Full Articles

Fragmental descriptors in QSPR: flash point calculations

*N. I. Zhokhova, I. I. Baskin, V. A. Palyulin, A. N. Zefirov, and N. S. Zefirov**

Department of Chemistry, M. V. Lomonosov Moscow State University, Leninskie Gory, 119992 Moscow, Russian Federation. Fax: +7 (095) 939 0290. E-mail: zhokhova@org.chem.msu.ru; zefirov@org.chem.msu.ru

Flash points of various classes of organic compounds were first studied using the fragmental approach in the framework of QSPR methodology. Fragmental descriptor based regression and neural network models for flash point prediction are proposed.

Key words: QSPR methodology, flash point, organic compounds, fragmental approach, fragmental descriptors.

The flash point (T_f) is an important characteristic of the flammability properties of organic substances, especially liquids and low-melting compounds.[1—3] This parameter gives the knowledge necessary for understanding the fundamental physical and chemical processes of combustion. Moreover, it is of practical importance for chemical industry, because it characterizes the extent of safety in handling and transporting a given compound. The flash point is defined as the lowest temperature at which a substance can be ignited in air with an initiator.[1—3]

A number of reliable experimental methods for determining T_f values have been elaborated (*e.g.*, the open cup and closed cup methods[1,3]) with an accuracy of ±(5—8) °C [1] and ±10 °C.[2,3] The T_f values are known for many compounds;[3] however, they are not reported in the literature even for industrially important compounds and the corresponding reference sources are also often not provided. What is more, for toxic, volatile, explosive, and radioactive compounds the experimental determination of T_f values is often difficult. Hence the development of theoretical methods for evaluating the flash points is needed. Various schemes for T_f calculations were proposed, including the QSPR methodology.[2,3]

Earlier,[4—15] we have pointed out the advantages of the fragmental approach and widely used the fragmental descriptors for prediction of the physicochemical characteristics of various classes of organic compounds, including the chromatographic retention indices,[11] boiling points,[11] and molecular polarizabilities.[13] In this work we report on the applicability of the fragmental approach to QSPR studies of flash points.

Calculation Procedure and Database Creation

All calculations were carried out using the QSAR/QSPR program package EMMA,[4—11] the NASAWIN neural network program package,[16—19] and the modified version of the FRAGMENT subroutine, which allows a large number of fragment types to be taken into account and provides tools for flexible classification of atoms[4,20—25] (the FRAGMENT module

Published in Russian in *Izvestiya Akademii Nauk. Seriya Khimicheskaya,* No. 9, pp. 1787—1793, September, 2003.

1066-5285/03/5209-1885 $25.00

Table 1. Experimental flash points (T_f^{exp}) of the compounds from Databases *1* and *2*, which differ by more than 10 °C, and predicted values (T_f^{pr}) obtained using the nine fragmental descriptor based models for Databases *1* (model 1), *1A* (model 2), *2* (model 3), *2A* (model 4), *3* (model 6), and *3A* (model 7)

Compound	T_f^{exp}/°C		ΔT_f^{exp}/°C*	T_f^{pr}/°C					
	Database *1*	Database *2*		Database *1*	Database *1A*	Database *2*	Database *2A*	Database *3*	Database *3A*
Ethyl methyl ether	−37.2	−9	−28.2	−35.7	−32.7	−41.7	−41.5	−33.3	−33.5
Isopropylamine	−17	−37	20	−9.0	−8.7	−27.1	−14.0	−8.5	−7.9
Methacrolein	−15	0	−15	−15.7	−14.1	−8.5	−18.4	−23.4	−23.7
Diethylamine	−23	−39	16	8.0	11.2	−13.6	−15.0	7.0	7.7
2-Methylbutan-2-ol	41	20	21	30.0	29.2	35.0	38.1	32.8	32.7
Hexan-2-one	35	23	12	28.2	33.6	24.3	23.7	26.3	26.4
Dipropyl ether	21.1	−28	49.1	5.9	10.6	−1.0	−1.2	5.5	5.2
Dipropylamine	17	3	14	34.4	37.9	13.6	11.9	32.9	33.5
Ethylcyclohexane	35	18	17	14.6	9.3	19.6	17.6	18.3	17.8
Hexan-2-ol	57.6	41	16.6	51.8	51.2	55.7	56.9	53.4	53.1
Ethyl vinyl ether	−18	−45	27	−13.8	−10.7	−32.5	−30.8	−26.0	−26.4
Benzyl acetate	90.6	102	−11.4	98.7	91.2	80.3	83.3	94.6	94.5

Note. The mean deviation of the predicted values from the experimental data for Database *1* (Database *2*) is 12.3 (20.6), 12.0 (22.8), 11.3 (15.1), 10.0 (16.8), 12.1 (20.1), and 12.4 (27.4) °C for models 1, 2, 3, 4, 6, and 7, respectively.
* Difference between Databases *1* and *2*.

was incorporated into the EMMA and NASAWIN programs). Databases were created using the MEOW and BASTET programs* developed for correcting the QSAR structural databases, including the manual input, sorting, and search for duplicates. These programs are convenient for joint use with the QSAR program packages EMMA and NASAWIN. The structural databases can be automatically converted into various file formats including the sdf format.

