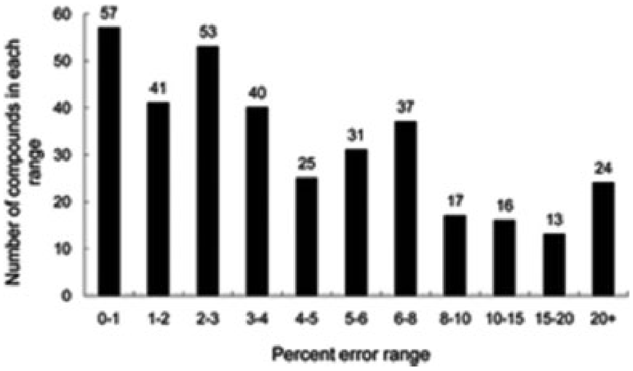

Figure 6. Percent errors of predicted LFL by Eq. 3, and the number of compounds in each range.

Results of UFL prediction

For the QSPR study of UFL, the dataset of 278 hydrocarbons on the UFL is randomly divided into a training set with 222 compounds and a test set with 56 compounds. By performing the GA-MLR procedure on the training set, an optimum subset of four descriptors is achieved. The types and definitions of these descriptors are presented in Table 7. The corresponding best MLR model is presented as following:

$$\text{UFL} = 17.084\ (\pm 0.302) - 16.819\ (\pm 2.527)\ \text{PW5} - 2.662\ (\pm 0.092)\ \text{CIC0} + 0.650\ (\pm 0.078)\ \text{Ui} - 3.496\ (\pm 0.346)\ \text{ARR} \quad (4)$$

$$\text{range: } 3.02 \text{ vol. \%} \leq \text{UFL} \leq 16.5 \text{ vol. \%}$$

$$R^2 = 0.904, Q^2_{\text{LMO}} = 0.896, s = 0.638, n = 222$$

where n is the number of compounds used in the model and s is the standard error of the model.

Equation 4 shows that the UFL can be predicted using four molecular descriptors. The physical meanings of these descriptors are presented as following:

PW5 is a topological descriptor related to molecular shape, such as the molecular geometry and molecular

Table 7. Descriptors Selected for the Presented Model for Prediction of UFL

Descriptor	Type	Definition
PW5	Topological descriptors	Path/walk 5—Randic shape index
CIC0	Information indices	Complementary information content (neighborhood symmetry of 0 order)
Ui	Molecular properties	Unsaturation index
ARR	Constitutional descriptors	Aromatic ratio

Table 8. The Main Statistical Parameters of the Developed Model (Eq. 4)

Statistical Parameters	Training Set	Test Set	The Whole Dataset
Squared correlation coefficient for fitting (R^2)	0.904	0.861	0.898
Squared correlation coefficient for LMO CVs (Q^2_{LMO})	0.896	–	–
AAE	0.488	0.503	0.491
RMSE	0.637	0.684	0.647
No. compounds (n)	222	56	278

branching. CIC0 is an information index, which is related to the differences in the atomic distribution and the molecular dimension of the considered molecules. It encodes the molecular complexity and quantifies the heterogeneity and redundancy of topological neighborhoods of atoms in molecules. Ui is a molecular property, which encodes the degree of molecular unsaturation. ARR is a constitutional descriptor obtained as a function of the ratio between number of aromatic bonds and the total bonds in the H-depleted molecule. It can roughly be related to the flexibility of the structure.[20]

All the four descriptors selected are mainly related to the dimensional features of the molecules like molecular shape, complexity, unsaturation, and flexibility. Therefore, the overall UFL property of hydrocarbons can be reasonably explained by their steric effects.

The developed model (Eq. 4) is then used to predict the UFL values of 56 hydrocarbons in the test set for external validation. The predicted UFL values are obtained and presented as the Supporting Information. The main statistical parameters of the model are shown in Table 8. A plot of the predicted UFL values vs. the observed ones for both the training and test sets is shown in Figure 7.

The obtained results presented in Table 8 showed that the prediction errors of Eq. 4 were as low as possible. Also, both the AAE and RMSE values were not only low but also as similar as possible for the training and test sets, which suggests that the proposed model also has both predictive ability and generalization performance to a certain extent.

