MATHEMATICS METHODS AND INFORMATION SYSTEMS IN CHEMICAL TECHNOLOGY

МАТЕМАТИЧЕСКИЕ МЕТОДЫ И ИНФОРМАЦИОННЫЕ СИСТЕМЫ В ХИМИЧЕСКОЙ ТЕХНОЛОГИИ

ISSN 2686-7575 (Online)

https://doi.org/10.32362/2410-6593-2020-15-6-84-103

UDC 541.6

RESEARCH ARTICLE

Structure–property models of organic compounds based on molecular graphs with elements of the spatial structures of the molecules

Nadezhda A. Shulaeva, Mariya I. Skvortsova[@], Nataliya A. Mikhailova

MIREA – Russian Technological University (M.V. Lomonosov Institute of Fine Chemical Technologies), Moscow, 119571 Russia

[@]Corresponding author, e-mail: skvorivan@mail.ru

Objectives. *This article aims to describe, elaborate, and test a general algorithmic method for constructing the structure–property models for organic compounds.*

Results. *The construction of the models is based on the statistical analysis of some sets of chemical structures of definite classes with known property values. These models have some forms of correlation equations. For the representation of chemical structures in this method, the special weighted molecular graphs (MGs) that reflect some peculiarities of the spatial structures of the corresponding molecules are used. The proposed method is realized in two steps. First, it is assumed that the required structure–property equation has a definite form and depends on several adjusted numerical parameters and two changeable functions of one variable. In this step, from some set of functions, the pair of functions that provide the best model is selected. In the second step, the best model (from the previous step) is modified. For this purpose, the classification of the vertices of MG by the chemical symbols of the corresponding atoms and their first-order environments is fulfilled. Further, the graph edges are classified according to the classes of the vertices which they connect. Furthermore, the numerical correction terms for the initial weights of the vertices and edges are introduced, and they improve the obtained model. The final result of the model-construction process is the equation of the definite form containing concrete numerical values of its parameters. Some examples of the application of the elaborated method for constructing the structure–property models for the concrete properties and classes of compounds are presented. The following classes of organic compounds and their physicochemical properties are considered: 1) the boiling point of alcohols, 2) the water solubility of alcohols, 3) the boiling point of sulfides, and 4) the retention indices of alkylphenols. The obtained results indicate the efficiency of the proposed approach and the significance of introducing the second step to the method.*

Conclusions. *In this work, a general algorithmic and computerized method for constructing the structure–property models of organic compounds is suggested. Examples of the application of this method demonstrated its high efficiency. The method is suitable for any class of organic compounds and properties, which are quantitatively measured. Owing to its high efficiency, the structure–property models obtained by this approach can be employed to calculate the properties of chemical compounds for which experimental data are unavailable.*

Keywords: *structure–property correlations, weighted molecular graphs, graph invariants, computer chemistry, mathematical chemistry, length of the chemical bond, covalent atomic radius*

For citation: Shulaeva N.A., Skvortsova M.I., Mikhailova N.A. Structure–property models of organic compounds based on molecular graphs with elements of the spatial structures of the molecules. *Tonk. Khim. Tekhnol. = Fine Chem. Technol.* 2020;15(6):84-103 (Russ., Eng.). https://doi.org/10.32362/2410-6593-2020-15-6-84-103

НАУЧНАЯ СТАТЬЯ

Модели связи «структура-свойство» органических соединений на основе молекулярных графов с элементами пространственного строения молекул

Н.А. Шулаева, М.И. Скворцова[@], Н.А. Михайлова

МИРЭА – Российский технологический университет (Институт тонких химических технологий имени М.В. Ломоносова), Москва, 119571 Россия
[@]Автор для переписки, e-mail: skvorivan@mail.ru

Цели. *Цель работы – разработать, описать и протестировать общий алгоритмический метод построения моделей связи «структура-свойство» для органических соединений.*
Результаты. *Вышеуказанные модели строятся на основе статистического анализа данных по выборкам структур и свойств химических соединений и имеют вид корреляционных уравнений. Химические структуры в предложенном подходе представляются в виде специальных молекулярных графов с весами вершин и ребер, отражающих определенные особенности пространственного строения соответствующих молекул. Реализация метода происходит в два этапа. На первом этапе предполагается, что искомое уравнение связи «структура-свойство» имеет определенный аналитический вид и зависит от ряда подгоночных параметров и двух функций одной переменной, которые могут варьироваться. На этом этапе происходит отбор пары функций из заданного множества функций, дающих наилучшую модель. На втором этапе происходит модификация полученной наилучшей модели. Для этой цели первоначально проводится классификация вершин молекулярного графа по химическим символам соответствующих атомов и картинам их первого окружения; проводится также классификация ребер графа в соответствии с классами вершин, которые они соединяют. На основе полученной классификации вводятся числовые «поправки» к исходным весам вершин и ребер молекулярных графов, что позволяет улучшить модель, полученную на первом этапе. Конечным результатом процесса построения модели служит уравнение определенного вида с конкретными числовыми значениями всех его параметров. Приведены примеры применения предложенного метода для построения моделей связи «структура-свойство» для конкретных свойств и классов соединений, показывающие его эффективность. Рассматривались следующие физико-химические свойства и классы органических соединений: 1) температура кипения спиртов; 2) растворимость спиртов в воде; 3) температура кипения сульфидов; 4) индексы удерживания алкилфенолов.*

Выводы. *Предложен общий алгоритмический метод построения корреляционных уравнений, связывающих структуру и свойства органических соединений. Приведены примеры его реализации. Метод может быть использован для любых классов органических соединений и любых их свойств, которые измеряются количественно. Модели, построенные на основе предложенного подхода, обладающие достаточно высоким качеством, могут быть использованы для расчета свойств соединений, для которых отсутствуют экспериментальные данные.*

Ключевые слова: *корреляции «структура-свойство», молекулярные графы, инварианты графа, компьютерная химия, математическая химия, длина химической связи, ковалентный радиус атома*

Для цитирования: Шулаева Н.А., Скворцова М.И., Михайлова Н.А. Модели связи «структура-свойство» органических соединений на основе молекулярных графов с элементами пространственного строения молекул. *Тонкие химические технологии.* 2020;15(6):84-103. https://doi.org/10.32362/2410-6593-2020-15-6-84-103

INTRODUCTION

One of the important problems of mathematical and computational chemistry is to determine the quantitative correlations between the structures and properties of chemical compounds [1–11]. The obtained relationships enable the prediction of the properties of given compounds (both real and hypothetical) from their structures through appropriate calculations, which can be employed for a targeted search for compounds with a predetermined set of properties. Notably, a large number of different chemical substances have been synthesized thus far. However, the experimental determination of their various properties for a targeted search for compounds is technically largely difficult; moreover, it requires significant finance and time. Hence, the development of various mathematical methods for modeling the correlations between the structure and properties of chemical compounds is an essential task.

Generally, a statistical approach that is based on the analysis of a given set of chemical structures with known values of the studied property is employed to construct models for the structure–property correlations. For the quantitative description of the structures of chemical compounds, a certain set of molecular parameters $x_1,..., x_n$ is first selected; as these parameters can be used, for example, any topological, electronic, or geometric characteristic of molecules. It is further assumed that the property, y, is related to these parameters through the function, f: $y = f(x_1,..., x_n)$. The analytic form of f is generally set by the researcher, e.g., f is a linear or quadratic function, but it depends on several selected parameters. These parameters are obtained from the known data of the initial (training) sample of chemical compounds so that the equation, $y = f(x_1,..., x_n)$, would be calculated with extensive accuracy (in a sense) for the initial data set.

To assess the accuracy of the approximation in the constructed model, a correlation is usually established between the calculated and experimental values of the studied properties of the training sample of the compounds. Therefore, the correlation coefficient, R, is determined, as well as the average relative error, δ (in %), which are subsequently employed to draw conclusions on the quality of the model. For example, in [12], the following characteristics of the quality of the model, which were determined by the value of R were proposed thus: $R \geq 0.990$ (outstanding), $R \geq 0.975$ (excellent), $R \geq 0.950$ (very good), $R \geq 0.925$ (good), and $R \geq 0.900$ (fair). Notably, these criteria, which express the acceptability of the model, can be selected differently in some cases.

In the studies of structure–property correlations, the methods for quantitatively describing the molecular structures are very crucial. One of the most common and conventional methods of representing the structure of a molecule is by a graph with numerical weights (or symbolic labels). The vertices and edges of such graph correspond to the atoms and bonds in the molecule, and their weights quantitatively characterize the peculiarities of atoms and bonds of different types. As topological molecular descriptors, $x_1,..., x_n$ in the structure–property correlation models, some numerical invariants of theses graphs are used [13–21].

