Full Articles

Artificial neural network and fragmental approach in prediction of physicochemical properties of organic compounds

*N. V. Artemenko, I. I. Baskin, V. A. Palyulin, 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: zefirov@org.chem.msu.su

An approach based on fragmental descriptors (occurrence number of structural fragments in chemical structures) in conjunction with the artificial neural network technique was developed for predicting the physicochemical properties of organic compounds. The construction of neural network models for predicting the viscosity, density, and saturated vapor pressure for various classes of organic compounds is discussed.

Key words: artificial neural networks, neural network modeling, viscosity, density, vapor pressure, physicochemical properties, fragmental descriptors.

Despite the fast development of quantum-chemical methods of molecular modeling, empirical approaches based on the use of various molecular structure descriptors still play the crucial role in predicting most of physicochemical properties.[1] Historically, additive schemes come first among the empirical approaches to approximation of physicochemical properties. In these techniques, the numbers of different simple fragments in the molecule were used as descriptors,[2—7] and the physicochemical properties were represented as the sums of values corresponding to pair interactions of atoms.[8—14] A number of studies[5,15—17] are underlain by the assumption that the numerical value for some property of compounds of a given series can be represented as the sum of contributions of various structural fragments. A fragment method for the calculation of physicochemical properties, in particular, the enthalpies of formation of saturated hydrocarbons has been developed.[6,7,18] It is based on a topological graph approach to structure consideration. In conformity with the general principles of graph theory, the structural formula of any chemical compound can be described in terms of the graph theory, atoms being regarded as the vertices of a molecular graph and chemical bonds being the graph edges. Chains of a particular length were proposed as fragments, carbon atoms being subdivided into primary, secondary, tertiary, and quaternary ones. The contributions of chains containing two or three vertices (atoms) were taken as corrections for the first- and second-order environment effects. This approach has been further developed as the EMMA program package,[19] which uses multiple linear regression (MLR) as the statistical method. Within the framework of this package, a program for calculating the fragmental descriptors (Fragment 1) was developed. The program provides generation

Published in Russian in *Izvestiya Akademii Nauk. Seriya Khimicheskaya,* No. 1, pp. 19—28, January, 2003.

1066-5285/03/5201-20 $25.00

of linear (one to nine atoms), cyclic (three to six-membered), and three types of branched fragments.[20,21] A modified version of this program,[22] which is also used in this study, is included in the NASAWIN package.[23] The BIBIGON program package developed previously[24] allows one to build linear structure—property relationships using fragmental descriptors. In this approach, $A_1...A_k$ chains of atoms are used for structure characterization. A set of labels is ascribed to each atom to encode the type of atom (chemical symbol, the number of nonhydrogen neighbors, chemical bond order, position of the atom in a ring or in a chain).

Fragmental approaches have been successfully used in studies dealing with the relationships between the chemical structures and biological activities of molecules. In the Free—Wilson method,[25—27] the biological activity (A) was described as the sum of contributions of substituents at a definite common fragment.

The next step in the progress of fragment-based computation methods was taken by introducing a rather simple language for describing the chemical structures (SSFN meaning substructure superposition fragment notation) based on the concept of descriptor centers.[28—31] According to this approach, so-called descriptor centers (atoms or atom groups) are distinguished in a chemical structure; they can be either interaction sites between a biologically active molecule and a target or reaction sites. Somewhat modified SSFN system underlies the PASS program package.[32] The first version of this program made it possible to estimate the probability that diverse organic compounds would exhibit any of 114 types of biological activity. As a further development of the PASS program, multilevel neighborhoods of atoms (MNA) were proposed as substructure descriptors.[33,34] The so-called logic-structural approach (LSA) meant for elucidating structure—activity relationships has been developed.[31] In terms of this method, the activity of a chemical compound is predicted judging by the presence of particular structural fragments called pharmacophores (promoting the display of activity) and pharmacophobs (preventing the display of activity). The activity of a structure was also estimated[35] using logical functions, which are conjunctions and disjunctions of predicates serving to indicate the presence of particular fragments in the structure.

The fragmental approach has been embodied in the CASE program (computer automated structure evaluation).[36—38] In this method, a molecular property is represented by simple summation of all the local values for atoms or linear fragments incorporated in the given structure. However, the individual contribution of each atom is determined by not only the nature of the atom but also by the nature of the environment. For successive operation of the program, the main groups of structural parameters were formed including the presence of heavy atoms with different hybridization types and functional groups. The latest versions of the CASE and MultiCASE programs are also of interest because they provide the possibility of automated identification of the structural fragments that exert pronounced positive or negative effect on the biological activity, so-called biophores and biophobs.[38] The structural fragments were used[39,40] to predict the mutagenicity using artificial neural networks (ANN). The occurrence numbers of simple monoatomic fragments in combination with ANN have been used[41] to predict the physicochemical properties and the biological activity of a number of organic compounds.

A concept of molecular hologram has been proposed;[42] it is actually a vector describing the occurrence of diverse fragments (specified in an explicit form as linear notation) in the structure of a chemical compound. Within the framework of this approach, the structure—property (biological activity) relationship for a chemical compound is identified using the partial least squares (PLS) method,[43] which makes possible handling a large number of correlated descriptors. This approach also provides the possibility of interpreting the resulting models by color coding of the molecular structure image. As a result, one can see the parts of the molecule that are either favorable or unfavorable for displaying the specified type of biological activity.

However, most of the described methods have a number of limitations: (1) too high generalization level in atom classification; (2) lack of flexibility in choosing the fragments; (3) use (in many cases) of the linear statistical method, which deteriorates the model quality and requires a too large number of correction factors.

In the previous study,[44] we considered the procedural aspects of using the fragmental descriptors as applied to construction of linear structure—property models. However, in many cases, the dependence of physicochemical properties on descriptors is essentially nonlinear, its general form being usually unknown beforehand. The use of the ANN procedure allows one to predict successfully the properties of organic compounds.[45—47]

The term artificial neural network is used to imply the set of computational mathematics techniques combined by the general idea of simulating the functioning of human brain in information processing.[48] It is generally believed that artificial neural networks consist of a set of simple computing units called "neurons" and a set of "synapses" connecting the neurons. Each synapse is characterized by a number called "synaptic weight". A neural network is learnable. Learning usually implies the adjustment of the synaptic weights in such a way as to minimize a particular error functional, which depends on the problem. In solving regression problems, a neural network is trained by adjusting its synaptic weights to minimize the error of prediction of the output vector of numbers on the basis of the input vector. In particular, when analyzing structure—property relationships, a neural network learns

to predict the output vector of the properties of chemical compounds on the basis of the input vector of the descriptor values.[46] The most frequently used architecture of ANN is the multilayer feed-forward error back-propagation network. Within the framework of this architecture, the input descriptors specify the values for activation of the input layer neurons. The predicted values of properties are taken from the output neurons, and for intermediate computations, hidden neurons are also used, the number of the latter being determined by the complexity of the correlation to be modeled.

