Available online at www.sciencedirect.com
ScienceDirect
Journal of Molecular Graphics and Modelling 26 (2007) 529–536
Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

QSPR modeling of flash points: An update

Alan R. Katritzky [a,*], Iva B. Stoyanova-Slavova [a], Dimitar A. Dobchev [a,c], Mati Karelson [b,c]

[a] Center for Heterocyclic Compounds, Department of Chemistry, University of Florida, Gainesville, FL 32611, USA
[b] Institute of Chemistry, Tallinn University of Technology, Ehitajate Tee 5, Tallinn 19086, Estonia
[c] Department of Chemistry, University of Tartu, Jakobi Street 2, Tartu 51014, Estonia

Received 21 December 2006; received in revised form 16 March 2007; accepted 19 March 2007
Available online 23 March 2007

Abstract

Quantitative structure–property relationship (QSPR) models for the flash points of 758 organic compounds are developed using geometrical, topological, quantum mechanical and electronic descriptors calculated by CODESSA PRO software. Multilinear regression models link the structures to their reported flash point values. We also report a nonlinear model based on an artificial neural network. The results are discussed in the light of the main factors that influence the property under investigation and its modeling.

Keywords: Flash point; QSPR; BLMR; CODESSA PRO; Artificial neural networks

1. Introduction

The flash point is the lowest temperature at which a liquid produces flammable vapor near its surface that ignites spontaneously when brought in contact with air and a spark or a flame [1].

Flash point is one of the most important and widely used flammability characteristics of liquids and low-melting substances. It provides a simple, convenient index of the flammability and combustibility of substances and is of importance, since it gives the knowledge needed for the handling and transporting of the compound in bulk quantities. It is a subject of interest in terms of understanding the fundamental chemical and physical processes in combustion chemistry.

Experimental flash point data are useful, but due to modern synthetic technology and virtual compound design, there is often a significant gap between the demand for such data and their availability. Moreover, for some toxic, explosive, or radioactive compounds the experimental determination of flash point can be very difficult. Hence, a reliable theoretical method for estimating flash points is desirable.

In our previous work [2b], we developed a prediction of flash points (FP) based on a multi-parameter regression technique of the CODESSA PRO [3] computer program, which utilizes solely descriptors calculated from chemical structure (eliminating the need for experimentally determined descriptors) and therefore can be used for the prediction of the flash points of unavailable or unknown compounds. Our previous work [2b], utilized 271 experimental flash points collected from the Acros and Aldrich catalogs to build several QSPR models. It was shown [2b] that the best correlations all included the boiling point (BP) as a descriptor; however, in place of experimentally determined BPs we used QSPR predicted BPs, as estimated by equation 1 in Ref. [2b].

$$\begin{aligned} T_{\mathrm{BP}} &= (64.87 \pm 1.46)\sqrt[3]{G_{\mathrm{b}}} + (44.92 \pm 4.46)\mathrm{HACA} \\ &\quad + (0.26 \pm 0.035)\Delta H_{\mathrm{f}} + (563.3 \pm 74.55)\mathrm{FHBCA} \\ &\quad + (45.95 \pm 7.70)N_{\mathrm{b}} - (152.40 \pm 12.30), \\ R^2 &= 0.892, R^2_{\mathrm{cv}} = 0.883, F = 436.87, s = 23.03 \end{aligned} \quad (1)$$

In Eq. (1), $\sqrt[3]{G_{\mathrm{b}}}$ is the cubic root of the gravitational index, HACA the hydrogen acceptor charged area relative to the total molecular surface area, T_{BP} the predicted boiling point, FHBCA the fractional hydrogen bond charged area, N_{b} the number of triple bonds and ΔH_{f} is the AM1-calculated heat of

* Corresponding author. Tel.: +1 352 392 0554; fax: +1 352 392 9199.
E-mail address: katritzky@chem.ufl.edu (A.R. Katritzky).

1093-3263/$ – see front matter
doi:10.1016/j.jmgm.2007.03.006

formation of the molecule.

$$T_{FP} = (0.67 \pm 0.014)T_{BP,pred} + (3.45 \pm 0.27)\text{DPSA} + (0.95 \pm 0.17)E_{e\text{-}n,C} - (161.8 \pm 31.07),$$
$$R_2 = 0.924, R^2_{cv} = 0.9217, F = 1086.59, s = 14.15 \quad (2)$$

In Eq. (2), T_{FP} is the flash point temperature, T_{BP} the predicted boiling point based on Eq. (1), DPSA the difference in charged partial positive surface area of the positively charged atoms and for the negatively charged atoms in the molecule, and $E_{e\text{-}n,C}$ is the minimum electron attraction for a C atom.

The statistical characteristics of (2) allow the FP to be predicted with an error of about 14 K, which is relatively small for such a large (271) and diverse data set. This model does not require experimental measurements or parameters to be used for prediction.