The starting point was database analysis and correction. First, using the available data,[2] we created a Database *1*, which contained a total of 400 structures (acyclic, cyclic, and aromatic hydrocarbons; halogen-containing compounds; alcohols; phenols; ethers and esters; aldehydes, ketones; carboxylic acids; amines; nitriles; isocyanates; nitro derivatives; amides; and sulfur-containing compounds). A thorough comparative study of Database *1* revealed a set of erroneous data: (i) a strongly underestimated T_f of −19 °C for octanoic (caprylic) acid (compound No. 160 in the test set, see Ref. 2; *cf.* +85 °C for homologous valeric acid), (ii) a strongly overestimated T_f of 110 °C for methyl formate (compound No. 161 in the training set, see Ref. 2; *cf.* −10.0 °C for a homologue, methyl acetate), and (iii) a too low flash point of 16 °C for nonadecane (*cf.* 168 °C in Ref. 3). Therefore, we excluded caprylic acid and methyl formate from Database *1* and corrected the data for nonadecane. Thus, the Database *1* contained a total of 398 compounds. In addition, a Database *1A* was created by excluding the data for a group of twelve compounds listed in Table 1 from Database *1* and by somewhat "modifying" the T_f values for a set of twenty compounds (the principles we were guided by are discussed below taking the data for Database *3A* as an example).

Second, we used the published data[3] to create the Database *2*, which contained a total of 271 compounds belonging to various classes of organic compounds. Analysis of the available data (see Table 1 in Ref. 3; the T_f values were converted into Celsius degrees) using the BASTET program revealed the following pairs of duplicates: structure Nos. 8 and 14, 143 and 163, and 100 and 235. In a private communication the authors of Ref. 3 confirmed the correctness of structure No. 8 and noted that structure No. 14 must correspond to a formula of ethyl methyl ether. The structure No. 163 was excluded. Besides, two structures (No. 100 and No. 235) corresponding to ethyl butyrate were characterized by different T_f values (26 and 19 °C, respectively). The structure No. 235 was excluded from the database, because this compound is characterized by a flash point of 24 °C in Database *1*. As a result, the number of structures in Database *2* decreased from 271 to 269. In addition, we created a Database *2A* by excluding the data for a total of twelve compounds listed in Table 1 from Database *2* (the principles we were guided by are discussed below taking the data for Database *3A* as an example).

Third, in order to extend the experimental data set and to check it for mutual consistency, we used the BASTET program to create a Database *3*, which combined the data of Databases *1* and *2*. Then, the duplicate structures characterized by the same experimental values were excluded. If the T_f values for the duplicate structures were different, preference was given to the data of Database *1*. Stereoisomeric structures were retained. Thus, the Database *3* comprised the data listed in Refs. 2 and 3 and contained a total of 525 structures.

Content analysis of Database *3* revealed a total of eighty duplicate structures characterized by (sometimes, significantly) different experimental values in Databases *1* and *2*. The experimental data were determined to an accuracy of ±5 °C [1] or 8—12 °C.[2,3] Therefore, we created a Database *3A* with the following principles in mind: (i) if the data in Refs. 2 and 3 (in Databases *1* and *2*, respectively) differed by at most 5 °C, the experimental value from Database *1* was used (for a total of

* The MEOW and BASTET programs and the databases are available on request.

48 compounds) and (ii) if the difference was 5 to 10 °C, the arithmetic mean of the values from Databases *1* and *2* was used (for a group of 20 compounds). However, for a set of 12 compounds the difference exceeded 10 °C (see Table 1). Since in this case it was impossible to assess the quality (reliability) of the experimental values, these structures were excluded from Database *3A*. Thus, the Database *3A* represented a corrected combined set of published data[2,3] and contained a total of 513 structures.

Results and Discussion

First of all, the results of the early QSPR studies[2,3] should be analyzed. This is necessary for comparing them with our results. In particular, a partial least squares (PLS) model and a neural network model for calculating the flash points were obtained.[2] In the latter case, the standard deviations (s) obtained using the model based on twenty-five descriptors characterizing the contributions of various types of atoms and functional groups were 10.8 °C for the training set (135 compounds), 14.1 °C for the test set (133 compounds), and 14.3 °C for the validation set (132 compounds). The PLS results were much worse, namely, the corresponding s values were 21, 25, and 23 °C. Noteworthy is that the PLS plot exhibited a remarkable curvature in the low temperature region, which may be indicative of some nonlinear dependence.