Evidently, the mathematical models for the structure–property correlations that were obtained within the framework of this approach depend significantly on the selected weights of the graphs representing the chemical structures since the

aforementioned graph invariants significantly depend on these weights. Generally, the selected weights of vertices and edges of a graph in a particular context do not depend on the considered class of compounds or properties. For example, for the molecular graphs (MGs) of alkanes (these graphs are constructed without considering the hydrogen atoms), it is generally assumed that the weights of all the vertices are equal to zero. Examples of weights of vertices and edges of the weighted graphs of heteroatomic molecules are presented in the paper [15]. These weights depend on atomic characteristics, such as the total number of electrons, the number of valence electrons, the number of neighboring hydrogen atoms, and the parameters that characterize the multiplicity of the chemical bonds. There are also examples of weights of vertices and edges of MGs in the literature based on the covalent atomic radii, degrees of vertices in the graph, and distances between the vertices [22–24].

Notably, during the construction of models connecting the structure and properties of chemical compounds, the questions about the best selection of the graph invariants, $x_1,..., x_n$ (molecular descriptors), approximating function f and weights of vertices and edges of MGs representing chemical structures arise. These problems are generally due to the lack of *a priori* information on what structural features and how the considered property depends for a given class of compounds, and an infinite number of variants to select the graph invariants, approximating function f, and the weights of a graph.

We shall now in more details describe the method for constructing the models of the structure–property correlations based on the optimal (in a sense) selection of the weights of the vertices in weighted MGs that represent the chemical structures described in [25]. Initially, some classification of the atoms present in the molecular structures of the studied compounds was performed. For example, the atoms can be divided into classes according to their chemical symbols taking into account the distribution of the bond types. A further detailed classification of the atoms can be obtained if, for this purpose, the pictures of their first-order environments are used. All the atoms of one, kth, class (k = 1,2,...) are assigned some weights z_k (k = 1,2,...) (the number of classes is unknown at this stage). For further constructions, all the atoms in each molecule are numbered. Thereafter, it is assumed that the dependence of the studied property, y, on the structure of the molecule has the following special form (Equation (1)):

$$y = \sum w_i w_j + c\,, \tag{1}$$

where w_i and w_j are the numerical weights of the atoms in the molecule with numbers i and j, which are determined by their class in the adopted classification, i.e., $w_i = z_k$, if the atom, i, is in the kth class. In Equation (1), the sum is for all the bonds (i,j) in the molecule, and c is a constant.

Further, the unknown weights of the classes, (k = 1,2,…), are selected so that the relation (1) would be as accurate as possible for a given sample of the compounds. Consequently, a nonlinear function of the k variables of the following form is introduced, as shown in Equation (2):

$$F(z_1,...,z_k) = \sum\nolimits_p (y_p^{\text{exp}} - y_p^{\text{calc}})^2\,, \tag{2}$$

for which we obtained the minimum and corresponding values of the variables, $z_1,..., z_k$. In Equation (2), y_p^{exp} is the experimental value of y of the pth compound, and y_p^{calc} is the analytical expression for calculating the property of the pth compound that was obtained from Equation (1) and depending on the parameters, $z_1,..., z_k$.

Thus, the resulting equation of the structure–property correlations has the form of Equation (1) in which $z_1,..., z_k$ are known and selected in the optimal method (in the above sense). Equation (1) can be employed afterward to calculate the property values of other compounds of the same class that are not present in the initial sample.

The following facts are the basis for selecting Equation (1) in the above studies. One of the most widely employed MG invariant for modeling structure–property correlations is the molecular connectivity index (the Randić index) χ, as shown in Equation (3):

$$\chi = \sum \left(v_i v_j\right)^{-\frac{1}{2}}\,, \tag{3}$$

where v_i, v_j are the degrees of vertices i and j in the graph, and the sum is the sum of all the edges (i,j) of the graph [26]. This index possesses the following property: for a tree graph with a fixed number of vertices, it takes its extreme values (the highest and lowest) on the most- and least-branched trees, i.e., chain and star graphs. Owing to this property, χ can serve as the quantitative measure of the degree of branching in an acyclic molecule. There are generalizations of χ for cases involving more general subgraphs (chains of different lengths; notably, χ corresponds to a chain length of 1, i.e., an edge of the graph), as well as cases, when in formula for χ instead of vertex degrees state some weights of vertices, depending on the characteristics of the corresponding atoms. The above modifications of χ also have wide applications in structure–property correlations [27, 28].

The replacement of the values of v_i with those of the form, $v_i + x$, in the formula for χ have been proposed when constructing the structure–property correlations [29]. Further, x are the selected parameters based on the chemical symbols of the corresponding atoms. As a specific example, it was demonstrated that a more accurate correlation of the structure–property type could be obtained with such a "variable" χ than with the original χ. Additionally, it was noted that for each property and class of compounds the set of optimal selected parameters is own set [29].

The idea of generalizing χ by introducing the selected atomic parameters was also implemented in [30–35] particularly for classes of compounds, such as alkanes, alcohols, and ethers, for their enthalpies of vaporization. The compounds were represented by graphs without considering the hydrogen atoms. In these works, a molecular descriptor of the following type was considered, as expressed in Equation (4):

$$^{0-3}\chi = {}^{0}\chi + \frac{{}^{1}\chi}{2} + \frac{{}^{2}\chi}{3} + \frac{{}^{3}\chi}{4}, \quad (4)$$

where

$$^{0}\chi = \sum\left(\ln \delta_i\right)^{-1}, {}^{1}\chi = \sum\left(\ln\left(\delta_i\delta_j\right)\right)^{-1},$$
$$^{2}\chi = \sum\left(\ln\left(\delta_i\delta_j\delta_k\right)\right)^{-1}, {}^{3}\chi = \sum\left(\ln\left(\delta_i\delta_j\delta_k\delta_l\right)\right)^{-1}. \quad (5)$$

In Equation (5) for $^{0}\chi$, the summation is for all the non-hydrogen atoms of the molecule, numbered $i = 1,2,...,n$. In Equation (5) for $^{1}\chi$, $^{2}\chi$, $^{3}\chi$, the summation is for all the chains of the atoms containing 1, 2, and 3 consecutive bonds, with the numbers of atoms in these chains corresponding to i,j, i,j,k, or i,j,k,l, respectively. The numbers of the kind, δ_i, were the selected parameters of the atoms depending on the method of their classification. Further, it was assumed that the dependence of the considered property on the parameter, $^{0-3}\chi$, has the following form (Equation (6)):

$$y = a \times {}^{0-3}\chi + b, \quad (6)$$

where the constants, a and b, were selected according to the least-squares method.

In [36], a modification of the classical Randić index according to the selection of the atomic parameters was considered. In that case, some of the parameters that were attributed to the carbon atoms were fixed (they were assumed to be equal to the degrees of the vertices in the corresponding graph), and the other part was selected.

In this study, we proposed and verified a new general method, which was developed within the framework of a statistical approach for solving such problems, to establish quantitative correlations between the structures and properties of organic compounds. The method is based on representing the structures of the studied compounds as weighted MGs with weights of vertices and edges reflecting the elements of spatial structure of corresponding molecules, with the subsequent corrections of these weights. The proposed approach is a generalized development of a previously described approach [25].

DESCRIPTION OF THE METHOD FOR CONSTRUCTING THE STRUCTURE–PROPERTY CORRELATION MODELS

Construction of weighted MGs of organic compounds

Let there be a certain sample with structural formulas of organic compounds of a certain class and known numerical values of some physicochemical property obtained experimentally and reported in the literature.

The construction of any structure–property correlation model within the framework of the above-described statistical approach presupposes the preliminary construction of special MGs for the compounds under consideration. In the given paper weighted MGs with vertices corresponding to the non-hydrogen atoms of the molecule, and edges corresponding to the chemical bonds between these atoms, are considered. The MG vertices were assumed to be numbered 1,2,...,n, and edge, which was formed by a pair of vertices that were numbered i and j ($i < j$), was denoted by the symbol (i,j). The weights of the vertices and edges of MG reflect the elements of the spatial structure of the molecule, namely, the vertex weights w_{ii} are taken as the covalent radii of the corresponding atoms, and edge weights w_{ij} are taken as the lengths of the corresponding bonds. The values of the atomic radii and bond lengths used in this work are presented in Tables 1 and 2, respectively [23], [37].

