Russian Chemical Reviews **75** (11) 961 – 979 (2006) © 2006 Russian Academy of Sciences and Turpion Ltd

DOI 10.1070/RC2006v075n11ABEH003616

Molecular similarity concept and its use for predicting the properties of chemical compounds

M I Skvortsova, I V Stankevich, V A Palyulin, N S Zefirov

Contents

I.	Introduction	961
II.	Interpretation of the 'molecular similarity' term in chemistry	962
III.	Graph theory representation of molecules and descriptions of molecular structures based thereon	963
IV.	The quantitative measures of the molecular similarity	966
V.	Methods for the prediction of properties of chemical compounds based on the molecular similarity concept	969
VI.	Formalisation of the postulate 'similar structures possess similar properties' and its analysis	970
VII.	Examples of application of the molecular similarity concept	971
VIII.	Conclusion	975

Abstract. Main methods for determination of the degree of molecular similarity used in modern theoretical chemistry are surveyed. Methods for predicting the properties of chemical compounds based on the statement 'similar structures possess similar properties' are described. Examples of using these approaches for the modelling of structure – property correlations for organic compounds are given. Bibliography includes 254 references.

I. Introduction

The similarity concept is very important in chemistry. Its essence can be expressed as follows: similar objects possess similar properties. The term 'similarity' is widely used when comparing different chemical objects (atoms, molecules, *etc.*) and reactions, chemical databases, molecular properties and so on. The popularity of the term 'similarity' in chemical literature is steadily increasing. Thus the number of abstracts in the *Chemical Abstracts* using the term 'similarity' has increased over 25 years from 233 (1967 – 1971) up to 1697 (from 1987 to the first half of 1992).[1] International conferences and workshops dealing with various aspects of the similarity concept in chemistry are being held (see, for example,[2, 3]); a number of books and reviews on this subject have been published.[2, 4 – 16] Commercial computer programmes based on the similarity methods of analysis have been developed.[17 – 19]

M I Skvortsova N D Zelinsky Institute of Organic Chemistry, Russian Academy of Sciences, Leninsky prosp. 47, 119991 Moscow, Russian Federation. Fax (7-495) 135 53 28, tel. (7-495) 135 90 89, e-mail: stan@ineos.ac.ru
I V Stankevich A N Nesmeyanov Institute of Organoelement Compounds, Russian Academy of Sciences, ul. Vavilova 28, 119991 Moscow, Russian Federation. Fax (7-495) 135 50 85, tel. (7-495) 135 79 98, e-mail: stan@ineos.ac.ru
V A Palyulin, N S Zefirov Department of Chemistry, M V Lomonosov Moscow State University, Leninskie Gory, 119992 Moscow, Russian Federation. Fax (7-495) 939 02 90, tel. (7-495) 939 39 69, e-mail: vap@org.chem.msu.su (V A Palyulin), tel. (7-495) 939 16 20, e-mail: zefirov@org.chem.msu.ru (N S Zefirov)

Received 11 November 2005
Uspekhi Khimii **75** (11) 1074 – 1093 (2006); translated by D S Yufit

The similarity concept in chemistry has a long-standing history and numerous interpretations. It is a scientific idea which is rather intuitive and is difficult to be clearly defined or formalised.[20, 21] However, formalisation of the main chemical terms and development of their quantitative description are required due to fast computerisation of the chemical studies, availability of enormous amounts of data on the structures, properties and chemical reactions, rapid development of mathematical chemistry, computer chemistry, combinatorial chemistry and medicinal chemistry. These main chemical terms include, in particular, the terms 'molecular structure', 'molecular similarity' and 'similarity of chemical reactions'.

The similarity concept is used for the solution of the following problems in chemistry:

— prediction of the physicochemical properties and biological activity of chemical compounds;[22 – 33]

— classification and ordering of data in chemical information databases (like the *Beilstein's Handbook of Organic Chemistry*);[34]

— planning of chemical experiments and the reactivity theory;[35 – 40]

— search for biologically active analogues in chemical databases in order to optimise the structures of already known active compounds;[41 – 51]

— clusterisation of databases and creation of representative subsets of compounds for further screening;[16, 52 – 59]

— creation of databases with the highest structural variety for further search for new key compounds.[18, 60, 61]

The present review discusses the concept of molecular similarity and the methods for predicting chemical properties based on this concept.

The so-called statistical approach is used most widely for determination of quantitative relationships between the structure and activity/property of compounds (QSAR/QSPR analysis).[62 – 64] Some limitations of this technique have induced the development of a method based upon the similarity concept. Successful realisation of the statistical method requires a training set of compounds with known values of the studied property. The structures of studied compounds are described by a set of some molecular descriptors $x_1, \ldots, x_N$ (topological, geometrical, electronic) and then the following equation can be written

$$y = f(x_1, \ldots, x_N),$$

where f is a function of N variables, y is a property value. Usually the type of function f (for example, linear, quadratic, *etc.*) is

specified in advance and depends on a series of fitting parameters, which are determined using the training subset of compounds. The linear function f is the most frequently used one. The obtained equation is called a linear regression model. In order to calculate the properties of new compounds on the basis of this model, initially the values of the descriptors $x_1, \ldots, x_N$ are calculated for the target compounds and then the value of the property y itself is calculated using the equation $y = f(x_1, \ldots, x_N)$.

However, another approach to predicting the properties of compounds also exists. It is based upon the following hypothesis: 'similar' structures possess 'similar' properties.[13,29,65] In order to estimate the property of a compound, another compound, 'similar' (in some sense) to the studied one but with known value of this property, has to be found in a dataset. Then it is assumed that the target properties of these compounds are the same (or just slightly different). The averaging of the properties of several 'nearest neighbours' is also possible. Some mathematical objects m_1 and m_2 (for example vectors, sequences, graphs, *etc*.) are assigned to the molecules M_1 and M_2 in order to make a quantitative comparison of their similarity. Then a scalar function $d(m_1,m_2)$ defined for the whole set of the objects is introduced. Calculating the values of this function for the pairs 'fixed compound M_1 – every other compound of the set' one can find the compound most similar to M_1 according to the measure $d(m_1,m_2)$.

There is a variety of methods for the description of molecular structures in mathematical terms. For each of them, in turn, a number of similarity may be defined. Thus, the similarity measures of molecules depend upon the method of their mathematical representation and on the selection of similarity criterion. It has to be noted that there is no universal similarity criterion, equally suitable for all properties and any class of compounds. Both the method of the description of compounds and the similarity criterion have to be specially selected for each individual problem. The result of a search for a 'similar' structure also depends on the dataset of compounds, through which the search is carried out.

While comparing the two above approaches to the modelling of structure – property relationships, one may notice that: (i) the latter provides more flexibility in the description of a molecular structure because molecules can be described not only by vectors of molecular descriptors $(x_1, \ldots, x_N)$ but also by graphs, sequences of symbols, sets, *etc*. Moreover, the number N of vectors $(x_1, \ldots, x_N)$ used in the former approach is usually not greater than the number of compounds in a dataset, while in the latter method, the value of N is unlimited. (ii) In the former method, the model takes into account the properties of all compounds of the training set; therefore, the value of a property of a new compound depends upon all the compounds, while in the latter method, this value depends only on the properties of a 'similar' compound. Both methods also have obvious shortcomings: in each case, there is a problem of adequate description of the molecular structure and of correct choice of the functions f and $d(m_1,m_2)$.

A huge number of papers based on the concept of molecular similarity have transformed this area into a new branch of theoretical chemistry. Analysis of these papers reveals that they differ by one or several of following points: (i) the studied subset of compounds; (ii) the studied properties; (iii) the method used for mathematical description of chemical structures; (iv) the quantitative criteria of molecular similarity; (v) the methods used to estimate the properties of compounds on the basis of the selected similarity criteria. The positions iii – v are the most methodologically interesting.

The aim of the present review is to discuss the most widespread methods of description and comparison of molecular structures and also the methods for prediction of properties of the chemical compounds based on the concept of molecular similarity. Some attention is also paid to the validity of the postulate 'similar structures possess similar properties' and to the possible ways of its formalisation. Besides, a series of examples illustrating the advantages and shortcomings of these methods are given and related problems are discussed.

II. Interpretation of the 'molecular similarity' term in chemistry

The term 'molecular similarity' is used in chemistry in a very broad meaning, because molecules may be characterised by a number of structural and physicochemical properties and may exhibit various biological activities. The areas of chemistry that used the molecular (or atomic) similarity concept in some form are summarised in a paper.[66] In particular, the definitions of such fundamental chemical terms as isomerism, stereoisomerism, homology, periodicity of properties of the chemical elements, *etc*., are based upon the similarity of some molecular (or atomic) properties.

Some examples of modern usage of the term 'molecular similarity' are given below.

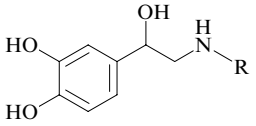
1. Classification of the organic compounds is based mainly upon the classical theory of chemical structure. The similarity of the chemical structures of molecules means some resemblance in the corresponding structural formulae and usually results in relatively small variations of some physicochemical properties and in similar methods of synthesis. However, it is obvious that there is no unique way for evaluation of molecular similarity based solely upon the structural formulae of molecules.

2. The molecules that possess identical skeleton and differ only by heteroatoms (for instance pyrrole and thiophene, benzene and pyridine) may be regarded as structurally related ones.

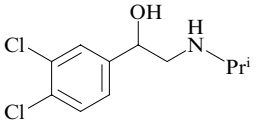
3. The members of homological series {for example, $CH_3-(CH_2)_n-CH_3$, $R^1\text{-}[\text{-}CH{=}CH\text{-}]_n\text{-}R^2$} are also regarded as structurally related molecules. The physicochemical properties of the molecules in such series are usually linear functions of n.

4. The ability of atomic groups to form chemical bonds is described by analogy principle proposed by Hoffmann.[67] The valence properties of chemical groups are determined by their frontier molecular orbitals (FMO), namely by their number, symmetry, shape and energy. Two chemical groups are called isolobal if the number and symmetry of their FMO are identical and the shape and energy of these FMO 'are reasonably close'. The principle of isolobal analogy states that isolobal groups are mutually replaceable in chemical compounds regardless of the type of atoms in them. This principle links the valence theory of classical organic chemistry with the chemistry of organoelement compounds and provides a possibility of a unified explanation of the structural variety of chemical compounds and of the similarity of their properties.

The given examples illustrate the fundamental character of the chemical similarity concept. However, the very same definition of the term 'similarity' is ambiguous and, usually, intuitive.[13,68] The prediction of physicochemical properties or biological activity of compounds on the basis of the postulate 'similar structures possess similar properties' is one of the most important areas of the similarity concept application. However, the postulate is not always valid in the case of visual comparison of the structural formulae. For example, in the series of close analogues (noradrenaline, adrenaline, isoproterenol), the α-adrenergic activity decreases, but the β-adrenergic activity increases. Moreover, the replacement of the phenolic hydroxy groups of isoproterenol by chlorine atoms converts this agonist of β-adrenergic receptors into an antagonist, dichloroisoproterenol.[69]

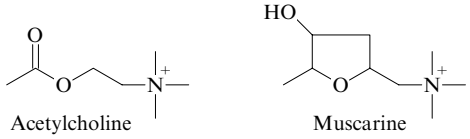


R = H (noradrenaline),
Me (adrenaline),
Pr^i (isoproterenol).



Dichloroisoproterenol

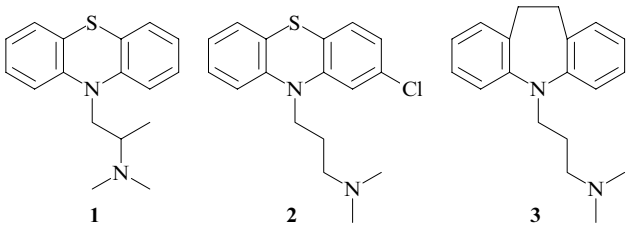
Yet another example is the poison muscarine, isolated from toadstools, and acetylcholine, which is produced in the human and animal bodies and takes part in the transfer of nerve impulses.[70]



Acetylcholine Muscarine

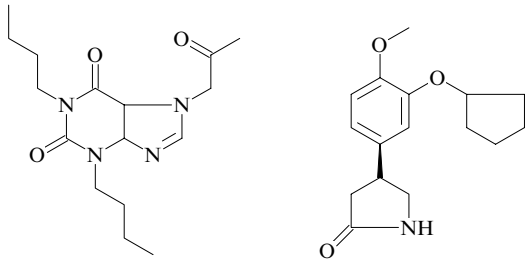
The structures of these compounds are similar and both of them affect the human cholinergic systems.

Sometimes, structurally similar compounds exhibit different types of biological activity. Thus, compound **1** is antihistamine, compound **2** is neuroleptic (antagonist of the dopamine receptors) and compound **3** is antidepressant.[71]



1 **2** **3**

At the same time structurally different compounds may exhibit identical biological activity, as denbufylline and rolipram do, which both are phosphodiesterase inhibitors.



Denbufylline Rolipram

The electrostatic potential of a molecule may be used for the description of similarity degree of these compounds.[69, 71]

These examples illustrate that in addition to the visual analysis of structural similarity, other methods of the molecular structure description are required and new approaches to determination of the molecular similarity degree have to be developed.

III. Graph theory representation of molecules and descriptions of molecular structures based thereon

1. Molecular graphs

According to the modern views, the molecular structure may be described at various levels of complexity. The simplest way of description is the molecular formula. The next step of complexity is the classical structural formula. Further steps include building of 3D molecular models and calculation of quantum-mechanical parameters.

However, there is yet another way of mathematical description of a molecular structure. A labelled molecular graph may represent a molecule. The vertices and edges of the graph will correspond to the atoms and bonds of a molecule and the labels of vertices and edges (numerical or alphabetical) mark the chemically different atoms and bonds. In fact, the classical structural formula is a particular type of labelled molecular graph, where the chemical symbols of corresponding atoms are used as the labels of the vertices and the edge labels correspond to bond orders. Generally, any atomic parameters (atomic charge, polarisability, atomic weight, atomic radius, *etc*.) may be used as vertex labels. Similarly, any bond parameters (bond length, bond order, obtained from the quantum-chemical calculations, *etc*.[29, 72]) may be used as edge labels. For example, so-called *be*-matrices (bond and electron matrices) have been proposed.[73] Their non-diagonal elements are equal to the bond orders and the diagonal elements, to the number of valence electrons of the corresponding atoms non-involved in bonding. A disconnected graph, *i.e*., a graph which consists only of isolated labelled vertices without edges, formally corresponds to a molecular formula.

It is also possible to create complete molecular graphs where each pair of vertices is linked by an edge. In this case, a label (or weight) is assigned also to each pair of non-bonded atoms. The topological (or geometrical) distance between these two vertices may be used as such a label. In the latter case, the graph will correspond to the well-known matrix of geometrical distances.

Let us assume that the spatial coordinates of the atoms are known. In this case, one can construct a graph comprising isolated vertices labelled with sequences of three numbers, namely, the atom coordinates. A physicochemical parameter of the corresponding atom can be used as a label for a graph vertex.

Thus, according to the choice of labels, the molecular graphs may be grouped into three major types: calculated on the basis of the structure of the graph itself, taken from literature (for example, the standard physicochemical properties of atoms and bonds) or obtained experimentally (for example, determined by the X-ray crystallography). The combination of these methods is also possible.

In some cases, so-called 'simple' molecular graphs are used. The vertices of these graphs are represented by dots and edges are represented by line segments. Although a simple graph does not have obvious labels, it is assumed that labels of the vertices are equal to zero. If the atoms are chemically bonded, the corresponding edge is labelled by 1, while for nonbonded atoms the edge label is equal to zero. Such graphs are usually used to represent hydrocarbons structures.