In this study, the efficiency of the proposed fragmental approach combined with the ANN system in modeling the physicochemical properties of organic compounds is evaluated by performing both the linear-regression and neural network modeling of the density (for liquids), viscosity, and saturated vapor pressure for organic compounds. These properties were chosen due to their practical importance. In particular, density prediction is needed to develop the energetic compounds. Viscosity must be known to optimize petrochemical processes. Prediction of the saturated vapor pressure allows one to estimate the evaporation and absorption rates and the maximum possible air concentrations of potential environmental pollutants.

Investigation procedure

The structure—property models were constructed and analyzed according to the following procedure. At the first stage, fragmental descriptors (occurrence numbers of structural fragments in the chemical structure) were calculated for each compound included in the database comprising information on the structures and properties of chemical compounds;[22] the maximum size of the fragments was varied from one to ten atoms. The fragments encountered for ≤1 % of compounds in the set and statistically identical fragments were excluded. For each descriptor D_i, nonlinear modifications were calculated including the square (D_i^2), square root ($D_i^{1/2}$), common logarithm ($\log D_i$, calculated only for those fragments that were found in all structures contained in the database), the ratio of the descriptor value to the number of nonhydrogen atoms in the molecule (D_i/N_a). At the next stage, some of the descriptors were rejected, so that none of the pair correlation coefficients r between the remaining descriptors exceeded 0.97.

The use of nonlinear modifications along with the fragmental descriptors themselves is quite justified. To study this aspect, preliminary comparative analysis of the linear-regression and neural network models (procedure for constructing the models is outlined below) was carried out for four sets of descriptors, both containing and not containing the above-listed modifications with maximum numbers of atoms equal to one and two. The analysis showed that statistical parameters of models constructed with both the descriptors and their linear modifications are much better than the corresponding parameters for models constructed without the nonlinear modifications of descriptors.

Subsequently the database was split into three sets, namely, training set (80% of compounds), validation set (10% of compounds), and prediction (test) set (10% of compounds). Splitting was done according to ten different patterns in such a way that each compound of the database was once encountered in each of the two last-mentioned sets. Then for each initial set of descriptors (of 10 or 13 sets for different databases, differing in the maximum size of the fragment) and each pattern of database splitting, selection of descriptors was carried out by stepwise multiple linear regression procedure according to which inclusion of every new descriptor was determined by a decrease in the prediction error for the validation set. Then, of the 10 or 13 initial sets for each splitting pattern of the database, the optimal descriptor was selected in terms of the average prediction error for the validation sets. The 10 sets of descriptors obtained were used subsequently in the study with multilayer back-propagation networks.[46,48]

At the next stage, five neural network models were constructed for each database splitting and for each number of hidden neurons, which varied from two to eight. Training was done using a "generalized delta-rule" (learning rate 0.25, momentum 0.9) until the minimum prediction error for the validation set was attained. After that, the optimal number of hidden neurons that ensured the smallest errors for the validation sets was identified, and the results of prediction in terms of half of the best (*i.e.*, providing the least error for the validation set) models for all compounds were averaged. This gave the following four parameters for each feature: correlation coefficient averaged over all training sets (R_{av}) and the root-mean square errors for the sets of three types. Since data for the third set did not participate in either construction or selection of models, the root-mean-square error for this type of set serves for objective evaluation of the predicting ability of the model.

Results and Discussion

Previously, density, viscosity, and saturated vapor pressure have been mainly predicted without using the fragmental approach. For example, the autocorrelation method was employed for predicting the saturated vapor pressures of alkanes and alkenes.[49] The components of autocorrelation vectors (descriptors in the analysis) were calculated from the surface area of 186 compounds using the Bondi method.[50] Previously, topostructural, topochemical, and geometric parameters[51] or quantum-chemical descriptors[52] were used as descriptors for modeling this property. The descriptors used to predict the density[53] were based on the ratio of the molecular weight to the volume for the given molecule. Neural network models for the description of physicochemical properties in terms of theoretical graph descriptors have been developed.[54] Topological descriptors have been used to predict the density of alkenes.[55,56] Viscosity was calculated by many researchers using either fragmental approaches[57—62] or methods for determining molecular parameters (refraction, dipole moment, critical temperature, molar magnetic susceptibility, and cohesion energy), which are taken as descriptors.[63,64]

Now we consider the application of the above-described procedure using the neural network modeling of

viscosity, density, and saturated vapor pressure as examples. The viscosity of organic compounds was modeled using a database[61] comprising 367 organic compounds of various classes: linear, branched, and mono- and bicyclic alkanes, alkenes, and alkynes, arenes, alcohols, ethers, esters, ketones, aldehydes, carboxylic acids, nitriles, imines, amines, amides, halogen- and sulfur-containing compounds, and nitro compounds. Two compounds were excluded from the set presented in an earlier publication,[61] because they had the same name but different viscosities (compounds 266 and 267). The database was split in ten different ways to form three sets: training (293 compounds), validation (37 compounds), and prediction sets (37 compounds). Using the stepwise linear regression procedure described above, the calculated set of descriptors was subjected to selection procedure for ten different patterns of database splitting. During construction of each linear-regression model, descriptors were successively included until the best predictive ability was attained for the validation set. The results of the linear-regression models obtained for 13 sets of descriptors with different maximum sizes of fragments (130 models) are presented in Table 1 and in Fig. 1.