Zefirov and coworkers [2a] used fragmental descriptors in two approaches for the QSPR modeling of the FP of 525 compounds, namely MLR and artificial neural networks. The datasets are divided in three databases from which were defined the training and test sets. They developed seven 9-descriptor MLR models with R^2 parameter between 0.872 and 0.935. In addition, they also developed an ANN model with architecture 25-2-1 which provided $R^2 = 0.959$ for the training set and RMS = 14.6 C.

Later, Catoire and Naudet developed an equation (equation 4a therein) [4] which was claimed to predict FP with a maximum absolute deviation of 10 °C, from three parameters: the experimental normal boiling point, the enthalpy of vaporization at 298.15 K and the number of carbon atoms in the molecule. However, the technique used to derive this equation was not mentioned. Equation 4a is limited to predictions in the range −100 to 200 °C. This study also stressed the need for reliable experimental values of the flash points.

The main goal of the current study is to build a reliable predictive QSPR model from a diverse set of a large number of organic compounds comprising reliable experimental data. Thus, the present experimental data set is considerably expanded compared with our previous work [2b]. In addition to the multilinear models we also developed nonlinear QSPR based on artificial neural networks (ANN).

2. New data set

In the present study we collected 758 experimental flash points from the literature, mainly data published after 2004, which are judged to be more reliable than those used in our previous work [2b]. The experimental flash point values utilized are shown in Table S1. These were taken from the following sources:

(1) International Chemical Safety Cards (ICSCs) on the internet [5].
(2) CRC Handbook of Chemical Physics and Physical Chemistry or other handbooks [6,7].
(3) US Bureau of Mines now Pittsburgh Research Center reports and compilations [8,9].
(4) Chemical manufacturer's MSDSs [10,11].
(5) Physical and Theoretical Chemistry Laboratory PTCL Oxford University chemical and other safety information [12].
(6) National Institute for Occupational Safety and Health (NIOSH) Pocket Guide NPG to Chemical Hazards [13].
(7) NFPA publication "Fire Protection Guide to Hazardous Materials" [14].

Most of the experimental data that we used were collected and reported in Ref. [4]. It is important to note that the FP values cited by ICSCs and published UNEP/ILO/WHO/EU are considered as a major source of reliable data since the data are updated when necessary. Thus, our experimental data were extended from the 271 used in Ref. [2b] to 758. Consequently, for QSPR modeling of the FP we chose the more recent experimental data [see supplementary material].

Dealing with such a large amount of data raises issue of reproducibility. Reproducibility of the experimental data is essential for dependable QSPR modeling and thus reliable data are needed. From our previous data in Ref. [2b], 134 compounds overlapped with the new data collected, and as shown in Fig. 1 the overlapped compounds frequently possess different experimental FP values. This difference can influence model correctness and prediction. Although the correlation coefficient between the previous data and the new set is high ($R^2 = 0.960$) for the 134 compounds common to both, some of the experimental discrepancies are greater than 20 K.

In all such cases we used the values provided in the new sources reported above since we considered them more reliable. During the reproducibility calculations, we attempted to verify of our previous model by obtaining exactly the same results as reported in Ref. [2b] for the same data. This verification procedure consisted of repeating the same calculations with the same software [15]. The calculations were performed in two ways, namely (i) individual calculations with the same

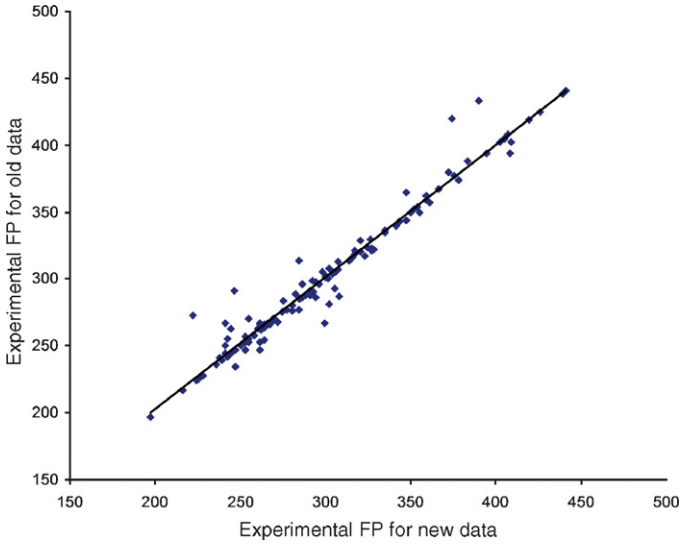

Fig. 1. Linear dependence between the experimental FP for the overlapped compounds.