2. Descriptions of molecular structures based on their theoretical graph representations

A labelled graph is determined by its topological matrix

$$A = (a_{ij}),$$

where a_{ii} is the label of vertex i, a_{ij} is the label of a pair of vertices (i, j). (In the case of a simple graph, matrix A is usually called an adjacency matrix.) Obviously, the change of the vertex numbering scheme results, generally speaking, in a different matrix. This is a significant disadvantage of the chemical structure description in matrix terms. This problem leads to the concept of 'graph invariant'. A graph invariant is a number (or a function if the graph labels are symbols) calculated from graph matrix A, which does not depend on the vertex numbering scheme. In theoretical chemistry the invariants of simple graphs are usually called topological indices (TI). Quite often, this definition is also used for invariants of labelled graphs, which can represent not only the topology, but also the elements of electronic or spatial structure. Invariants of any subgraph of a graph can be determined by a similar way. They are called local TI. Topological indices are widely used in theoretical organic chemistry for quantitative characterisation of the molecular structure in a search for structure–property correlations, in the computer-aided synthesis and in creation and analysis of structure–property databases. The most well-known TI and their applications in chemistry are described in the reviews.[63, 64, 74–76] The simplicity of calculations of TI, the possibility to take into account the electronic and spatial features of the structures and also numerous examples of successful property–TI correlations are the reasons for the popularity of this method of molecular structure description. At the same time, this approach has an obvious disadvantage: it does not differentiate between molecular configurations (*e.g*. stereoisomers) and does not take into account their conformational features.

Some examples of the most common named TI are given below.[63, 64]

The Randić index χ is defined according to the following formula

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

where v_i and v_j are the degrees of graph vertices i and j and the sum is taken for all the graph edges. The index χ has been generalised by Kier and Hall, who introduced the definition of m-order connectivity indices.

$${}^{m}\chi^{v} = \sum \left(\delta_i \delta_j \delta_k \ldots\right)^{-1/2}, \quad m \geqslant 1, \quad \delta_i = \frac{Z_i^{v} - h_i}{Z_i - Z_i^{v} - 1},$$

where Z_i is a total number of electrons, Z_i^{v} is a number of valence electrons of atom i (corresponding to vertex i); h_i is a number of hydrogen atoms bonded to atom i; the summation takes place along all the chains which consist of m graph edges; i, j, k are the numbers of vertices forming the chain.

The Wiener index $\boldsymbol{W}$ is defined as a sum over all pairs of vertices

$$\boldsymbol{W} = 0.5 \sum d_{ij},$$

where d_{ij} is the shortest distance between vertices i and j. Matrix $D = (d_{ij})$ is called the matrix of topological distances of the graph.

The Hosoya index $\boldsymbol{Z}$ is defined as

$$\boldsymbol{Z} = \sum p_k,$$

where p_k is a number of ways to choose k nonadjacent edges ($k \geqslant 1$) in a graph.

The Balaban index $\boldsymbol{B}$ [for acyclic graphs (trees)] is defined according to the following formula

$$\boldsymbol{B} = \sum \boldsymbol{r_i^2},$$

where r_i is a number of the tree vertices cut out at step i of the 'tree pruning' procedure. This procedure consists of consecutive removal of the 1st degree vertices of the initial tree and the edges incident to them.

During the construction of informational topological indices,[77] a set of fragments (for example, a set of graph vertices) is selected and split into a number of equivalent classes according to some feature (for example, according to the first coordination type). Let assume that n is a number of graph vertices, n_i is a number of vertices in the equivalent class i, $i = 1, \ldots, m$. The main informational TI is defined as

$$\mathrm{IC} = -\sum \frac{n_i}{n} \log_2 \frac{n_i}{n}.$$

This TI characterises the inhomogeneity of the graph vertices set and has extreme values for the least homogeneous ($m = n$) and the most homogeneous ($m = 1$) sets of vertices. Instead of the set of vertices, a set of any other fragments can be used.

Another large class of TI is the spectral-type TI.[64] Let us call λ_i and $\boldsymbol{C}^{(j)} = [C_1^{(j)} \ldots, C_n^{(j)}]$ the eigenvalues and the eigenvectors of the matrix of a graph with n vertices. They can be used for the construction of various TI, for example such as

$$J_1 = \max\left(\lambda_i\right),$$

$$J_2 = \lambda_r - \lambda_m,$$

where λ_r is the minimum positive and λ_m is the maximum negative eigenvalues, or

$$q_i = \sum C_i^{(k)} C_i^{(k)},$$

$$p_{ij} = \sum C_i^{(k)} C_j^{(k)}$$

(the summation in two latter formulae is over those eigenvectors to which the positive eigenvalues correspond). Spectral-type topological indices appear naturally in some quantum-chemical methods and usually have a quite clear physicochemical meaning.

It should be noted that even for a single fixed labelling scheme of a molecular graph an infinite variety of TI exists. For instance, the multiplication or summation of two TI results in a new TI. Therefore, it is impossible in principle to describe all TI, which can be constructed and used in the QSAR/QSPR analysis. The problem of reasonable choice of just a few TI out of infinite number of variants always arises during the construction of each particular model of structure – property correlation. This happens because it is often difficult to give a relatively clear physicochemical or structural interpretation of TI and to correlate it with the studied property. Moreover, some TI are related to each other by exact or approximate mathematical equations. These facts, as well as a common sense, prompt that for any finite set of the weighted graphs a finite set of TI $J_1, J_2, \ldots, J_N$ should exist. This set is called a basic set. Then, any TI may be represented as

$$\mathrm{TI} = F(J_1, J_2, \ldots, J_N),$$

where $F(J_1, J_2, \ldots, J_N)$ is a function of N variables. The type of the function depends on the given topological index. It was rigorously proved [78] that such basis always exists. The number of occurrences of some structural fragments in the graph may be taken as the basis invariants and F may be a linear function. Some examples of the expansion to a basis of various invariants of molecular graphs have been reported.[79] These results may be regarded as a mathematical foundation of the well-known principle of additive schemes, used for approximation of the physicochemical properties of compounds.[80, 81]

One more approach to the problem of selection of basic TI has been suggested.[82] These basic TI are constructed relying on specially selected subgraphs and depend on variable parameters. Another, more QSAR/QSPR-friendly definition of the basis is also given in the same paper. The suggested method was successfully used for the estimation of various physicochemical properties of the hydrocarbons of different classes.

The number of occurrences of a structural fragment in the structure of a molecule is a particular case of TI. The simplest structural fragments have been used in the theoretical chemistry for a long time for the estimation of physicochemical properties by additive schemes.[80, 81] The QSAR/QSPR studies use more complicated fragments as well.[62, 83–85]

It is easy to notice that virtually all TI are somehow related to subgraphs. Fragmental descriptors are of great importance for QSAR/QSPR methods. Below we discuss some of these descriptors.

Let us assume that a chain of $L-1$ bonds (*i.e.* containing L atoms) 'links' together atoms i and j (with Y_i or Y_j codes). Such a chain fragment of the molecule is usually called 'an atomic pair (i, j)' and is coded by a symbol $\mathrm{Y}_i - (L-1) - \mathrm{Y}_j$.[86] The code Y_i (or Y_j) contains information about the type of the chemical element, the number of connected atoms (without hydrogens) and the number of π-electrons. As an example, Table 1 gives the 'atomic pairs' and the number of their occurrences in the structural formula of the acetone molecule.[69] The symbols C and O (see Table 1) correspond to carbon and oxygen atoms, the symbol X_n ($n = 1, 2, 3$) indicates that this particular atom is connected to n

Table 1. Atomic pairs in the acetone molecule.

Atom	Atomic pairs	Number of occurrences of the fragment in the structure
1	$\mathrm{CX_1-1-C.X_3}$	2
2	$\mathrm{C.X_3-1-O.X_1}$	1
3	$\mathrm{CX_1-2-O.X_1}$	2
4	$\mathrm{CX_1-2-CX_1}$	1

atoms (hydrogen atoms are ignored), the symbol '.' indicates that this atom contains one π-electron. The number $L-1$ in this example is equal to 1 or 2. A different method of the description of atomic pairs based upon physicochemical parameters of the atom has also been suggested.[87]

Below are examples of the descriptors, labelled MNA/k, which are linked to the molecule fragments that are k-order neighbours of atoms of various types ($k = 0, 1, ...$).[69,88] The symbols '−' in these descriptors are given to acyclic atoms; the zero order fragment is the atom itself; the first order fragment is the atom with the notification (in the parentheses) of its first-order neighbour, *etc*. The fragments MNA/0 and MNA/1 of the structural formula of phenol for three phenol atoms (labelled 1, 2 and 3) are given in Table 2 as examples.[69]

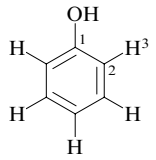
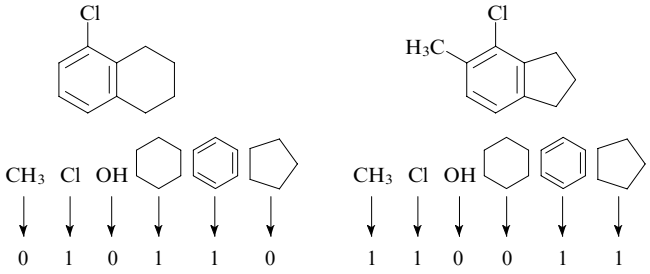


Table 2. Fragments MNA/0 and MNA/1 for atoms 1, 2 and 3 in the phenol molecule.

Atom	MNA/0	MNA/1
1	C	C(CC−O)
2	C	C(CC−H)
3	−H	−H(C)

The atom-centred fragments of various orders have been previously used a number of times.[65,77,80,81] They have also been discussed in other publications.[89,90]

Usually the vectors $\boldsymbol{X} = (X_1, X_2, ...)$, where $X_i = 0$ if the fragment i is absent in the structure and $X_i = 1$ if the fragment i is present, are composed on the basis of the selected fragment set. The number of occurrences of a fragment i (*i.e*., an atom chain) in the structure also may be used as X_i.[91] (Originally, this approach was suggested for alkanes.[92]) Such (0,1) vectors $\boldsymbol{X} = (X_1, X_2, ...)$ are also called 'a bit string'. Some relevant examples are given below.[69]



Various TI X_i may be used for the construction of vectors $\boldsymbol{X} = (X_1, X_2, ...)$. Particularly, if the components of vector $\boldsymbol{X}$ are non-negative and $\sum X_i \neq 0$, then another vector $\boldsymbol{P} = (P_1, ..., P_k)$, where $P_i = X_i / \sum X_i$, may be constructed.[9]

Examples of vector construction have been reported.[69,93] Two matrices P and C are introduced for a structure containing N atoms. P is a rectangular matrix of the $N \times 3$ size, where row i is given by $(1, p_i', q_i')$ (p_i', q_i' are specially normalised electronegativity and effective charge of atom i). The C is an (0,1)-adjacency matrix. Then a family of square 3×3 matrices TM_n, called electrotopological matrices, is defined:

$$TM_n = \frac{P^* C^n P}{N}, \; n = 1, 2, ..., n_{\max},$$

where P^* is the transposed P matrix; C^n is the n-degree of matrix C; $n_{\max}$ is the longest sequence in the corresponding graph. These matrices do not depend on the labelling of the vertices of the molecular graph; therefore, their elements may be used as descriptors called electrotopological indices. Vectors consisting of elements of these matrices for $n = 1, 2, ..., n_{\max}$ may be constructed on the basis of these elements.

It was suggested[94,95] to use specially formulated rectangular matrices dependent on fragment descriptors to construct vectors describing a molecular structure.

The vectors may also be constructed from quantum-chemical[96] and any other descriptors[97] and also on the basis of physicochemical properties of the compounds.[98]

A molecular graph with indication of various sets of fragments, described by a sequence of symbols, may be used for characterisation of a molecular structure.[9] Linear nomenclatures are examples of such graphs. In such a nomenclature the structural formula is coded by a string of symbols. Each symbol, in turn, determines the structural fragments. The works of Wisswesser are at the origin of the linear nomenclatures.[99] For instance, for the compounds $CH_3-CH_2-O-CH_2-CH_2-CH_3$ and $H_2N-CH_2-CH_2-C(O)-OH$ the Wisswesser codes are (2O3) and (Z2VQ) respectively.[62] The symbol 'O' indicates the oxygen atoms, not bonded to H atoms; 'Q' is a hydroxy group; 'V' is a carbonyl group; 'Z' is a nitrogen atom of an NH_2 group. Number 2 (or 3) indicates an unbranched chain of two (or three) CH_2 groups.

Yet another linear nomenclature is called SMILES.[100,101] It is used, particularly, in search for the structural similarity of compounds. The examples of linear codes obtained by the FCSS (fragmentary code of substructure superposition) language have been reported.[85] The SLN code (SYBYL Line Notation) has been described.[102] This code has been used[103] for protein similarity estimation.

Simple graphs may be coded with the minimum binary code. In order to do that, all possible ($n!$, where n is a number of graph vertices) adjacency matrices are constructed. Then each matrix is converted into one string of 0 and 1. This string may be regarded as a binary number. A minimum lexicographic sequence corresponding to the minimum decimal number is selected as a code.[64]

There is one more method of sequences construction. Let us consider a set of N molecules $\{M_i\}$, $i = 1, ..., N$, and assume that for each pair M_1 and M_2 of this set a symmetrical function $d(M_1,M_2) = d(M_2,M_1)$ is given. This function is regarded as a 'distance' between M_1 and M_2. Then for each molecule M_k of set $\{M_i\}$ a sequence $\{d(M_k,M_i)\}$, $i = 1, ..., N$, $i \neq k$ may be composed,[9,29,104] as well as sequences consisting of local graph invariants, *i.e*., the powers of graph vertices.[105] In contrast to the vectors, which are used for chemical structure descriptions, the sequences may be of various lengths.

Functions are also used for the description of a molecular structure. Let us assume that a molecule is described by a simple graph G and d_{ij} is a topological distance between vertices i and j. The function called topological transformation of graph G

$$t(s) = \frac{k(G) \sum \sum Z_i Z_j [\sin(s d_{ij})]}{s d_{ij}},$$

where s is real parameter, $k(G)$ is a normalising constant such that $\int [t(s)]^2 ds = 1$, Z_i and Z_j are charges of nuclei of corresponding atoms has been considered previously.[106,107] A similar formula has also been suggested[108] for 3D structures, when d_{ij} is a geometrical distance between atoms i and j. In the latter case, functions of several variables, for example, charge densities, electrostatic potentials *etc*., are also used.[13,109]

An algorithm for the description and analysis of the molecular shape, based on quantum-chemical calculations of charge densities or electrostatic potentials, has been suggested.[110–115] A study of the molecular surface is the starting point of the procedure. The charge density surfaces of various levels are calculated. A set of

these equal level surfaces are regarded as topological objects. The objects are characterised by homology groups, called shape groups. The shape groups are used for the description of molecular shapes and for the analysis of their similarity. This method may be successfully combined with the description of stereoisomers in terms of the knot theory and is easily programmable, which provides an opportunity for automatic non-visual analysis of the molecular similarity degree.