As can be seen in Fig. 1, the use of fragmental descriptors based on more than two or three atoms does not improve the quality of regression models. The neural network models provided best statistical characteristics for the set of descriptors with a maximum fragment size equal to three. The optimum set of descriptors was chosen based on the root-mean-square error for the validation set. In-

Table 1. Averaged statistical characteristics of linear-regression models upon variation of the maximum size of descriptors for the simulation of viscosity of organic compounds*

Number of atoms	n	n_d	MLR			
			R_{av}	RMS_{tr}	RMS_{val}	RMS_{test}
1	146	38±20	0.9204	0.2172	0.2366	0.2407
2	531	53±12	0.9740	0.1260	0.1857	0.1853
3	1757	46±16	0.9794	0.1113	0.1950	0.2119
4	1974	42±22	0.9593	0.1336	0.2079	0.2341
5	2183	34±21	0.9531	0.1470	0.2113	0.2330
6	2413	36±21	0.9681	0.1307	0.1960	0.2207
7	2566	33±19	0.9662	0.1302	0.2088	0.2392
8	2649	35±22	0.9656	0.1337	0.2075	0.2305
9	2703	33±20	0.9652	0.1348	0.2077	0.2322
10	2732	35±22	0.9658	0.1330	0.2081	0.2316
11	2945	35±22	0.9657	0.1331	0.2044	0.2297
12	2759	35±22	0.9657	0.1331	0.2044	0.2297
13	2770	35±22	0.9657	0.1331	0.2044	0.2297

* Designations: n is the total number of descriptors, n_d is the average number of selected descriptors, MLR is multiple linear regression; R_{av} is the correlation coefficient; RMS_{tr}, RMS_{val}, and RMS_{test} are the root-mean-square errors for the training, validation, and prediction (test) sets, respectively.

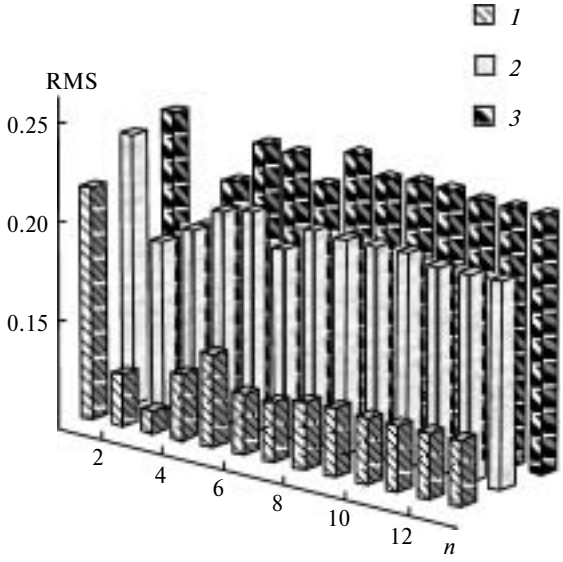

Fig. 1. Averaged root-mean-square error (RMS) *vs.* he maximum size (n) of fragmental descriptors of viscosity for the training (*1*), validation (*2*), and prediction sets (*3*).

deed, it is improper to be guided by either the minimum error for the training set (because this would give rise to overfitted models) or the minimum error for the prediction set (because data for this set are to be used only for estimating the predictive ability and not for model construction).

The mere existence of an optimal maximum size of the generated fragments, which ensures the best predictive ability of the models, is not obvious and, therefore, it deserves special attention. This is due to the fact that an increase in the size of the fragments entails a sharp increase in the number of their types, and hence, the number of fragmental descriptors. However, several mathematical theories (see below) imply that the predictive ability of a statistical model is deteriorated upon extension of the initial number of descriptors used for selection, all other things being equal (*i.e.*, for the same error for the training set and equal numbers of selected descriptors). Indeed, according to the Vapnik—Chervonenkis statistical prediction theory,[65] the minimum size of the set of compounds needed to attain the specified prediction quality depends on both the number of descriptors selected and the initial number of descriptors. In the latter case, for binary descriptors (so-called features), the minimum size of the set was shown to follow a logarithmic dependence on the logarithm of the number of initial descriptors. Thus, for an invariable size of the set, the model quality decreases as the initial number of descriptors increases. Thus, the effective number of descriptors in the statistical model (the Vapnik—Chervonenkis dimension) in not equal, in the general case, to the number of selected descriptors. It also depends on the initial number of descriptors, out of which the selection is carried out. The theory of inductive infer-

ences leads to the same conclusions.[48] The expected error of a statistical model for data not included in the training set is known[66] to be determined by the degree of data compression in terms of this model. The shorter the total length of data definition in the model and definition of the model itself, the smaller the error of prediction in terms of this model. The length of model M definition is equal to the quantity of information needed to choose this model out of a set with an *a priori* probability distribution $P(M)$, which can be approximated as $-\log P(M)$. It can be seen that the greater the initial number out of which descriptors are selected, the lower the *a priori* probability of the obtained model and, hence, the longer the model definition and the greater the expected prediction error. This leads to a conclusion that is highly important for the structure—property models. One cannot infinitely extend the set of input descriptors for statistical procedures expecting that the required descriptors would be anyhow automatically selected, because this would increase the probability of erroneous selection. Thus, for constructing models with the best predictive ability, one should optimize not only sets of selected descriptors but also the initial sets of descriptors inputted to automated selection procedures; this is demonstrated in the present study.

One more interesting item reflected in Table 1 is that the dependence of the average number of descriptors selected during modeling on the maximum size of generated descriptors (*i.e.*, the number of atoms in the fragment) passes through maxima. The positions of maxima in this dependence exactly coincide with the positions of minima of prediction errors. To interpret this fact, one can use the following regularity, which can be easily observed, at least, for physicochemical properties: as the size of the fragment increases, the fraction of fragmental descriptors valuable for modeling in the total number of descriptors decreases. The precise number of valuable descriptors is difficult to determine. In addition, they can be correlated with one another. Therefore, this tendency can be approximately estimated from the dependence of the average number of selected descriptors on the total number of generated fragments with a definite maximum size (Fig. 2). When the fragments are small, the benefit from the addition of valuable descriptors surpasses the loss caused by the increased probability of erroneous selection. Therefore, the number of selected descriptors increases and the error of prediction decreases. As the fragments become larger, the fraction of valuable descriptors diminishes in parallel with the increase in their total number; as a consequence, the loss caused by the increased probability of erroneous selection starts to surpass the benefit caused by the addition of valuable descriptors. Therefore, the model quality expressed as the predictive ability is deteriorated, *i.e.*, the prediction errors increase. Since in the stepwise selection procedure we used descriptors are included in the model as it reaches the best

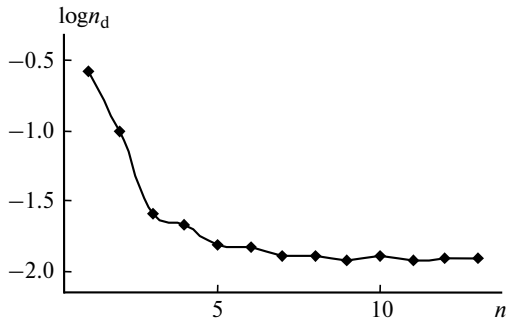

Fig. 2. Logarithm of the fraction of descriptors selected in the model (n_d) *vs.* the total number (n) in modeling of viscosity.

predictive ability, which is estimated from the validation set, the predictive ability is deteriorated in parallel with the decrease in the number of descriptors included in the model. In other words, as the size of fragments increases, some valuable descriptors become "noised" due to the increase in the total number of descriptors, and they can no longer be automatically selected; this results in a decrease in the number of selected descriptors.