Analysis of flash points using the CODESSA program allowed one to construct a three-parameter model with the following statistical characteristics: square of correlation coefficient, $r^2 = 0.9020$; square of cross-validated correlation coefficient, $r^2_{cv} = 0.8985$; the Fisher test $F = 820$, and $s = 16.1$.[3] It is noteworthy that good correlations were found when using either the experimental or calculated boiling points as descriptors. Yet another interesting feature is the presence of a rather large number of "out-lying" compounds ("outliers;" especially, small molecules) despite a satisfactory overall statistics.

The first step of our study was to reproduce the results obtained in Ref. 2 using the fragmental approach and the same training and test sets. The linear regression models constructed for Databases *1* and *1A* using the stepwise inclusion of the calculated fragmental descriptors are listed in Table 2 (models 1 and 2, respectively). The statistical characteristics of the nine-descriptor model 1 are comparable with the parameters of the PLS model (mean absolute error was 20.6 °C for the training set and 23.3 °C for

Table 2. Statistical parameters (correlation coefficient (r), standard deviation (s), and Fisher test (F)) for the fragmental descriptor based linear regression QSPR models constructed for Databases *1*, *1A*, *2*, and *2A*

Parameter	Number of descriptors								
	1	2	3	4	5	6	7	8	9
Model 1, Database *1*									
Training set (compounds 1—266)									
r^2	0.6075	0.7155	0.7762	0.8095	0.8246	0.8405	0.8524	0.8656	0.8724
s	32.5	27.7	24.6	22.8	21.9	20.1	20.1	19.3	18.8
F	409	331	303	277	245	227	213	207	195
Maximum error*/°C	119.0 (122)	117.2 (244)	83.3 (81)	79.2 (244)	78.1 (244)	77.6 (244)	76.3 (244)	73.0 (214)	73.9 (214)
Test set (compounds 267—398)									
r^2 for prediction	0.6130	0.7122	0.7442	0.7561	0.7534	0.7737	0.7980	0.8307	0.8334
Average error of prediction/°C	24.9	21.6	20.5	19.4	19.5	18.3	17.5	15.3	15.2
Maximum error of prediction*/°C	97.0 (381)	73.2 (347)	59.2 (372)	65.8 (397)	66.9 (397)	69.4 (397)	67.7 (397)	66.8 (397)	64.2 (299)
Model 2, Database *1A*									
*Training set (compounds 1—266)***									
r^2	0.5958	0.7077	0.7732	0.8082	0.8235	0.8399	0.8524	0.8643	0.8712
s	32.8	28.1	24.8	22.8	21.9	20.9	20.2	19.4	18.9
F	382	308	287	265	234	219	205	197	186
Maximum error*/°C	119.0 (122)	117.2 (244)	83.3 (81)	79.2 (244)	78.1 (244)	75.5 (244)	74.3 (244)	71.8 (214)	72.9 (244)
Test set (compounds 267—398)									
r^2 for prediction	0.6094	0.7082	0.7402	0.7527	0.7500	0.7706	0.7933	0.8264	0.8292
Average error of prediction/°C	25.1	21.8	20.70	19.6	19.7	18.5	17.7	15.6	15.3
Maximum error of prediction*/°C	97.0 (381)	73.2 (347)	59.2 (372)	65.8 (397)	66.9 (397)	68.8 (397)	67.4 (397)	66.6 (397)	59.7 (397)

(*to be continued*)

Table 2 (*continued*)

Parameter	Number of descriptors								
	1	2	3	4	5	6	7	8	9
Model 3, Database *2*									
Training set (compounds 1—269)									
r^2	0.5414	0.7964	0.8439	0.8694	0.8847	0.8983	0.9128	0.9237	0.9317
s	34.8	23.3	20.4	18.7	17.6	16.6	15.4	14.4	13.7
F	315	520	478	439	404	386	390	394	392
Maximum error*/°C	131.6 (197)	132.3 (82)	115.2 (82)	72.2 (82)	72.4 (82)	69.6 (82)	68.4 (191)	72.2 (82)	55.9 (191)
Model 4, Database *2A*									
*Training set (compounds 1—269)***									
r^2	0.5354	0.7983	0.8471	0.8765	0.8924	0.9113	0.9255	0.9317	0.9352
s	35.2	23.2	20.2	18.2	17.0	15.5	14.2	13.7	13.3
F	294	503	467	447	416	428	442	423	396
Maximum error*/°C	130.6 (197)	132.0 (82)	69.0 (82)	67.0 (82)	63.9 (82)	69.1 (191)	66.9 (82)	53.0 (191)	53.2 (191)
Model 5, Database *2*									
Training set (179 compounds)									
r^2	0.5169	0.7768	0.8340	0.8603	0.8823	0.8974	0.9085	0.9136	0.9195
s	35.4	24.2	20.9	19.2	17.7	16.6	15.7	15.3	14.8
F	189	306	293	268	259	251	243	225	214
Maximum error*/°C	132.2 (196)	132.1 (82)	71.7 (82)	73.7 (82)	69.7 (82)	69.6 (82)	49.2 (190)	51.9 (190)	52.3 (190)
*Test set (89 compounds)****									
r^2 for prediction	0.5867	0.8316	0.8414	0.8598	0.8949	0.8929	0.9007	0.9184	0.9314
Average error of prediction/°C	27.1	17.2	16.4	15.0	13.2	13.0	11.7	11.0	9.9
Maximum error of prediction*/°C	109.3 (240)	59.1 (120)	50.3 (45)	52.2 (198)	48.5 (84)	48.6 (186)	47.2 (186)	45.7 (186)	45.4 (186)