Spatial geometrical models of molecules may be constructed on the basis of structural formulae, using the molecular mechanics or quantum-chemical methods.[62] These models are potential source of new molecular descriptors. Thus, for example, if the orthogonal coordinates (X_{i1}, X_{i2}, X_{i3}) of a molecule, containing N atoms of masses m_i $(i = 1, \ldots, N)$, are known, then the formulae

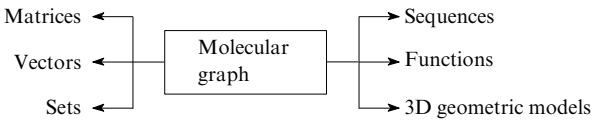
$$U_j = \frac{1}{M}\sum_{i=1} m_i X_{ij},$$

where $j = 1, 2, 3$; $M = \sum_{i=1}^{N} m_i$ is the molecular mass, determine the coordinates of the centre of mass. Let us denote the inertia tensor as $R = (r_{jk})$

$$r_{jk} = \sum_{i=1}^{N} m_i (X_{ij} - U_j)(X_{ik} - U_k),$$

and the eigenvalues of matrix R as X, Y, Z. These eigenvalues and some of their functions, such as X/Y, X/Z, *etc*., were suggested [62] as molecular geometrical descriptors. Some authors use the van der Waals volume or the surface area of the molecule for the description of the molecular spatial structure. The details of the calculations of these descriptors can be found in the literature.[62]

Usually, methods for molecular structure description are subdivided into 2D and 3D methods. The bond angles and the Euclidean interatomic distances are not taken into account in the former case. Therefore, the methods, based on the matrices of topological distances of the molecular graph are 2D methods and the invariants determined by these methods are 2D descriptors. The methods based on the matrices of Euclidean distances are 3D methods and the invariants determined by these methods are 3D-descriptors. The degree of molecular similarity discussed below usually depends only on the mathematical representation of the molecules (graphs, sets, matrices, vectors, sequences, *etc*.) and does not depend on the information coded by these mathematical models. Therefore, we do not discuss separately the 2D and 3D methods of analysis of molecular similarity. The methods of representing the molecular structure, where those based on the graph theory are most important and the other ones are just its derivatives, may be illustrated in a form of the following Scheme.



IV. The quantitative measures of the molecular similarity

The quantitative measure of similarity of two molecules M_1 and M_2 is a scalar function $d(S_1,S_2)$ of two variables S_1 and S_2. Its type depends on the method of description of the structures S_1 and S_2 of molecules M_1 and M_2. The value of $d(S_1,S_2)$ for a pair of the structures is a measure of their similarity. Let us assume that the function $d(S_1,S_2)$ is such that:

$d(S_1,S_2) \geqslant 0$;

$d(S_1,S_2) = 0$ then and only then, when $S_1 = S_2$;

$d(S_1,S_2) = d(S_2,S_1)$ (symmetry);

$d(S_1,S_2) \leqslant d(S_1,S_3) + d(S_2,S_3)$ (the triangle inequality).

In this case the function $d(S_1,S_2)$ is called a metric. Most of the measures of the molecular similarity are metrics. In this Section we will describe some similarity measures, developed for various mathematical methods of molecular representation, which are listed in the Scheme at the end of Section III.[9]

1. Let us assume that the structures S_1 and S_2 are represented by n-dimensional vectors $\boldsymbol{X} = (x_1, x_2, \ldots, x_n)$ and $\boldsymbol{Y} = (y_1, y_2, \ldots, y_n)$. The components of these vectors are some invariants of molecular graphs. The following similarity measures are described in literature:

$$d_1(\boldsymbol{X},\boldsymbol{Y}) = \sum |x_i - y_i| \ (L_1 \text{ is metric}),$$

$$d_2(\boldsymbol{X},\boldsymbol{Y}) = \left(\sum |x_i - y_i|^p\right)^{1/p} \ (L_p \text{ is metric});$$

for $p = 2$, d_2 is a distance between points X and Y in the Euclidean space R^n. (Euclidean distances are widely used in the estimation of the molecular similarity degree [24, 65, 116–122].)

$$d_3(\boldsymbol{X},\boldsymbol{Y}) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\left|\sum (x_i - \bar{x}) \sum (y_i - \bar{y})\right|^{1/2}},$$

$$\bar{x} = \frac{1}{n}\sum x_i, \quad \bar{y} = \frac{1}{n}\sum y_i$$

(correlation coefficient);

$$d_4(\boldsymbol{X},\boldsymbol{Y}) = \frac{\sum x_i y_i}{\sum x_i^2 + \sum y_i^2 - \sum x_i y_i}$$

(Tanimoto coefficient).

The Tanimoto coefficient is one of the most common similarity measures in QSAR/QSPR studies as well as in the analysis of chemical databases.[2, 53, 122–127] (0,1)-Vectors are usually used as vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$. The component values of these vectors depend on the presence (1) or absence (0) of some fragments in the structures S_1 and S_2. In this case, the Tanimoto coefficient has the form

$$d_4(\boldsymbol{X},\boldsymbol{Y}) = \frac{N_{xy}}{N_x + N_y - N_{xy}},$$

where N_x and N_y are the numbers of components in each of the vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ equal to 1; N_{xy} is the number of 'common' components of these vectors equal to 1.

$$d_5(\boldsymbol{X},\boldsymbol{Y}) = \frac{\sum \min(x_i, y_i)}{\sum x_i^2 + \sum y_i^2 - \sum \min(x_i, y_i)} \quad \text{(Ref. 88)},$$

$$d_6(\boldsymbol{X},\boldsymbol{Y}) = \sum (x_i - y_i) \log_2 \frac{x_i}{y_i} \quad (x_i,\ y_i > 0),$$

$$d_7(\boldsymbol{X},\boldsymbol{Y}) = \frac{\sum x_i y_i}{\left(\sum x_i^2\right)^{1/2} + \left(\sum y_i^2\right)^{1/2}}$$

(cosine of the angle between the vectors),[91, 121, 126, 127]

$$d_8(\boldsymbol{X},\boldsymbol{Y}) = \frac{\sum \min(x_i, y_i)}{0.5\left(\sum x_i + \sum y_i\right)}.$$

The metric $d_8(\boldsymbol{X},\boldsymbol{Y})$ was introduced for a case where the components of vector $\boldsymbol{X}$ (or $\boldsymbol{Y}$) are equal to the number of occurrences of the special fragment, namely, an atomic pair, in the structure S_1 (or S_2) (see Section III).[86] The method for estimation of the similarity based on atomic pairs is sometimes called the 'atomic pairs method'.

The measure considered in a study [69]

$$d_9(\boldsymbol{X},\boldsymbol{Y}) = \left[1 + \frac{1}{n}\sum_i (X_i - Y_i)^2\right]^{-1},$$

was defined for vectors constructed from the elements of electrotopological matrices [70, 92] (see Section III).

2. Let us assume that the structures S_1 and S_2 are represented by finite sets D_1 and D_2 and denote the number of elements in the set D as $|D|$. The following similarity measures are described:[9]

$$d_{10}(D_1, D_2) = |D_1| + |D_2| - 2|D_1 \cap D_2| = |D_1 \backslash D_2| + |D_2 \backslash D_1|$$

(Hamming metric [128]),

$$D_{11}(D_1, D_2) = \frac{|D_1 \cap D_2|}{|D_1 \| D_2|}$$

(Ohia coefficient).

3. Let us assume that the structures S_1 and S_2 are represented by the labelled graphs G_1 and G_2; E is a set of graph vertices, $|G|$ is the total number of vertices and edges of the graph G; $|\mathrm{MACS}(G_1, G_2)|$ is the total number of vertices and edges in the maximum common subgraph G_1 and G_2. The following measures were introduced for the graphs:[9]

$$d_{12}(G_1,G_2) = |E_2 \backslash E_1| + |E_1 \backslash E_2|,$$

$$d_{13}(G_1, G_2) = |G_1| + |G_2| - 2|\mathrm{MACS}(G_1,G_2)|,$$

$$d_{14}(G_1,G_2) = \frac{|\mathrm{MACS}(G_1, G_2)|^2}{|G_1 \| G_2|}.$$

The methods of finding the maximum common subgraph have been reviewed.[129, 130] It should be noted, that the subgraph may also be found for the edge graphs derived from these graphs.[131] Eight similarity measures based on the maximum common subgraph have been reported.[132]

In order to define the measure $d_{15}(G_1,G_2)$, let us choose a subgraph (fragment) H. Let $n(G) = 1$ if H belongs to the graph G, and $n(G) = 0$ if H does not belong to G. Let us consider the measure

$$d_{15}(G_1,G_2) = |n(G_1) - n(G_2)|.$$

Obviously, $d_{15}(G_1,G_2) = 1$ if one of the graphs contains the fragment H and another one does not (the graphs G_1 and G_2 are 'dissimilar'); and $d_{15}(G_1,G_2) = 0$ if H either belongs to both graphs or to none of them (G_1 and G_2 are 'similar'). This measure allows one to sort the structures of a set into two classes: those containing and not containing the fragment H. The application of this measure in QSAR results in a search for the fragment H present in all active compounds of a set and absent in the structures of inactive compounds.[62, 85]

4. Let us assume that the structures S_1 and S_2 are represented by the sequences l_1 and l_2; $|l|$ is the length of the sequence; $|\mathrm{LCS}(l_1, l_2)|$ is the length of the longest common sequence in l_1 and l_2. For the sequences, the following similarity measures were suggested:[9]

$$d_{16}(l_1, l_2) = |l_1| + |l_2| - 2|\mathrm{LCS}(l_1, l_2)|,$$

$$d_{17}(l_1, l_2) = \frac{2|\mathrm{LCS}(l_1, l_2)|^2}{|l_1 \| l_2|},$$

$$d_{18}(l_1, l_2) = \frac{1 - d_{17}(l_1, l_2)}{|l_1 \| l_2|}.$$

Also, let us define the elementary 'editing' operations of the sequences: the removal of the element x, the addition of the element x, the replacement of x by y in the particular sequence. The similarity measure of two sequences l_1 and l_2 may be introduced as the minimum number of such operations required to transform l_1 into l_2.[14]

5. Let us assume that the functions f_1 and f_2 correspond to the structures S_1 and S_2. The following similarity measures may be defined on this basis:

$$d_{19}(f_1,f_2) = \left(\int |f_1(x) - f_2(x)|^2 \mathrm{d}x\right)^{1/2},$$

$$d_{20}(f_1,f_2) = \int |f_1(x) - f_2(x)| \mathrm{d}x,$$

$$d_{21}(f_1,f_2) = \frac{\int f_1 f_2 \mathrm{d}x}{\int |f_1|^2 \mathrm{d}x \int |f_2|^2 \mathrm{d}x}.$$

Similar formulae may also be written for the functions of several variables.

There are several 'named' similarity measures associated with the functions. Let us consider the measure [133] specified as

$$d(f_1,f_2) = \int\int f_1(r_1)\Omega(r_1, r_2)f_2(r_2)\mathrm{d}r_1\mathrm{d}r_2,$$

where f_1 and f_2 are the functions of charge density, Ω is a positive definite operator. The measure $d(f_1,f_2)$ was used to specify other measures, for example:

$$d_{\mathrm{C}}(f_1,f_2) = \frac{d(f_1,f_2)}{d(f_1,f_1)^{1/2} d(f_2,f_2)^{1/2}}$$

(Carbo's index [134–140]),

$$d_{\mathrm{HR}}(f_1,f_2) = \frac{2\,d(f_1,f_2)}{d(f_1,f_1) + d(f_2,f_2)}$$

(Hodgkin–Richards index [109, 134–136, 141]),

$$d_{\mathrm{P}}(f_1,f_2) = \frac{d(f_1,f_2)}{\max[d(f_1,f_1), d(f_2,f_2)]}$$

(Petke's index [142, 143]).

Two sets of similarity indices, dependent on the parameters x and k, were suggested[133]

$$D_1(f_1,f_2) = \left[\frac{k[d(f_1,f_1) + d(f_2,f_2)]}{2} - xd(f_1,f_2)\right]^{1/2},$$

$$D_2(f_1,f_2) = \frac{(k-x)d(f_1,f_2)}{D_1^2(f_1,f_2)}.$$

The indices D_1 and D_2 become equal to the Carbo and Hodgkin–Richards indices at some particular k and x values. The similarity measures listed above are usually called the quantum-chemical ones, because the functions f_i (charge density, electrostatic potential, *etc.*) are calculated by quantum-chemical methods.[133, 135, 142, 144–171]

It was suggested [172] to consider three- or four-atomic molecules similar if they have equal numbers of electrons. It was found that in this case the values of some physicochemical parameters (for example, the atomisation heats) of similar molecules are comparable. However, for other properties (for instance, for the ionisation potentials) this similarity measure does not work.

Various modifications of the above-mentioned quantum-chemical similarity measures have been proposed.[173–175] These give rise to new, nonsymmetrical measures d'_{ij} for some sets of compounds. For example, so-called 'stochastic transformation' d'_{ij} of the similarity measure d_{ij} is suggested:

$$d'_{ij} = d_{ij}(d_{i1} + d_{i2} + \ldots + d_{iN})^{-1},$$

where N is the number of compounds in a set. The quantity d'_{ii} called the self-similarity measure, is regarded as a molecular descriptor and is used for the modelling in QSAR/QSPR analysis.[176]

6. Let us assume that the structures S_1 and S_2 are represented by matrices $A = (a_{ij})$ and $B = (b_{ij})$ of the same order k. In this case the measures $d_{22}(A,B)$, $d_{23}(A,B)$, $d_{24}(A,B)$ may be used for the comparison of the structures S_1 and S_2.[9, 177, 178]

Let us define the measure $d_{22}(A,B)$ as follows:

$$d_{22}(A,B) = \frac{1}{\sqrt{k(k-1)}}\left(\sum (a_{ij} - b_{ij})^2\right)^{1/2}.$$

Obviously, the $d_{22}(A,B)$ depends on the atomic numbering scheme in the structures S_1 and S_2. However, it is possible to design a numbering independent similarity measure $d_{23}(A,B)$. Keeping the numbering of atoms in the structure S_1 fixed, all possible numbering schemes of the structure S_2 are considered and all possible values of $d_{22}(A,B)$ are calculated. The minimum obtained value is regarded as a similarity measure $d_{23}(A,B)$ of the structures S_1 and S_2.[9, 177]

The measure $d_{24}(A,B)$ is often used to compare isomers described in terms of so-called *be*-matrices.[73] This measure is called 'the minimal chemical distance' and is determined by the following formula [178]

$$\mathrm{MCD}(A,B) = d_{24}(A,B) = \min\sum_{i,j} |a_{ij} - b_{p(i)p(j)}|.$$

Here the summation takes place over the indices i and j; p are all possible combinations of n symbols, where n is the matrix order. Some combinatorial problems appear during the calculation of the MCD(A,B) value. Methods for tackling these problems have been proposed.[178]

A method of estimation of molecular similarity based on the matrices of the geometrical distances has been described.[179, 180]

7. Let us discuss the similarity measures for the molecules represented by three-dimensional geometrical objects Γ_1 and Γ_2.

In order to calculate these measures, the 3D molecular structures are spatially superimposed and a function of several variables is minimised. This function depends on the mutual orientation of the molecules. For instance, the following function may be used

$$R(\Gamma_1,\Gamma_2) = \left[\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{3}\left(x_{ij}^{(1)} - x_{ij}^{(2)}\right)^2\right]^{1/2},$$

where N is a number of atoms in the structures Γ_1 and Γ_2, $x_{ij}^{(1)}$ [or $x_{ij}^{(2)}$] are the coordinates of atoms in Γ_1 (or Γ_2). The minimum $R(\Gamma_1,\Gamma_2)$ of all possible rotations and translations of the structure Γ_2 in the space R^3 (with a fixed position of the structure Γ_1) may be regarded as a similarity measure $d_{25}(\Gamma_1,\Gamma_2)$ of the structures Γ_1 and Γ_2.

The following function may be used to evaluate the similarity of the structures Γ_1 and Γ_2 instead of the function $R(\Gamma_1,\Gamma_2)$

$$R'(\Gamma_1,\Gamma_2) = \left[\frac{1}{N}\sum_{i=1}^{N} m_i \sum_{j=1}^{3}\left(x_{ij}^{(1)} - x_{ij}^{(2)}\right)^2\right]^{1/2},$$

where m_i is the mass of the atom i. As in the previous case, the minimum of this function, determined by relative positions of the structures Γ_1 and Γ_2, is calculated. The value of this minimum is regarded as a similarity measure $d_{26}(\Gamma_1,\Gamma_2)$ of the structures Γ_1 and Γ_2.