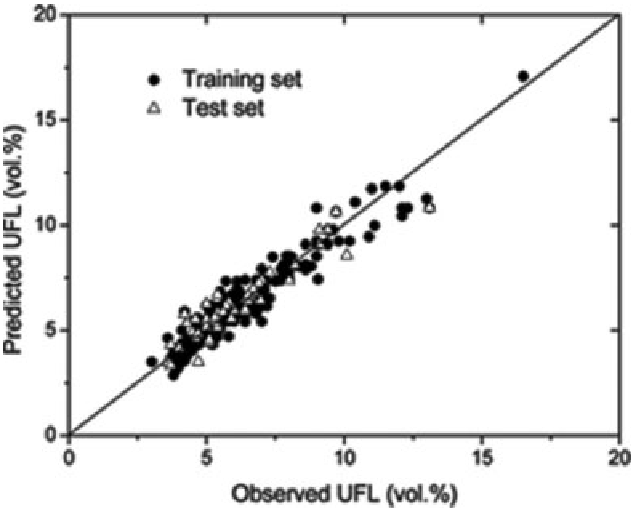

Figure 7. Correlation between the predicted and observed UFL values for both the training and test sets.

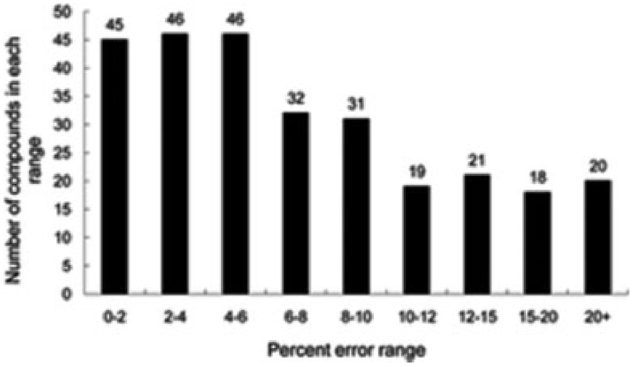

Figure 8. Percent errors of predicted UFL by Eq. 4, and the number of compounds in each range.

The predicted percentage error of all the 278 compounds was also calculated. The obtained average percentage error for these compounds was 7.90%, while the maximum percentage error was 40.66%. The results were shown in detail in Figure 8. A few compounds are above the 20% error range, as can be seen in Figure 8. For these compounds, the model prediction errors are higher and should be used with caution.

Model stability validation and results analysis

All the presented four models were tested for chance correlation to further analyze the model stability. The Y-randomization test method was used, which is a widely used technique to ensure the robustness of a QSPR model. In this test, the dependent-variable vector (Y vector) is randomly scrambled, and a new QSPR model is developed using the original independent-variable matrix. This process is repeated 50–100 times. It is expected that the resulting QSPR models should generally have low R^2 and low Q^2_{LMO} values. It is likely that sometimes high R^2 and Q^2_{LMO} values may be obtained due to a chance correlation. If all QSPR models obtained in the Y-randomization test have relatively high R^2 and Q^2_{LMO}, it implies that an acceptable QSPR model cannot be obtained for the given dataset by the current modeling method.[28] In this study, the Y-randomization test was performed on the training set for each flammability property. As expected, all the models generated had

Table 9. The Obtained Maximum R^2 and Q^2_{LMO} Values of the Generated Models vs. the Ones of the Presented Models for Each Flammability Property

	The Generated Models		The Presented Models	
Flammability Characteristics	Maximum R^2	Maximum Q^2_{LMO}	R^2	Q^2_{LMO}
FP	0.055	0.030	0.989	0.988
AIT	0.117	0.031	0.873	0.860
LFL	0.078	0.041	0.966	0.951
UFL	0.061	0.035	0.904	0.896

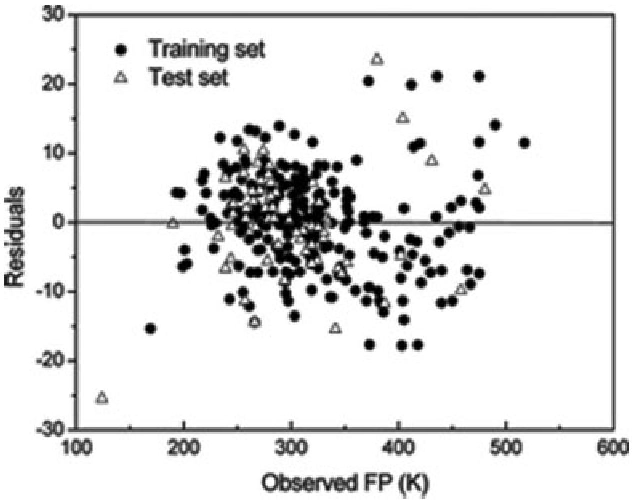

Figure 9. Plot of the residuals vs. the observed FP values for the presented model (Eq. 1).

produced low R^2 and low Q^2_{LMO} values. The obtained maximum R^2 and Q^2_{LMO} values of the generated models for each property were presented in Table 9.