structures used in Ref. [2b] and (ii) calculations carried out based on the original data storage used in Ref. [2b]. Thus, for equation 2 in Ref. [2b] we selected the same descriptors reported therein and rebuilt the models. However, our present final results from predictions (i) and (ii) do not match those reported in Ref. [2b]. The R^2 values for the flash points we now obtained were: (i) $R^2 = 0.841$, (ii) $R^2 = 0.863$ whereas 0.924 was reported in Ref. [2b]. We believe that a main reason for these discrepancies could be hidden in the descriptor values. In Eq. (2) the BP descriptor is related to other descriptors via Eq. (1), and we therefore also checked Eq. (1) descriptors. A comparison of the descriptor values for each compound indicated that they differ according calculation procedure (i). For calculation procedure (ii) we compared the reported descriptor ranges in Ref. [2b] with those included in their original storage. The differences in the ranges for these descriptors were significant. Two possible reasons for the discrepancies in our present results as compare with those of Ref. [2b] are: (a) some wrongly reported descriptor values in Ref. [2b], proved by calculations (i) and (ii) and (b) errors of typographical or general character (mistaken descriptor names or general errors in the calculations). The attempt to verify the above investigation motivated us to build a completely new QSPR model based on the extended experimental data. The statistical characteristics between the current QSAR model and the previous developed model are compared in Table 1. As described in Table 1, we also extended the QSAR investigation with nonlinear modeling developed by artificial neural networks (ANN).

3. Methodology

3.1. Molecular geometry

Three-dimensional conversions and pre-optimization were performed for all molecules using the molecular mechanics (MM+) implemented in the HyperChem 7.5 package [16].

Final geometry optimization of the molecules was carried out using the semi-empirical quantum-mechanical AM1 parameterization [17]. The optimized geometries were loaded into CODESSA PRO software. Overall, more than 1000 theoretical descriptors were calculated. These descriptors can be classified into several groups: (i) constitutional, (ii) topological, (iii) geometrical, (iv) thermodynamic, (v) quantum chemical and (vi) charge-related descriptors.

3.2. Multilinear approach

An important stage of the multilinear regression QSAR methodology is the search for the best multilinear equation among a given pool of descriptors. In other words, Eq. (3) correlates the dependent variable (T_{FP}) with a certain number n of molecular descriptors (D_i) weighted by the regression coefficients b_i:

$$T_{FP} = b_0 + \sum_{i=1}^{n} b_i D_i \quad (3)$$

Table 1
Comparison between our previous and current models

Number of descriptors	Descriptor name	R^2	R^2_{cv}	s^2	Number of compounds	Comments
Previous work (MLR)						
3	BP_{pred}, d1; DPSA, d2; $E_{e\text{-}nC}$, d3	0.924	0.922	14.15	271	Heuristic method was used
Current work (MLR)						
4	BP, d1; HA dependent HDCA-1/TMSA (Zefirov PC), d2; HASA-1/TMSA (Zefirov PC) (all), d3; Relative number of triple bonds, d4	0.849	0.846	18.9	758	BMLR technique was used. Model obtained after reproducibility investigation of the results and addition of new compounds

Number of descriptors	Descriptor name	R^2	Relative RMS training set	Relative RMS validation	Number of compounds	Comments
Current work (ANN)						
4	BP, d1; HA dependent HDCA-2/SQRT(TMSA) (Zefirov PC), d2; HASA-1/TMSA (Zefirov PC) (all), d3; Balaban J index (based on topological distance), d4	0.878	0.670	0.940	600	Heuristic method was used. For selection of the input units 158 compounds were used for validation

The Best Multilinear Regression method (BMLR) [18] encoded in CODESSA PRO software was used to select significant descriptors for building multilinear QSPR models. The treatment started with the reduction of the number of molecular descriptors. If two descriptors were highly correlated, then only one descriptor was selected; descriptors with insignificant variance were also rejected. This helps to speed up the descriptor selection and reduces the probability of including irrelevant descriptors by chance.

The strategy used to develop physically meaningful multilinear QSAR equations from the very large pool of descriptors is a combination of the multilinear regression and forward selection procedures, discussed elsewhere [18–20,30].

A major decision in developing successive QSAR is when to stop adding descriptors to the model during the stepwise regression procedure. A simple technique to control the model expansion is the so-called ''breaking point'' in the improvement of the statistical quality of the model, by analyzing the plot of the number of descriptors involved in the models obtained versus the squared correlation coefficient R^2 (or/and cross-validation correlation coefficient R^2_{cv}) values corresponding to those models. Frequently, the statistical improvement of the regression model is less significant ($\Delta R^2 < 0.02$–0.05) after a certain number of independent variables in the model (''breaking point''). Consequently, the model corresponding to the breaking point is considered the best/optimum model.