Now we will consider the selected descriptors. When one models the viscosity, the most important descriptors (here we access the importance from the number of models that include the given descriptor) are the contribution of the number of nonhydrogen atoms in the molecule (N_a), the ratio of the number of Me groups attached to C atoms to the number of nonhydrogen atoms $N(\mathrm{Me—C})/N_a$, and the ratio of the number of *n*-propyl groups to the number of nonhydrogen atoms $N(\mathrm{Pr})/N_a$. In addition, one should note the importance of some other descriptors such as the number of amino groups $N(NH_2)$, the number of N atoms at the double bond $N(=\mathrm{N})/N_a$, the number of chains containing hydroxy groups $N(C_{sp^3}—C_{sp^3}—C_{sp^3}—OH)/N_a$, and the numbers of halogen atoms and amide groups. The first three descriptors might be related to the van der Waals interaction between molecules and the other, to the electrostatic interaction (including hydrogen bonding). Descriptors related to aromatic bonds were not taken into account, because this decreased the prediction quality.

After constructing a series of neural network models (350 models) with variation of the number of hidden neurons from two to eight, the optimal number of hidden neurons was chosen to be seven (Table 2), although the statistical parameters of models obtained with different numbers of hidden neurons did not differ much.

Figure 3 presents the dependence of the root-mean-square error of the prediction set (testing set) RMS_{test} and the root-mean-square error for the validation set RMS_{val} for 50 models constructed with seven hidden neurons. The slope ratio of the straight line corresponding to a linear approximation of the dependence (see Fig. 3) of the root-mean-square errors for this sets is positive. This

Table 2. Averaged RMS* value *vs.* the number of hidden neurons

Number of neurons in the hidden layer	RMS_{tr}	RMS_{val}	RMS_{test}
2	0.110	0.193	0.226
3	0.106	0.191	0.222
4	0.108	0.192	0.220
5	0.106	0.192	0.219
6	0.105	0.191	0.219
7	0.105	0.189	0.219
8	0.105	0.191	0.220

* For designations, see Table 1.

type of variation is observed for any ensembles of models with different numbers of hidden neurons. In some cases, negative slope angles can also be found (ensembles of models for the density of liquid organic compounds with numbers of hidden neurons other than the optimal number). The sign of the slope angle of this straight line point to either underfitted or overfitted model. Thus, when such a slope angle is obtained, there appears the possibility of improving the quality, *i.e.*, the predictive ability. Thus, of the 50 models constructed with the given number of hidden neurons, a half with lower root-mean-square errors for the validation set RMS_{val} should be chosen. Table 3 presents the statistical parameters for each stage of the modeling.

The final results (see Table 3) are much better than those averaged over linear-regression models and are more statistically valid, as they take into account a more extensive ensemble of models. Correlation of the models averaged over the whole array of calculated data for all sets with experimental results is presented in Fig. 4.

Figure 5 shows the distribution of the prediction errors for viscosity. It is noteworthy that the root-mean-square error RMS_{test} for the predicted values of the property considered is 0.141 logarithmic units for the whole set. The compounds whose viscosity proved to be predicted with great errors are mainly high-polarity compounds (some of them can form hydrogen bonds):

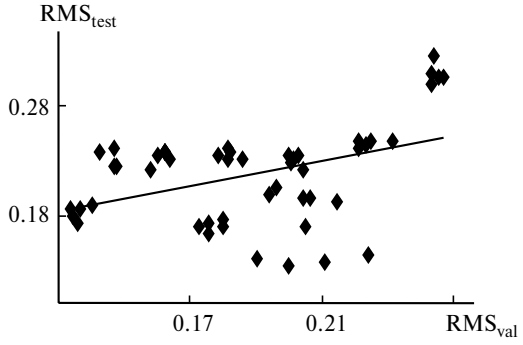

Fig. 3. Averaged RMS_{test} value *vs.* RMS_{val} for viscosity.

Table 3. Statistical parameters* of the modeling at different stages of neural network investigations

Stage	*R*	RMS_{tr}	RMS_{val}	RMS_{test}
MLR	0.9794	0.111	0.195	0.212
Averaging over 50 models	0.9815	0.105	0.189	0.219
Calculation from individual contributions for 50 models	0.9904	0.078	0.177	0.208
Averaging over 25 best models	0.9814	0.106	0.161	0.212
Calculation from individual contributions for 25 best models	0.9885	0.084	0.104	0.141

* For designations, see Table 1.

2-methylpentane-2,4-diol ($\Delta = 0.76$), glycerol trinitrate ($\Delta = -0.72$), formic acid ($\Delta = -0.58$), dibutyl *o*-phthalate ($\Delta = -0.60$), cyclohexanol ($\Delta = -0.64$), 2-methoxyethanol ($\Delta = 0.58$), acrylic acid ($\Delta = 0.54$), trifluoroacetic acid ($\Delta = 0.54$), 4-hydroxy-4-methylpentan-2-one (0.52), ethoxybenzene ($\Delta = 0.49$), 2-methylbutan-2-ol ($\Delta = -0.48$), butane-1,3-diol ($\Delta = -0.48$), dibutyl maleate ($\Delta = 0.45$), glycerol ($\Delta = 0.43$), and *N*,*N*-dimethylformamide ($\Delta = 0.42$).