As can be seen from Table 9, for each property, the obtained maximum R^2 and Q^2_{LMO} values were much lower than the ones calculated when the dependent variables were not scrambled. It can be, thus, concluded that only the correct dependent variables can be used to generate reasonable models and the chance correlation had little or no effect in the presented four models.

Also, the residuals of the predicted values of FP, AIT, LFL, and UFL vs. the observed ones for the developed models were shown in Figures 9–12, respectively. For each property, because most of the calculated residuals are distributed on both sides of the zero line, one may conclude that there is no systematic error in the development of the presented models.

All the results discussed above showed that all the four presented models are valid models and can be effectively used to predict the flammability characteristics of hydrocarbon compounds.

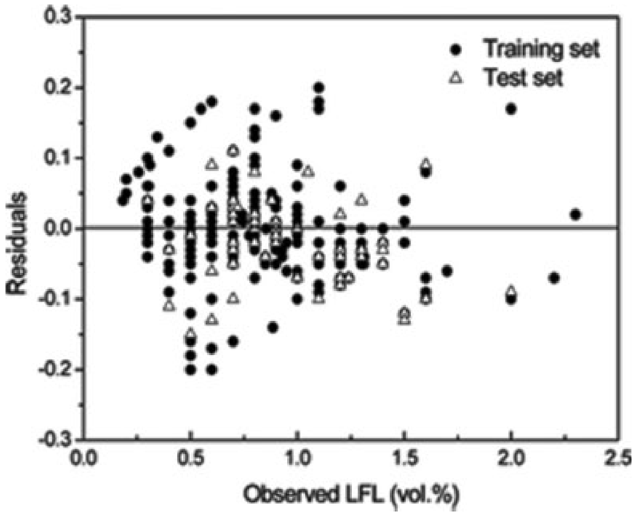

Figure 11. Plot of the residuals vs. the observed LFL values for the presented model (Eq. 3).

This work demonstrates that the complex flammability properties can be modeled by QSPR approach using only theoretical descriptors, derived solely from the molecular structures. The results showed that the prediction ability of the QSPR models are excellent, which can predict the flammability characteristics of hydrocarbons with accuracies that can approach the accuracies of experimental determinations. Considering the limited number of experimental data available and the complex nature of the flammability characteristics, it was not possible to further improve the models predictions beyond the current results.

Also, the QSPR approach can learn about the inherent relationships between the flammability characteristics and the molecular structures and can probe the structural characteristics that have significant contribution to each property of

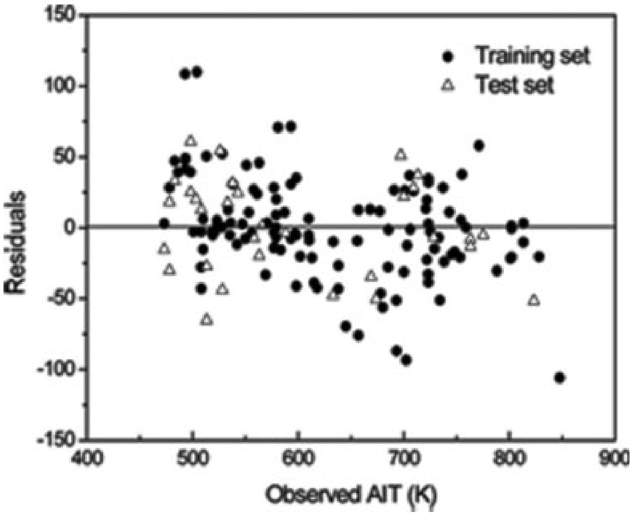

Figure 10. Plot of the residuals vs. the observed AIT values for the presented model (Eq. 2).