In order to determine the Meyer's similarity index $S(\Gamma_1,\Gamma_2)$ (Ref. 181), a three-dimensional lattice with relatively small distances between the nodes (for example, 0.02 nm) is constructed in the space R^3. The geometrical body formed by a system of atom-centred van der Waals spheres is regarded as a model of the molecule. Molecules Γ_1 and Γ_2 are placed in this three-dimensional latticed space and the numbers of the nodes $T(\Gamma_1)$, $T(\Gamma_2)$ and $T(\Gamma_1,\Gamma_2)$ are counted. These are the numbers of the nodes in the van der Waals areas of each molecule and in their overlapping region, respectively. Then the Meyer's index is calculated according to the following formula [181]

$$0 \leqslant S(\Gamma_1,\Gamma_2) = \frac{T(\Gamma_1,\Gamma_2)}{\left[T(\Gamma_1)T(\Gamma_2)\right]^{1/2}} \leqslant 1.$$

The superimposion is performed to ensure the highest possible value of $S(\Gamma_1,\Gamma_2)$.

The value of $d(\Gamma_1,\Gamma_2)$, calculated according to the following formula, is used as a similarity measure of two molecules

$$d(\Gamma_1,\Gamma_2) = \frac{V_{\Gamma_1\Gamma_2}}{V_{\Gamma_2}},$$

where V_{Γ_2} is the volume of the molecule Γ_2; $V_{\Gamma_1\Gamma_2}$ is the volume of the overlapping area of the molecules Γ_1 and Γ_2.[182] It should be noted that this measure is better to be used when the position of one of the molecules is fixed.

The elementary operations required to transform structure Γ_1 into structure Γ_2 during estimation of the molecular similarity are documented.[183] These operations include the addition and removal of atoms, the change of atomic labels, *etc*. The special measure of similarity of the structures Γ_1 and Γ_2, which depends on the minimum number of operations for the transformation of one molecule into another, is introduced. The review of the methods for superimposition of three-dimensional structures is given. Another method of optimal superimposition of three-dimensional molecular structures was suggested in a study,[184] where the Carbo and Meyer indices were used for the molecular similarity measuring. The similarity measures based on the spatial superimposing of the molecules were also used [185] for the construction of 3D QSAR models.

The biological activity of molecules is closely related to their three-dimensional structure. As a result, the increased attention is attracted nowadays to the analysis of similarity of such structures.[41, 42, 46, 186–202] The correct choice of the biologically active conformation of the molecule is the main problem of 3D analysis of similarity. Usually either the conformation obtained by X-ray structural study of the complexes of the ligand with the receptor or the conformation calculated on the basis of two-dimensional structure by a special converter program is taken. It was supposed that the biologically active conformation of the molecule corresponds to a global minimum of its potential energy. However, it was shown that it is not the case.[69] Therefore, the consideration of only one main conformation in the analysis of similarity of the compounds may produce erroneous results.

Most of the 3D databases contain just one conformation for each compound. Therefore the 3D methods of similarity analysis probably will not provide a significant advantage over the 2D ones without taking into account a conformational flexibility of the molecule. It requires either to generate and keep a set of conformations for each molecule or to perform conformational calculations in the process of similarity analysis, which significantly slows it down.[69]

Superimposing (or overlaying) of two molecules in the similarity analysis is a limiting step, which restricts the application of 3D methods for the analysis of large databases. The superimposition procedure may be speeded up in several ways.[202, 203] A 4D similarity measure, which depends on the conformation of the molecule, has been considered.[204] The protein similarity measure [205] is based upon their 2D representation.

Alongside with 2D, 3D, 4D representations, 1D representation is also in use.[206] In this case the molecule is represented by a chain graph with labelled vertices, which correspond to the atoms of the molecule. The labels of the vertices include the atomic type, hybridisation and bonding of the corresponding atoms. The methods of the comparison of these chain graphs are identical to the methods for comparison of symbol sequences.

A comparative analysis of 2D and 3D methods was performed during the classification of four sets of compounds into the active and inactive ones.[207] The possibility of application of 2D and 3D methods for the estimation of similarity of various experimentally measured and calculated physicochemical properties related to such interactions as hydrophobic, dispersion, electrostatic, steric and donor–acceptor ones has been discussed.[186] The best property predictions were made on the basis of 2D fragment descriptors.

In a study,[208] various sets of 2D and 3D descriptors have been used for the classification of compounds according to their biological activity. It appeared that only 2D fragments provide effective sorting of the compounds to the active and inactive ones and, therefore, they are preferable for the analysis of similarity of chemical compounds.

The results of comparison of the 2D method of atomic pairs and its 3D modifications in the search for active analogues of morphine, cyclopamine, diazepam, apomorphine, serotonin, haboxadol, diethylstilbestrol in the database of 30 000 compounds are reported.[50] It is shown that the three-dimensional approach does not surpass the 2D method in the number of obtained biologically active analogues. At the same time, some active compounds, topologically different from the corresponding standard compound, were found only by 3D methods. A comparison of 2D and 3D descriptors for identification of active and inactive compounds was also carried out in other studies;[209] both methods provide similar results. The comparison of the similarity measure defined by the Tanimoto coefficient and the measures based on the superposition of the structures has also demonstrated [210] that the efficiency of 2D and 3D methods is approximately the same.

The method of superposition of the structures has performed better in the search for analogues in the non-homogeneous 3D databases, while for more homogeneous databases the best results were obtained by the maximum common subgraph method.[68]

A comparison of several similarity measures during the search for compounds in databases was performed;[211–216] methods for determination of molecular resemblance based on 22 similarity measures were suggested.[216]

The problem of analytical description of an arbitrary similarity measure specified for some finite set of the labelled graphs was discussed.[217,218] The general analytical formula depending on two matrices A_1 and A_2 was obtained. These results imply that in the prediction of properties of chemical compounds by methods based upon a similarity measure the best matrices A_1 and A_2 have to be obtained. In order to do that, the measure should be selected by optimal way from the sample subset of the compounds. This method of measure selection is a universal one and may be programmed (in contrast to the usual approaches where the measure is often 'guessed'). The method was applied [219] for the prediction of lipophilicity ($\log P$) of compounds, where P is a distribution coefficient in the n-octanol–water system. The advantages of this method were discussed.[13,220]

Thus, there is a number of various molecular similarity measures. The new measures may be constructed on their base. The review by Sheridan and Kearsley [221] discusses the reasons for invention of such a variety of measures and their application in theoretical and computer chemistry.

In the conclusion of this Section it has to be noted that quantitative similarity measures used in the theoretical chemistry are commonly divided in the 'similarity measures' and 'dissimilarity measures'. In the former case, the higher is the similarity (intuitive) of the molecules, the greater is the value of the corresponding similarity measure. In the latter case, the situation is opposite. By inversing the similarity measure $d(M_1,M_2)$ of the molecules M_1 and M_2 and calculating the value of $1-d(M_1,M_2)$, the dissimilarity measure of these molecules may be obtained. A new molecular descriptor of the molecule M may be constructed on the basis of any similarity measure $d(M_1,M_2)$: $d(M) = d(M,M_0)$, where M_0 is a fixed molecule.[9] Various methods of determination of the molecular 'diversity' are discussed, for example, in the papers.[12,222–234]

V. Methods for the prediction of properties of chemical compounds based on the molecular similarity concept

1. Method of *k* nearest neighbours

Let us take a set of N chemical compounds $M_1, \ldots, M_N$ with known values $y_1, \ldots, y_N$ of a chemical property. Then, according to the method of k Nearest Neighbours (kNN),[24,69,235,236] the property y_0 of a compound M_0 may be calculated as follows:

$$y_0 = \frac{1}{k}\sum W(M_j)y_j,$$

where $W(M_j) = 1$ if M_j is among the nearest k neighbours and $W(M_j) = 0$ in the opposite case. Generally, $W(M_j)$ may be replaced by any other weighted functions, particularly a similarity measure $d(M,M_j)$ may be used instead. The advantages and limitations of the kNN method and some of its modifications have been discussed.[235,236] The correct choice of k is a key problem of this method. For example, the optimal k was selected out of the series of values 1, 2, 3, 4, 5 in the prediction of lipophilciity;[23] the $k = 1$ was used for the estimation of boiling points and mutagenicity;[25,69] the $k = 1, \ldots, 10$; 15, 20, 25 were also applied.[24] The measure used for identification of the 'nearest neighbours' also affects the results of the kNN method. The composition of a set is also very important (there is a possibility that there are no 'good' neighbours in a set).

2. The estimation of a property on the basis of a system of equations

Zefirov and co-workers [219] suggested a method based on the optimal selection of a similarity measure $d(M_i,M_j)$ by the system of equations of the following type

$$|y_i - y_j| \approx d(M_i,M_j).$$

It was suggested to determine the property y_0 of a new structure M_0 on the basis of a system of equations in order to avoid the problem of selecting a correct number of neighbours k in kNN method:

$$|y_0 - y_i| \approx d(M_0,M_i),$$

where $\{M_i\}$ is a training set. In fact, this method is based on the assumption that the obtained model 'similar structures–similar properties' is also valid for the new compound M_0.

3. The methods of pattern recognition theory

The background of the theory of pattern recognition (TPR) [62,76,85,237–239] are as follows. There is a set of objects, each of them belongs to one of the classes (patterns). The main task of TPR is to produce the rules (usually written as inequalities) according to which the studied objects may be assigned to corresponding classes. If the class of an object is *a priori* known, then the problem is the learning strategy of pattern recognition. In the absence of the *a priori* knowledge the system has to train itself to recognise the classes. The quality of the rules is determined by the number of errors in the classification of a training set.

Let us discuss the geometrical interpretation of the TPR. In a number of cases the chemical objects are described as vectors $\boldsymbol{X} = (x_1, \ldots, x_n)$, *i.e.*, any object is represented by a point in the Euclidean space R^n. The Euclidean distance $d(\boldsymbol{X}_1,\boldsymbol{X}_2) = [\sum(x_{1i}-x_{2i})^2]^{1/2}$ may be regarded as a measure of proximity of these points. If the choice of the recognition criteria is adequate, then the points of the same class occupy a compact area in the space R^n (so-called cluster) and the clusters of different classes do not intersect each other. The term 'cluster' is a fundamental one for the self-training procedure. It may be determined as a set of points {A}, where any point is closer to the point of the same set

rather than to a point that does not belong to {A}. In the simplest case, the recognition rule may be as follows: an object described by vector $\boldsymbol{X}$ belongs to the class that is located at the shortest distance from the object.

The TRP may also be described in the probability terms. In this case the selection rules are described in terms of the probability theory.

The theory of pattern recognition is widely used in the prediction of biological activity of the compounds. In the simplest case, the compounds are sorted into two classes: active and inactive ones. The vectors based on the descriptors are used for the description of the compounds.

The linear regression method, used for the quantitative description of properties, may be regarded as a specific case of TPR. Indeed, the range of the property values may be split into the non-intersecting intervals and the chemical compounds may be assigned to the corresponding classes. The regression equation determines the selection rule for the classification of the compounds of a training set. However, the method of linear regression is not usually regarded as one of the methods of molecular similarity in the prediction of properties of chemical compounds.

4. The methods of quantitative correlations between the similarity of compounds and their biological activity or property (QSiAR/QSiPR analysis)

The following type of correlations is established in these methods:

$$y_j = \sum a_i S(i,j) + a_0 ,$$

where $S(i,j)$ is a similarity coefficient between the compounds i and j, a_i, a_0 are regression coefficients. The principal component analysis or partial least squares method are used for the determination of significant variables in the correlation equation (*i.e.*, for determination of i values). Examples of applications of the QSiAR/QSiPR analysis can be found in the literature.[29, 144, 240–243]

VI. Formalisation of the postulate 'similar structures possess similar properties' and its analysis

As it was mentioned above, the application of a concept of molecular similarity for the prediction of properties of chemical compounds is based on the statement 'similar structures possess similar properties' (SSPSP), which is accepted as an axiom, without being proved.

Naturally, the questions are: what is the correct interpretation of this postulate, how well is it justified, how to formalise it?

The hypothesis that similar structures possess similar properties originated from numerous experimental correlations between the properties of chemical compounds and the presence and number of some structural fragments in their molecules. Two structures are called similar if they have several common fragments, which mainly determine the value of a property, and differ by the fragments, which do not influence the value of a property. In this case one can expect that really 'similar' structures will possess 'similar' properties.

Nowadays very abstractive and complicated similarity measures, not directly related to the studied properties (see chapter IV), are often used for the estimation of molecular similarity degree. In spite of this fact, the SSPSP statement is still applied; however, its interpretation and validity in these cases is not entirely clear.

Johnson [235] reports the results of an empirical study of the validity of SSPSP postulate for various particular cases. Sixteen properties (both physicochemical properties and biological activity types), 16 databases and 72 quantitative similarity measures based on the vector representation of the molecular structures were investigated. The kNN method was used for the calculation of properties. It was shown that the measures based on the correlation coefficients are the most successful ones. A vector whose components are equal to the number of appearances of some structural fragments in the molecule is the most preferable one for the description of the molecular structure.

Thus, the validity of SSPSP statement significantly depends upon the choice of the chemical descriptors, the similarity measures, the class of the chemical compounds, the studied properties and the method of property estimation.[235] As the application of this postulate for the resolving of the particular problems of theoretical organic chemistry may turn out to be not always correct, the success of predictions of the properties of chemical compounds by the abovementioned methods is, generally, accidental. Therefore, a theoretical confirmation of the validity of SSPSP postulate is required.

In order to analyse the postulate, let us consider the most general case when the chemical structures $\{M_i\}$, $i = 1, \ldots, N$, are represented by some mathematical objects $\{S_i\}$, $i = 1, \ldots, N$, and for the pairs S_i and S_j of these objects a symmetrical function $d(S_i,S_j) \geqslant 0$ [such, that $d(S_i,S_j) = 0$ only when $S_i = S_j$], the similarity measure of the corresponding structures M_i and M_j, is specified. Besides, let us assume that the numbers $\{y_i\}$, *i.e.*, the values of some property of the structures $\{M_i\}$, $i = 1, \ldots, N$, are also known. It is reasonable to propose, as a similarity measure, a quantity $|y_i - y_j|$. Let us assume that the numerical criteria of similarity of properties and structures are *a priori* known. The criteria are the numbers $\varepsilon \geqslant 0$ and $\delta \geqslant 0$, such that if $d(S_i,S_j) \leqslant \delta$, then the structures are regarded as similar and if $|y_i - y_j| \leqslant \varepsilon$, then the properties are also regarded as similar. Obviously, the number ε is always selected by a researcher according to the particular problem to be solved and $d(S_i,S_j)$ and δ are variable characteristics. Then the postulate SSPSP may be formulated as follows: if for any structures M_i and M_j $d(S_i,S_j) \leqslant \delta$, then $|y_i - y_j| \leqslant \varepsilon$.

For the analysis of the validity of the postulate, let us make the following simple statement: for any set of structures $\{M_i\}$ ($i = 1, \ldots, N$), any property y and any similarity measure $d(S_i,S_j) \geqslant 0$ [$d(S_i,S_j) = 0$ only in the case where $S_i = S_j$] the following inequalities are valid

$$ad(S_i,S_j) \leqslant |y_i - y_j| \leqslant bd(S_i,S_j),$$

where a and b are constants depending on the measure, properties and the set of structures, namely

$$a = \min_{i,j,\, i \neq j} \frac{|y_i - y_j|}{d(S_i, S_j)} \geqslant 0,$$

$$b = \max_{i,j,\, i \neq j} \frac{|y_i - y_j|}{d(S_i, S_j)} \geqslant 0.$$

Several conclusions may be drawn from this statement.