Note. Descriptors 1—9 for model 1 are as follows: an "arbitrary atom," OH, N, "a chain of three arbitrary atoms linked by simple bonds," C—S, C—C—C—Hal, C—C=O, Me, $C_{Ar}H \div C_{Ar}H \div C_{Ar}R \div C_{Ar}H$ (÷ is an aromatic bond); for model 2: an "arbitrary atom," OH, N, "a chain of three arbitrary atoms linked by simple bonds," C—S, C—C—C—Hal, C—C=O, Me, —O—CR=O; for model 3: an "arbitrary atom," OH, N, Me, CH_2=CHR, C—C=O, C—Hal, RC_{Ar}, C—C—N; for model 4: an "arbitrary atom," OH, N, Me, C—Hal, C—C=O, $C_{Ar} \div C_{Ar} \div C_{Ar} \div C_{Ar} \div C_{Ar}$, C_{sp^3}—N, =CH_2; and for model 5: an "arbitrary atom," OH, N, "two arbitrary atoms linked by a simple bond," C—Hal, CR_2=O, C_{sp^3}—N, Me, and CH_2=CH—.

* The number of compound in the corresponding database is given in parentheses.

** Twelve compounds were excluded.

*** Each third compound.

the test set).[2] The QSPR equation for calculating T_f values has the form

$$T_f^{calc} = -0.826 + 0.285f_1 + 0.497f_2 + 0.151f_3 - 6.718f_4 + \\ + 0.208f_5 + 0.130f_6 - 1.87f_7 + 4.50f_8 + 0.369f_9, \quad (1)$$

and is characterized by $r^2 = 0.8724$, $s = 18.8$, and $F = 195$ for $n = 266$. The mean absolute error of prediction was 15.2 °C and the maximum error was 73.9 °C. In Eq. (1), the descriptors f_i denote the number of atoms or particular types of molecular fragments, namely, the N atoms (f_1), the OH groups (f_2), arbitrary atoms (f_3), the Me (f_4), C—S (f_5), and C—C=O (f_6) groups, a chain of three arbitrary atoms linked by simple bonds (f_7), and the $C_{Ar}H \div C_{Ar}H \div C_{Ar}R \div C_{Ar}H$ (f_8; ÷ is an aromatic bond) and C—C—C—Hal (f_9) fragments.

An increase in the number of fragmental descriptors to 25 allows the linear regression model to be improved to nearly compete with the neural network model.[2] The descriptors used were the number of halogen, nitrogen, oxygen, and sulfur atoms and the number of di- and triatomic fragments containing different (double, triple, and aromatic) bonds, that is, the I (f_1), F (f_2), Br (f_3), S (f_4), and N (f_5) atoms, the OH group (f_6), an arbitrary atom (f_7), the C=O (f_8), $MeNR_2$ (f_9), CH_2Hal (f_{10}), =CR—NHR (f_{11}), =CR—OH (f_{12}), and Me—C_{sp^3} (f_{13}) groups, and the $HC_{Ar} \div C_{Ar}R \div C_{Ar}$ (f_{14}), C—C=O (f_{15}), =CR—C_{sp^3}—Cl (f_{16}), CH_2—CH_2—C≡ (f_{17}), C—C_{sp^3}—Cl (f_{18}), =C—$C_{Ar} \div C_{Ar}$—OH (f_{19}), C—C—C—N (f_{20}), $C_{Ar} \div C_{Ar} \div C_{Ar} \div C_{Ar}$—N ($f_{21}$), C—C—C—S—C ($f_{22}$), C—C—C—C—C—O ($f_{23}$), $MeC_{Ar}(\div C_{Ar}H)_2$ (f_{24}), and Hal—C(—C)$_2$ (f_{25}) fragments. Figure 1 presents the cor-

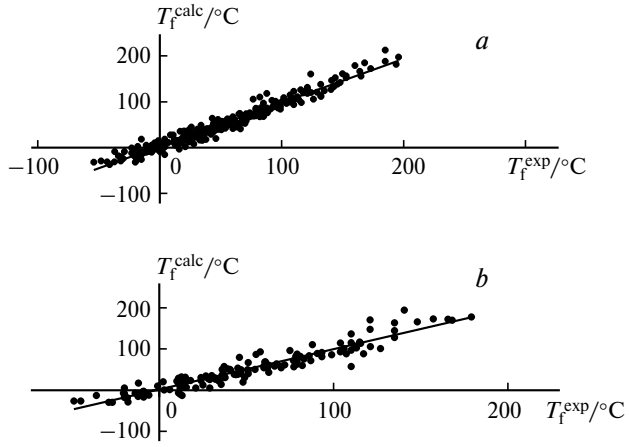

Fig. 1. Correlation between the calculated (T_f^{calc}) and experimental (T_f^{exp}) flash point values obtained for the training set (*a*) and the test set (*b*) of compounds from Database *1* using the twenty-five fragmental descriptor based linear regression model.