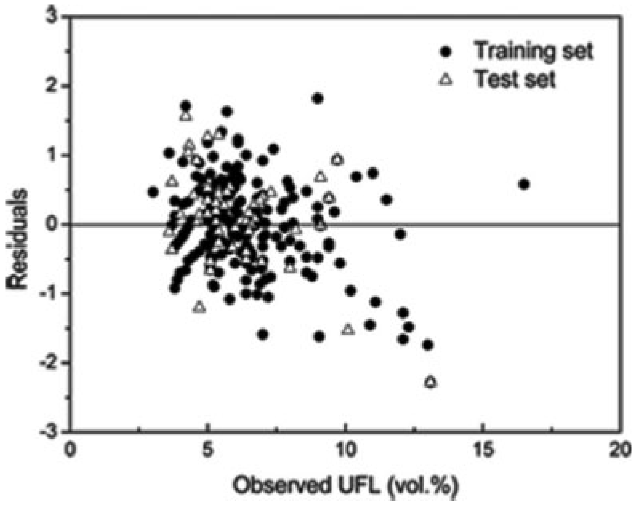

Figure 12. Plot of the residuals vs. the observed UFL values for the presented model (Eq. 4).

Table 10. Comparisons of the Results Obtained by the SGC Method and by This Work for the Whole Dataset

Property		R^2	*AAE*	Maximum Error	Average % Error	Maximum % Error	*n*	Property Range
FP	Albahri[2]	0.980	5 K	35 K	1.68	13.8	287	165–511 K
	This work	0.988	5.41 K	25.48 K	1.76	20.55	314	169.15–517 K
AIT	Albahri[2]	0.846	28 K	166 K	4.2	31	131	473–828 K
	This work	0.882	25.65 K	109.94 K	4.24	22.02	153	473.15–847.59 K
LFL	Albahri[2]	0.865	0.04 vol %	5.6 vol %	4.1	85	454	0.11–6 vol %
	This work	0.967	0.043 vol %	0.2 vol %	6.07	39.15	354	0.185–2.3 vol %
UFL	Albahri[2]	0.922	1.25 vol %	16 vol %	11.8	65	464	4–100 vol %
	This work	0.898	0.491 vol %	2.28 vol %	7.90	40.66	278	3.02–16.5 vol %

the molecules, such as molecular steric, electrostatic, skew, and inductive effects, which are usually unknown.

Furthermore, the QSPR approach is based on the molecular structure, which is always known. Once properly developed, the QSPR models can provide predictions of flammability characteristics quickly and conveniently. As can be seen from Eqs. 1–4, all the descriptors selected in the developed models are 2D descriptors, which could also be calculated from the Simplified Molecular Input Line Entry System. Also, these descriptors are freely accessible from the Milano Chemometrics and QSAR research group website (http://michem.disat.unimib.it/mole_db/).

However, attention should be paid for the applicability range of the presented models, because only compounds with defined range of flammability characteristic values have been used to develop the models. In this article, for the presented model of FP prediction, the applicability range is qualitatively defined as the chemicals with FP values between 169.15 K and 517 K. That is to say, only the predictions for chemicals with FP values between 169.15 K and 517 K can be considered reliable and not model extrapolations. The properties of AIT, LFL, and UFL are the same. For the presented models of AIT, LFL, and UFL predictions, the corresponding applicability ranges are defined as the chemicals with AIT values between 473.15 K and 847.59 K, the chemicals with LFL values between 0.185 vol % and 2.3 vol %, and the chemicals with UFL values between 3.02 vol % and 16.5 vol %, respectively. The four developed models can be expected to reliably predict the FP, AIT, LFL, and UFL properties, respectively, only for the chemicals falling within the corresponding applicability ranges.

Comparison with previous works

To our best knowledge, there is no QSPR study available in the literature for predicting the flammability characteristics of hydrocarbons from the knowledge of only the molecular structures. Recently, Albahri[2] developed a few models to predict the flammability characteristics of hydrocarbons using a structural group contribution (SGC) method. These SGC models were reported to be able to predict the FP, AIT, LFL, and UFL of pure hydrocarbons with higher accuracy and can be applied with less difficulty using only the molecular structures. The reported results of these models on the whole dataset were summarized in Table 10.

In this article, for the purpose of verifying the validity of the presented QSPR approach, a general comparison between the presented work and the work of Albahri[2] was performed. Considering that these two works were carried out based on different dataset and different methods, it is suggested that not only the prediction results but also more other important characteristics of models should be taken into account and analyzed, such as the model applicability efficiency and applicability range. Consequently, a detailed comparison between the two works is presented as follows.

First, regarding the input parameters used in the prediction models, the SGC method of Albahri[2] used a set of 35, 20, 19, and 19 structural groups as input parameters for the models of FP, AIT, LFL, and UFL, respectively, while the presented QSPR approach used only several molecular descriptors (3–5 descriptors) as input parameters. The QSPR models are, thus, considered to be more simple and robust. Moreover, the descriptors used in the QSPR models have definite physical meanings, which are useful to probe the structure characteristics that have significant contribution to the overall flammability properties of hydrocarbons.

