

Prediction of Physical Properties of Organic Compounds Using Artificial Neural Networks within the Substructure Approach

N. V. Artemenko, I. I. Baskin, V. A. Palyulin, and Academician N. S. Zefirov

Received July 30, 2001

In spite of the rapid development of quantum-chemical molecular simulation methods, the central place in the prediction of the majority of physicochemical properties is occupied by empirical approaches based on the use of different molecular structure descriptors [1]. One of the first empirical approaches for the approximation of physicochemical properties was based on linear additive schemes where the numbers of the simplest fragments in the molecule were used as descriptors [2–4]. However, in many cases the dependence of physicochemical properties on descriptors is substantially nonlinear, and, commonly, its general form is not known in advance. In this case, the use of the artificial neural network technique provides an efficient solution to the problem of predicting properties of organic compounds [5–7].

Previously [8], we demonstrated that any molecular graph invariant (any property of a chemical compound) can be unambiguously represented as a linear combination of the number of occurrence of some structure fragments (connected or disconnected) or as a polynomial of the number of occurrence of some connected substructures. According to the Kolmogorov–Arnold theorem [9], any continuous function (including any polynomial) can be approximated using a three-layer neural network; therefore, any of the properties that are rather insensitive to stereoisomerism (which include the majority of physicochemical properties) can be approximated by output values of a multilayer neural network that employs the numbers of occurrence of

connected fragments in the molecular graph as input values. These concepts formed the basis of the approach to predicting physicochemical properties of organic compounds considered in this work.

In this paper, we demonstrate that the combined use of substructure descriptors and artificial neural networks can be considered as a powerful tool for predicting physicochemical properties of organic compounds.

Within this approach, substructure descriptors are chains of atoms containing 1 to 15 vertices (denoted by $p1-pf$), 3- to 15-membered cycles ($c3-cf$), branched fragments containing 4 to 6 vertices ($s4-s6$), bicycles containing 6 to 15 vertices ($b0-bd$), and tricycles containing 12 to 15 vertices ($t0-te$).

In addition, within this method a special internal hierarchical classification of atomic types was developed. Three symbols are used for coding an atom corresponding to some element. The most comprehensive classification (including valence, hybridization type, and atomic charge) was proposed for organogenic elements (H, C, N, O, S, Se, P, As, Si, F, Cl, Br, and I; see Table 1).

The classification is based on the following principle: each subsequent symbol refines the previous. Thus, there are several generalization levels, which can be demonstrated with an example of the carbon atom in the methyl group. Classification levels are presented below (generalization level decreases from left to right):

Level no.	Level 1	Level 2	Level 3	Level 4
Atomic type	•	C	C_{sp^3}	$-\text{CH}_3$
Designation of atomic type	---	C_{--}	CA_{-}	CA1

In addition, special types including several different classes of either the same atoms or atoms of one subgroup are additionally formed if necessary. The first classification level corresponds to the “generalized”

atomic type; types of chemical elements are included at the second level. The third and fourth classification levels additionally include the character of hybridization, bond environment of the atom, its formal charge, and the number of hydrogen neighbors. The most detailed classification of atoms corresponding to the fourth level is presented in Table 1. Note that the classification for the Se atom is similar to that for the S atom; for As, it is