1. The postulate will always be valid (in the abovementioned form) if $\delta = \varepsilon/b$ is selected.

2. Let us assume that for a given set, not all the values of the properties are close to each other, *i.e.*, that there is a pair S_i and S_j, such that $|y_i - y_j| > \varepsilon$. Therefore, if $\delta = \max_{i,j} d(S_i, S_j)$ is selected, then the postulate will not be valid for the given set.

3. Earlier, we assumed ε to be specified. However, if ε is selected to be very large, then the postulate will be valid for any δ and $d(S_i,S_j)$.

4. From the obtained double inequalities, it follows that the lower is the value of $d(S_i,S_j)$, the smaller is the quantity of $|y_i - y_j|$, thus for the 'very similar structures' the properties are also 'very similar'. This qualitative conclusion, drawn from the rigorous mathematical argumentation, is, in fact, a proof for the non-formal postulate typically used in theoretical chemistry.

Thus, we have shown that the validity of the postulate depends on the choice of ε and δ values if the measure $d(S_i,S_j)$ is fixed. However, the similarity measure of the properties is very seldom selected during the practical application of the molecular similarity concept. Therefore, without this criterion, it is very difficult to

judge if the structures really possess similar properties. Usually the experimental and calculated properties are listed in the QSAR/QSPR papers; however, the allowed error of prediction is never given. Therefore, a careful reader have to judge for himself the quality of methods applied.

An alternative method of selection of the measure $d(S_i, S_j)$ has been suggested.[219] This measure is selected using a set of structures on the basis of the conditions $d(S_i, S_j) \approx |y_i - y_j|$ at all i and j. In this case, the measure is represented by a special equation, depending on the variable square symmetrical matrix A. This equation is a scalar product of vectors AX and X, where X is a vector composed from certain descriptors. This method provides the validity of the SSPSP postulate for a given set with any values of ε and $\delta = \varepsilon$ and seems to be more efficient than 'guessing' of the measure.

Let us discuss the applicability of SSPSP postulate for the prediction of properties of the compounds that do not belong to the initial set. If this postulate is valid for the initial set with some δ, then for another set it cannot be valid with the same value of δ. Thus, it is obvious that the application of the postulate for estimating the properties of other compounds may be erroneous.

A 'training set' of compounds with known properties is always required for the estimation of the δ value. However, the standard regression equation may be written if the set is present and the problem of property prediction may be solved by other methods. Both these methods (regression analysis and similarity method) require the selection of the way of mathematical description of the molecular structure, the molecular descriptors, the type of the dependence between the descriptors and property or the similarity measure, and also the training set. The possibility of the analysis in the case of a non-linear structure–property correlation or even without a correlation at all is an advantage of the similarity method (the regression analysis usually uses the linear models). Moreover, the similarity method looks more attractive (in comparison with the regression analysis) because of the illusion that it does not requires complicated computation and a training set. Measures that compare the structures as a whole and do not require any descriptors are also possible. For example, the maximum common subgraph and its characteristics (the number of vertices, edges, *etc.*) may be found. However, such simplification of the similarity method results in arbitrary character of the predictions.

VII. Examples of application of the molecular similarity concept

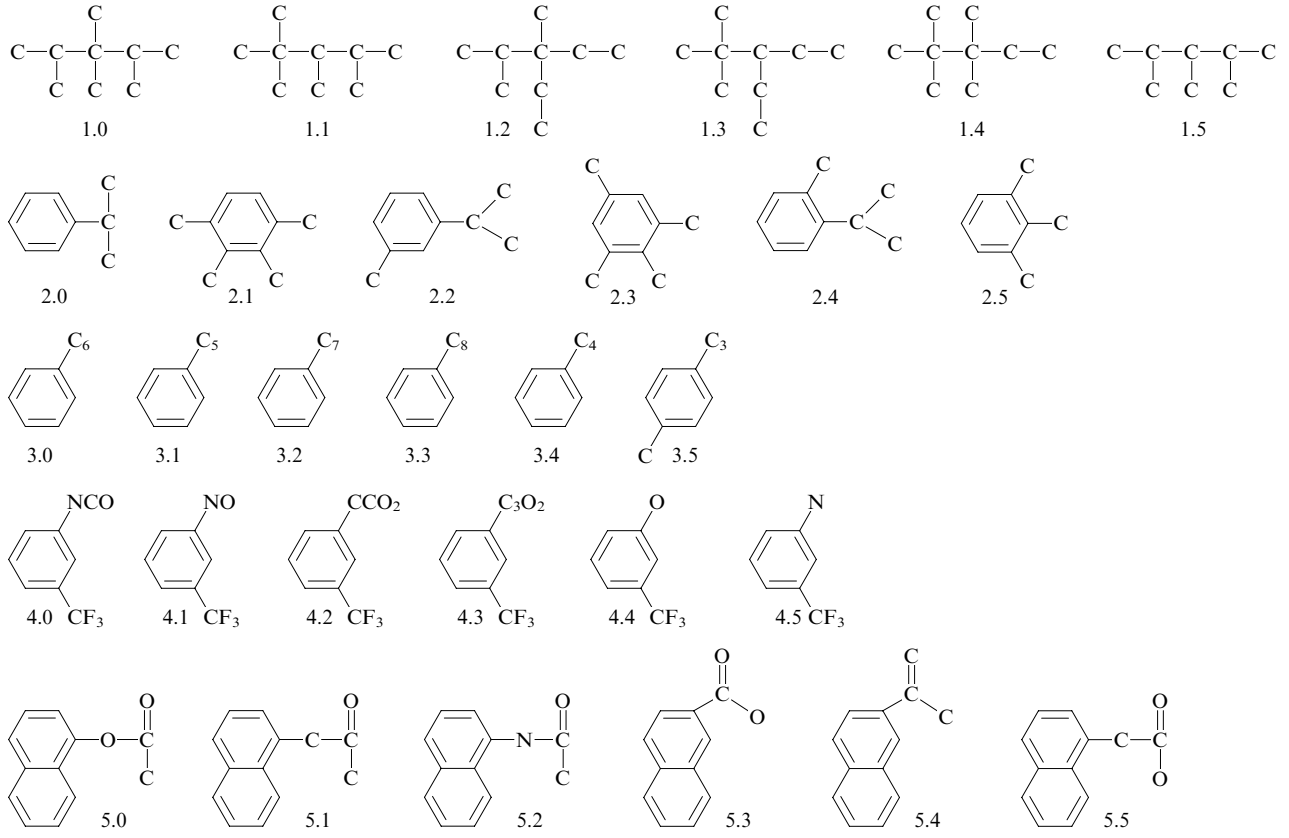
1. Determination of the structural similarity in terms of topological indices

The values of 90 topological indices (including the Randić index, the Wiener index, the series of informational indices, connectivity indices, *etc.*) have been calculated [243] for 3692 organic compounds of various classes. Ninety-dimensional vectors were composed for each of the studied structures on the basis of these TI. Ten new TI $X_1, \ldots, X_{10}$, which are linear combinations of the initial ones, were constructed as well, because some of the used indices were correlated by linear equations. Let us assume that $\boldsymbol{X}^{(i)} = (X_1^{(i)}, \ldots, X_{10}^{(i)})$ $(i = 1, 2)$ are two such vectors corresponding to a pair of structures S_1 and S_2. The similarity measure d of the structures S_1 and S_2 was determined [244] as the Euclidean distance between two points in the space R^{10}

$$d[\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}] = \left\{ \sum \left[X_i^{(1)} - X_i^{(2)} \right]^2 \right\}^{1/2}$$

and the applicability of this measure in the search for structural analogues of some compounds in a specific database was studied. Five compounds were randomly selected and for each of them five nearest {according to the measure $d[\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}]$} neighbours were found. The results of this procedure are shown in Scheme 1; indices 1.0, 2.0, 3.0, 4.0, 5.0 correspond to the selected compounds and indices of the i,j $(i, j = 1, \ldots, 5)$ type, to the found nearest

Scheme 1



'neighbours'. The compounds are listed according to increasing values of the measure $d[\boldsymbol{X}^{(1)},\boldsymbol{X}^{(2)}]$. The compounds found are in agreement with the intuitive chemical ideas about the structural similarity of the molecules. In order to find 'good' compounds in a database, it is necessary, at least, to have them in the database.[244]

2. The dependence of the result of clusterisation of a set of compounds on the method of chemical structure description

The correlation between the toxic activity for mice and the structure has been studied[245] for 25 ethers R^1-O-R^2 (R^1 = Me, Et, Pr^n, Pr^i, $CH_2{=}CH$; R^2 = Me, Et, Pr^n, Pr^i, Bu and other). The electrotopic descriptor ε (Ref. 245) and valence-weighted connectivity index ${}^m\chi^v$ at $m = 1$ were used for the classification of compounds.

Clusters in the descriptor–property planes for ε and χ^v are shown in Fig. 1. Obviously, the quality of clusterisation depends on the selection of a descriptor and ε works better than χ^v.

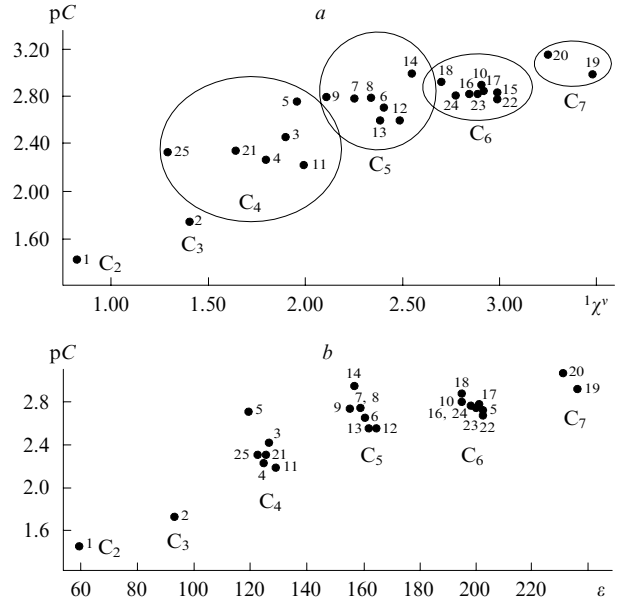
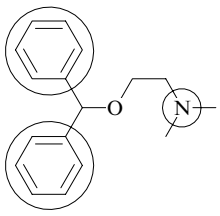


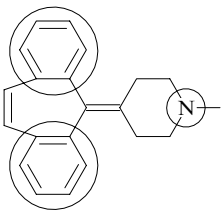
Figure 1. The clusterisation of the set of 25 esters, containing from 2 to 7 carbon atoms, on the planes descriptor $^1\chi^v$–property (*a*) and descriptor ε–property (*b*);[245]
p*C* is a toxicity.

3. The search for similar structural fragments of compounds exhibiting similar properties

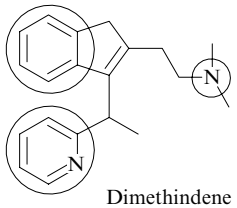
It appeared[246] that the structural similarity of antihistamine agents (diphenhydramine, methapyrilene, cyproheptadine, dimethindene, promethazine) may be represented by a topological triangle of two aromatic rings and the tertiary nitrogen atom.



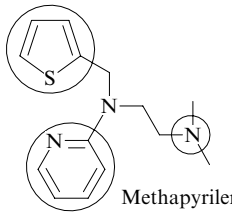
Diphenhydramine



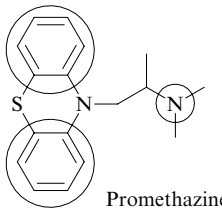
Cyproheptadine



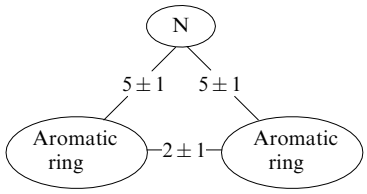
Dimethindene



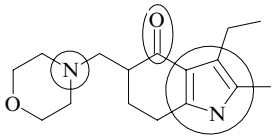
Methapyrilene



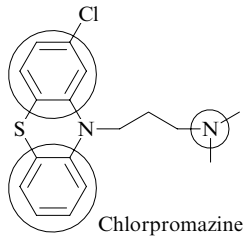
Promethazine



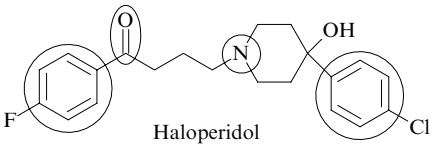
A common 'triangle' with some functional groups in the vertices has also been found for antipsychotic agents (molindone, chlorpromazine, haloperidol, pimozide, chlorprothixene, sulpiride).



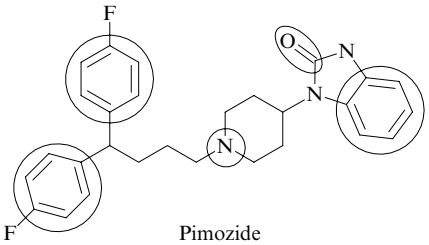
Molindone



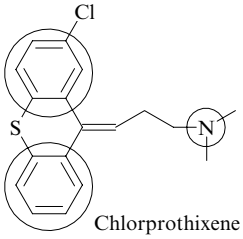
Chlorpromazine



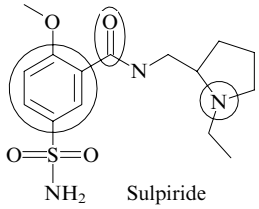
Haloperidol



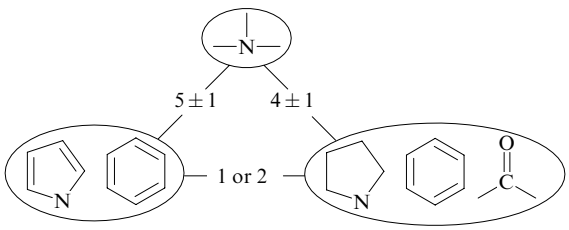
Pimozide



Chlorprothixene



Sulpiride



4. Dependence of the result of property prediction by the *k*NN method on the choice of the similarity measure and number *k* of nearest neighbours

Two molecular similarity measures have been studied[23] for the estimation of lipophilicity (log*P*) of 76 compounds of various classes. One of them was determined according to the following formula

$$d' = \frac{2C}{T_1 + T_2},$$

where T_1, T_2 are the numbers of fragments that are atomic pairs in the first and second structures, respectively; C is the number of atomic pairs shared by these two molecules (see Sections III and IV). The second similarity measure was determined according to the following formula

$$d'' = \left\{ \sum \left[X_i^{(1)} - X_i^{(2)} \right]^2 \right\}^{1/2},$$

where $\boldsymbol{X}^{(j)} = [X_1^{(j)}, \ldots, X_9^{(j)}]$, $j = 1, 2$ are some vectors from the Euclidean space R^9 obtained by the main component analysis from 90 various TI (see part 1 of this Section).

Five nearest neighbours of each compound of the set have been found according to the measures d' and d'' and the values of log*P* have been calculated by *k*NN method in five variants (at $k = 1, 2, 3, 4, 5$). Thus, ten various estimations of the log*P* were obtained. The correlation coefficients r between the observed and calculated values of log*P* are listed in Table 3. The best results were obtained when the measure d' with $k = 4$ ($r = 0.874$) was used. According to a publication,[23] both methods provide good results.

Table 3. Correlation coefficients between experimental and theoretical log*P* values calculated by ten methods.[23]

k	Correlation coefficient (r)	
	for d'	for d''
1	0.774	0.788
2	0.854	0.821
3	0.869	0.814
4	0.874	0.846
5	0.852	0.845

Another method of the similarity analysis based on the optimal selection of similarity measure d, has been applied for the same data.[219] A set of structural fragments has been selected for the description of the structures. It was established that according to some criteria, the obtained model is more precise than that used in the other publication.[23] Even more accurate model may be obtained by varying the set of fragments, however, no such studies were performed in the discussed paper.