Regarding the statistical parameters of the prediction models, the comparison between the two works was presented in Table 10. As can be seen from Table 10, the presented models of AIT and UFL predictions by the proposed QSPR approach show obvious superiority over those by the SGC method. While for the FP and LFL predictions, the statistical parameters of the presented QSPR models were very close to those of the SGC models. Moreover, although the SGC models were developed based on larger number of compounds in the dataset than the presented QSPR models for the LFL (472 vs. 354) and UFL (475 vs. 278) predictions, much more compounds have been used in the test set for model external validation in this work for each property. The external test sets used by SGC method for model validation contained only 4%, 5%, 4%, and 3% compounds of the datasets for FP, AIT, LFL, and UFL predictions, respectively, whereas those used in this work all contained 20% (12 vs. 63 for FP, 7 vs. 31 for AIT, 18 vs. 70 for LFL, and 13 vs. 56 for UFL, respectively). This is to say, the presented QSPR models have been externally validated more rigorously and have the fewer probability of being suffered from chance correlation.

Finally, regarding the applicability efficiency and applicability range of the models, the SGC models are conceptually simple and easy to apply. However, the applicability range of the SGC method is too related to the studied dataset, and the chemicals with functional groups not included in those used for the model development will be out of the model applicability range and, thus, will not be predicted. For example, for the compound of "2, 6-dimethyloctane," the FP value cannot be obtained using the SGC model. In fact, for many other hydrocarbons like 2, 6-dimethyloctane, the FP values can also not be calculated by the SGC model. The reasons are analyzed and presented as follows. For the set of groups selected to represent the FP values of the dataset, 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 of Albahri,[2] the author had 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 FP calculation of hydrocarbons with long HC chains, such as 2, 6-dimethyloctane with a ">CH—" group in the sixth 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. Hence, the SGC method has obvious limitations in applications. Moreover, the SGC models also provide a weak ability in distinguishing the isomeric compounds. Meanwhile, for the proposed QSPR approach, because only theoretical descriptors derived solely from the molecular structure is involved, the presented QSPR models should theoretically be applicable to any hydrocarbons. For example, for the compounds of 2, 6-dimethyloctane" and 2,2,4,4,6,8,8-heptamethylnonane, the FP values of whichcan not be obtained using the SGC model. However, for the proposed QSPR approach, because the molecular structures of these compounds are well known, the involved descriptors can be directly calculated and the FP values of them can be successfully predicted. Thus, the presented QSPR models can be expected to reliably predict the flammability characteristics for any hydrocarbons falling within the applicability ranges discussed above.

Conclusions

In this work, four multilinear equations for the prediction of the flammability characteristics of pure hydrocarbons (FP, AIT, the LFLs, and the UFLs) were developed via QSPR studies. The parameters of the obtained models are molecular descriptors, which can be calculated from the knowledge of only the molecular structures. Model validation was performed to check the stability and predictive capability of the presented models. The results showed that the presented models are valid models and can be effectively used to predict the flammability characteristics of hydrocarbons with accuracies that can approach the accuracies of experimental determinations. A general comparison between the presented work and the literature work[2] was also performed. The results showed that the presented models possess some obvious superiority despite of the different composition of the studied datasets. Thus, it can be reasonably concluded that the proposed models would be expected to predict the flammability characteristics for new hydrocarbons or for other hydrocarbons for which experimental values are unknown. In addition, the presented models could also identify and provide some insight into what structural features are related to the FP, the AIT, and the LFL and UFL of hydrocarbons, respectively.

Acknowledgments

This research was supported by National Natural Science Fund of China (50774048), Research Fund for the Doctoral Program of Higher Education of China (200802910007), and Program for New Century Excellent Talents in University (NCET-05-0505).