To validate the models internally, the parent data set was divided into three subsets (*a*, *b* and *c*): the 1st, 4th, 7th, etc. data points go into the first subset (*a*), the 2nd, 5th, 8th, etc. into the second subset (*b*), and the 3rd, 6th, 9th, etc. into the third subset (*c*). Then, three training sets *A*, *B* and *C* were prepared as the combinations of two subsets (*a* and *b*), (*a* and *c*), and (*b* and *c*), respectively. The remaining subsets (*c*, *b* and *a*, respectively) become the corresponding test sets.

For each of the training sets the correlation equation was derived with the same descriptors. Then, the equation obtained was used to predict FP values for the compounds from the corresponding test set.

The efficiency of QSAR models to predict FP value was estimated using the cross-validation (leave-one-out method) [21].

3.3. Nonlinear approach

Artificial neural networks (ANNs) [22–24] have become an important modeling technique for QSAR and QSPR. They have been applied in numerous application areas of chemistry and pharmacy [25–27]. The mathematical adaptability of ANN commends them as a powerful tool for pattern classification and building predictive models. A particular advantage of ANNs is their inherent ability to incorporate nonlinear dependencies between the dependent and independent variables without using an explicit mathematical function.

In this work, a back-propagation fully connected network was developed and used to obtain a nonlinear QSPR model. For precise mathematical definition of the back-propagation ANN the reader is referred to [28,29]. Topologically, it consists of input, hidden, and output layers of neurons or units connected by weights. Each input layer node corresponds to a single independent variable (molecular descriptor). Similarly, each output layer node corresponds to a different dependent variable (property under investigation).

In this work one fourth of the data were used as a validation set in order to control and monitor the ANN prediction. These data were not used in the training process of the network.

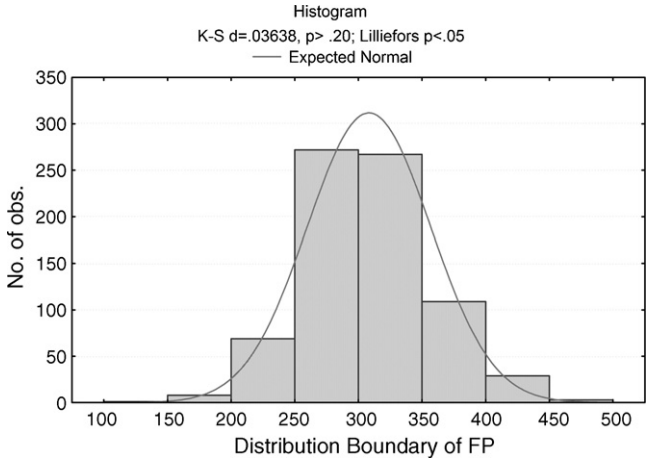

Fig. 2. Distribution of the experimental flash point values.

4. Results and discussion

4.1. Multilinear QSPR model based on 758 compounds

In order to build statistically reliable QSPR model, a large number of compounds and a good algorithm for descriptor selection when exploring a huge descriptor space are needed. We used the Best Multilinear Regression method (BMLR) as implemented in CODESSA PRO to obtain the main QSPR model shown in Table 2. In addition, the distribution of the experimental FP values should be close to the normal one. As can be seen from Fig. 2 this requirement is fulfilled for our 758 data points.

Table 2
The main multilinear QSPR model obtained for 758 organic compounds

No.	X	$\pm\Delta X$	t-Test	R^2	R^2_{cv}	Descriptor
0	30.185	4.490	6.72238			Intercept
1	0.651	0.011	60.643	0.633	0.631	BP, d1
2	4948.53	194.41	25.454	0.799	0.797	HA dependent HDCA-1/TMSA (Zefirov PC), d2
3	60.704	5.467	11.103	0.830	0.827	HASA-1/TMSA (Zefirov PC) (all), d3
4	300.245	30.508	9.842	0.849	0.846	Relative number of triple bonds, d4

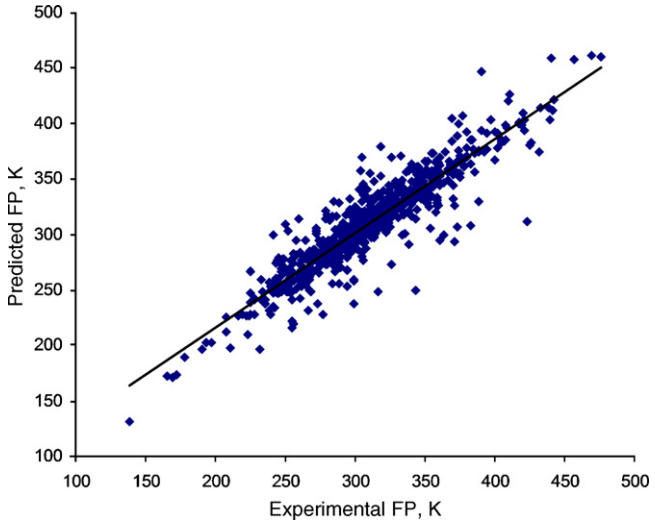

Fig. 3. Linear fit between experimental and predicted FP according to the multilinear model in Table 2.