Table 1. Covalent radii of the atoms (Å) [23]

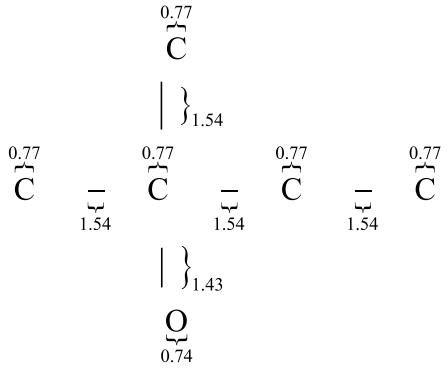
No.	Atom	Atom radius, Å
1	Csp^3	0.77
2	Csp^2	0.67
3	Csp	0.60
4	Nsp^3	0.74
5	Nsp^2	0.62
6	Nsp	0.55
7	Osp^3	0.74
8	Osp^2	0.62
9	F	0.72
10	Psp^3	0.10
11	Psp^2	1.00
12	Ssp^3	1.04
13	Ssp^2	0.94
14	Cl	0.99
15	Br	1.14
16	I	1.33

Table 2. Average bond lengths in the organic molecules (Å) [37]

No.	Bond	Bond length, Å	No.	Bond	Bond length, Å
1	C4–C4	1.54	23	C2–C1	1.63
2	C4–C3	1.52	24	C3=C3	1.34
3	C4–C2	1.46	25	C3=C2	1.31
4	C4–N3	1.47	26	C3=N2	1.32
5	C4–N2	1.47	27	C3=O1	1.22
6	C4–O2	1.43	28	C2=C2	1.28
7	C4–S2	1.81	29	C2=N2	1.32
8	C4–F	1.40	30	C2=O1	1.16
9	C4–Cl	1.76	31	N3=O1	1.24
10	C4–Br	1.94	32	N2=O1	1.22
11	C4–I	2.14	33	N2=N2	1.25
12	C3–C3	1.46	34	C2≡C2	1.20
13	C3–C2	1.45	35	C2≡N1	1.16
14	C3–N3[1]	1.40	36	C3=C3[2]	1.40
15	C3–N2	1.40	37	C3=N2[2]	1.34
16	C3–O2	1.36	38	C3–F[2]	1.32
17	C3–F	1.33	39	C3–Cl[2]	1.71
18	C3–Cl	1.73	40	C3–Br[2]	1.89
19	C2–C2	1.38	41	C3–I[2]	2.10
20	C2–N3	1.33	42	C3–O2[2]	1.36
21	C2–O2	1.33	43	C3–N3[2]	1.48
22	C2–F	1.30	44	C3–S2[2]	1.81

In Table 2, the numbers that are placed after the symbols of atoms are those of the neighboring atoms of a given atom in a molecule; the superscript, 1, in No. 14 implies that the bond length of the N–C=O group is 1.32; superscript, 2, in Nos. 36–44 implies that C3 is a carbon atom in the benzene ring.

An example of the described weighted MG, which corresponded to the structural formula of 2-methyl-2-butanol, is shown in the figure (the MG vertices are not numbered).



Weighted molecular graph of 2-methyl-2-butanol.

Description of the initial structure–property correlation model and the method of selecting its parameters

In the proposed approach to modeling the structure–property correlations, it was initially assumed that the corresponding equation that connects the property, y, to some structural characteristics of the molecule (or its MG) has the following form (Equation (7)):

$$y = a\sum f_1\left(w_{ij}\right) + b\sum f_2\left(w_{ii}\times w_{jj}\right) + c, \qquad (7)$$

where w_{ij} is the weight of the edge (i,j) in MG; w_{ii} is the weight of the vertex, i, of MG, the summation in each sum is for all edges (i,j) in MG. The functions, $f_1(x)$ and $f_2(x)$, are some fixed functions of one variable. Both functions were selected independently from any given set of functions. The following set of functions was considered in this paper: $f(x) = \ln x$, $f(x) = x^k$, where k takes the values, ±1 and ±0.5, i.e., five functions were considered. Notably, this set of functions could be extended or changed. Thus, the model was determined by specifying a pair of

functions, $f_1(x)$ and $f_2(x)$. The parameters, a, b, and c, were selected by the least-squares method according to a given (training) sample of compounds for each pair of $f_1(x)$ and $f_2(x)$. Therefore, 25 different models could be built in this case.

To assess the accuracy of each of these 25 models (its quality) for a training set of compounds, R and the standard deviation, s, were calculated for the correlation between the experimental and calculated values of y. From the obtained set of models based on the parameters, R and s, one (the most accurate model) was selected. Moreover, this model can be further improved.

Notably, to assess the accuracy of the model on the initial sample of compounds, the average and maximum absolute (or relative) errors in calculating y (Δ_{avg}, Δ_{max}, δ_{avg}, δ_{max}, respectively) could be used.

Correction of the model based on the modifications of weights of vertices and edges of MGs

To improve the constructed model (with a selected pair of functions, $f_1(x)$ and $f_2(x)$), we performed the corrections of w_{ii} and w_{ij} of MG. This adjustment was performed as follows: initially, the vertices and the edges (i,j) of MG were divided into some classes according to certain criteria.

The classification of the MG vertices was performed according to the chemical symbols of the corresponding atoms and the pictures of their first-order environment. The classes of the vertices were numbered arbitrarily as $1,2,\ldots,n_1$. The class with the number, p, was initially assigned some indefinite numerical "weight," x_p ($p = 1,2,\ldots,n_1$). The bonds were classified according to the classes of atoms that they connect; the bond classes were also numbered arbitrarily with numbers $1,2,\ldots,n_2$. The class with the number, q, was initially assigned some undefined numerical "weight," z_q ($q = 1,2,\ldots,n_2$).

Instead of the initial w_{ii}, the weights with the form, $w_{ii} + x_p$, were considered, where p is the number of the class, to which the ith atom belongs. Further, instead of w_{ij}, the weights were similarly corrected to the form, $w_{ij} + z_q$, where q is the class number to which the edge (i,j) belongs.

Further, these corrections, as well as the coefficients, a, b, and c, which are assumed to be unknown, were selected simultaneously by minimizing a function of many variables of the following form, as shown in Equation (8):

$$F\left(a,b,c,x_1,x_2,x_3,\ldots,z_1,z_2,z_3,\ldots\right) = \\ = \sum_k \left(y_{\text{exp},k} - y_{\text{calc},k}\right)^2, \quad (8)$$

where $y_{exp,k}$ is the experimental value of y of the kth compound, $y_{calc,k}$ is the calculated value of y of this compound that is obtained by Equation (7), where instead of the initial weights, w_{ij} and w_{ii} are their corrected expressions; the summation is for all the compounds of the training set.

For the optimal selection of the parameters, a, b, c, x_1, x_2, x_3,..., in this task, the add-in, "Load the Solver," of the Microsoft Excel program could be used.

Notably, to refine the model, only w_{ii} can be adjusted, while w_{ij} remains unchanged.

To evaluate the quality of the model on the training set of compounds, R and s were determined for the correlation between the experimental and calculated values of the studied property, as well as the average and maximum absolute (or relative) errors (Δ_{avg}, Δ_{max}, δ_{avg}, δ_{max}, respectively) in calculating the property value.

To assess the predictive power of the model, a new set of chemical compounds of the same class and known values of y was used (test sample). For these compounds, the obtained formula was employed to calculate y and determine R_1 and s_1 for the correlation between the experimental and calculated values of y, as well as the parameters, $\Delta_{avg,1}$, $\Delta_{max,1}$, $\delta_{avg,1}$, $\delta_{max,1}$, similar to the above parameters for the training sample.

EXAMPLES OF CONSTRUCTING THE STRUCTURE–PROPERTY CORRELATIONS ON THE BASE OF THE PROPOSED METHOD

This section offers examples for the application of the developed method in the construction of models of the structure–property correlations for a range of physicochemical properties of organic compounds of some classes, as well as the analysis of the results.

In each example, a certain sample of compounds is considered and divided into training and test samples. N is the number of compounds in the initial sample (N_1 and N_2 are the numbers of compounds in the training and test samples, respectively). The initial data are presented as a table containing the names of the compounds and values of their considered physicochemical property, which were culled from the literature. The compounds included in the test set are marked as *.

Following the method described herein, the weighted MGs of the compounds under consideration were constructed at the first stage of research in each example. Thereafter, for the training set, the unknown constants, a, b, c, were calculated for each pair of the selected functions, $f_1(x)$ and $f_2(x)$, by the least-squares method, as well as the parameters, R and s. The best model was selected from the obtained models by the parameters, R and s, i.e., the best pair of $f_1(x)$ and $f_2(x)$ was selected. For this model, R, s, δ_{avg}, and δ_{max},

as well as similar parameters, R_1, s_1, $\delta_{avg,1}$, $\delta_{max,1}$, were determined, thereby characterizing the qualities of the model on the training and test samples, respectively.

The obtained best model was refined in the second stage of the research. Therefore, the MG vertices were initially classified according to the chemical symbols of the corresponding atoms and the patterns of their first-order environments. The classification results are summarized in a table containing the structural fragments that define the classes and their descriptions.

Further, based on the obtained classification of the vertices, corrections x_1, x_2,… were introduced to the initial w_{ii}. The obtained values of corrections x_1, x_2,… and the coefficients a, b, c for the improved model are summarized in another table. The final result of the model-building process is an equation of a certain form with the specific numerical values of all its parameters, which enables the calculation of y for any compound of a given class. For this equation, R, s, δ_{avg}, δ_{max} and R_1, s_1, $\delta_{avg,1}$, and $\delta_{max,1}$ were also determined, thus characterizing the qualities of the model on the training and test sets, respectively.