Moscow State University, Vorob'evy gory, Moscow,
119899 Russia

Table 1. Classification of atomic types

$-\text{CH}_3$ CA1	>CH_2 CA2	>CH CA3	$-\overset{ }{\underset{ }{\text{C}}}-$ CA4	$=\text{CH}_2$ CD1	>CH CD2
$=\text{C}<$ CD3	$=\text{C}=\text{}$ CD4	>CH CB1*	$\text{>C}-$ CB2*	>CH CH1**	$\text{>C}-$ CH2**
$\equiv\text{CH}$ CT1	$\equiv\text{C}-$ CT2	$\equiv\text{C}-$ CN-	$-\text{NH}_2$ NA1	>NH NA2	$\text{>N}-$ NA3
$=\text{NH}$ ND1	>N ND2	$\equiv\text{N}$ NT-	$\equiv\text{N}^-$ NN-	$-\text{NH}_3^+$ NC1	>NH_2^+ NC2
>NH^+ NC3	>N^+ NC4	>NH^+ NE1	$=\text{N}^+=$ NE2	$\text{>N}^+=$ NE3	$\equiv\text{N}^+$ NE4
$\text{>N}^+=$ NHC**	>NH NH1**	$\text{>N}-$ NH2**	>N NHD**	>N NB2*	$\text{>N}^+=$ NBC*
$-\text{N}\equiv$ N5-***	$-\text{OH}$ O1-	$-\text{O}-$ O2-	$=\text{O}$ OD-	$-\text{O}^-$ ON-	>O^+ OC-
>O OH1**	>O^+ OBC*	>S SH1*	>S^+ SBC*	$-\text{SH}$ S1-	$-\text{S}-$ S2-
$=\text{S}$ SD1	$=\text{S}=\text{}$ SD2	$\text{>S}=\text{}$ SD3	$\text{>S}\equiv$ SD4	$-\text{S}^-$ SN-	>S^+ SC-
$\text{>S}<$ S6-	$\text{>P}-$ P3-	$\text{>P}=\text{}$ P5-	>P^+ PC-	>P PH-	>Si IB-
$-\text{SiH}_3$ IA1	>SiH_2 IA2	$\text{>SiH}-$ IA3	$-\overset{ }{\underset{ }{\text{Si}}}-$ IA4	$\text{>Si}=\text{}$ ID-	$-\text{F}$ HF-
$-\text{Cl}$ HL1	$-\text{Cl}=\text{}$ HL2	>Cl^+ HLC	$-\text{Cl}\equiv$ HL3	$\text{>Cl}\equiv$ HL4	

* In six-membered aromatic cycles or heterocycles.

** In five-membered aromatic cycles or heterocycles.

*** The formal designation of the nitrogen atom in the nitro group (independently of its structural representation in databases).

similar to that for P; and for Br and I, it is similar to that for Cl. Alkali and alkaline earth elements are denoted by M** (three generalization levels). All other elements can be denoted by ** using only two generalization levels: particular element and any atom.

The proposed scheme of fragment classification was implemented in the computer program Fragment (developed using the Delphi programming language), which provides the determination of the number of occurrences of each of the substructures in each of the structures of the studied series of compounds in the prediction of physical properties and biological activity.

The possibilities of this approach can be demonstrated in problems of predicting formation enthalpy,

polarizability, refraction index, boiling point, density, viscosity, and saturated vapor pressure for different organic compounds. The three former properties are rather adequately predicted using additive schemes; therefore, here we dealt with the four latter properties. In all cases, studies were performed by the following scheme. At the first step, for all compounds from the database including information on the structures of chemical compounds and their properties, fragment descriptors (numbers of occurrences of structure fragments in the chemical structure) were calculated; the maximum size of fragments was varied from 1 to 10 atoms. Next, three or four nonlinear modifications (square, square root, logarithm, and ratio to the number

Table 2. Parameters of neural-network and linear-regression models

Parameters of models	Boiling point, °C [10]	Viscosity, $\log \eta$ [Pa s] [11]	Density, g/cm ³ [10]	Saturated vapor pressure, $\log(VP)$ [Pa] [12]
Number of compounds	510	367	803	349
Number of descriptors	4–71	21–62	19–90	18–62
Neural-network model				
R_{av}	0.9899	0.9885	0.9935	0.9981
RMS_{train}	9.54	0.084	0.020	0.087
RMS_{valid}	15.02	0.104	0.026	0.167
$RMS_{predict}$	18.10	0.141	0.038	0.219
MLR				
R_{av}	0.9730	0.9794	0.9897	0.9941
RMS_{train}	14.56	0.111	0.036	0.151
RMS_{valid}	19.42	0.195	0.055	0.257
$RMS_{predict}$	21.21	0.212	0.067	0.276
Types of fragments	$p1, p2, p3$	$p1, p2, p3$	$p1, p2, p3, p4, p5, c4, c5, s4, s5$	$p1, p2, p3, p4, p5, p6, c5$

Note: MLR is multiple linear regression; R_{av} is the correlation coefficient; RMS_{train} , RMS_{valid} , and $RMS_{predict}$ are the root-mean-square errors on the training set, validation set, and set for estimating the predictive ability, respectively.