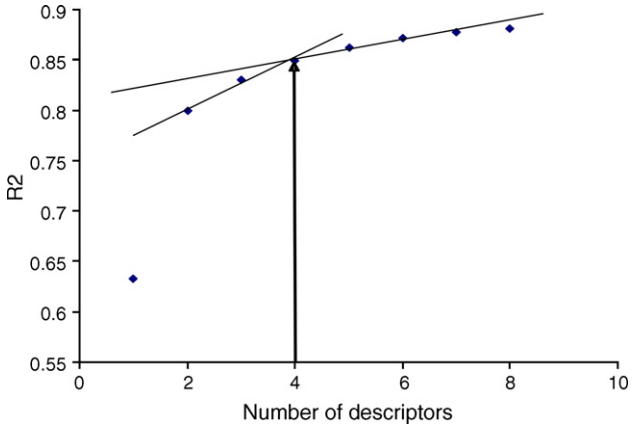

Fig. 4. Defining of the optimum number of descriptors based on the breaking point rule.

By using the BMLR, we obtained a model that possesses good statistical characteristics bearing in mind the large number of compounds (758) i.e. $R^2 = 0.849$, $R^2_{cv} = 0.846$, $F = 1058.3$, $s = 18.9$ K and the average absolute error of 13.9 K. In these notations R^2, R^2_{cv}, F and s are the coefficient of determination, the squared cross-validation correlation coefficient (leave-one-out method), Fisher criterion and the standard deviation of the model, respectively. A graphical presentation of the relationship between experimental and predicted flash points of the model is shown in Fig. 3. In Table 2, X and ΔX are regression coefficients of the QSPR equation and their standard errors, respectively.

An important step for model building was to define the number of independent variables in the main QSPR equation. This step ensures the over-parameterization of the model and prevents to some extent the chance correlations between the descriptors. As described in Section 3 this procedure is based on the breaking rule showing the critical improvement of the R^2 (and R^2_{cv}) over the number of the descriptors added in the equation. Based on the BMLR method we built consecutively several equations with different number of descriptors up to seven independent variables. Then, it was indicated that the maximum improvement of R^2 was at four descriptors (breaking point Fig. 4).

A main issue for a QSPR model is the statistical significance of the descriptors included. According to the t-test of the descriptors in Table 2 the significance (at probability level $p < 0.005$) of the descriptors is in the following order d1 > d2 > d3 > d4. The statistical significance is connected to the relative errors of the regression coefficients of the respective descriptors, i.e. the bigger the t-test value the less the relative error of the regression coefficient.

For building the multilinear QSPR model, we included the calculated boiling point (BP) of the compounds as an external descriptor. These calculated boiling points were obtained by using Eq. (4) developed in our earlier work [18] (model in Table 3 therein), based on only two descriptors and developed using 298 experimental data with $R^2 = 0.954$, i.e.

$$\mathrm{BP\,(K)} = -170.7 + 65.8\sqrt[3]{G_I} + 18470\mathrm{HDSA}(2) \quad (4)$$

In Eq. (4), $\sqrt[3]{G_I}$ is the cubic root of the gravitational index, and HDSA is the area-weighted surface charge of the hydrogen-bonding donor atoms. We used (4) due to its reliability and high predictive capability. Also, based on our experience and knowledge, we expected that the BP would be a significant descriptor for the FP (as can be seen from Table 2), since similar processes may govern the boiling point and the flash point phenomena.

This conjecture was also supported by other authors [4]. Therefore, in order to investigate more precisely the FP and BP connection we constructed several functions from BP as exponential, reciprocal, cubic root, squared root and thus included in the whole CODESSA PRO descriptor space. As can be seen from Table 2, the BMLR algorithm reveals that the BP descriptor is the most significant variable according to the t-test.

The second statistically significant descriptor is HA dependent HDSA-2/SQRT(TMSA) (Zefirov PC), d2. This descriptor accounts for the area weighted surface charge of hydrogen bonding donor atoms HDSA2 and is defined by the

Table 3
Internal three-fold validation of the multilinear model in Table 2

Training set	N	R^2 (fit)	R^2_{cv} (fit)	S (fit)	Test set	N	R^2 (pred.)	s (pred.)
A + B	506	0.872	0.871	17.5	C	252	0.806	17.64
A + C	506	0.853	0.843	18.3	B	252	0.814	18.02
B + C	504	0.858	0.849	18.6	A	254	0.803	18.57
Average		0.861	0.854	18.13			0.808	18.08

following equation:

$$\mathrm{HDSA}(2) = \sum_{D} \frac{q_D \sqrt{s_D}}{\sqrt{S_{\mathrm{tot}}}}, \quad D \in H_{\text{H-donor}} \tag{5}$$

where S_D is the solvent accessible surface area of H-bonding donor atoms, q_D the partial charge on H bonding donor atoms, and S_{tot} is the total solvent accessible molecular surface area [30]. This descriptor accounts for the hydrogen bonding ability of the molecule. Quite naturally, the larger the hydrogen bonding ability of the compounds, the higher the flash point temperature, as indicated by the respective value of the respective regression coefficient in our QSPR model (Table 2).