Finally, the results were analyzed at the third stage of the research. Thus, all the statistical characteristics of the four cases (initial/refined models and training/test samples) are summarized in one table, after which these characteristics were compared in the initial and refined models, and a conclusion was made about the advisability of introducing corrections to w_{ii} of MG to improve the model. Moreover, the quality of the final, refined model was assessed.

a) Boiling points of alcohols

A sample of alcohols with their boiling points, $N = 31$, $N_1 = 21$, $N_2 = 10$ [29] was considered as the initial data for constructing the model. The initial data are presented in Table 3.

The best model selected at the first stage was an equation of the following form (9):

$$y = a\sum\left(w_{ij}\right)^{-1} + b\sum\left(w_{ii}\times w_{jj}\right)^{-\frac{1}{2}} + c\,. \qquad (9)$$

Further, the MG vertices were classified, and the results are presented in Table 4.

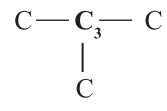
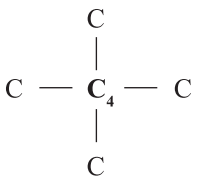
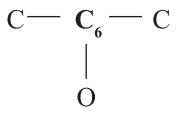
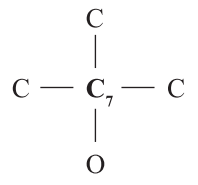
Table 3. Alcohols and their boiling points [29]

No.	Compound	Boiling point, °C
1	Ethanol	78.0
2	Propanol	97.1
3	2-Propanol	82.4
4	Butanol*	117.6
5	2-Methyl-1-propanol	108.1
6	2-Butanol*	99.5
7	2-Methyl-2-propanol	82.4
8	Pentanol*	138.0
9	3-Methyl-1-butanol	131.0
10	2-Methyl-1-butanol	128.0
11	2-Pentanol*	119.3
12	3-Pentanol	116.2
13	3-Methyl-2-butanol*	112.9
14	2-Methyl-2-butanol	102.3
15	Hexanol	157.6
16	3-Methyl-1-pentanol	153.0
17	4-Methyl-1-pentanol	151.9
18	2-Methyl-1-pentanol*	149.0
19	2-Ethyl-1-butanol	147.0
20	2,3-Dimethyl-1-butanol*	144.5
21	3,3-Dimethyl-1-butanol	143.0

Table 3. Continued

No.	Compound	Boiling point, °C
22	2-Hexanol*	140.0
23	2,2-Dimethyl-1-butanol	136.5
24	3-Hexanol	135.0
25	3-Methyl-2-pentanol*	134.3
26	4-Methyl-2-pentanol	131.6
27	2-Methyl-3-pentanol*	126.5
28	3-Methyl-3-pentanol	122.4
29	2-Methyl-2-pentanol	121.1
30	3,3-Dimethyl-2-butanol	120.4
31	2,3-Dimethyl-2-butanol	118.4

Table 4. Classification of vertices of MG, corresponding to alcohol molecules

No.	Structural fragment that defines the class of atoms	Description of the atom class
1	$\mathbf{C_1}$— C	The central C_1 atom is bonded to one C atom
2	C— $\mathbf{C_2}$— C	The central C_2 atom is bonded to two C atoms
3		The central C_3 atom is bonded to three C atoms
4		The central C_4 atom is bonded to four C atoms
5	C — $\mathbf{C_5}$ — O	The central C_5 atom is bonded to one C and one O atoms
6		The central C_6 atom is bonded to two C and one O atoms
7		The central C_7 atom is bonded to three C and one O atoms
8	C — **O**	The central O atom is bonded to one C atom

Based on this classification, corrections x_1–x_8 (correction parameters) were introduced to the weights of the MG vertices. The determined corrections x_1, x_2, …, x_8 and the coefficients a, b, and c are presented in Table 5.

Thus, the improved model is expressed as Equation (10):

$$y = a\sum\left(w_{ij}\right)^{-1} + b\sum\left(\left(w_{ii}+x_n\right)\times\left(w_{jj}+x_m\right)\right)^{-\frac{1}{2}} + c, \quad (10)$$

where n and m are the numbers of the classes of vertices to which the corresponding vertices i and j

belong. The values of the corrections, x_n and x_m, as well as the coefficients a, b, and c were obtained from Table 5. The summation in each sum is for all edges (i,j) in MG.

Table 6 presents some statistical parameters of the constructed models for the four cases (initial/refined models; training/test samples).

As presented in Table 6, the introduction of corrections to w_{ii} of MG significantly improved the initial model (on all characteristics). Thus, an equation of the form of Equation (10) was more accurate than Equation (9) in expressing the relationship between the structures and boiling points of alcohols.

b) Solubilities of alcohols in water

A sample of alcohols, $N = 30$, $N_1 = 20$, $N_2 = 10$ [38], with their water-solubility values, $-\log X$ (X is the molar fraction of a substance in solution), was considered as the initial data (Table 7) for constructing the model.

Table 5. Values of the correction parameters, x_1–x_8, and the coefficients a, b, and c for the improved model

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	a	b	c
0.0348	0.0003	0.0112	0.0170	0.0001	0.0117	0.0174	0.00303	15898	7936.4	−549.4

Table 6. Some statistical parameters of the initial/modified models for the training/test samples for the boiling point of alcohols

Statistical parameters of the model	Initial model (without the corrections of w_{ii} of MG)		Modified model (with the corrections of w_{ii} of MG)	
	Training sample	Test sample	Training sample	Test sample
s	13.19	10.09	1.82	2.04
R	0.838	0.796	0.997	0.992
δ_{avg}	9.23%	6.18%	1.21%	1.39%
δ_{max}	25.43%	13.25%	3.96%	3.05%

The best model that was selected at the first stage is an equation of the following form (11):

$$y = a\sum\left(w_{ij}\right)^{\frac{1}{2}} + b\sum\left(w_{ii}\times w_{jj}\right)^{-\frac{1}{2}} + c\,. \qquad (11)$$

Further, the MG vertices were classified, and the results are presented in Table 4. Corrections x_1–x_8 (correction parameters) were introduced to the weights of the MG vertices based on this classification. The obtained corrections, $x_1, x_2, \ldots, x_8$, and the coefficients a, b, and c are presented in Table 8.

Thus, the improved model is expressed as Equation (12):

$$y = a\sum\left(w_{ij}\right)^{\frac{1}{2}} + \\ +b\sum\left(\left(w_{ii}+x_n\right)\times\left(w_{jj}+x_m\right)\right)^{-\frac{1}{2}} + c, \qquad (12)$$

where n and m are the numbers of vertex classes to which the corresponding vertices, i and j, belong. The values of corrections x_n and x_m, as well as the coefficients a, b, and c were obtained from Table 8. The summation in each sum is for all edges (i,j) of MG.

Table 9 presents some statistical parameters of the constructed models for the four cases (initial/refined models; training/test samples).

According to Table 9, the introduction of corrections to w_{ii} of MG significantly improved the initial model (on all characteristics). Thus, an equation of the form of Equation (12) was more accurate than Equation (11) in reflecting the relationship between the structures of alcohols and their water solubility.

To further improve the model, a classification of edges in MG (or bonds in structural formula) based on mentioned above classification of the MG vertices, was fulfilled. Therefore, the following edge classes were determined:

1) C_1–C_2, 2) C_1–C_3, 3) C_1–C_4, 4) C_1–C_5, 5) C_1–C_6, 6) C_1–C_7, 7) C_2–C_2, 8) C_2–C_3, 9) C_2–C_4, 10) C_2–C_5, 11) C_2–C_6, 12) C_2–C_7, 13) C_3–C_5, 14) C_3–C_6, 15) C_3–C_7, 16) C_4–C_5, 17) C_4–C_6, 18) O–C_5, 19) O–C_6, 20) O–C_7.

Table 7. Alcohols and their water-solubility values [38]

No.	Compound	Solubility in water, −log*X*
1	1-Butanol	1.750
2	2-Methyl-1-propanol	1.743
3	2-Butanol*	1.724
4	1-Pentanol	2.332
5	3-Methyl-1-butanol	2.254
6	2-Methyl-1-butanol*	2.207
7	2-Pentanol	2.025
8	3-Pentanol*	1.961
9	3-Methyl-2-butanol	1.926
10	2-Methyl-2-butanol	1.608
11	2,2-Dimethyl-1-propanol	2.030
12	1-Hexanol*	2.957
13	2-Hexanol	2.612
14	3-Hexanol	2.542
15	3-Methyl-3-pentanol*	2.109
16	2-Methyl-2-pentanol	2.233
17	2-Methyl-3-pentanol*	2.445
18	3-Methyl-2-pentanol*	2.458
19	2,3-Dimethyl-2-butanol	2.118
20	3,3-Dimethyl-1-butanol	2.870
21	3,3-Dimethyl-2-butanol	2.359
22	4-Methyl-1-pentanol	2.737
23	4-Methyl-2-pentanol*	2.534
24	2-Ethyl-1-butanol	2.956
25	1-Heptanol	3.554
26	2-Methyl-2-hexanol	2.820
27	3-Methyl-3-hexanol*	2.729
28	3-Ethyl-3-pentanol	2.579
29	2,3-Dimethyl-2-pentanol	2.615
30	2,3-Dimethyl-3-pentanol*	2.588

Table 8. Values of the correction parameters, x_1–x_8, and the coefficients *a*, *b*, and *c* for the improved model

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	*a*	*b*	*c*
0.0005	0.0011	0.0011	0.0011	0.0001	0.0002	−0.0003	−0.0002	213.25	−203.61	15.14

By this classification, corrections z_1–z_{20} to w_{ij} were introduced and obtained, and new values of the parameters, *a*, *b*, *c*, and x_1–x_8 were simultaneously obtained. Table 10 shows some statistical parameters of the constructed modified models for the four cases (model with corrections to the weights of vertices/the weights of vertices and edges; training/test samples).