Literature Cited

1. Chang YM, Lee JC, Chen JR, Liaw HJ, Shu CM. Flammability characteristics studies on toluene and methanol mixtures with different vapor mixing ratios at 1 atm and 150°C. *J Therm Anal Calorim.* 2008;93:183–188.
2. Albahri TA. Flammability characteristics of pure hydrocarbons. *Chem Eng Sci.* 2003;58:3629–3641.
3. Liaw HJ, Chen CT, Gerbaud V. Flash-point prediction for binary partially miscible aqueous-organic mixtures. *Chem Eng Sci.* 2008; 63:4543–4554.
4. AIChE/CCPS. *Guidelines for Engineering Design for Process Safety*. New York, NY: American Institute of Chemical Engineers, 1993.
5. ASTM. *ASTM Standard Test Method E659-78*. West Conshohocken: American Society for Testing and Materials, 2000.
6. Mitchell BE, Jurs PC. Prediction of autoignition temperatures of organic compounds from molecular structure. *J Chem Inform Comput Sci.* 1997;37:538–547.
7. ASTM *Annual Book of Standards*. Philadelphia, PA: American Society for Testing and Materials, 2002.
8. Taskinen J, Yliruusi J. Prediction of physicochemical properties based on neural network modeling. *Adv Drug Deliv Rev.* 2003; 55:1163–1183.
9. Vidal M, Rogers WJ, Holste JC, Mannan MS. A review of estimation methods for flash points and flammability limits. *Process Saf Progr.* 2004;23:47–55.
10. Pan Y, Jiang JC, Wang R, Cao HY, Cui Y. Predicting the auto-ignition temperatures of organic compounds from molecular structure using support vector machine. *J Hazard Mater.* 2009;164:1242–1249.
11. Brauner N, Cholakov GSt, Kahrs O, Stateva RP, Shacham M. Linear QSPRs for predicting pure compound properties in homologous series. *AIChE J.* 2008;54:978–990.
12. Cholakov GSt, Stateva RP, Shacham M, Brauner N. Prediction of properties in homologous series with a shortcut QS2PR method. *AIChE J.* 2007;53:150–159.
13. Bernazzani L, Duce C, Micheli A, Mollica V, Sperduti A, Starita A, Tiné MR. Predicting physical-chemical properties of compounds from molecular structures by recursive neural networks. *J Chem Inform Model.* 2006;46:2030–2042.
14. Brauner N, Shacham M, Cholakov GSt, Stateva RP. Property prediction by similarity of molecular structures-practical application and consistency analysis. *Chem Eng Sci.* 2005;60:5458–5471.
15. Shacham M, Brauner N, Cholakov GST, Stateva RP. Property prediction by correlations based on similarity of molecular structures. *AIChE J.* 2004;50:2481–2492.
16. Dyekjaer JD, Jonsdottir SO. QSPR models based on molecular mechanics and quantum chemical calculations, Part 2: Thermodynamic properties of alkanes, alcohols, polyols and ethers. *Ind Eng Chem Res.* 2003;42:4241–4259.
17. Katritzky AR, Maran U, Lobanov V, Karelson M. Structurally diverse quantitative-structure property relationship correlations of technologically relevant physical properties. *J Chem Inform Comput Sci.* 2000;40:1–18.
18. Katritzky AR, Lobanov VS, Karelson M. QSPR: the correlation and quantitative prediction of chemical and physical properties from structure. *Chem Soc Rev.* 1995;24:279–287.
19. Rowley RL, Wilding WV, Oscarson JL, Yang Y, Zundel NA. *DIPPR Data Compilation of Pure Chemical Properties Design Institute for Physical Properties*. Provo, Utah: Brigham Young University, 2006. Available at http//dippr.byu.edu.
20. Todeschini R, Consonni V. *Handbook of Molecular Descriptors*. Weinheim: Wiley-VCH, 2000.
21. Karelson M, Lobanov VS, Katritzky AR. Quantum-chemical descriptors in QSAR/QSPR studies. *Chem Rev.* 1996;96:1027–1043.

22. Todeschini R, Consonni V, Mauri A, Pavan M. *DRAGON User Manual*. Milano, Italy: Talete srl, 2006.
23. Holland JH. *Adaptation in Natural and Artificial Systems*. Ann Arbor: University of Michigan Press, 1975.
24. Leardi R, Boggia R, Terrile M. Genetic algorithms as a strategy for feature selection. *J Chemometr*. 1992;6:267–281.
25. Gramatica P. Principles of QSAR models validation: internal and external. *QSAR Comb Sci*. 2007;26:694–701.
26. Tetteh J, Suzuki T, Metcalfe E, Howells S. Quantitative structure-property relationships for the estimation of boiling point and flash point using a radial basis function neural network. *J Chem Inform Comput Sci*. 1999;39:491–507.
27. Suzuki T. Quantitative structure-property relationships for auto-ignition temperatures of organic compounds. *Fire Mater*. 1994;18:81–88.
28. Tropsha A, Gramatica P, Gombar VK. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb Sci*. 2003;22:69–77.

Manuscript received Apr. 2, 2009, and revision received Jun. 7, 2009.