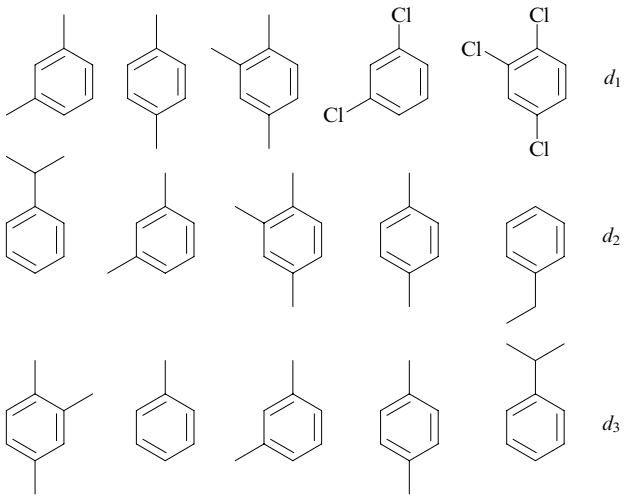
5. The selection of structural analogues by various molecular similarity measures

Three molecular similarity measures d_1, d_2 and d_3 have been examined[43] and five structural analogues of *o*-xylene have been found according to these measures. Seven physicochemical properties (including molar volume, log*P*, boiling point, *etc.*) were used for the description of molecules during the composing of measure d_1. On the basis of the principal component analysis of these seven parameters, five new parameters were composed. These new parameters are linear combinations of the original ones. By coding the structures with the corresponding five-dimensional vectors, the measure d_1 was determined as a standard Euclidean metric in the space R^5.

The measure d_2 is similar to d_1 and is based on 81 TI. Ten new descriptors were obtained from these TI by the principal component analysis. Therefore, d_2 is defined as an Euclidean metric in the space R^{10}.

The d_3 measure is based on the atomic pairs and has been discussed earlier.

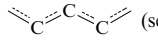
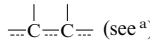
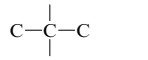
The structural formulae of the analogues of *o*-xylene selected with the help of the discussed similarity measures are listed below (the structures are arranged according to the decrease in the measure values). The result of the search for 'similar' structures depends on the measure selection. However, there are examples of coincidence between the structures found by using different similarity measures.



6. Revealing of the similarity of musk compounds by the pattern recognition theory method

A set of 300 compounds has been studied.[62] Among these compounds 60 are musk odorants and 240 possess another fragrance (camphoric, floral, ester, mint, *etc.*). Also, the musk odorants belonged to various structural types: there were 23 macrocyclic compounds, 19 polynitrobenzenes, 11 steroids, 5 γ-butirolactones and two compounds of other types. Initially 68 descriptors (geometrical and fragmental) were used, further 13 of the most important among them were selected by TPR for

Table 4. The main descriptors used for the classification of compounds into musk and non-musk odorants.[62]

Descriptor	Rate of correct classification
Number of O atoms	84.3
Number of double bonds	82.3
Number of aromatic bonds	80.0
Maximum eigenvalue of the inertia tensor	86.7
Minimum eigenvalue of the inertia tensor	80.0
Number of the fragments	
 (see [a])	80.0
 (see [a])	80.0
	86.0
	90.3
	80.7
	80.0
	80.0
	83.0

[a] ---- is an aromatic bond.

classification of these compounds into musk and non-musk odorants (Table 4). For testing of the predictive ability of the obtained descriptors, a test set of compounds has been used. This set contained 121 musk and 120 non-musk odorants not present in the initial set. As a result, 109 out of 121 musk odorants (90.1%) and 117 out of 120 non-musk odorants (97.5%) were recognised correctly.

The problem of correlation of the structure and electronic configuration with an ambergris fragrance has been studied.[247, 248] The authors have used a combined approach based upon a combination of pattern recognition methods and quantum-chemical calculations with the precise electron-topological analysis of the potentially active parts of molecules. It appeared that the ambergris fragrance is associated with the presence in the molecular structure of so-called ambergris triangles.

7. Comparison of the accuracy of different methods of similarity analysis in the prediction of compound properties

The comparison of the prediction accuracy for various properties of chemical compounds by six methods (abbreviated as PCs, TIu, TIs, AP, TEI, MNA) of similarity analysis has been carried out.[69, 88] The two initial sets consisted of 139 hydrocarbons with known boiling points and 15 nitrosamines with known values of mutagenicity. Three similarity methods (PCs, TIu, TIs) used the Euclidian distances as the similarity measures. In these cases various sets of the topological indices were used. These indices were obtained from 96 TI by various procedures, decreasing the dimensionality of the data.[25] The fourth method (AP) is based on the atomic pairs and the corresponding similarity measure (see Sections III and IV). Other methods (TEI and MNA) used the vectors composed from electrotopological indices and MNA descriptors based on the atom-centred fragments [see the measures $d_9(\boldsymbol{X},\boldsymbol{Y})$ and $d_5(\boldsymbol{X},\boldsymbol{Y})$ in Section IV and also the Section III]. The kNN method with $k = 1$ has been used for the estimation of properties.

The data of Tables 5 and 6 indicate that in the first case, the method TIs, which used the topological indices selected by the VARCLUS procedure, was the best one.[24] The informational indices and the connectivity indices of Kier and Hall of various orders were present in the selected set of ten TI (see Section III). This may be explained [69] by the fact that the boiling point mainly depends on the shape and the degree of branching of molecules. Both these characteristics are described well by topological indices. In the case of mutagenicity prediction (see Table 6), the MNA and AP methods based on the fragments performed better and the TEI method was at the third place. In this particular case, the TEI method is better than TI-based methods because the prediction of mutagenicity requires additional information about physicochemical characteristics of the molecules and exactly this information is contained in the electrotopological indices.[69]

Table 5. The comparison of efficiency of six methods of similarity analysis in the prediction of hydrocarbons boiling points.[69]

Method	r	s
TIs	0.993	19.34
AP	0.986	26.26
MNA	0.985	27.89
TIu	0.983	29.36
PCs	0.966	41.29
TEI	0.959	45.37

Note. Here and in the Table 6, r is a correlation coefficient, s is a mean square deviation.

Table 6. Comparison of the efficiency of six methods of similarity analysis in the prediction of mutagenicity of nitrozoamines.[69]

Method	r	s
MNA	0.945	1.25
AP	0.944	1.26
TEI	0.931	1.43
TIu	0.923	1.47
PCs	0.830	2.33
TIs	0.740	2.67

8. Construction of QSiAR models

The QSiAR model of affinity of fourteen *N*-substituted 3-(3-hydroxyphenyl)piperidines (Table 7) towards a σ-receptor (pIC_{50}) has been created.[104] The Meyer index S_{AB} (see Section IV) has been used as a measure of similarity of the structures A and B, where A is the initial (unsubstituted) compound, B are all other compounds of the set. The *R* and *S* enanthiomers of the corresponding compounds were also taken into account. Two models have been composed, each of them on the basis of seven sets of compounds:

for *R* isomers ($r^2 = 0.750$)

$$pIC_{50} = -0.04\,S_{AB} + 2.29,$$

for *S* isomers ($r^2 = 0.974$)

$$pIC_{50} = -10.21\,S_{AB} + 9.30.$$

Table 7. The values of affinity towards σ-receptor (pIC_{50}) and of the Meyer indices (S_{AB}) for the series of *N*-substituted 3-(3-hydroxyphenyl)-piperidines.[104]

Compound		S_{AB}	pIC_{50} (see [a])
substituent	configuration		
H	*R*	1.000	−0.66
	S	1.000	−1.19
Me	*R*	0.952	0.43
	S	0.950	−0.28
Et	*R*	0.917	0.95
	S	0.912	−0.01
Pr^n	*R*	0.881	1.52
	S	0.877	0.81
Pr^i	*R*	0.884	0.61
	S	0.877	0.68
Bu^n	*R*	0.849	2.05
	S	0.845	1.51
EtC_6H_4	*R*	0.792	2.10
	S	0.786	1.80

[a] IC_{50} is expressed in μmol litre^{-1}.

9. Study of the efficiency of the similarity analysis methods in the prediction of biological activity

The cross-validation method has been used for the estimation of efficiency of the similarity analysis methods in the prediction of biological activity.[69] One compound (numbered k) was excluded from the database at each step of the procedure, and the probability $P(k)$ of this compound to be active was estimated:

$$P(k) = \frac{A_k}{A_k + B_k},$$

$$A_k = (n-1)^{-1}\sum S_i^1,$$

$$B_k = (m-1)^{-1}\sum S_j^0,$$

where S_i^1 and S_j^0 are the values of similarity measure between the compound k and the active compound i and the inactive compound j, respectively; n and m are the numbers of active and inactive compounds in the training set, respectively. The summation in the expression for A_k takes place over all i values from 1 up to n (except for $i = k$). Similarly, the summation in the formula of B_k takes place over all j from 1 to m (except for $k = j$).

Let us assume that for each pair of compounds M_i and M_j, where M_i is an active compound and M_j is an inactive compound ($i = 1, \ldots, n$; $j = 1, \ldots, m$), the values of probabilities $P(i)$ and $P(j)$ are calculated, and N is the number of pairs, such that $P(i) > P(j)$. The quantity IAP $= N/mn$ ($0 \leqslant$ IAP $\leqslant 1$) was used for the estimation of the prediction quality. A set of 17124 compounds with known values of 16 types of biological activity were selected from the database MDDR 96.2 (MDL Information System) as a training set for similarity analysis methods. Two modifications of similarity methods (using the Tanimoto coefficient as a similarity measure), based upon the electrotopological indices (TEI) and on the substructural descriptors (MNA), have been studied. The average accuracy of the prediction in the first case is equal to 86%, and in the second case, to 96%.

10. The optimal choice of k value in the kNN method

The efficiency of the clusterisation of molecules of various chemical databases based upon the method of k nearest neighbours depends on the choice of the parameter k.

The boiling points of 267 haloalkanes containing from 1 to 4 carbon atoms have been studied.[249] The molecular similarity (dissimilarity) was estimated in terms of Euclidian metrics in eight-dimensional space of topological descriptors composed by the principal component analysis from a set of 59 TI. The correlation coefficients between the experimental and calculated values are in the range of 0.85 – 0.94. The chosen k values were equal to 1 – 10, 15, 20, 25. It was shown that the best standard deviation corresponds to $k = 5$ and equals to 23.7 °C. Similar results were obtained in other studies.[23, 250]

11. Comparison of the efficiency of 2D and 3D descriptors in the estimation of properties of compounds

Three new databases were created on the basis of Medchem 94b database:[186] (i) 8651 compounds with known $\log P_{\text{octanol/water}}$; (ii) 762 compounds with known $\log P_{\text{cyclohexane/water}}$; (iii) 8416 compounds with known pK_a. Both 2D and 3D descriptors (four types) were used for coding of the structures. Particularly, the vectors, components of which were equal to the numbers of occurrences of fixed small fragments in the structures (in one case 153 fragments called MACCS, in another case 960 fragments called SSKEYS) were used. In order to estimate the properties of compounds, the preliminary clusterisation of corresponding databases has been performed by various methods and using the similarity measures. The property was calculated as an average value of all the property values of the corresponding cluster. The cross-validation procedure was employed for the estimation of the prediction quality. For this purpose, the compounds were taken out of the database one by one, and then the value of a property for this compound was calculated. These experimental and calculated property values were compared and then the correlation coefficient and the root-mean-square deviation of corresponding regression were calculated. It was found that 2D descriptors MACCS performed better in all cases. Also, the comparison of the results of calculations of $\log P$ for 7054 structures (out of 8651) was carried out by the above-mentioned methods and by a special programme CLOGP. The methods of molecular similarity yielded the values of $r = 0.95$ and the root-mean-square deviation $rmse = 0.60$, which is almost as accurate as the results of the calculations by the well known program CLOGP ($r = 0.96$, $rmse = 0.49$).

VIII. Conclusion

The molecular similarity concept ('similar compounds possess similar properties') is very popular in theoretical and experimental chemistry. One of the areas of its application, discussed in the present review, is the prediction of properties of chemical compounds on the basis of the 'molecular similarity' concept. The widespread computerisation of the chemical studies requires the formalisation of this concept and the development of methods for quantitative description of the degree of molecular similarity. Nowadays there is a number of various similarity measures, which usually deal with the structures only and are not related to any physicochemical properties or biological activity of the molecules. However, the measures may be optimally tuned to a specific task.

An important problem is the formalisation and approval of the validity of concept 'similar compounds possess similar properties'. As it was shown, the validity of this statement depends on the class of compounds, on the studied property, on the selected similarity measure, *etc*. The problem of formalisation of the concept is discussed in the present review for the first time. It was found that under certain conditions, this postulate is always valid, and a qualitative proof of the unformalised postulate was given.

On the basis of the analysis of published data, one can draw a conclusion that the molecular similarity method used for the prediction of properties of chemical compounds quite often yields satisfactory results. In particular, one of the advantages of the method is that the *a priori* knowledge of the type of the structure – property functional dependence is not required. Besides, the whole structural formula (a molecular graph) is used for the determination of some similarity measures and there is no need to calculate any molecular descriptors. The discovery of fullerenes is an example of usefulness of the similarity concept in chemistry: these unusual allotropes of carbon were first obtained in 1985,[251] while their existence had been predicted approximately 15 years earlier [252, 253] on the basis of a hypothesis about similarity of electronic properties of planar polycondensed aromatic hydrocarbons and spherical polyhedral polycyclic conjugated molecules. This led to successful application to fullerenes of the mathematical models based on the resonance theory and on the Hückel method.