Table 9. Some statistical parameters of the initial/modified models for the training/test samples for the water solubility of alcohols

Statistical parameters of the model	Initial model (without the corrections of w_{ii} of MG)		Modified model (with the corrections of w_{ii} of MG)	
	Training sample	Test sample	Training sample	Test sample
s	0.301	0.239	0.078	0.091
R	0.794	0.798	0.988	0.973
δ_{avg}	9.25%	6.31%	2.40%	3.43%
δ_{max}	28.53%	17.49%	6.29%	17.40%

Table 10. Some statistical parameters of the modified models with the correction parameters to vertex weights and to weights of vertices and edges for the training and test samples for the water solubility of alcohols

Statistical parameters of the model	Modified model (with corrections to the vertex weights of MG)		Modified model (with corrections to the weights of vertices and edges of MG)	
	Training sample	Test sample	Training sample	Test sample
s	0.078	0.091	0.026	0.099
R	0.988	0.973	0.999	0.968
δ_{avg}	2.40%	3.43%	0.69%	3.43%
δ_{max}	6.29%	17.40%	2.48%	18.86%

Table 11. Sulfides and their boiling points [39]

No.	Compound	Boiling point, °C
1	Methyl ethyl sulfide	66.6
2	Methyl propyl sulfide	95.5
3	Diethyl sulfide	92.0
4	Methylisopropyl sulfide	84.4
5	Ethylisopropyl sulfide	107.4
6	Methyl butyl sulfide*	123.2
7	Methyl isobutyl sulfide	112.5
8	Ethylpropyl sulfide*	118.5
9	Methyl *tert*-butyl sulfide	101.5
10	Methylamyl sulfide	145.0
11	Ethyl butyl sulfide	144.2
12	Dipropyl sulfide*	142.8
13	Propylisopropyl sulfide	132.0
14	Ethyl isobutyl sulfide	134.2
15	Methylisoamyl sulfide	137.0

Table 11. Continued

No.	Compound	Boiling point, °C
16	Methyl 2-methylbutyl sulfide*	139.0
17	Ethyl *sec*-butyl sulfide	133.6
18	Ethyl *tert*-butyl sulfide*	120.4
19	Diisopropyl sulfide	120.0
20	Methyl 1-ethylpropyl sulfide	137.0
21	Methyl *sec*-butyl sulfide*	114.5
22	Methyl *tert*-amyl sulfide	128.3
23	Methyl-1,2-dimethylpropyl sulfide	133.0
24	Methylhexyl sulfide*	171.0
25	Propyl butyl sulfide	166.0
26	Propyl isobutyl sulfide	155.0
27	Isopropyl isobutyl sulfide*	145.0
28	Ethyl 2-methylbutyl sulfide*	159.0
29	Propyl *tert*-butyl sulfide	138.0
30	Isopropyl *sec*-butyl sulfide*	142.0

Notably, the introduction of additional parameters to the correlation equation improved all the model characteristics on the training set. However, in the test sample, the model became slightly worse. In this regard, it was not sensible to further classify edges of MG.

c) Boiling point of sulfides

A sample of sulfides, N = 30, N_1 = 20, N_2 = 10 [39], with their boiling points was considered as the initial data (Table 11) for constructing the model.

The best model that was selected at the first stage is an equation of the following form (13):

$$y = a\sum\left(w_{ij}\right)^{-1} + b\sum\ln\left(w_{ii}\times w_{jj}\right) + c\,. \qquad (13)$$

Further, the MG vertices were classified, as presented in Table 12.

Corrections x_1–x_8 to w_{ii} of MG were introduced based on this classification. The obtained corrections,

Table 12. Classification of vertices of MG, corresponding to sulfide molecules

No.	Structural fragment that defines the class of atoms	Description of the atom class
1	$\mathbf{C_1}$— C	The central C_1 atom is bonded to one C atom
2	C— $\mathbf{C_2}$— C	The central C_2 atom is bonded to two C atoms
3	C— $\mathbf{C_3}$— C | C	The central C_3 atom is bonded to three C atoms
4	$\mathbf{C_4}$— S	The central C_4 atom is bonded to one S atom
5	C — $\mathbf{C_5}$ — S	The central C_5 atom is bonded to one C and one S atoms
6	C — $\mathbf{C_6}$ — S | C	The central C_6 atom is bonded to two C and one S atoms

Table 12. Continued

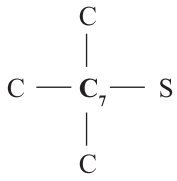
No.	Structural fragment that defines the class of atoms	Description of the atom class
7		The central C_7 atom is bonded to three C and one S atoms
8		The central S atom is bonded to two C atoms

Table 13. Values of the correction parameters, x_1–x_8, and the coefficients a, b, and c for the improved model

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	a	b	c
0.00016	0.00014	0.00002	0.00017	0.00006	0.00006	0.00021	0.00006	7284.7	9091.2	4055.93

x_1, x_2, …, x_8, and the coefficients a, b, and c are shown in Table 13.

Thus, the best model is expressed as Equation (14):

$$y = a\sum\left(w_{ij}\right)^{-1} + \\ +b\sum\ln\left(\left(w_{ii}+x_n\right)\times\left(w_{jj}+x_m\right)\right)+c, \quad (14)$$

where n and m are the numbers of the vertex classes to which the corresponding vertices, i and j, belong. The values of the corrections, x_n and x_m, as well as the coefficients a, b, and c were obtained from Table 13. The summation in each sum is for all edges (i,j) in MG.

Table 14 presents some statistical parameters of the models for the four cases (initial/refined models; training/test samples).

According to Table 14, the introduction of corrections to w_{ii} of MG significantly improved the initial model (on all characteristics). Thus, an equation of the form of Equation (14) was more accurate than Equation (13) in reflecting the correlation between the structures and boiling points of sulfides.

d) Alkylphenol retention indices

A sample of alkylphenols, N = 30, N_1 = 20, N_2 = 10 [40], with their values of retention indices was considered as the initial data (Table 15) for constructing the model.

The best model that was selected at the first stage is an equation of the following form (15):

$$y = a\sum\left(w_{ij}\right)^{-1} + b\sum\left(w_{ii}\times w_{jj}\right)^{\frac{1}{2}} + c\,. \quad (15)$$

Further, the MG vertices were classified, as presented in Table 16. The C atoms included in the benzene ring are denoted as C^b.

Corrections x_1–x_9 to w_{ii} were introduced based on this classification. The obtained corrections, x_1, x_2, …, x_9, and the coefficients a, b, and c are presented in Table 17.

Table 14. Some statistical parameters of the initial/modified models of the training/test samples for the boiling point of sulfides

Statistical parameters of the model	Initial model (without the corrections of w_{ii} of MG)		Modified model (with the corrections of w_{ii} of MG)	
	Training sample	Test sample	Training sample	Test sample
s	7.76	10.47	2.73	2.64
R	0.954	0.846	0.994	0.991
δ_{avg}	4.41%	6.77%	1.83%	1.81%
δ_{max}	12.48%	10.72%	5.60%	4.07%

Table 15. Alkylphenols and their retention indices [40]

No.	Compound	Retention indices
1	Phenol	1281
2	2-Methylphenol	1354
3	3-Methylphenol*	1386
4	4-Methylphenol	1385
5	2-Ethylphenol	1430
6	3-Ethylphenol	1483
7	4-Ethylphenol*	1473
8	2,3-Dimethylphenol	1495
9	2,4-Dimethylphenol	1456
10	2,5-Dimethylphenol	1453
11	2,6-Dimethylphenol*	1416
12	3,5-Dimethylphenol*	1489
13	3,4-Dimethylphenol	1530
14	4-Isopropylphenol	1527
15	2-*n*-Propylphenol*	1502
16	3-*n*-Propylphenol	1565
17	4-*n*-Propylphenol	1563
18	2-Ethyl-4-methylphenol*	1523
19	2-Ethyl-5-methylphenol	1529
20	2-Ethyl-6-methylphenol	1485
21	3-Ethyl-5-methylphenol*	1581
22	4-Ethyl-2-methylphenol*	1539
23	4-Ethyl-3-methylphenol	1608
24	2,3,4-Trimethylphenol	1638
25	2,3,5-Trimethylphenol	1593
26	2,3,6-Trimethylphenol*	1551
27	2,4,5-Trimethylphenol*	1593
28	3,4,5-Trimethylphenol	1667
29	4-*sec*-Butylphenol	1612
30	2-*n*-Butylphenol	1600