relation between the calculated and experimental T_f values obtained for the training set and for the test set of compounds from Database *1* using the twenty-five fragmental descriptor based model (r^2 = 0.9557; s = 11.4; F = 199; the average error of prediction was 11.8 °C, the RMS error for the training set (RMS_{tr}) was 10.9 °C; and the RMS error of prediction (RMS_{pr}) was 15.8 °C).

The largest errors of prediction (in °C) were obtained for 2-phenylphenol (36.8), fluorobenzene (33.1), dodecane-1-thiol (30.4) (training set) and for *p*-nonylphenol (53.5), ethyl cyanoacetate (52.6), and anthracene (49.3) (test set).

A decrease in the number of compounds in Database *1* due to the exclusion of twelve structures insignificantly deteriorates the quality of the models constructed for Database *1A* (*cf.* models 1 and 2 in Table 2). Here, the descriptors used in the models are by and large identical in nature, except for replacement of the $C_{Ar}H \div C_{Ar}H \div C_{Ar}R \div C_{Ar}H$ fragment by the —O—CR=O fragment.

Then, we used the fragmental approach to construct models for Databases *2* and *2A* (see Table 2, models 3 and 4, respectively). As earlier,[3] the models obtained using the training set, which contained all compounds from Database *2*, were as good as the models 1 and 2 and had better statistical characteristics than the model from Ref. 3 (see above). For instance, the statistical characteristics of the twenty-five descriptor model constructed for Database *2* are as follows: r^2 = 0.9566, s = 11.2, F = 206, and RMS_{tr} = 10.7 °C.

In contrast to the above-mentioned sets of compounds (Databases *1* and *1A*), the exclusion of twelve compounds from Database *2* somewhat improves the parameters of the corresponding models (*cf.* models 3 and 4 in Table 2). The fragments in the descriptor sets used in these models are also different. Probably, in this case the quality of QSPR models is determined by the accuracy of determination of experimental data for the structures from this database, so a simple increase in the number of noisy data does not improve the QSPR model.

The predictive power of the QSPR model for Database *2* was assessed taking the training (179 compounds) and test (89 compounds) sets as examples. The nine-descriptor model had a rather high predictive power, r^2_{pr} = 0.9315, the mean absolute error of prediction was 9.9 °C (see Table 2, model 5).

It was also of interest to consider the results of prediction of T_f values for the compounds listed in Table 1. First of all, two predictions were made, using the nine-descriptor models 1 and 2 (see Table 2, the models for Databases *1* and *1A*, respectively). These results are also listed in Table 1, from which it follows that the mean deviation of the T_f values predicted using model 1 from the experimental data for Database *1*, is 12.3 °C, which is smaller than the difference between the predicted and experimental values for Database *2* (on the average, 20.6 °C). However, it should be noted that this is not necessarily true and, moreover, the predicted value often lies between two different experimental values.

Prediction based on model 2 (a true prediction, because this model was constructed using Database *1A*) gave a value of 12.0 °C for the mean deviation of the predicted values from the experimental data for Database *1*. This is much smaller than the difference between the predicted values and experimental data for Database *2* (on the average, 22.8 °C).

Thus, the results of the predictions carried out for the compounds listed in Table 1 using Databases *1* and *1A* are in better agreement with the data of Database *1*. This brings up the question concerning the results of the predictions made using Databases *2* and *2A*. Two predictions were made taking the nine fragmental descriptor based models 3 and 4 (see Table 2, the models for Databases *2* and *2A*, respectively) as examples. The results obtained are also listed in Table 1. As can be seen, the mean deviation of the values obtained using model 2 from the experimental data for Database *1* is 11.3 °C, which is smaller than the difference between the calculated values and experimental data in Database *2* (on the average, 15.1 °C).

For the prediction made using model 4 (note that this is a true prediction, because model 4 was constructed using Database *2A*), the mean deviation of the calculated values from the experimental data in the two above-mentioned databases even more increases and becomes equal to 10.0 °C for Database *1*, which, again, is much smaller than for Database *2* (on the average, 16.8 °C).