The next significant descriptor in the main QSPR model of Table 2 is HASA-1/TMSA (Zefirov-all), d3. The descriptor [31] is defined by the following equation:

$$\mathrm{d3} = \frac{\mathrm{HASA\text{-}1}}{\mathrm{TMSA}} \tag{6}$$

where TMSA (Zefirov) is the total solvent-accessible molecular surface area and HASA-1 is given as follows:

$$\mathrm{HASA\text{-}1} = \sum_{A} S_A \tag{7}$$

In (7) S_A is the solvent-accessible surface area of H-bonding acceptor atoms. The sum is carried out over all atoms. This descriptor is connected to the hydrogen bonding-acceptor ability of the molecule.

The last descriptor in the model of Table 2 is the relative number of triple bonds. It was noted that most of the molecules with triple bonds were outliers. Hence, this descriptor appears as a correction to the equation regarding the compounds with triple bonds.

4.2. Internal validation of the multilinear QSPR model

The efficiency of QSPR models to predict FP was also estimated using the internal three-fold cross-validation. The procedure is described in Section 3. The results of validation are shown in Table 3. As can be noted from this table, the average values of the statistical parameters are quite similar and close to the parameters of the multilinear model in Table 2. This fact shows that the model is statistically stable and its predictive power reserves with such division of the data.

4.3. Nonlinear QSPR model

In this study, we also used artificial neural network (ANN) methodology for prediction of the FP values. In this attempt we pursued maximum predictivity of the flash point. Thus, it was possible to build a general nonlinear QSPR model based on all the experimental data. To do this, all the experimental data (758 data points) for FP were divided into training (600) and validation subsets (158). The validation set was used during the training stage in which the weights of the ANN were adjusted according to the output prediction error by using the back-propagation algorithm. The validation set error (and also R^2) was monitored in order to avoid the over-training of the ANN and to stop the training process.

The descriptors featured in input neurons for the ANN model were selected according to the Heuristic method encoded in CODESSA PRO. This algorithm selected a four-descriptor equation ($R^2 = 0.83$) whose descriptors were used as inputs. We constructed several ANN architectures trying to obtain the one showing lowest RMS and with the least as possible hidden units. Thus, we found that the architecture 4-3-1 is the best of nine.

After optimizing the ANN on the training set the root-mean-squared (relative RMS error) error for the training and validation data is 0.67 and 0.94, respectively. The predicted values of FP obtained are given in Table S1. Graphical presentation as a fitting plot for the training set is given in Fig. 5. Also, in Fig. 6 is given the corresponding plot for the validation set of 158 compounds.

The maximum squared correlation coefficient for the training set was $R^2 = 0.878$ for 600 experimental data points. The corresponding validation set (which was not used to train the ANN model) had a maximum $R^2 = 0.824$ at which the training of the network was stopped in order to avoid overfitting. The result for the validation set is shown in Fig. 6 where experimental and predicted values are linearly presented.

As can be seen from Table S1 and Figs. 3 and 5, the ANN model gave superior prediction over the multilinear QSAR model in Table 2. This conjecture is supported by the statistical (i.e. R^2 and the number of compounds) characteristics of the two models.

The ANN model included the following descriptors used as inputs BP, HA dependent HDCA-2/SQRT(TMSA) (Zefirov PC), HASA-1/TMSA (Zefirov PC) (all), and Balaban J index (based on topological distance) [30].

The descriptors in Table 2 and the ANN model show considerable similarity since the Heuristic and the BMLR are