Table 16. Classification of vertices of MG, corresponding to alkylphenol molecules

No.	Structural fragment that defines the class of atoms	Description of the atom class
1	C^b—$\mathbf{C_1^b}$—C^b	The central C_1^b atom is bonded to two C^b atoms
2	C^b—$\mathbf{C_2^b}$(—C)—C^b	The central C_2^b atom is bonded to two C^b and one C atoms
3	C^b—$\mathbf{C_3}$	The central C_3 atom is bonded to the C^b atom

Table 16. Continued

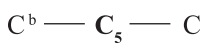
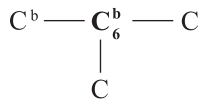
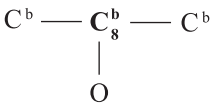
No.	Structural fragment that defines the class of atoms	Description of the atom class
4		The central C_4 atom is bonded to the C atom
5		The central C_5 atom is bonded to one C^b and one C atoms
6		The central C_6^b atom is bonded to one C^b and two C atoms
7		The central C_7 atom is bonded to two C atoms
8		The central C_8^b atom is bonded to two C^b and one O atoms
9		The central O atom is bonded to one C^b atom

Table 17. Values of the corrections parameters, x_1–x_9, and the coefficients a, b, and c for the improved model

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	a	b	c
0.6198	0.5934	0.7245	0.2346	0.4256	0.5157	0.3836	0.9956	0.9777	780.9	1055.5	28.2

Thus, the best model is expressed as Equation (16):

$$y = a\sum\left(w_{ij}\right)^{-1} + \\ +b\sum\left(\left(w_{ii}+x_n\right)\times\left(w_{jj}+x_m\right)\right)^{\frac{1}{2}} + c, \quad (16)$$

where n and m are the numbers of the vertex classes containing the corresponding vertices, i and j. The values of the corrections, x_n and x_m, as well as the coefficients a, b, and c were obtained from Table 17. The summation in each sum is for all edges (i,j) in MG.

Table 18 presents some statistical parameters of the models for the four cases (initial/refined models; training/test samples).

According to Table 18, the introduction of the corrections to w_{ii} of MG significantly improved the initial model (on all characteristics). Thus, an equation of the form of Equation (16) was more accurate than Equation (15) in reflecting the correlation between the structures and retention indices of alkylphenols.

Table 18. Some statistical parameters of the initial/modified models of the training/test samples for the retention indices of alkylphenols

Statistical parameters of the model	Initial model (without the corrections of w_{ii} of MG)		Modified model (with the corrections of w_{ii} of MG)	
	Training sample	Test sample	Training sample	Test sample
s	36.92	30.37	15.03	16.18
R	0.934	0.904	0.989	0.974
δ_{avg}	1.94%	2.01%	0.72%	1.00%
δ_{max}	5.86%	5.42%	2.16%	3.17%

CONCLUSIONS

In this study, a general algorithmic method for constructing the mathematical models for the structure–property correlations of organic molecules was developed. The abovementioned models were built based on the statistical analysis of the data of the samples of structures and properties of chemical compounds of some classes, and they have the forms of correlation equations. The method is suitable to any classes of organic compounds and any properties, which can be measured quantitatively.

To represent chemical structures within the framework of the proposed method, special weighted MGs reflecting some elements of the spatial structure of the corresponding molecules were used.

The method is realized in several stages. At the first stage, it was assumed that the desired equation for the structure–property correlation has a well-defined form and depends on several adjustable numerical parameters and selected pair of function on one variable. Additionally, the best pair of functions (according to certain criteria) from a given set, i.e., the best model, is selected at this stage. In addition, the preassigned set of functions can be extended, that promote the increasing of the accuracy of the best selected model.

The second stage is only required if the construction of a more accurate model was desired.

At the second stage, the obtained best model undergoes some modifications, which are aimed at improving its accuracy. Therefore, the MG vertices are initially classified according to the chemical symbols of the corresponding atoms and the patterns of the first-order environment, as well as the edges according to the vertex classes that they connect. Based on the obtained classification of vertices and edges, several numerical "corrections" were introduced to the initial weights of vertices and edges. The final result of the model-building process is an equation of a certain form containing the specific numerical values of all its parameters, thereby allowing the calculation of the value of *y* for any compound of a given class.

Notably, at the second stage, the modeling can be restricted only to the classification of the MG vertices, because the further classification of the edges is only required if the resulting model (based on the classification of the vertices) is not sufficiently accurate. In this case, the new model might be better than the previous one on the training set and worse on the test set. Let us indicate one more method of improving the model obtained at the second stage through the classification of the MG vertices described above—the more detailed classification of the vertices, e.g., according to the pictures of the second-order environment.

Additionally, the paper presents examples of the application of the developed method for constructing models of the structure–property correlations of specific properties and classes of compounds, thereby demonstrating its effectiveness. Moreover, the feasibility of introducing the second stage in this method was also analyzed.

Models of structure–property correlations (built according to the above method), which have sufficiently high quality, can be employed to calculate the properties of compounds for which there are no experimental data.

Financial support

This study did not have any financial support from outside organizations.

Authors' contribution

All authors equally contributed to the research work.

The authors declare no conflicts of interest.

REFERENCES

1. Raevskii O.A. *Modelirovanie sootnoshenii "struktura–svoistvo"* (Modeling the Structure–Property Relationships). Moscow: Dobrosvet; 2015. 288 p. (in Russ.). ISBN 978-5-7913-0103-11

2. Stuper A.J., Brügger W.E., Jurs P.C. Computer-Assisted Studies of Chemical Structure and Biological Function. New York: John Wiley & Sons; 1979. 220 p.

3. Rozenblit A.B., Golender I.E. *Logiko-kombinatornye metody v konstruirovanii lekarstv* (Logico-combinatorial Methods in Drug Design). Riga: Zinatne; 1983. 352 p. (in Russ.).

4. Cherkasov A., Muratov E.N., Fourches D., Varnek A., Baskin I.I., Cronin M., Dearden J., Gramatica P., Martin Y.C., Todeschini R., Consonni V., Kuz'min V.E., Cramer R., Benigni R., Yang C., Rathman J., Terfloth L., Gastaiger J., Richard A., Tropsha A. QSAR Modeling: Where have you been? Where are you going to? *J. Med. Chem.* 2014;57(12):4977-5010. https://doi.org/10.1021/jm4004285

СПИСОК ЛИТЕРАТУРЫ

1. Раевский О.А. Моделирование соотношений «структура-свойство». М.: Добросвет; 2015. 288 с. ISBN 978-5-7913-0103-11

2. Stuper A.J., Brügger W.E., Jurs P.C. Computer-Assisted Studies of Chemical Structure and Biological Function. New York: John Wiley & Sons; 1979. 220 p.

3. Розенблит А.Б., Голендер И.Е. Логико-комбинаторные методы в конструировании лекарств. Рига: Зинатне, 1983. 352 с.

4. Cherkasov A., Muratov E.N., Fourches D., Varnek A., Baskin I.I., Cronin M., Dearden J., Gramatica P., Martin Y.C., Todeschini R., Consonni V., Kuz'min V.E., Cramer R., Benigni R., Yang C., Rathman J., Terfloth L., Gastaiger J., Richard A., Tropsha A. QSAR Modeling: Where have you been? Where are you going to? *J. Med. Chem.* 2014;57(12):4977-5010. https://doi.org/10.1021/jm4004285

5. Nizhnii S.V., Epshtein N.A. Quantitative "chemical structure–biological activity" relations. *Russ. Chem. Rev.* 1978;47(4):383-400.
http://dx.doi.org/10.1070/RC1978v047n04ABEH002225

6. Papulov Yu.G., Chernova T.I., Smolyakov V.M., Polyakov M.N. The use of topological indices in structure–property correlations. *J. Phys. Chem.* 1993;67(2):182-188.

7. Filimonov D.A., Poroikov V.V., Karaicheva E.I., Kazaryan R.K., Budunova A.P., Mikhailovsky E.M., Rudnizkikh A.V., Goncharenko L.V., Burov Yu.V. Computer prediction of the spectrum of biological activity of chemical compounds by their structure formulae: PASS system. *Eksperimentalnaya i klinicheskaya farmakologiya = Russian Journal of Experimental and Clinical Pharmacology.* 1995; 58(2):56-62 (in Russ.).