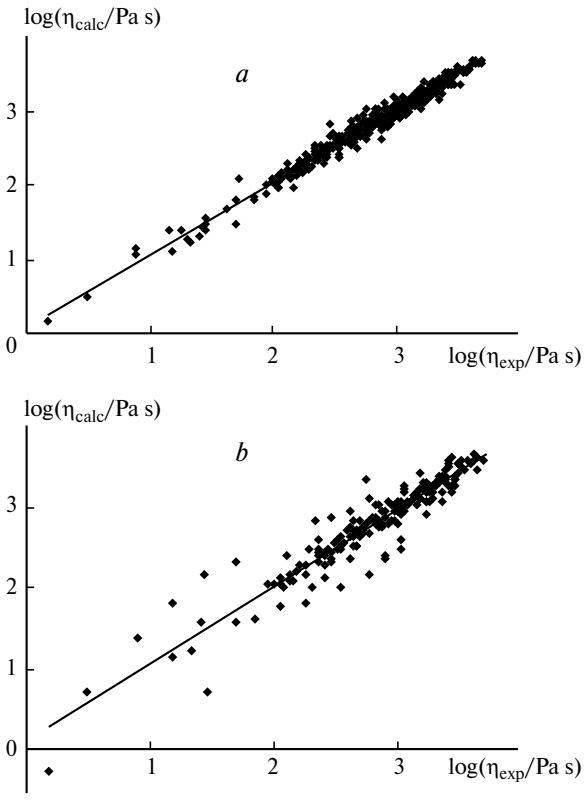

Fig. 4. Results of viscosity (log(η/Pa s)) modeling: training set (*a*), prediction set (*b*).

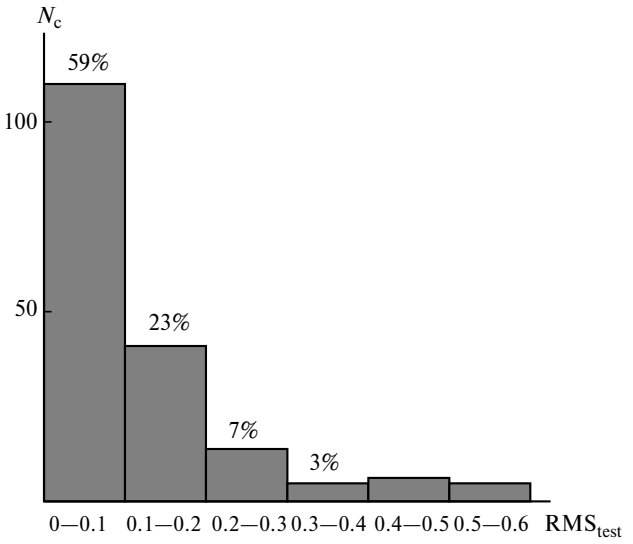

Fig. 5. Distribution of prediction errors (RMS_{test}) for the viscosity of organic compounds ($\log(\eta/\text{Pa s})$); N_c is the number of compounds.

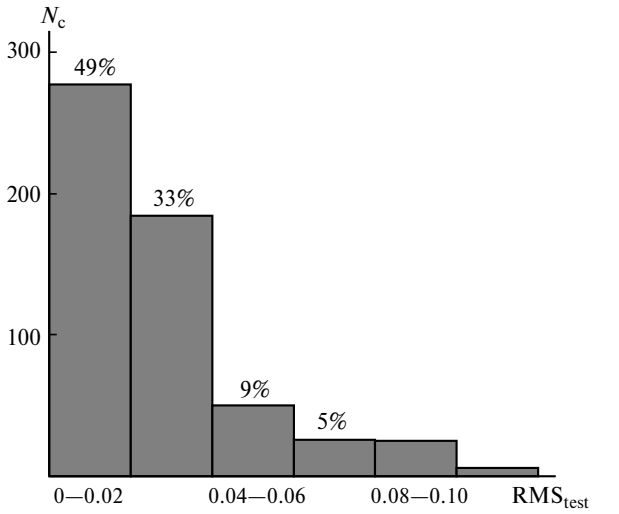

Fig. 6. Distribution of prediction errors (RMS_{test}) for the density of liquid (d^{20}/g cm^{-3}); N_c is the number of compounds.

A similar procedure was used to model the density[67] of liquid organic compounds (using a database comprising 803 compounds: alkanes, alkenes, alkynes, arenes, allenes, alcohols, ethers, esters, nitro compounds, aldehydes, carboxylic acids, ketones, nitriles, amines, imines, amines, compounds with heteroatoms, and mono-, bi-, and tricyclic structures) and the saturated vapor pressure[68] (using a database comprising 352 hydrocarbons: linear, branched, and cyclic alkanes, alkenes, arenes, and halo derivatives).

The parameters important for modeling the density of organic compounds include the numbers of sp^3- and sp^2-hybridized C atoms ($N(C_{sp^3})/N_a$ and $N(H_2C=)/N_a$) and the relative numbers of various heteroatoms (in particular, halogens, oxygen, nitrogen, silicon, sulfur, *etc.*), which can be due to different weights and covalent and van der Waals radii of different atoms. Diverse corrections are matched by descriptors such as the number of triple bonds C≡C and the descriptors characterizing the degree of branching. Descriptors corresponding to aromatic bonds were not taken into account in the modeling, because this did not improve the prediction quality. The root-mean-square prediction error RMS_{test} for the whole set amounted to 0.051 g cm^{-3}. Compounds whose density (g cm^{-3}) tends to be predicted with a substantial error were those that contained fragments and heteroatoms seldom encountered in the database, namely, selenophenol ($\Delta = 0.47$), iodomethyl(trimethyl)silane ($\Delta = -0.29$), 1,4-diiodobutane ($\Delta = -0.27$), *R*(−)-propylene glycol ($\Delta = 0.25$), germanium(IV) chloride ($\Delta = 0.20$), 1,2-dithiobis(trimethylsilyl)ethane ($\Delta = -0.20$), and some other. Figure 6 shows the distribution of the prediction errors for density for a number of liquid organic compounds over the resulting ensemble of the 25 best models. Correlation of the calculated data averaged over the ensemble of models with the experimental density values for liquid organic compounds is shown in Fig. 7.