At the same time, it should be noted that the similarity method has some problems. Among them are the correct choice of the mathematical description of the structure, the selection of the similarity measure and the choice of the method for estimation of the properties of a compound on the basis of a similarity measure. Analysis of literature shows that there are no strict and approved methods for the solution of these problems and the selection is intuitive. Probably, here is the right place to quote a remark given by R Feynman in his Nobel lecture: 'I think equation guessing might be the best method to proceed to obtain the laws for the part of physics which is presently unknown'.[254]

References

1. S R Heller *J. Chem. Inf. Comput. Sci.* **32** 578 (1992)
2. D H Rouvray *J. Chem. Inf. Comput. Sci.* **34** 446 (1994)
3. C Jochum, M G Hicks, J Sunkel *J. Chem. Inf. Comput. Sci.* **32** 577 (1992)
4. R Carbo-Dorca, P G Mezey (Eds) *Advances in Molecular Similarity* Vol. 2 (London: JAI Press, 1998)
5. R Carbo (Ed.) *Molecular Similarity and Reactivity. From Quantum Chemical to Phenomenological Approaches* (Dordrecht: Kluwer Academic, 1995)
6. P M Dean (Ed.) *Molecular Similarity in Drug Design* (London: Blackie Academic and Professional, 1995)
7. M A Johnson, M Maggiora, M S Lajiness, J B Moon, J D Petke, D C Roher, in *Advanced Computer-Assisted Techniques in Drug Discovery* (Ed. H van de Waterbeemd) (New York: VCH, 1994) p. 89

8. H van de Waterbeemd, B Testa, G Folkers (Eds) *Computer-Assisted Lead Finding and Optimization* (Basel: VHCA, Weinheim: Wiley-VCH, 1997)
9. M A Johnson *J. Math. Chem.* **3** 117 (1989)
10. M A Johnson, G M Maggiora (Eds) *Concepts and Applications of Molecular Similarity* (New York: Wiley, 1990)
11. V Monev *MATCH — Commun. Math. Comput. Chem.* (51) 7 (2004)
12. Y C Martin *J. Combinat. Chem.* **3** 231 (2001)
13. N Nikolova, J Jaworska *QSAR Combinat. Sci.* **22** 1006 (2003)
14. M Randić, in *Mathematical Methods in Contemporary Chemistry* (Ed. S I Kuchanov) (Amsterdam: Gordon and Breach, 1996) p. 1
15. M Johnson, S Basak, G Maggiora *Math. Modell. Sci. Techn.* **11** 630 (1988)
16. P Willett *Similarity and Clustering in Chemical Information Systems* (Letchworth: Research Studies Press, 1987)
17. G Grethe, T E Moock *J. Chem. Inf. Comput. Sci.* **30** 511 (1990)
18. W A Warr *Persp. Drug Discov. Design* **7–8** 115 (1997)
19. O A Raevsky, S V Trepalin, H P Trepalina, V A Gerasimenko, O E Raevskaja *J. Chem. Inf. Comput. Sci.* **42** 540 (2002)
20. D Horvath, C Jeandenans *J. Chem. Inf. Comput. Sci.* **43** 680 (2003)
21. D Horvath, C Jeandenans *J. Chem. Inf. Comput. Sci.* **43** 691 (2003)
22. S C Basak, S Bertelsen, G D Grunwald *J. Chem. Inf. Comput. Sci.* **34** 270 (1994)
23. S C Basak, G D Grunwald *New J. Chem.* **19** 231 (1995)
24. S C Basak, G D Grunwald *J. Chem. Inf. Comput. Sci.* **35** 366 (1995)
25. S C Basak, G D Grunwald *SAR QSAR Environ. Res.* **2** 289 (1994)
26. A C Good, S J Peterson, W G Richards *J. Med. Chem.* **36** 2929 (1993)
27. A C Good, in *Molecular Similarity in Drug Design* (Ed. P M Dean) (London: Blackie Academic and Professional, 1995) p. 24
28. H Kubinyi, in *Computer-Assisted Lead Finding and Optimization* (Eds H van de Waterbeemd, B Testa, G Folkers) (Basel: VHCA, Weinheim: Wiley-VCH, 1997) p. 7
29. O Ivanciuc *Rev. Roum. Chim.* **46** 787 (2001)
30. B Bertoša, B Kojić-Prodić, R C Wade, M Ramek, S Piperaki, A Tsantili-Kakoulidou, S Tomić *J. Chem. Inf. Comput. Sci.* **43** 1532 (2003)
31. R D Cramer, R J Jilek, S Guessregen, S J Clark, B Wendt, R D Clark *J. Med. Chem.* **47** 6777 (2004)
32. R P Sheridan, B P Feuston, V N Maiorov, S K Kearsley *J. Chem. Inf. Comput. Sci.* **44** 1912 (2004)
33. D Pan, M Iyer, J Z Liu, Y Li, A J Hopfinger *J. Chem. Inf. Comput. Sci.* **44** 2083 (2004)
34. M G Hicks *J. Chem. Inf. Comput. Sci.* **32** 631 (1992)
35. J Gasteiger, W D Ihlenfeldt, R Fick, J R Rose *J. Chem. Inf. Comput. Sci.* **32** 700 (1992)
36. A J Lawson *J. Chem. Inf. Comput. Sci.* **32** 675 (1992)
37. R Ponec *J. Chem. Inf. Comput. Sci.* **36** 554 (1996)
38. R Ponec, M Strnad *J. Chem. Inf. Comput. Sci.* **32** 693 (1992)
39. R Ponec, M Strnad *Collect. Czech. Chem. Commun.* **55** 2583 (1990)
40. G Sello *J. Chem. Inf. Comput. Sci.* **32** 713 (1992)
41. P J Artymiuk, P A Bath, H M Grindley, C A Pepperrell, A R Poirrette, D W Rice, D A Thorner, D J Wild, P Willett, F H Allen, R Taylor *J. Chem. Inf. Comput. Sci.* **32** 617 (1992)
42. P A Bath, A R Poirrette, P Willett, F H Allen *J. Chem. Inf. Comput. Sci.* **34** 141 (1994)
43. S C Basak, G D Grunwald *Math. Modell. Sci. Comput.* **2** 735 (1993)
44. G M Downs, P Willett *Rev. Comput. Chem.* **7** 1 (1996)
45. W Fisanick, A H Lipcus, A Rusinco III *J. Chem. Inf. Comput. Sci.* **34** 130 (1994)
46. C M R Ginn, D B Turner, P Willett, A M Ferguson, T W Heritage *J. Chem. Inf. Comput. Sci.* **37** 23 (1997)
47. M S Lajiness, in *Structure–Property Correlations in Drug Research* (Ed. H van de Waterbeemd) (New York: Academic Press, 1996) p. 179
48. C A Pepperrell, A R Poirrette, P Willett, R Taylor *Pestic. Sci.* **33** 97 (1991)
49. N C Perry, V J van Geerestein *J. Chem. Inf. Comput. Sci.* **32** 607 (1992)
50. R P Sheridan, M D Miller, D J Underwood, S K Kearsley *J. Chem. Inf. Comput. Sci.* **36** 128 (1996)
51. D B Turner, P Willett, A M Ferguson, T W Heritage *SAR QSAR Environ. Res.* **3** 101 (1995)
52. J M Barnard, G M Downs *J. Chem. Inf. Comput. Sci.* **32** 644 (1992)
53. R D Clark *J. Chem. Inf. Comput. Sci.* **37** 1181 (1997)
54. G M Downs, P Willet, in *Applied Multivariate Analysis in SAR and Environmental Studies* (Eds J Devillers, W Karcher) (Dordrecht: Kluwer Academic, 1991) p. 247
55. J D Holliday, S S Ranadle, P Willett *Quantum Struct.–Act. Relat.* **14** 501 (1995)
56. M Johnson, M Lajiness, G Maggior, in *QSAR: Quantitative Structure–Activity Relationships in Drug Design* (Ed. J L Fauchere) (New York: Alan R. Liss Inc., 1989) p. 167
57. J W Godden, J R Furr, L Xue, F L Stahura, J Bajorath *J. Chem. Inf. Comput. Sci.* **44** 21 (2004)
58. L Xue, J W Godden, F L Stahura, J Bajorath *J. Chem. Inf. Comput. Sci.* **44** 1275 (2004)
59. L Xue, F L Stahura, J Bajorath *J. Chem. Inf. Comput. Sci.* **44** 2032 (2004)
60. V J Gillet, P Willett, J Bradshaw *J. Chem. Inf. Comput. Sci.* **37** 731 (1997)
61. P Willett *Persp. Drug Discov. Design* **7–8** 1 (1997)
62. A Stupper, W Brugger, P Jurs *Computer Assisted Studies of Chemical Structure and Biological Function* (New York: Wiley, 1979)
63. O A Raevsky *Usp. Khim.* **68** 555 (1999) [*Russ. Chem. Rev.* **68** 505 (1999)]
64. M I Stankevich, I V Stankevich, N S Zefirov *Usp. Khim.* **57** 337 (1988) [*Russ. Chem. Rev.* **57** 191 (1988)]
65. L H Hall, L B Kier, B B Brown *J. Chem. Inf. Comput. Sci.* **35** 1074 (1995)
66. D H Rouvray *J. Chem. Inf. Comput. Sci.* **32** 580 (1992)
67. R Hoffmann *Angew. Chem., Int. Ed. Engl.* **21** 711 (1982)
68. C Pepperrell, in *Computers and Chemical Structure Information Series* Vol. 3 (Ed. D Bawden) (Taunton, Somerset, England: Research Studies Press Ltd., 1994)
69. Yu V Borodina, Candidate Thesis in Biological Sciences, Research Institute of Biomedical Chemistry, Russian Academy of Medical Sciences, Moscow, 1998
70. S N Golikov *Yady i Protivoyadiya* (Toxins and Antidotes) (Moscow: Znanie, 1968)
71. F E Blaney, P Finn, R Phippen, M Wyatt *J. Mol. Graphics* **11** 98 (1993)
72. O Ivanciuc, S L Taraviras, D Cabrol-Bass *J. Chem. Inf. Comput. Sci.* **40** 126 (2000)
73. I Ugi, M Wochner, E Fontain, J Bauer, B Gruber, R Karl, in *Concepts and Applications of Molecular Similarity* (Eds M A Johnson, G M Maggiora) (New York: Wiley, 1990) p. 239
74. D H Rouvray, in *Chemical Applications of Topology and Graph Theory* (Ed. R King) (Amsterdam: Elsevier, 1983) p.159
75. J Devillers, A T Balaban (Eds) *Topological Indices and Related Descriptors in QSAR and QSPR* (Amsterdam: Gordon and Breach, 1999) p. 811
76. P Jurs, T Isenhour *Chemical Applications of Pattern Recognition* (New York: Wiley, 1975)
77. V R Magnuson, D K Harriss, S C Basak, in *Chemical Applications of Topology and Graph Theory* (Ed. R King) (Amsterdam: Elsevier, 1983) p. 178
78. I I Baskin, M I Skvortsova, I V Stankevich, N S Zefirov *J. Chem. Inf. Comput. Sci.* **35** 527 (1995)
79. M I Skvortsova, I I Baskin, L A Skvortsov, V A Palyulin, N S Zefirov, I V Stankevich *J. Mol. Struct. (THEOCHEM)* **466** 211 (1999)
80. N F Stepanov, M E Erlykina, T G Filippov *Metody Lineinoi Algebry v Fizicheskoi Khimii* (Methods of Linear Algebra in Physical Chemistry) (Moscow: Khimiya, 1976)
81. Yu G Papulov, V M Smoljakov, in *Mathematical Methods in Contemporary Chemistry* (Ed. S I Kuchanov) (Amsterdam: Gordon and Breach, 1996) p. 143
82. M I Skvortsova, K S Fedyaev, V A Palyulin, N S Zefirov *Izv. Akad. Nauk, Ser. Khim.* 1527 (2004)
83. P Willett *Curr. Opin. Biotechnol.* **11** 8588 (2000)

84. N S Zefirov, V A Palyulin *J. Chem. Inf. Comput. Sci.* **42** 1112 (2002)
85. A B Rozenblit, V E Golender *Logiko-kombinatornye Metody v Konstruirovanii Lekarstv* (Logical Combinatorial Methods in Drug Design) (Riga: Zinatne, 1983)
86. R E Carhart, D H Smith, R Venkataraghavan *J. Chem. Inf. Comput. Sci.* **25** 64 (1985)
87. S K Kearsley, S Sallamack, E M Fluder, J D Andose, R T Mosley, R P Sheridan *J. Chem. Inf. Comput. Sci.* **36** 118 (1996)
88. D Filimonov, V Poroikov, Yu Borodina, T Gloriozova *J. Chem. Inf. Comput. Sci.* **39** 666 (1999)
89. A Bender, H Y Mussa, R C Glen, S Reiling *J. Chem. Inf. Comput. Sci.* **44** 1708 (2004)
90. A Bender, H Y Mussa, R C Glen, S Reiling *J. Chem. Inf. Comput. Sci.* **44** 170 (2004)
91. P Willett, V Winterman *J. Chem. Inf. Comput. Sci.* **26** 36 (1986)
92. E A Smolenskii *Zh. Fiz. Khim.* **38** 1288 (1964) [a]
93. Yu V Borodina, D A Filimonov, V V Poroikov *Quantum Struct. – Act. Relat.* **17** 459 (1998)
94. R D Hull, S B Singh, R B Nachbar, R P Sheridan, S K Kearsley, E M Fluder *J. Med. Chem.* **44** 1177 (2001)
95. R D Hull, S B Singh, R B Nachbar, R P Sheridan, S K Kearsley, E M Fluder *J. Med. Chem.* **44** 1185 (2001)
96. M Karelson, V S Lobanov, A R Katritzky *Chem. Rev.* **96** 1027 (1996)
97. L Xue, J W Godden, F L Stahura, J Bajorath *J. Chem. Inf. Comput. Sci.* **43** 1218 (2003)
98. G M Downs, P Willett, W Fisanick *J. Chem. Inf. Comput. Sci.* **34** 1094 (1994)
99. W J Wisswesser *The Wisswesser Line-Formula Chemical Notation*. (New York: McGraw-Hill, 1968)
100. A Drefahl, M Reinhard *J. Chem. Inf. Comput. Sci.* **33** 886 (1993)
101. D Weiniger *J. Chem. Inf. Comput. Sci.* **30** 237 (1990)
102. S Ash, M A Cline, R W Homer, T Hurst, G B Smith *J. Chem. Inf. Comput. Sci.* **37** 71 (1997)
103. A Schuffenhauer, P Floersheim, P Acklin, E Jacoby *J. Chem. Inf. Comput. Sci.* **43** 391 (2003)
104. A Sery-Levy, S West, W G Richards *J. Med. Chem.* **37** 1727 (1994)
105. F Harary *Graph Theory* (Reading: Addison-Wesley, 1969)
106. Z Gabányi, P Surján, G Náray-Szabó *Eur. J. Med. Chem.* **17** 307 (1982)
107. L J Soltzberg, C L Wilkins *J. Am. Chem. Soc.* **99** 439 (1977)
108. J H Schuur, P Selzer, J Gasteiger *J. Chem. Inf. Comput. Sci.* **36** 334 (1996)
109. E E Hodgkin, W G Richards *J. Chem. Soc., Chem. Commun.* 1342 (1986)
110. P G Mezey *J. Chem. Inf. Comput. Sci.* **32** 650 (1992)
111. P G Mezey *J. Chem. Inf. Comput. Sci.* **34** 244 (1994)
112. P G Mezey *Shape in Chemistry* (New York: VCH, 1993)
113. P G Mezey, Z Zimpel, P Warburton, P D Walker, D G Irvine, S L Dixon, B Greenberg *J. Chem. Inf. Comput. Sci.* **36** 602 (1996)
114. P G Mezey, in *Encyclopedia of Computational Chemistry* Vol. 4 (Eds P R Schleyer, N L Allinger, T Clark, J Gastaiger, P R Kollman, H F Schaefer III, P R Schreiner) (Chichester: Wiley, 1999) p. 2582
115. P D Walker, P G Mezey *J. Am. Chem. Soc.* **116** 1202 (1994)
116. A Golbraikh *J. Chem. Inf. Comput. Sci.* **40** 414 (2000)
117. M Randić *J. Chem. Inf. Comput. Sci.* **36** 1092(1996)
118. M Randić *J. Chem. Inf. Comput. Sci.* **32** 686 (1992)
119. M Randić *J. Chem. Inf. Comput. Sci.* **35** 373 (1995)
120. W Zheng, A Tropsha *J. Chem. Inf. Comput. Sci.* **40** 185 (2000)
121. I Gutman, L Türker, I Bálint *Indian J. Chem., Sect. A* **34** 1 (1995)
122. X Chen, C H Reynolds *J. Chem. Inf. Comput. Sci.* **42** 1407 (2002)
123. D M Bayada, H Hamersma, V J van Geerestein *J. Chem. Inf. Comput. Sci.* **39** 1 (1999)
124. D R Flower *J. Chem. Inf. Comput. Sci.* **38** 379 (1998)
125. J W Godden, L Xul, J Bajorath *J. Chem. Inf. Comput. Sci.* **40** 163 (2000)
126. E Abrahamian, P C Fox, L Narum, I T Christensen, H Thogersen, R D Clark *J. Chem. Inf. Comput. Sci.* **43** 458 (2003)
127. J D Holliday, S P Jelfs, P Willett, P Gedeck *J. Chem. Inf. Comput. Sci.* **43** 406 (2003)
128. M Randić, M Razinger *J. Chem. Inf. Comput. Sci.* **35** 594 (1995)
129. J W Raymond, P Willett *J. Comput.-Aid. Mol. Design* **16** 521 (2002)
130. D M Bayada, R W Simpson, A P Johnson *J. Chem. Inf. Comput. Sci.* **32** 680 (1992)
131. J W Raymond, E J Gardiner, P Willett *J. Chem. Inf. Comput. Sci.* **42** 305 (2002)
132. J W Raymond, P Willett *J. Comput.-Aid. Mol. Design* **16** 59 (2002)
133. D Robert, R Carbo-Dorca *J. Chem. Inf. Comput. Sci.* **38** 469 (1998)
134. R Benigni, M Cotta-Ramusino, F Giorgi, G Gallo *J. Med. Chem.* **38** 629 (1995)
135. G Boon, W Langenaeker, F De Proft, H De Winter, J P Tollenaere, P Geerlings *J. Phys. Chem. A* **105** 8805 (2001)
136. P E Bowen-Jenkins, W G Richards *Int. J. Quantum Chem.* **30** 763 (1986)
137. R Carbo, B Calabuig *J. Mol. Struct. (THEOCHEM)* **254** 517 (1992)
138. R Carbo, L Leyda, M Arnau *Int. J. Quantum Chem.* **17** 1185 (1980)
139. R Carbo, B Calabuig, L Vera, E Besalu *Adv. Quantum Chem.* **25** 253 (1994)
140. F Manaut, F Sanz, J Jose, M Milesi *J. Comput.-Aid. Mol. Design* **5** 1 (1991)
141. E E Hodgkin, W G Richards *Int. J. Quantum Chem., Quantum Biol. Symp.* **14** 105 (1987)
142. C Podlipnic, J Koller *Acta Chim. Slov.* **48** 325 (2001)
143. J D Petke *J. Comput. Chem.* **14** 928 (1993)
144. L Amat, D Robert, E Besalu, R Carbo-Dorca *J. Chem. Inf. Comput. Sci.* **38** 624 (1998)
145. P Bultinck, R Carbo-Dorca, C Van Alsenoy *J. Chem. Inf. Comput. Sci.* **43** 1208 (2003)
146. A Gallegos, X Girones *J. Chem. Inf. Model.* **45** 321 (2005)
147. A G Saliner, L Amat, R Carbo-Dorca, T W Schultz, M T D Cronin *J. Chem. Inf. Comput. Sci.* **43** 1166 (2003)
148. P Bultinck, T Kuppens, X Girones, R Carbo-Dorca *J. Chem. Inf. Comput. Sci.* **43** 1143 (2003)
149. X Girones, R Carbó-Dorca, R Ponec *J. Chem. Inf. Comput. Sci.* **43** 2033 (2003)
150. E Besalu, X Girones, L Amat, R Carbo-Dorca *Acc. Chem. Res.* **35** 289 (2002)
151. U A Chaudry, P L A Popelier *J. Phys. Chem. A* **107** 4578 (2003)
152. U A Chaudry, P L A Popelier *J. Org. Chem.* **69** 233 (2004)
153. G Boon, C Van Alsenoy, F De Proft, P Bultinck, P Geerlings *J. Phys. Chem. A* **107** 11120 (2003)
154. C Amovilli, R McWeeny *J. Mol. Struct. (THEOCHEM)* **227** 1 (1991)
155. N L Allan, D L Cooper *J. Chem. Inf. Comput. Sci.* **32** 587 (1992)
156. S E O'Brien, P L A Popelier *J. Chem. Soc., Perkin Trans. 2* 478 (2002)
157. G Boon, F De Proft, W Langenaeker, P Geerlings *Chem. Phys. Lett.* **295** 122 (1998)
158. S E O'Brien, P L A Popelier *J. Chem. Inf. Comput. Sci.* **41** 764 (2001)
159. S E O'Brien, P L A Popelier *Can. J. Chem.* **77** 28 (1999)
160. L Cooper, N L Allan *J. Am. Chem. Soc.* **114** 4773 (1992)
161. R Carbo, E Besalu, L Amat, X Fradera *J. Math. Chem.* **19** 47 (1996)
162. R Carbo, B Calabuig *J. Chem. Inf. Comput. Sci.* **32** 600 (1992)
163. R Carbo, B Calabuig *Int. J. Quantum Chem.* **42** 1681 (1992)
164. J Cioslowski, E D Fleischmann *J. Am. Chem. Soc.* **113** 64 (1991)
165. X Fradera, L Amat, E Besalu, R Carbo-Dorca *Quantum Struct. – Act. Relat.* **16** 25 (1997)
166. A C Good, S-S So, W G Richards *J. Med. Chem.* **36** 433 (1993)
167. P L A Popelier, U A Chaudry, P J Smith *J. Chem. Soc., Perkin Trans. 2* 1231 (2002)
168. P L Popelier *J. Phys. Chem. A* **103** 2883 (1999)
169. R Ponec, M Strnad *Croat. Chim. Acta* **66** 123 (1993)
170. D Robert, R Carbo-Dorca *J. Chem. Inf. Comput. Sci.* **38** 620 (1998)