By and large, the data set provided by Database *1* seems to be the most appropriate for making predictions for the compounds listed in Table 1 despite the presence

Table 3. Statistical parameters (correlation coefficient (r), standard deviation (s), and Fisher test (F)) for the fragmental descriptor based linear regression QSPR models constructed for Databases *3* and *3A*

Parameter	Number of descriptors								
	1	2	3	4	5	6	7	8	9
Model 6, Database *3*									
Training set (compounds 1—266 and 399—525)									
r^2	0.5634	0.7309	07956	0.8271	0.8439	0.8619	0.8720	0.8801	0.8884
s	35.6	27.9	24.5	22.5	21.4	20.2	19.4	18.8	18.2
F	504	529	493	453	409	392	375	352	339
Maximum error*/°C	119.0 (122)	117.2 (244)	83.3 (81)	79.2 (244)	78.1 (244)	77.6 (244)	82.7 (214)	83.0 (214)	75.7 (214)
Test set (compounds 267—398)									
r^2 for prediction	0.6092	0.7017	0.7269	0.7834	0.7968	0.8021	0.8081	0.8272	0.8502
Average error of prediction/°C	27.9	21.5	19	18	17.3	17	16.9	15.9	14.5
Maximum error of prediction*/°C	97.0 (381)	73.2 (347)	59.2 (372)	65.8 (397)	66.9 (397)	69.4 (397)	63.4 (347)	64.1 (397)	61.5 (397)
Model 7, Database *3A*									
*Training set (compounds 1—266 and 399—525)***									
r^2	0.5589	0.7269	0.7956	0.8271	0.8439	0.8619	0.8713	0.8796	0.8883
s	35.9	28.3	24.5	22.5	21.4	20.2	19.5	18.9	18.2
F	484	507	493	453	409	392	363	342	330
Maximum error*/°C	119.0 (122)	117.2 (244)	83.3 (81)	79.2 (244)	78.1 (244)	77.6 (244)	76.3 (244)	73.0 (214)	61.2 (214)
Test set (compounds 267—398)									
r^2 for prediction	0.6050	0.6979	0.7269	0.7834	0.7968	0.8021	0.8051	0.8249	0.8469
Average error of prediction/°C	24.9	21.7	21.1	18	17.3	17	17	16.1	14.7
Maximum error of prediction*/°C	97.0 (381)	73.2 (347)	59.2 (372)	65.8 (397)	66.9 (397)	69.4 (397)	67.7 (397)	66.8 (397)	64.2 (299)

Note. Descriptors 1—9 for models 6 and 7 are as follows: an "arbitrary atom," OH, N, Me, $=CH_2$, $C_{sp^3}—C_{sp^3}$, S, C—C—Hal, and $C_{sp^3}—C=O$.

* The number of compound in the corresponding database is given in parentheses.

** Twelve compounds were excluded.

of some "outliers" due to insufficient accuracy of experimental determination of T_f values.

The influence of the set quality and size on the statistical characteristics of the QSPR models for flash point calculations can also be assessed by comparing the parameters of the regression models constructed for Databases *3* and *3A* (Table 3, models 6 and 7, respectively) and the above-mentioned models for Databases *1* and *2* (see Table 2, models 1, 3, and 5). In the case of Databases *1* (n = 266) and *3* (n = 392), an increase in the number of structures in the set does not improve the model, as could be expected from general considerations (*cf.* model 1 in Table 2 and model 6 in Table 3). A similar situation was described above for Databases *2* and *2A*.

Exclusion of twelve structures from Database *3* causes no significant changes in the parameters of the models obtained for Database *3A* (*cf.* models 6 and 7). The statistical characteristics of the best model constructed for Database *3A* using a set of fifteen descriptors (r^2 = 0.9241, s = 15.2, F = 279, the average error of prediction was 11.9 °C, RMS_{tr} = 14.8 °C, and RMS_{pr} = 15.5 °C) are worse than those of the best regression model obtained for Database *1* (see above). The largest absolute errors of T_f prediction (in °C) were obtained for the following compounds of the training set: 4-methylpentan-2-one (73.9), 2-propenal (62.5), nitroethane (62.1) and 3-methylbut-1-ene (50.8) and for the following compounds of the test set: isobutylaldehyde (56.3), allyl acetate (49.1) and dibutyl sulfide (46.8).

Thus, in this case the statistical characteristics of the regression models are to a great extent determined by not only the sample size (common situation in QSPR studies) but also the accuracy of determination of the experimental data for the structure set under study.

Next, we used an artificial neural network (ANN)[26] to construct "structure—flash point" models using the fragmental approach. A three-layer, feed-forward backpropagation neural network implemented in the NASAWIN program package[27] was employed. The input layer of the network contained twenty-five or fifteen neurons and one

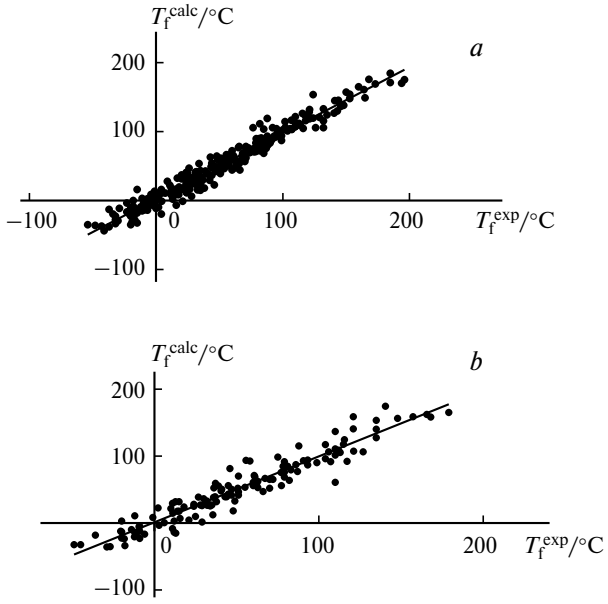

Fig. 2. Correlation between the calculated (T_f^{calc}) and experimental (T_f^{exp}) flash point values obtained for the training set (*a*) and the test set (*b*) of compounds from Database *1* using the twenty-five fragmental descriptor based neural network model.