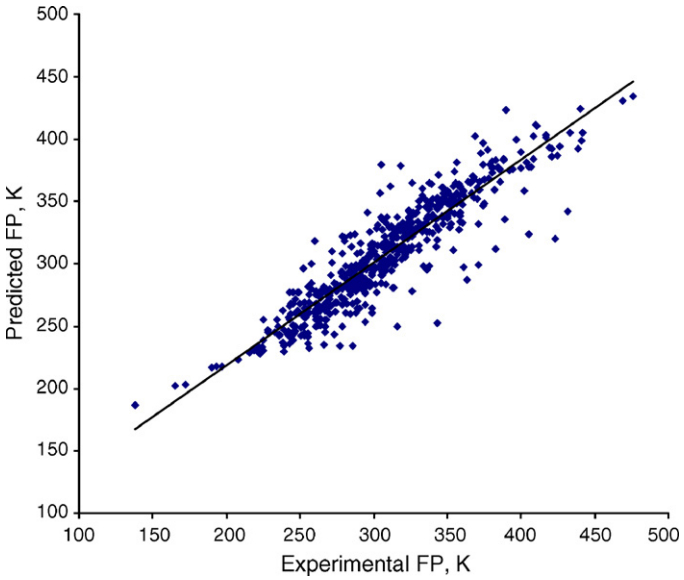

Fig. 5. Experimental and predicted FP values according to the ANN model for the training set of 600 compounds.

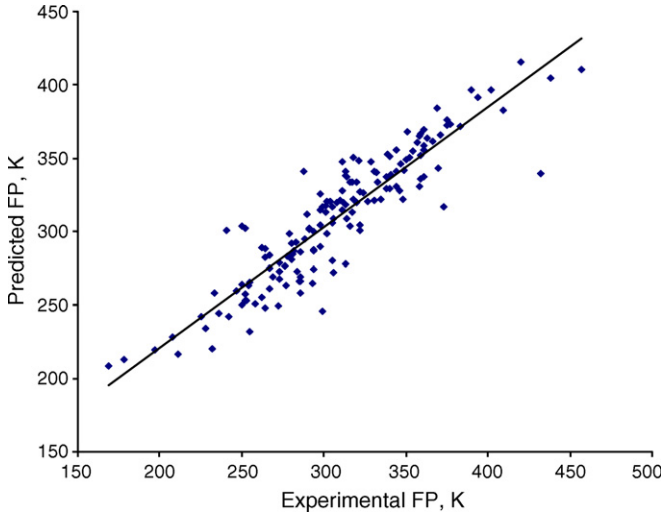

Fig. 6. Experimental and predicted FP values according to the ANN model for the validation set of 158 compounds.

similar techniques for the selection of best descriptors for a given data set. However, the ANN model has better predictive ability of FP due to the nonlinear nature of the backpropagation algorithm.

5. Conclusions

Our current attempt to correlate the FP of various classes of compounds with theoretically calculated molecular descriptors has led to good QSPR models that relate this chemical property to structural characteristics. Notably, all descriptors appearing in the main four-parameter multilinear regression equation as well as those employed in the ANN model have been derived from theoretical molecular calculations. Also, the inclusion of a theoretically calculated external descriptor BP improves the correlation with the FP. Thus, to make a prediction of FP, the experimental boiling points of the investigated compounds are not required.

The current computational power available for chemical research allows the calculation of molecular descriptors involved in the model for large data sets of 758 compounds in realistic time. Thus, in principle, the QSPR model developed in our present work can be used for the prediction of flash points for a wide range of organic compounds with an average error of 13.9 K. The descriptors appearing in this model also have physical meaning, being primarily related to electrostatic and hydrogen bonding interactions as well as to the molecular shape. Notably, our present model is much improved compared to the QSPR model for flash points developed mainly on the basis of more limited data [2b]. In addition to the multilinear model we have also used ANN methodology for better prediction. The ANN model gave better statistical characteristics ($R^2 = 0.978$ and average error of 12.6 K) for prediction.

A general comparison can be made between the current study and the work of Zefirov and coworkers [2a]. Regarding the four-descriptor MLR models, current equation is slightly better in terms of R^2 for four out of the seven equations in Ref. [2a]. In addition our model is based on larger number of compounds for the training sets (758 versus 269). However, the MLR models of work [2a] were externally validated with test sets whilst our model was only internally validated. Regarding to the ANN models, our model possesses a smaller number of adjustable parameters (weights and biases) than the model of work [2a], which is related to the generality of the ANN prediction. Also, the input neurons (descriptors) are less: four in the current, 25 in Ref. [2a] based on comparison of both architectures (current 4-3-1, work [2a] 25-2-1). The current ANN model is based on larger training (600 versus 266) and test sets (158 versus 131). However, the ANN model of Zefirov et al. is based on fragmental descriptors which describe better the structural features of the compounds important for the FP phenomenon and thus this model had better R^2 than the current one (0.959 versus 0.878).

In conclusion, both models, the multilinear and nonlinear, were validated by (i) three-fold internal validation and (ii) validation using an external control, respectively. Finally the results of the present work confirm the utility of the theoretical molecular descriptors and the QSPR model derived on their basis for the effective prediction of flash points. The QSPR model developed herein can serve as a first tool to determine and predict FP of unknown or not yet synthesized organic compounds.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2007.03.006.