171. M Sola, J Mestres, J M Oliva, M Duran, R Carbo *Int. J. Quantum Chem.* **58** 361 (1996)
172. R Hefferlin, M T Matus *J. Chem. Inf. Comput. Sci.* **41** 484 (2001)
173. X Girones, R Carbo-Dorca *J. Chem. Inf. Comput. Sci.* **42** 317 (2002)
174. X Girones, R Carbo-Dorca *J. Chem. Inf. Comput. Sci.* **42** 1185 (2002)
175. P Bultinck, R Carbo-Dorca *J. Chem. Inf. Comput. Sci.* **43** 170 (2003)
176. L Amat, E Besalu, R Carbo-Dorca, R Ponec *J. Chem. Inf. Comput. Sci.* **41** 978 (2001)
177. J Danzinger, P M Dean *J. Theor. Biol.* **116** 215 (1985)
178. E Fontain *J. Chem. Inf. Comput. Sci.* **32** 748 (1992)
179. M G Bures, Y C Martin, P Willett *Top. Stereochem.* **21** 467 (1994)
180. M J Sippl, H Stegbuchner *Comput. Chem.* **15** 73 (1991)
181. A Y Meyer, W G Richards *J. Comput.-Aid. Mol. Design* **5** 427 (1991)
182. D E Walters, A J Hopfinger *J. Mol. Struct. (THEOCHEM)* **134** 317 (1986)
183. X Wang, J T L Wang *J. Chem. Inf. Comput. Sci.* **40** 442 (2000)
184. T Kotani, K Higashiura *J. Chem. Inf. Comput. Sci.* **42** 58 (2002)
185. G Bringmann, C Rummey *J. Chem. Inf. Comput. Sci.* **43** 304 (2003)
186. R D Brown, Y C Martin *J. Chem. Inf. Comput. Sci.* **37** 1 (1997)
187. J W Raymond, P Willett *J. Chem. Inf. Comput. Sci.* **43** 908 (2003)
188. N Rhodes, P Willett, A Calvet, J B Dunbar, C Humblet *J. Chem. Inf. Comput. Sci.* **43** 443 (2003)
189. F Boeckler, H Lanig, P Gmeiner *J. Med. Chem.* **48** 694 (2005)
190. V Consonni, R Todeschini, M Pavan *J. Chem. Inf. Comput. Sci.* **42** 682 (2002)
191. V Consonni, R Todeschini, M Pavan, P Gramatica *J. Chem. Inf. Comput. Sci.* **42** 693 (2002)
192. R D Clark, A M Ferguson, R D Crame *Persp. Drug Discov. Design* **9–11** 213 (1998)
193. A C Good, W G Richard *Persp. Drug Discov. Design* **9–11** 321 (1988)
194. B B Goldman, W T Wipke *J. Chem. Inf. Comput. Sci.* **40** 644 (2000)
195. S Handschuh, M Wagener, J Gastaiger *J. Chem. Inf. Comput. Sci.* **38** 220 (1998)
196. T W Heritage, A M Ferguson, D B Turner, P Willett *Persp. Drug Discov. Design* **9–11** 381 (1998)
197. H Kubinyi *Persp. Drug Discov. Design* **9–11** 225 (1998)
198. R S Pearlman, K M Smith *Persp. Drug Discov. Design* **9–11** 339 (1998)
199. B D Silverman, D E Platt *J. Med. Chem.* **39** 2129 (1996)
200. D A Thorner, D J Wild, P Willett, P M Wrigh *Persp. Drug Discov. Design* **9–11** 301 (1998)
201. R Todeschini, P Gramatica *Persp. Drug Discov. Design* **9–11** 355 (1998)
202. D J Wild, P Willett *J. Chem. Inf. Comput. Sci.* **36** 159 (1996)
203. P Constans, R Carbo *J. Chem. Inf. Comput. Sci.* **35** 1046 (1995)
204. J C Duca, A J Hopfinger *J. Chem. Inf. Comput. Sci.* **41** 1367 (2001)
205. B C P Allen, G H Grant, W G Richards *J. Chem. Inf. Comput. Sci.* **43** 134 (2003)
206. S L Dixon, K M Merz *J. Med. Chem.* **44** 3795 (2001)
207. R D Brown, Y C Martin *J. Chem. Inf. Comput. Sci.* **36** 572 (1996)
208. H Matter *J. Med. Chem.* **40** 1219 (1997)
209. V J Gillet, P Willett, J Bradshaw *J. Chem. Inf. Comput. Sci.* **43** 338 (2003)
210. M Thimm, A Goede, S Hougardy, R Preissner *J. Chem. Inf. Comput. Sci.* **44** 1816 (2004)
211. M Whittle, P Willett, W Kraffke, P van Noort *J. Chem. Inf. Comput. Sci.* **43** 449 (2003)
212. J D Holliday, N Salim, M Whittle, P Willett *J. Chem. Inf. Comput. Sci.* **43** 819 (2003)
213. M Cotta-Ramusino, R Benigni, L Passerini, A Giuliani *J. Chem. Inf. Comput. Sci.* **43** 248 (2003)
214. M Whittle, V J Gillet, P Willett, A Alex, J Loesel *J. Chem. Inf. Comput. Sci.* **44** 1840 (2004)
215. S F B Abolmaali, C Ostermann, A Zell *J. Mol. Model.* **9** 66 (2003)
216. N Salim, J D Holliday, P Willett *J. Chem. Inf. Comput. Sci.* **43** 435 (2003)
217. M I Skvortsova, I V Stankevich, I I Baskin, V A Palyulin, N S Zefirov *Dokl. Akad. Nauk* **350** 786 (1996) [b]
218. M I Skvortsova, I I Baskin, I V Stankevich, V A Palyulin, N S Zefirov *J. Chem. Inf. Comput. Sci.* **38** 785 (1998)
219. A V Pasyukov, M I Skvortsova, V A Palyulin, N S Zefirov *Dokl. Akad. Nauk* **374** 786 (2000) [b]
220. P J Peruzzo, D J G Marino, E A Castro, A A Toropov *Internet Electron. J. Mol. Des.* **2** 334 (2003); http://www.biochempress.com
221. R P Sheridan, S K Kearsley *Drug Discov. Today* **17** 903 (2002)
222. E J Videlock, V K Chung, M A Mohan, T M Strok, D J Austin *J. Am. Chem. Soc.* **126** 3730 (2004)
223. C Qin, X Bu, X Wu, Z Guo *J. Combinat. Chem.* **5** 353 (2003)
224. G M Makara *J. Med. Chem.* **44** 3563 (2001)
225. Y-K Kim, M A Arai, T Arai, J O Lamenzo, E F Dean, N Patterson, P A Clemons, S L Schreiber *J. Am. Chem. Soc.* **126** 14740 (2004)
226. D K Agrafiotis, D N Rassokhin *J. Chem. Inf. Comput. Sci.* **42** 117 (2002)
227. D K Agrafiotis *J. Chem. Inf. Comput. Sci.* **41** 159 (2001)
228. S J Yu, S M Keenan, W Tong, W J Welsh *Chem. Res. Toxicol.* **15** 1229 (2002)
229. C H Reynolds, A Tropsha, L B Phahler, R Druker, S Chakravorty, G Ethiraj, W Zheng *J. Chem. Inf. Comput. Sci.* **41** 1470 (2001)
230. A M M Jorgensen, J T Pedersen *J. Chem. Inf. Comput. Sci.* **41** 338 (2001)
231. C T Clein, D C Kaiser, G Ecker *J. Chem. Inf. Comput. Sci.* **44** 200 (2004)
232. S Monteiro, M A Picarra-Pereira, P R Mesquita, V B Loureiro, A Teixeira, R B Ferreira *J. Agric. Food Chem.* **49** 3999 (2001)
233. S V Trepalin, V A Gerasimenco, A V Kozyukov, N P Savchuk, A A Ivaschenko *J. Chem. Inf. Comput. Sci.* **42** 249 (2002)
234. W H B Sauer, M K Schwarz *J. Chem. Inf. Comput. Sci.* **43** 987 (2003)
235. M Johnson, in *Chemical Information Systems: Beyond the Structure Diagram* (Eds D Bawden, E Mitchell) (New York: Ellis Horwood, 1990) p. 149
236. H Friedman, B W Silverman *Technometrics* **31** 3 (1989)
237. V N Vapnik, A Ya Chervonenkis *Teoriya Raspoznavaniya Obrazov* (The Theory of Pattern Recognition) (Moscow: Nauka, 1974)
238. A A Gorelik, V A Skripkin *Metody Raspoznavaniya* (The Methods of Recognition) (Moscow: Vysshaya Shkola, 1984)
239. N G Zagoruiko *Prikladnye Metody Analiza Dannykh i Znanii* (Applied Methods of Data and Knowledge Analysis) (Novosibirsk: Novosibirsk Institute of Mathematics, 1999)
240. W C Herndon, G Rum, in *QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications* (Eds F Sanz, J Giraldo, F Manant) (Barcelona: Prous Science, 1995) p. 380
241. O Mekenyan, J Ivanov, S Karabunarliev, B Hansen, G Ankley, S Bradbury, in *Quantitative Structure — Activity Relationships in Environmental Sciences-VII* (Eds F Chen, G Schürmann) (Pensacola, FL: Setac Press, 1997) p. 299
242. W G Richard, in *QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications* (Eds F Sanz, J Giraldo, F Manant) (Barcelona: Prous Science, 1995) p. 354
243. S Tomić, R P Gabdoulline, B Kojić-Prodić, R Wide *J. Comput.-Aid. Mol. Design* **12** 63 (1998)
244. S C Basak, in *Practical Applications of Quantitative Structure–Activity Relationships (QSAR) in Environmental Chemistry and Toxicology* (Eds W Karcher, J Devillers) (Dordrecht, Boston: Kluwer Academic, 1990) p. 83
245. M Randić, A Sablić, S Nicolić, N Trinajstić *Int. J. Quantum Chem., Quantum Biol. Symp.* **15** 267 (1988)

246. J Takahashi, M Sukekawa, S Sasaki *J. Chem. Inf. Comput. Sci.* **32** 639 (1992)
247. M Yu Gorbachev, A S Dimoglo, I B Bersuker, P F Vlad, M N Koltsa *Teoret. Eksp. Khim.* **22** 355 (1986) [c]
248. I B Bersuker, A S Dimoglo, M Yu Gorbachev, P F Vlad *Nouv. J. Chim.* **9** 211 (1985)
249. S C Basak, B D Gute, G D Grunwald *Croat. Chim. Acta* **69** 1159 (1996)
250. S C Basak, G D Grunwald *Chemosphera* **31** 2529 (1995)
251. H W Kroto, J R Heath, S C O'Brien, R F Curl, R E Smalley *Nature (London)* **318** 162 (1985)
252. D A Bochvar, E G Gal'pern *Dokl. Akad. Nauk SSSR* **209** 610 (1973) [b]
253. E Osawa *Kagaku (Kyoto)* **25** 854 (1970); *Chem. Abstr.* **74** 75698 (1971)
254. R Feynman *The Character of Physical Law* (London: Cox and Wyman, 1965)

[a] — *Russ. J. Phys. Chem. (Engl. Transl.)*
[b] — *Dokl. Chem. (Engl. Transl.)*
[c] — *Theor. Exp. Chem. (Engl. Transl.)*