bias pseudoneuron to match the number of descriptors that were pre-selected using the stepwise linear regression analysis. The hidden layer contained two neurons, because a larger number of neurons leads to strong "overtraining" of this database, while the model constructed with a smaller number of neurons is in essence identical to the linear regression models. The output layer contained one neuron, which corresponded to one property to be predicted. The generalized delta rule[26] was used as the learning algorithm. The learning constant and momentum were equal to 0.25 of 0.9, respectively. The training process was stopped after reaching the smallest error of prediction using the test set.

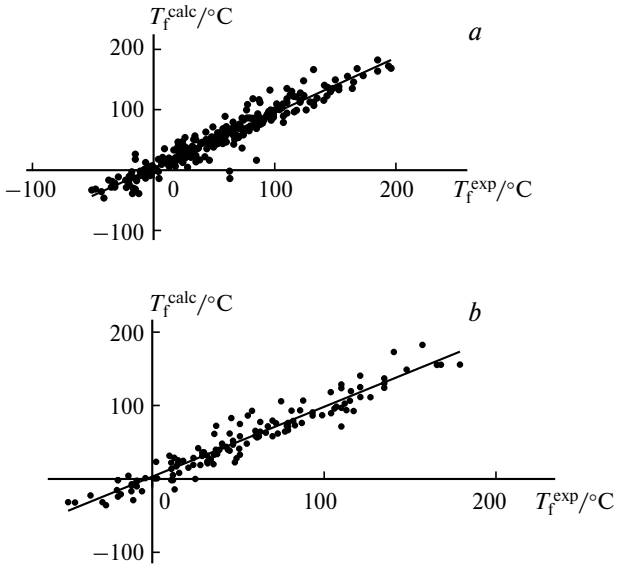

Fig. 3. Correlation between the calculated (T_f^{calc}) and experimental (T_f^{exp}) flash point values obtained for the training set (*a*) and the test set (*b*) of compounds from Database *3A* using the fifteen fragmental descriptor based neural network model.

As a result, we obtained two neural network models with better predictive power as compared to both the corresponding linear regression models (see above) and the neural network models reported in the literature.[2] The twenty-five fragmental descriptor based neural network model obtained for Database *1* has the following statistical characteristics: $r^2 = 0.9590$ and $RMS_{pr} = 14.6$ °C (*cf.* 15.8 °C for the corresponding linear model constructed using the same descriptors). The parameters of the fifteen-descriptor neural network model obtained for Database *3A* are $r^2 = 0.9343$ and $RMS_{pr} = 14.0$ °C (*cf.* 15.5 °C for the corresponding linear model). Figures 2 and 3 present the correlations between the calculated and experimental flash point values for the training sets (*a*) and test sets (*b*), which were obtained using the neural network models constructed for Databases *1* and *2*, respectively.

Thus, we constructed a number of fragmental descriptor based linear regression and neural network models, which can predict the flash points with an accuracy, which can sometimes approach the accuracy of experimental flash point determination.

References

1. S. F. Evlanov, *Zh. Prikl. Khim.*, 1991, **64**, 832 [*J. Appl. Chem. USSR*, 1991, **64** (Engl. Transl.)].
2. J. Tetteh, S. Takahiro, E. Metcalfe, and S. Howells, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 491.
3. A. R. Katritzky, R. Petrukhin, and R. M. Jain Karelson, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 1521.
4. N. S. Zefirov and V. A. Palyulin, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 1112.
5. D. E. Petelin, V. A. Palyulin, and N. S. Zefirov, *Dokl. Akad. Nauk*, 1992, **324**, 1019 [*Dokl. Chem.*, 1992 (Engl. Transl.)].
6. T. S. Pivina, D. V. Sukhachev, and L. K. Maslova, *Dokl. Akad. Nauk*, 1993, **330**, 468 [*Dokl. Chem.*, 1993 (Engl. Transl.)].
7. D. V. Sukhachev, T. S. Pivina, V. A. Shlyapochnikov, E. A. Petrov, V. A. Palyulin, and N. S. Zefirov, *Dokl. Akad. Nauk*, 1993, **328**, 188 [*Dokl. Chem.*, 1993 (Engl. Transl.)].
8. D. V. Sukhachev, T. S. Pivina, N. I. Zhokhova, N. S. Zefirov, and S. I. Zeman, *Izv. Akad. Nauk, Ser. Khim.*, 1995, 1653 [*Russ. Chem. Bull.*, 1995, **44**, 1585 (Engl. Transl.)].
9. D. V. Sukhachev, T. S. Pivina, N. I. Zhokhova, and N. S. Zefirov, *Izv. Akad. Nauk, Ser. Khim.*, 1995, 1657 [*Russ. Chem. Bull.*, 1995, **44**, 1589 (Engl. Transl.)].
10. D. V. Sukhachev, T. S. Pivina, N. I. Zhokhova, N. S. Zefirov, and S. I. Zeman, *Izv. Akad. Nauk, Ser. Khim.*, 1995, 1661 [*Russ. Chem. Bull.*, 1995, **44**, 1594 (Engl. Transl.)].