8. Fujita T. The Application of Classical QSAR to Agrochemical Research. *International Journal of Quantitative Structure–Property Relationships (IJQSPR).* 2017;2(1):1-18. https://doi.org/10.4018/IJQSPR.2017010101

9. Torrens F., Gastellano G. QSPR Prediction of Retention Times of Methylxanthines and Cotinine by Bioplastic Evolution. *International Journal of Quantitative Structure–Property Relationships (IJQSPR).* 2018;3(1):74-87.
https://doi.org/10.4018/IJQSPR.2018010104

10. Dearden J.C. The History and Development of Quantitative Structure–Activity Relationships (QSARs): Addendum. *International Journal of Quantitative Structure–Property Relationships (IJQSPR).* 2017;2(2):36-46.
https://doi.org/10.4018/IJQSPR.2017070104

11. Salman M., Ahmed S., Nandy S. QSAR and Anticancer Drug Design on Benzothienopyrimidinones as Promising Pim Kinase Inhibitors Utilizing Structural Descriptors. *International Journal of Quantitative Structure–Property Relationships (IJQSPR).* 2019;4(2):82-99.
https://doi.org/10.4018/IJQSPR.2019040104

12. Randić M. Comparative Regression Analysis. Regressions Based on a Single Descriptor. *Croat. Chem. Acta.* 1993;66(2):289-312.

13. Devillers J., Balaban A.T. (Eds.). Topological Indices and Related Descriptors in QSAR and QSPR. Amsterdam: Gordon and Breath Science Publishers; 1999. 811 p.

14. Todeschini R., Consonni V. Handbook of Molecular Descriptors. Weinheim: Willey-VCH; 2000. 668 p.

15. Stankevich M.I., Stankevich I.V., Zefirov N.S. Topological indices in organic chemistry. *Russ. Chem. Rev.* 1988;57(3):191-208.

16. Raevsky O.A. Molecular structure descriptors in the computer-aided design of biologically active compounds. *Russ. Chem. Rev.* 1999;68(6):505-524.
https://doi.org/10.1070/RC1999v068n06ABEH000425

17. King R.B. (Ed.). Chemical Application of Topology and Graph Theory. Amsterdam: Elsevier; 1983. 494 p.

18. Balaban A.T. (Ed.). Chemical Applications of Graph Theory. London: Academic Press; 1976. 389 p.

19. Trinajstić N. (Ed.) Chemical Graph Theory (2nd Edition). Boca Raton, CRC Press, 1992; 352 p.

20. Basak S.C., Magnuson V.R., Niemi G.I., Regal R.R., Veith G.D. Topological Indices: Their Nature, Mutual Relatedness and Applications. *Mathematical Modelling.* 1987;8(C):300-305.
https://doi.org/10.1016/0270-0255(87)90594-X

21. Randić M. Topological Indices. In: Schleyer P.v.R., Allinger N.L., Clark T., Gasteiger J., Kollman P.A., Schaefer H.F. III, Schreiner P.R. (Eds.). The Encyclopedia of Computational Chemistry. Chichester: John Wiley & Sons; 1998. P.3018-3032.

5. Нижний С.В., Эпштейн Н.А. Количественные соотношения «химическая структура-биологическая активность». *Успехи химии.* 1978;47(4):739-772.

6. Папулов Ю.Г., Чернова Т.И., Смоляков В.М., Поляков М.Н. Использование топологических индексов при построении корреляций «структура-свойство». *Журн. физ. химии.* 1993;67(2):203-209.

7. Филимонов Д.А., Поройков В.В., Караичева Е.И., Казарян Р.К., Будунова А.П., Михайловский Е.М., Рудницких А.В., Гончаренко Л.В., Буров Ю.В. Компьютерное прогнозирование спектра биологической активности химических соединений по их структурной формуле: система PASS. *Эксперим. и клин. фармакол.* 1995;58(2):56-62.

8. Fujita T. The Application of Classical QSAR to Agrochemical Research. *International Journal of Quantitative Structure-Property Relationships (IJQSPR).* 2017;2(1):1-18. https://doi.org/10.4018/IJQSPR.2017010101

9. Torrens F., Gastellano G. QSPR Prediction of Retention Times of Methylxanthines and Cotinine by Bioplastic Evolution. *International Journal of Quantitative Structure–Property Relationships (IJQSPR).* 2018;3(1):74-87.
https://doi.org/10.4018/IJQSPR.2018010104

10. Dearden J.C. The History and Development of Quantitative Structure–Activity Relationships (QSARs): Addendum. *International Journal of Quantitative Structure–Property Relationships (IJQSPR).* 2017;2(2):36-46.
https://doi.org/10.4018/IJQSPR.2017070104

11. Salman M., Ahmed S., Nandy S. QSAR and Anticancer Drug Design on Benzothienopyrimidinones as Promising Pim Kinase Inhibitors Utilizing Structural Descriptors. *International Journal of Quantitative Structure–Property Relationships (IJQSPR).* 2019;4(2):82-99.
https://doi.org/10.4018/IJQSPR.2019040104

12. Randić M. Comparative Regression Analysis. Regressions Based on a Single Descriptor. *Croat. Chem. Acta.* 1993;66(2):289-312.

13. Devillers J., Balaban A.T. (Eds.). Topological Indices and Related Descriptors in QSAR and QSPR. Amsterdam: Gordon and Breath Science Publishers; 1999. 811 p.

14. Todeschini R., Consonni V. Handbook of Molecular Descriptors. Weinheim: Willey-VCH; 2000. 668 p.

15. Станкевич М.И., Станкевич И.В., Зефиров Н.С. Топологические индексы в органической химии. *Успехи химии.* 1988;57(3):337-366.

16. Raevsky O.A. Molecular structure descriptors in the computer-aided design of biologically active compounds. *Russ. Chem. Rev.* 1999;68(6):505-524.
https://doi.org/10.1070/RC1999v068n06ABEH000425

17. Кинг Р. Химические приложения топологии и теории графов. М.: Мир; 1987. 560 с.

18. Balaban A.T. (Ed.). Chemical Applications of Graph Theory. London: Academic Press; 1976. 389 p.

19. Trinajstić N. (Ed.). Chemical Graph Theory (2nd Edition). Boca Raton, CRC Press, 1992; 352 p.

20. Basak S.C., Magnuson V.R., Niemi G.I., Regal R.R., Veith G.D. Topological Indices: Their Nature, Mutual Relatedness and Applications. *Mathematical Modelling.* 1987;8(C):300-305.
https://doi.org/10.1016/0270-0255(87)90594-X

21. Randić M. Topological Indices. In: Schleyer P.v.R., Allinger N.L., Clark T., Gasteiger J., Kollman P.A., Schaefer H.F. III, Schreiner P.R. (Eds.). The Encyclopedia of Computational Chemistry. Chichester: John Wiley & Sons; 1998. P.3018-3032.

22. Pogliani L. How Far Are Molecular Connectivity Descriptors from IS Molecular Pseudoconnectivity Descriptors? *J. Chem. Inf. Comput. Sci.* 2001;41(3):836-847. https://doi.org/10.1021/ci000142c

23. Kier L. B. Indexes of Molecular Shape from Chemical Graphs. *Med. Res. Rev.* 1987;7:417-440.

24. Antipin I.S., Arslanov N.A., Palyulin V.A., Konovalov A.I., Zefirov N.S. Prediction of nonspecific solvation enthalpy for organic non-electrolites. *Doklady Akademii Nauk.* 1993;331(2):173-176 (in Russ.).

25. Yakovenko Yu.Yu., Skvortsova M.I., Mikhailova N.A. Modeling the relation between the structure and physicochemical properties of organic compounds on the basis of optimal atom parameters. *Vestnik MITHT = Fine Chemical Technologies.* 2012;7(6):110-113 (in Russ.).

26. Randić M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* 1975;97(23):6609-6615. https://doi.org/10.1021/ja00856a001

27. Kier L.B., Hall L.H. Molecular Connectivity in Structure–Activity Analysis. N.Y.: Research Studies Press Ltd., John Wiley and Sons Inc.; 1986. 262 p.

28. Kier L.B., Hall L.H. Molecular Connectivity in Chemistry and Drug Research. N.Y.: Academic Press; 1976. 257 p.

29. Randić M., Pompe M. The Variable Connectivity Index $^{1}\chi^{f}$ Versus the Traditional Molecular Descriptors: A Comparative Study of $^{1}\chi^{f}$ Against Descriptors of CODESSA. *J. Chem. Inform. Comput. Sci.* 2001;41(3):631-638. https://doi.org/10.1021/ci000119r

30. Krasnykh E.L. Prediction of vaporization enthalpy based on modified Randič indices. Alkanes. *J. Struct. Chem.* 2008;49(6):986-983. https://doi.org/10.1007/s10947-008-0170-9

31. Krasnykh E.L. Prediction of vaporization enthalpy based on modified Randič indices. I. Monohydric alcohols. *J. Struct. Chem.* 2009;50(3):556-560. https://doi.org/10.1007/s10947-009-0084-1

32. Krasnykh E.L. Prediction of vaporization enthalpy based on modified Randič indices. II. Polyatomic alcohols. *J. Struct. Chem.* 2009;50(4):599-605. https://doi.org/10.1007/s10947-009-0094-z

33. Krasnykh E.L. Prediction of vaporization enthalpy based on modified Randič indices. III. Carbonic acides. *J. Struct. Chem.* 2010;51(2):217-222. https://doi.org/10.1007/s10947-010-0034-y