References

[1] National Fire Protection Agency, Fire Protection Guide on Hazardous Materials, 10th ed., NFPA, Quincy, MA, 1991.
[2] (a) N.I. Zhokhova, I.I. Baskin, V.A. Palyulin, A.N. Zefirov, N.S. Zefirov, Fragmental descriptors in QSPR: flash point calculations, Russ. Chem. Bull. Int. Ed. 52 (9) (2003) 1885–1892;
(b) A.R. Katritzky, R. Petrukhin, R. Jain, M. Karelson, QSPR analysis of flash points, J. Chem. Inf. Comput. Sci. 41 (2001) 1521–1530.
[3] A.R. Katritzky, V.S. Lobanov, M. Karelson, CODESSA Reference Manual Version 2.0, Gainesville, 1996.
[4] L. Catoirea, V. Naudet, A unique equation to estimate flash points of selected pure liquids application to the correction of probably erroneous flash point values, J. Phys. Chem. Ref. Data 33 (2004) 1083–1111.
[5] http://www.inchem.org/pages/icsc.html.
[6] J.A. Dean (Ed.), Lange's Handbook of Chemistry, 13th ed., McGraw-Hill, New York, 1985.
[7] J.A. Dean, Handbook of Organic Chemistry, McGraw-Hill, New York, 1987.
[8] M.G. Zabetakis, U.S. Bureau of Mines Bulletin, vol. 114, 1965, p. 627.
[9] J.M. Kuchta, U.S. Bureau of Mines Bulletin, vol. 71, 1985, 680.
[10] Matheson TRI''GAS MSDS available at: http://www.matheson-trigas.-com/mathportal/msds/.
[11] Where to find MSDS On the Internet? at http://www.ilpi.com/msds/.
[12] The Physical and Theoretical Chemistry Laboratory Oxford University Chemical and Other Safety Information available at: http://physchem.ox.-ac.uk/MSDS.
[13] NIOSH Pocket Guide (NPG) to Chemical Hazards available at: www.cdc.gov/niosh/npg/npg.html.
[14] A.B. Spencer, G.R. Colonna (Eds.), NFPA Fire Protection Guide to Hazardous Materials, 13th ed., NFPA Publication, Quincy, 2002.

[15] A.R. Katritzky, V.S. Lobanov, M. Karelson, CODESSA, User's Manual, University of Florida & Tartu University, 2000.
[16] Hyperchem, v. 7.5, Hypercube Inc., Gainesville, FL.
[17] M.J.S. Dewar, E.G. Zoebisch, E.F. Healy, J.J.P. Stewart, J. Am. Chem. Soc. 107 (1985) 3902.
[18] A.R. Katritzky, L. Mu, V.S. Lobanov, M. Karelson, J. Phys. Chem. 100 (1996) 10400.
[19] A. Beteringhe, A. Balaban, QSAR for toxicities of polychlorodibenzofurans, polychlorodibenzo-1,4-dioxins, and polychlorobiphenyls, Arkivoc xii (2004) 1081.
[20] B. Lucic, I. Bašic, D. Nadramija, A. Milicevic, N. Trinajstic, T. Suzuki, R. Petrukhin, M. Karelson, A.R. Katritzky, Correlation of liquid viscosity with molecular structure for organic compounds using different variable selection methods, Arkivoc iv (2002) 45–59.
[21] M. Stone, P. Jonathan, Statistical thinking and technique for QSAR and related studies. 1. Genaral theory, J. Chemometr. 7 (1993) 455–475.
[22] S. Goll, P. Jurs, J. Chem. Inf. Comput. Sci. 39 (1999) 1081.
[23] J. Tetteh, T. Suzuki, E. Metcalfe, S. Howells, J. Chem. Inf. Comput. Sci. 39 (1999) 491.
[24] J. Zupan, J. Gasteiger, Neural Networks for Chemists: An Introduction, VCH–Verlag, Weinheim, 1993, pp. 213–228.
[25] J.A. Burns, G. Whitesides, Chem. Rev. 93 (1993) 2583.
[26] A.R. Katritzky, D.A. Dobchev, D.C. Fara, M. Karelson, Bioorg. Med. Chem. 13 (2005) 6598–6608.
[27] A.R. Katritzky, D.A. Dobchev, D.C. Fara, E. Hur, K. Taemm, L. Kurunczi, M. Karelson, A. Varnek, V.P. Solov'ev, J. Med. Chem. 49 (2006) 3305–3314.
[28] S. Haykin, Neural Networks. A Comprehensive Foundation, Pearson ed., 1999, pp. 156–256.
[29] T. Masters, Practical Neural Network Recipes in C++, Academic Press Inc., 1993, pp. 77–116.
[30] M. Karelson, Molecular Descriptors in QSAR/QSPR Analysis, Wiley & Sons, New York, 2000.
[31] D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci. 32 (1992) 306.