11. N. S. Zefirov and V. A. Palyulin, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 1022.
12. N. S. Zefirov, V. A. Petelin, V. A. Palyulin, and J. McFarland, *Dokl. Akad. Nauk*, 1992, **327**, 504 [*Dokl. Chem.*, 1992 (Engl. Transl.)].
13. N. I. Zhokhova, V. A. Palyulin, I. I. Baskin, A. N. Zefirov, and N. S. Zefirov, *Izv. Akad. Nauk, Ser. Khim.*, 2003, 1005 [*Russ. Chem. Bull., Int. Ed.*, 2003, **52**, 1061 (Engl. Transl.)].
14. I. I. Baskin, M. I. Skvortsova, I. V. Stankevich, and N. S. Zefirov, *J. Chem. Inf. Comput. Sci.*, 1995, **35**, 527.
15. N. V. Artemenko, I. I. Baskin, V. A. Palyulin, and N. S. Zefirov, *Izv. Akad. Nauk, Ser. Khim.*, 2003, 19 [*Russ. Chem. Bull., Int. Ed.*, 2003, **52**, 20 (Engl. Transl.)].
16. I. I. Baskin, V. A. Palyulin, and N. S. Zefirov, in *QSAR and Molecular Modelling: Concepts, Computational Tools, and Biological Applications*, Eds. F. Sanz, J. Giraldo, and F. Manaut, Prous Science Publishers, Barcelona, 1995, 30.
17. N. M. Halberstam, I. I. Baskin, V. A. Palyulin, and N. S. Zefirov, *Proc. Int. Symp. CACR-96*, Moscow, 1996, 37.
18. I. I. Baskin, V. A. Palyulin, and N. S. Zefirov, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 715.
19. N. M. Halberstam, I. I. Baskin, V. A. Palyulin, and N. S. Zefirov, *Mendeleev Commun.*, 2002, 185.
20. I. I. Baskin, V. A. Palyulin, and N. S. Zefirov, *Tez. dokl. Mezhvuz. konf. "Molekulyarnye grafy v khimicheskikh issledovaniyakh" [Abstrs. Higher School Conf. "Molecular Graphs in Chemical Research"]*, Kalinin, 1990, 5 (in Russian).
21. I. I. Baskin, V. A. Palyulin, and N. S. Zefirov, *Tez. dokl. 1-i Vsesoyuz. konf. po teoreticheskoi organicheskoi khimii* [*Abstrs. 1st All-Union Conf. on Theoretical Organic Chemistry*], Volgograd, 1991, 557 (in Russian).
22. V. A. Palyulin, I. I. Baskin, D. E. Petelin, and N. S. Zefirov, *Abstrs. 10th Eur. Symp. on Structure—Activity Relationships: QSAR and Molecular Modelling*, Barcelona, 1994, B257.
23. V. A. Palyulin, I. I. Baskin, D. E. Petelin, and N. S. Zefirov, in *QSAR and Molecular Modelling: Concepts, Computational Tools, and Biological Applications*, Eds. F. Sanz, J. Giraldo, and F. Manaut, Prous Science Publishers, Barcelona, 1995, 51.
24. V. A. Palyulin, E. V. Radchenko, I. I. Baskin, A. Yu. Zotov, and N. S. Zefirov, *Abstrs. 11th Eur. Symp. on QSAR: Computer Assisted Lead Finding and Optimization*, Lausanne, 1996, 31A.
25. N. V. Artemenko, I. I. Baskin, V. A. Palyulin, and N. S. Zefirov, *Dokl. Akad. Nauk*, 2001, **381**, 203 [*Dokl. Chem.*, 2001 (Engl. Transl.)].
26. J. Zupan and J. Gasteiger, *Neural Networks in Chemistry*, Wiley—VCH, Weinheim, 1999, 380.
27. I. I. Baskin, N. M. Halberstam, N. V. Artemenko, V. A. Palyulin, and N. S. Zefirov, *Abstrs. 14th Eur. Symp. on Quantative Structure—Activity Relationships*, Bournemouth, 2002, P173.

Received February 26, 2003;
in revised form May 8, 2003