34. Krasnykh E.L. Prediction of vaporization enthalpy based on modified Randič indices. Ethers. *J. Struct. Chem.* 2012;53(2):383-387. https://doi.org/10.1134/S0022476612020266

35. Krasnykh E.L. Prediction of vaporization enthalpy based on modified Randič indices. Aldehydes and ketones. *J. Struct. Chem.* 2013;54(4):792-796. https://doi.org/10.1134/S0022476613040203

36. Nesterova T.N., Nesterov I.A. *Kriticheskie temperatury i davleniya organicheskikh soedinenii. Analiz sostoyaniya bazy dannykh i razvitie metodov prognozirovaniya* (Critical Temperatures and Pressures of Organic Compounds. The Analysis of Data Base and Development of Prediction Methods). Samara: Samarskii nauchnyi tsentr RAN Publ.; 2009. 580 p. (in Russ.). ISBN 978-5-93424-424-9

37. Estrada E. Spectral Moments of the Edge-Adjacency Matrix of Molecular Graphs. 2. Molecules Containing Heteroatoms and QSAR Applications. *J. Chem. Inf. Comput. Sci.* 1997;37(2):320-328. https://doi.org/10.1021/ci960113v

22. Pogliani L. How Far Are Molecular Connectivity Descriptors from IS Molecular Pseudoconnectivity Descriptors? *J. Chem. Inf. Comput. Sci.* 2001;41(3):836-847. https://doi.org/10.1021/ci000142c

23. Kier L. B. Indexes of Molecular Shape from Chemical Graphs. *Med. Res. Rev.* 1987;7:417-440.

24. Антипин И.С., Арсланов Н.А., Палюлин В.А., Коновалов А.И., Зефиров Н.С. Прогнозирование энтальпий неспецифической сольватации органических неэлектролитов. *Доклады АН.* 1993;331(2):173-176.

25. Яковенко Ю.Ю., Скворцова М.И., Михайлова Н.А. Моделирование связи между структурой и физико-химическими свойствами органических соединений на основе оптимальных атомных параметров. *Вестник МИТХТ*. 2012;7(6):110-113.

26. Randić M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* 1975;97(23):6609-6615. https://doi.org/10.1021/ja00856a001

27. Kier L.B., Hall L.H. Molecular Connectivity in Structure-Activity Analysis. N.Y.: Research Studies Press Ltd., John Wiley and Sons Inc.; 1986. 262 p.

28. Kier L.B., Hall L.H. Molecular Connectivity in Chemistry and Drug Research. N.Y.: Academic Press; 1976. 257 p.

29. Randić M., Pompe M. The Variable Connectivity Index $^{1}\chi^{f}$ Versus the Traditional Molecular Descriptors: A Comparative Study of $^{1}\chi^{f}$ Against Descriptors of CODESSA. *J. Chem. Inform. Comput. Sci.* 2001;41(3):631-638. https://doi.org/10.1021/ci000119r

30. Красных Е.Л. Прогнозирование энтальпий испарения на основе модифицированных индексов Рандича. Алканы. *Жур. структурн. химии.* 2008;49(6):1026-1033.

31. Красных Е.Л. Прогнозирование энтальпий испарения на основе модифицированных индексов Рандича. Одноатомные спирты. *Журн. структурн. химии.* 2009;50(3):557-561.

32. Красных Е.Л. Прогнозирование энтальпий испарения на основе модифицированных индексов Рандича. Многоатомные спирты. *Журн. структурн. химии.* 2009;50(4):631-637.

33. Красных Е.Л. Прогнозирование энтальпий испарения на основе модифицированных индексов Рандича. III. Карбоновые кислоты. *Журн. структурн. химии.* 2010;51(3)557-561. (2):231-236.

34. Красных Е.Л. Прогнозирование энтальпий испарения на основе модифицированных индексов Рандича. Простые эфиры. *Журн. структурн. химии*. 2012;53(2):399-403.

35. Красных Е.Л. Прогнозирование энтальпий испарения на основе модифицированных индексов Рандича. Альдегиды и кетоны. *Журн. структурн. химии.* 2013;54(4):746-750.

36. Нестерова Т.Н., Нестеров И.А. Критические температуры и давления органических соединений. Анализ состояния базы данных и развитие методов прогнозирования. Самара: Изд-во Самарский научный центр РАН; 2009. 580 с. ISBN 978-5-93424-424-9

37. Estrada E. Spectral Moments of the Edge-Adjacency Matrix of Molecular Graphs. 2. Molecules Containing Heteroatoms and QSAR Applications. *J. Chem. Inf. Comput.* Sci. 1997;37(2):320-328. https://doi.org/10.1021/ci960113v

38. Skvortsova M.I., Fedyaev K.S., Palyulin V.A., Zefirov N.S. Molecular Design of Chemical Compounds with Prescribed Properties from QSAR Models Containing the Hosoya Index. *Internet Electronic Journal of Molecular Design (IEJMD).* 2003;2(2):70-85. URL: http://www.biochempress.com/av02_0070.html

38. Skvortsova M.I., Fedyaev K.S., Palyulin V.A., Zefirov N.S. Molecular Design of Chemical Compounds with Prescribed Properties from QSAR Models Containing the Hosoya Index. *Internet Electronic Journal of Molecular Design (IEJMD).* 2003;2(2):70-85.
URL: http://www.biochempress.com/av02_0070.html

39. Zefirov N.S., Palyulin V.A. QSAR for Boiling Points of "Small" Sulfides. Are the "High-Quality Structure-Property-Activity Regressions" the Real High Quality QSAR Models? *J. Chem. Inf. Comput. Sci.* 2001;41(4):1022-1027.
https://doi.org/10.1021/ci0001637

40. Zefirov N.S., Palyulin V.A. Fragmental Approach in QSPR. *J. Chem. Inf. Comput. Sci.* 2002;42(5):1112-1122.
https://doi.org/10.1021/ci020010e

39. Zefirov N.S., Palyulin V.A. QSAR for Boiling Points of "Small" Sulfides. Are the "High-Quality Structure-Property-Activity Regressions" the Real High Quality QSAR Models? *J. Chem. Inf. Comput. Sci.* 2001;41(4):1022-1027.
https://doi.org/10.1021/ci0001637

40. Zefirov N.S., Palyulin V.A. Fragmental Approach in QSPR. *J. Chem. Inf. Comput. Sci.* 2002;42(5):1112-1122.
https://doi.org/10.1021/ci020010e

About the authors:

Nadezhda A. Shulaeva, Student, Department of Higher and Applied Mathematics, M.V. Lomonosov Institute of Fine Chemical Technologies, MIREA – Russian Technological University (86, Vernadskogo pr., Moscow, 119571, Russia). E-mail: akacija@inbox.lv. https://orcid.org/0000-0002-5974-5213

Mariya I. Skvortsova, Dr. of Sci. (Physics and Mathematics), Associate Professor, Head of the Department of Higher and Applied Mathematics, M.V. Lomonosov Institute of Fine Chemical Technologies, MIREA – Russian Technological University (86, Vernadskogo pr., Moscow, 119571, Russia). E-mail: skvorivan@mail.ru. Scopus Author ID 6603801652, https://orcid.org/0000-0002-6179-8875

Nataliya A. Mikhailova, Senior Lecturer, Department of Higher and Applied Mathematics, M.V. Lomonosov Institute of Fine Chemical Technologies, MIREA – Russian Technological University (86, Vernadskogo pr., Moscow, 119571, Russia). E-mail: essen.05@mail.ru, https://orcid.org/0000-0002-8888-0190

Об авторах:

Шулаева Надежда Александровна, студентка кафедры высшей и прикладной математики Института тонких химических технологий им. М.В. Ломоносова ФГБОУ ВО «МИРЭА – Российский технологический университет» (119571, Россия, Москва, пр. Вернадского, д. 86). E-mail: akacija@inbox.lv. https://orcid.org/0000-0002-5974-5213

Скворцова Мария Ивановна, доктор физико-математических наук, доцент, заведующий кафедрой высшей и прикладной математики Института тонких химических технологий им. М.В. Ломоносова ФГБОУ ВО «МИРЭА – Российский технологический университет» (119571, Россия, Москва, пр. Вернадского, д. 86). E-mail: skvorivan@mail.ru. Scopus Author ID 6603801652, https://orcid.org/0000-0002-6179-8875

Михайлова Наталия Александровна, старший преподаватель кафедры высшей и прикладной математики Института тонких химических технологий им. М.В. Ломоносова ФГБОУ ВО «МИРЭА – Российский технологический университет» (119571, Россия, Москва, пр. Вернадского, д. 86). E-mail: essen.05@mail.ru. https://orcid.org/0000-0002-8888-0190

The article was submitted: January 10, 2020; approved after reviewing: February 24, 2020; accepted for publication: October 16, 2020.

Translated from Russian into English by S. Durakov
Edited for English language and spelling by Enago, an editing brand of Crimson Interactive Inc.