of non-hydrogen atoms) were calculated for each descriptor. After this, the database was divided into three sets: training set (80% of compounds), validation set (10% of compounds), and a set for estimating the predictive ability of the model (10% of compounds). Partitioning was performed in 10 different ways, so that each compound from the database occurred once in each of the two latter sets. Next, for each initial set of descriptors (differing in the maximum size of fragments) and each partitioning of the database, descriptors were selected using the stepwise multiple linear regression procedure. After this, the optimal descriptor set was selected among 10 initial sets according to the average prediction error on validation sets, and selected descriptor sets were further used in the study using multilayer neural networks with back propagation of errors. Next, for each partitioning of the database, five neural-network models with different numbers of hidden neurons (varied from 2 to 8) were constructed; learning was performed using the generalized δ -rule (rate parameter 0.25, momentum 0.9) until the minimum prediction error on the validation set was attained. After this, the optimal number of hidden neurons providing the smallest errors on validation sets was determined, and prediction results of a half of the best (i.e., providing the smallest error on the validation set) models for all compounds were averaged. As a result, for each property, we obtained the following four parameters: the average correlation coefficient R_{av} for training sets and root-mean-square errors on all three types of sets. Because information from the third set was involved neither in the construction of models nor in their selection, it was the root-mean-square error on this set that served as an adequate estimate of the predictive ability

of constructed models. Table 2 presents parameters of obtained regression and neural-network models for the above physicochemical properties.

From Table 2, it is easily seen that the predictive ability of neural-network models (which is most correctly estimated by the value of $RMS_{predict}$, i.e., by the root-mean-square error on the set for estimating the predictive ability) is higher than the analogous characteristics of regression models. In addition, constructed neural-network models in some characteristics are superior to the best model published previously. In particular, the accuracy of the prediction of boiling temperature for a nonuniform set was the best among the published models of this series (see [13]). Results of the prediction of the density of liquids for compounds of different types were close to the best among the published models (see [14]); however, our model was constructed based on a much more representative set. The model for predicting the viscosity of liquid organic compounds constructed in this work is superior to the best published models in all characteristics (see [11, 15]). The accuracy of the prediction of saturated vapor pressure by our model was comparable to that for the Jurs model [12] and significantly higher than the accuracy for the other published models [14].

REFERENCES

1. Pogliani, L., *Chem. Rev.*, 2000, vol. 100, no. 10, pp. 3827–3858.
2. Bernstein, H.J., *J. Chem. Phys.*, 1952, vol. 20, no. 2, pp. 263–269.
3. Benson, S.W., Cruickshank, F.R., Golden, D.M., *et al.*, *Chem. Rev.*, 1969, vol. 69, no. 3, pp. 279–324.

4. Smolenskii, E.A., *Zh. Fiz. Khim.*, 1964, vol. 38, no. 5, pp. 1288–1291.
5. Baskin, I.I., Ait, A.O., Halberstam, N.M., *et al.*, *Dokl. Akad. Nauk*, 1997, vol. 357, no. 1, pp. 57–59 [*Dokl. Phys. Chem.* (Engl. Transl.), vol. 357, nos. 1–3, pp. 353–355].
6. Zupan, J. and Gasteiger, J., *Neural Networks for Chemists. An Introduction*. Weinheim; N.Y.; Chichester; Brisbane; Singapore; Toronto, Wiley-VCH, 1993, vol. 1.
7. Baskin, I.I., Palyulin, V.A., and Zefirov, N.S., *Dokl. Akad. Nauk*, 1993, vol. 332, no. 6, pp. 713–716.
8. Baskin, I.I., Skvortsova, M.I., Stankevich, I.V., and Zefirov, N.S., *J. Chem. Inf. Comput. Sci.*, 1995, vol. 35, no. 3, pp. 527–531.
9. Kolmogorov, A.N., *Dokl. Akad. Nauk SSSR*, 1957, vol. 114, no. 5, pp. 953–956.
10. *Flukalog Database*. Fluka Chemie AG, 1995.
11. Ivanciuc, O., Ivanciuc, T., Filip, P.A., and Cabrol-Bass, D., *J. Chem. Inf. Comput. Sci.*, 1999, vol. 39, no. 3, pp. 515–524.
12. Goll, E.S. and Jurs, P.C., *J. Chem. Inf. Comput. Sci.*, 1999, vol. 39, no. 6, pp. 1081–1089.
13. Tetteh, J., Suzuki, T., Metcalfe, E., and Howells, S., *J. Chem. Inf. Comput. Sci.*, 1999, vol. 39, no. 3, pp. 491–507.
14. Katritzky, A.R., Maran, U., Lobanov, V.S., and Karelson, M., *J. Chem. Inf. Comput. Sci.*, 2000, vol. 40, no. 1, pp. 1–18.
15. Katritzky, A.R., Chen, K., Wang, Y.L., *et al.*, *J. Phys. Org. Chem.*, 2000, vol. 13, no. 1, pp. 80–86.