The descriptors that proved to be most significant for modeling the saturated vapor pressures of organic compounds include the number of carbon atoms squared $N^2(C)$; the logarithm of the total number of nonhydrogen atoms $\log N_a$; the number of halogen atoms attached to a carbon atom incorporated in a six-membered aromatic

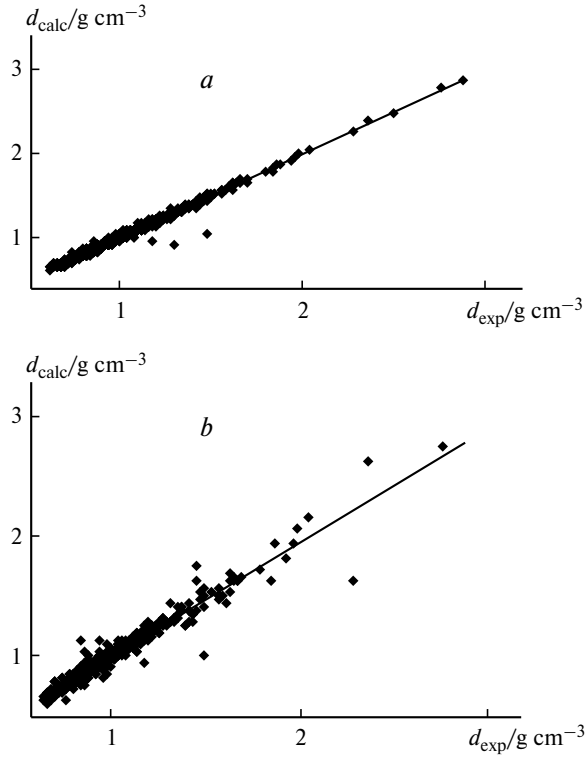

Fig. 7. Results of density (d^{20}/g cm^{-3}) modeling: training set (*a*), prediction set (*b*).

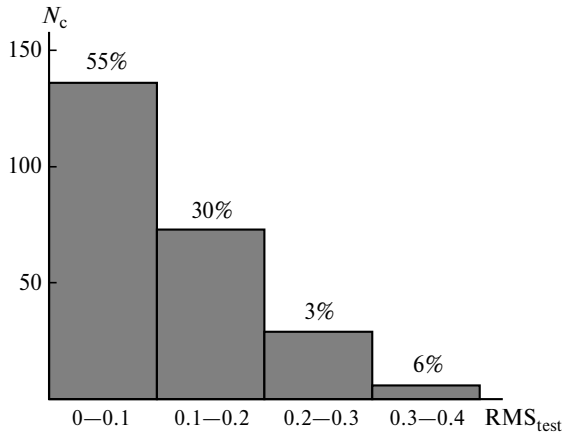

Fig. 8. Distribution of prediction errors (RMS_{test}) for the saturated vapor pressure of hydrocarbons and halohydrocarbons (P_{sat}/Pa).

ring $N[C_{Ar}—Hal]$; the number of methylene groups attached to a carbon atom incorporated in a six-membered aromatic rings $N[C_{Ar}—CH_2]$; the square root of the number of fluorine atoms $\sqrt{N[F]}$; the number of C—C single bonds $N(C—C)/N_a$; the number of diatomic fragments of an aromatic system $N[—C_{Ar}—C_{Ar}]$, and other. This particular set of the most important descriptors is, apparently, due to the predominant role of the van der Waals interactions in this case. The predictive ability of the model for the saturated vapor pressure of organic compounds is rather high, RMS_{test} amounts to 0.152 logarithmic units. However, for some compounds, the logarithms of the vapor pressure are predicted with great errors; these are iodobenzene ($\Delta = -0.81$), dibromodifluoromethane ($\Delta = -0.56$), bromotrifluoromethane ($\Delta = 0.42$), and 1,1,2-trifluoroethane ($\Delta = -0.38$), *i.e.*, mainly unsaturated and polar compounds containing halogen atoms. In addition, spatial effects play an important role for these structures and they cannot always be taken into account using fragmental descriptors. Figure 8 shows the distribution of the prediction errors for saturated vapor pressure

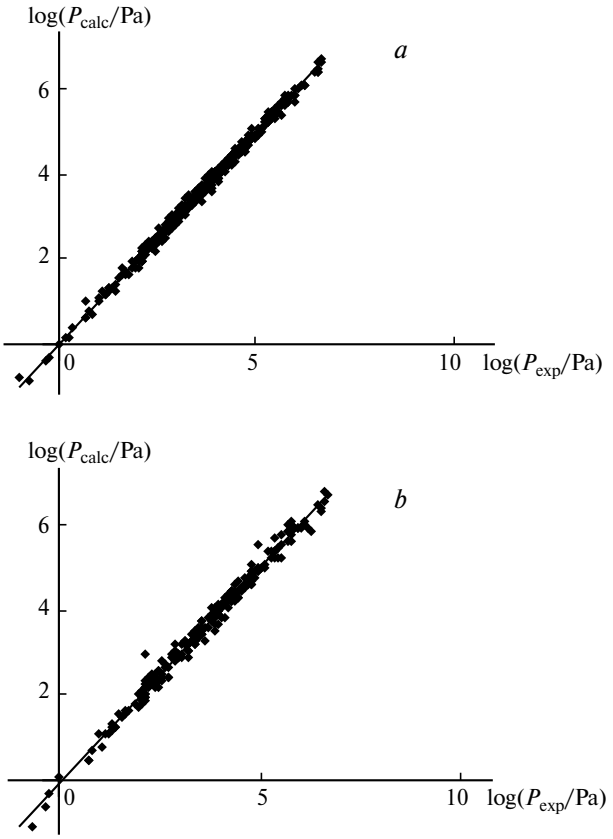

Fig. 9. Results of modeling of the saturated vapor pressure (P_{sat}/Pa): training set (*a*), prediction set (*b*).

of hydrocarbons and halogenated hydrocarbons averaged over the 25 best models.

Correlation of the calculated data for the saturated vapor pressure averaged over the array of models for all sets with the experimental values is shown in Fig. 9.

The summary table (Table 4) presents the resultant statistical parameters of the obtained linear-regression and neural network models for the above physicochemical properties. The predictive ability of the neural network

Table 4. Parameters[a] of neural network and linear-regression models (numbers in parentheses) for a series of physicochemical properties

Property	N_c[b]	n_d[c]	R_{av}	RMS_{tr}	RMS_{val}	RMS_{test}	Fragment types[22]
Viscosity[58], log(η/Pa s)	367	46±16	0.9885	0.084	0.104	0.141	p1, p2, p3
			(0.9794)	(0.111)	(0.195)	(0.212)	
Density[65], d^{20}/g cm^{-3}	803	69±21	0.9980	0.021	0.046	0.051	p1, p2, p3, p4, p5, c4, c5, s4, s5
			(0.9885)	(0.038)	(0.055)	(0.067)	
Saturated vapor pressure[66], $\log(P_{sat}/Pa)$	352	56±9	0.9971	0.090	0.122	0.152	p1, p2
			(0.9902)	(0.198)	(0.248)	(0.258)	

[a] For designations, see Table 1.
[b] The number of compounds.
[c] The average number of selected descriptors.

models (which is evaluated most correctly based on the RMS_{test} value, *i.e.*, on the root-mean-square error for the prediction set) exceeds similar characteristics of the regression models. Moreover, the constructed neural network models surpass in some parameters the best of the published models. In particular, the results of prediction of the density of liquids for compounds of various classes proved to be close to the best of the published models:[69] for 303 compounds, $R = 0.9874$ and $s = 0.0458$. However, our model is based on a much more representative data set. The model we obtained for predicting the viscosity (Pa s) of liquid organic compounds substantially surpasses in all parameters the best of the published models.[61,62] In particular, the Ivanciuc model[61] for 337 compounds provides a cross-validation root mean square error of 0.38 logarithmic units, while the Katritzky model[62] for 361 compounds gives a standard deviation of 0.22 logarithmic units (in our model, this is 0.14 for the prediction set). The accuracy of prediction of the saturated vapor pressure (Pa) in our model is better than in the Jurs model[68] (RMS_{test} for the prediction set is 0.209 logarithmic units, whereas in our case, for evaluation of 25 best models, this is 0.152 logarithmic units; in the case of 10 best models, the predictive ability further increases, the value being 0.142 logarithmic units) and much better than in the other published models: for 476 compounds, Basak obtained a model with the characteristics $R = 0.9182$ and $s = 0.29$ logarithmic units; for the Layang model based on 479 compounds, $R = 0.9798$ and $s = 0.534$ logarithmic units, and in the Katritzky model, constructed using the data for 411 compounds, $R = 0.9742$ and $s = 0.331$ logarithmic units.[69]

Thus, the use of fragmental descriptors in combination with the ANN system provides high-accuracy models for predicting a number of physicochemical properties of organic and organoelement compounds resorting only to topological parameters.

References

1. L. Pogliani, *Chem. Rev.,* 2000, **100**, 3827.
2. H. J. Bernstein, *J. Chem. Phys.,* 1952, **20**, 263.
3. H. J. Bernstein, *Trans. Faraday Soc.*, 1962, **58**, 2285.
4. S. W. Benson, F. R. Cruickshank, D. M. Golden, G. R. Haugen, H. E. O'Neal, A. S. Rodgers, R. Shaw, and R. Walsh, *Chem. Rev.,* 1969, **69**, 279.
5. V. M. Tatevskii, *Teoriya fiziko-khimicheskikh svoistv molekul i veshchestv* [*Theory of Physico-Chemical Properties of Molecules and Substances*], MGU, Moscow, 1987, 239 pp. (in Russian).
6. E. A. Smolenskii, *Zh. Fiz. Khim.*, 1964, **38**, 1288 [*J. Phys. Chem. USSR*, 1964, **38** (Engl. Transl.)].
7. E. A. Smolenskii, *Dokl. Akad. Nauk SSSR*, 1976, **230**, 373 [*Dokl. Chem.*, 1976 (Engl. Transl.)].
8. C. T. Zahn, *J. Chem. Phys.*, 1934, **2**, 671.
9. M. Sounders, Jr., C. S. Matthews, and C. O. Hurd, *Ind. Eng. Chem.*, 1949, **41**, 1037.
10. M. Sounders, Jr., C. S. Matthews, and C. O. Hurd, *Ind. Eng. Chem.*, 1949, **41**, 1048.
11. J. L. Franklin, *Ind. Eng. Chem.*, 1949, **41**, 1070.
12. J. L. Franklin, *J. Chem. Phys.*, 1953, **21**, 2029.
13. T. L. Allen, *J. Chem. Phys.*, 1959, **31**, 1039.
14. A. J. Kalb, A. L. H. Chung, and T. L. Allen, *J. Am. Chem. Soc.*, 1966, **88**, 2938.
15. V. M. Tatevskii, *Khimicheskoe stroenie uglevodorodov i zakonomernosti v ikh fiziko-khimicheskikh svoistvakh* [*Chemical Structure of Hydrocarbons and Regular Features in Their Physicochemical Properties*], MGU, Moscow, 1953, 320 pp. (in Russian).
16. V. M. Tatevskii, V. A. Benderskii, and S. S. Yarovoi, *Zakonomernosti i metody rascheta fiziko-khimicheskikh svoistv parafinovykh uglevodorodov* [*Regularities and Methods of Calculation of Physicochemical Properties of Paraffin Hydrocarbons*], MGU, Moscow, 1960, 114 pp. (in Russian).
17. V. M. Tatevskii, *Klassicheskaya teoriya stroeniya molekul i kvantovaya mekhanika* [*Classical Theory of Molecular Structure and Quantum Mechanics*], Khimiya, Moscow, 1973, 516 pp. (in Russian).
18. E. A. Smolenskii and L. V. Kocharova, *Dokl. Akad. Nauk SSSR*, 1982, **264**, 112 [*Dokl. Chem.*, 1982 (Engl. Transl.)].
19. D. E. Petelin, V. A. Palyulin, N. S. Zefirov, and J. U. McFarland, *Dokl. Akad. Nauk*, 1992, **327**, 508 [*Dokl. Chem.*, 1992 (Engl. Transl.)].
20. I. I. Baskin, V. A. Palyulin, and N. S. Zefirov, *Conf. "Molecular Graphs in the Chemical Investigations," Abstrs.*, Kalinin, 1990, 5 (in Russian).
21. I. I. Baskin, V. A. Palyulin, and N. S. Zefirov, *1-st Vsesoyuz. konf. po teoreticheskoi organicheskoi khimii* [*1-st All-Union Conf. on Theor. Org. Chem.*], *Abstrs.,* Volgograd, 1991, 557 (in Russian).
22. N. V. Artemenko, I. I. Baskin, V. A. Palyulin, and N. S. Zefirov, *Dokl. Akad. Nauk*, 2001, **381**, 203 [*Dokl. Chem.*, 2001 (Engl. Transl.)].
23. I. I. Baskin, N. M. Galberstam, V. A. Palyulin, and N. S. Zefirov, *VII Vseros. konf. "Neirokomp'yutery i ikh primenenie" NKP-2001 s mezhdunarodnym uchastiem* [*VII All-Russian Conf. "Neurocomputers, their Applications" NKP-2001 with Int. Participation*], *Proc.*, Ed. A. I. Galushkin, V. A. Trapeznikov Institute of Management Problems of the RAS, Moscow, 2001, 419 (in Russian).
24. M. I. Kumskov, L. A. Ponomareva, E. A. Smolenskii, D. F. Mityushev, and N. S. Zefirov, *Izv. Akad. Nauk, Ser. Khim.*, 1994, 1391 [*Russ. Chem. Bull.*, 1994, **43**, 1317 (Engl. Transl.)].
25. S. M. Free and J. W. Wilson, *J. Med. Chem.*, 1964, **7**, 395.
26. A. Cammarata, *J. Med. Chem.*, 1972, **15**, 573.
27. T. Fujita and T. Ban, *J. Med. Chem*, 1971, **14**, 148.
28. V. V. Avidon, *Khim.-Farm. Zhurn.*, 1974, **8**, 22 [*Pharm. Chem. J.*, 1974, **8** (Engl. Transl.)].
29. V. V. Avidon, V. S. Arolovich, S. P. Kozlova, and L. A. Piruzyan, *Khim.-Farm. Zhurn.*, 1978, **12**, 88 [*Pharm. Chem. J.*, 1978, **12** (Engl. Transl.)].
30. V. V. Avidon, I. A. Pomerantsev, A. B. Rozenblit, and V. E. Golender, *J. Chem. Inf. Comp. Sci.*, 1982, **22**, 207.
31. A. B. Rozenblit and V. E. Golender, *Logical Combinatorial Algorithms for Drug Design, Research Studies Press,* Wiley and

Sons, New York—Chihester—Brisbane—Toronto, 1983, 352 pp.
32. Yu. V. Borodina, D. A. Filimonov, and V. V. Poroikov, *Pharm. Chem. J.*, 1996, **30**, 760.
33. D. Filimonov, V. Poroikov, Yu. Borodina, and T. Gloriozova, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 666.
34. V. V. Poroikov, D. A. Filimonov, Yu. V. Borodina, A. A. Lagunin, and A. Kos, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 1349.
35. S. K. Kotovskaya, L. A. Tyurina, E. Yu. Chernova, G. A. Mokrushina, O. N. Chupakhin, A. P. Novikova, and V. I. Il´enko, *Khim.-Farm. Zhurn.*, 1989, **22**, 310 [*Pharm. Chem. J.*, 1989, **22** (Engl. Transl.)].
36. G. Klopman, *J. Am. Chem. Soc.*, 1984, **106**, 7315.
37. G. Klopman and H. S. Rosenkranz, *Mutat. Res.*, 1994, **305**, 33.
38. A. R. Cunningham, G. Klopman, and H. S. Rosenkranz, *Mutat. Res.*, 1998, **405**, 9.
39. M. Brinn, P. T. Walsh, M. P. Payne, and B. Bott, *SAR QSAR Environ. Res.,* 1993, 1, 169.
40. M. W. Brinn, M. P. Payne, and P. T. Walsh, *Chem. Eng. Res. Des.*, 1993, **71**(A3), 337.
41. F. R. Burden, *Quant. Struct.-Act. Relat.*, 1996, **15**, 7.
42. T. Hurst and T. Heritage, *Thes. 213th ACS Natl. Meeting*, San Francisco, CA, 1997, CINF 019.
43. A. Hoskuldsson, *J. Chemometrics*, 1988, **2**, 211.
44. N. S. Zefirov and V. A. Palyulin, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 1112.
45. I. I. Baskin, A. O. Ait, N. M. Galberstam, V. A. Palyulin, M. V. Alfimov, and N. S. Zefirov, *Dokl. Akad. Nauk*, 1997, **357**, 57 [*Dokl. Chem.*, 1997 (Engl. Transl.)].
46. J. Zupan and J. Gasteiger, *Neural Networks for Chemists. An Introduction,* Wiley—VCH Publishers, Weinheim—New York—Chichester—Brisbane—Singapore—Toronto, 1993, **1**, 244 pp.
47. I. I. Baskin, V. A. Palyulin, and N. S. Zefirov, *Dokl. Akad. Nauk*, 1993, **332**, 713 [*Dokl. Chem.*, 1993 (Engl. Transl.)].
48. A. A. Ezhov and S. A. Shumskii, *Neirokomp'yuter i ego primenenie v ekonomike* [*Neurocomputer and its Use in the Economy*], MIFI, Moscow, 1998, 57 (in Russian).
49. M. Chastrette, D. Cretin, and F. Tiyal, *C. R. Acad. Sci.,* 1994, **318**, 1059.
50. A. Bondi, *J. Phys. Chem.*, 1964, **68**, 441.
51. S. C. Basak, B. D. Gute, and G. D. Grunwald, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 651.
52. C. Liang and D. A. Gallagher, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 321.
53. M. Karelson and A. Perkson, *Comput. Chem.*, 1999, **23**, 49.
54. A. A. Gakh, E. G. Gakh, B. G. Stumper, and D. W. Noid, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 832.
55. R. Zhang, S. Liu, M. Liu, and Z. Hu, *Comput. Chem.*, 1997, **21**, 335.
56. S. Liu, R. Zhang, M. Liu, and Z. Hu, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 1146.
57. K. G. Joback and R. C. Reid, *Chem. Eng. Commun.*, 1987, **57**, 233.
58. P. Škubla, *Collect. Czech. Chem. Commun.*, 1985, **50**, 1907.
59. C. Hansch and T. Fujita, *J. Am. Chem. Soc.*, 1964, **86**, 1616.
60. T. Fujita, J. Iwasa, and C. Hansch, *J. Am. Chem. Soc.*, 1964, **86**, 5175.
61. O. Ivanciuc, T. Ivanciuc, P. A. Filip, and D. Cabrol-Bass, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 515.
62. A. R. Katritzky, K. Chen, Y. L. Wang, M. Karelson, B. Lucic, N. Trinajstic, T. Suzuki, and G. Schuurmann, *J. Phys. Org. Chem.*, 2000, **13**, 80.
63. T. Suzuki, K. Ohtaguchi, and K. Koide, *Comput. Chem. Eng.*, 1996, **20**, 161.
64. T. Suzuki, R.-U. Ebert, and G. Schuurmann, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 1122.
65. V. E. Vapnik and A. Ya. Chervonenkis, *Teoriya raspoznavaniya obrazov* [*Image Recognition Theory*], Nauka, Moscow, 1979, 237 pp. (in Russian).
66. J. Rissanen, in *Complexity, Entropy and the Physics of Information*, Ed. W. H. Zurek, Addison-Wesley, California, Redwood City, 1990, 117.
67. *Flukalog Database*, Fluka Chemie AG, 1995.
68. E. S. Goll and P. C. Jurs, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 1081.
69. A. R. Katritzky, U. Maran, V. S. Lobanov, and M. Karelson, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 1.

Received February 26, 2002;
in revised form May 23, 2002