CHAPTER 1

Fragment Descriptors in SAR/QSAR/QSPR Studies, Molecular Similarity Analysis and in Virtual Screening

IGOR BASKIN[a] AND ALEXANDRE VARNEK[b]

[a] Department of Chemistry, Moscow State University, Moscow 119992, Russia; [b] Laboratoire d'Infochimie, UMR 7177 CNRS, Université Louis Pasteur, 4, rue B. Pascal, Strasbourg 67000, France

1.1 Introduction

Chemoinformatics[1–5] is an emerging science that concerns the mixing of chemical information resources to transform data into information, and information into knowledge. It is a branch of theoretical chemistry based on its molecular model, and which uses its own basic concepts, learning approaches and areas of application. Unlike quantum chemistry, which considers molecules as ensemble of electrons and nuclei, or force field molecular mechanics or dynamics simulations based on a classical molecular model ("atoms" and "bonds"), chemoinformatics represents molecules as objects in a chemical space defined by molecular descriptors. Among thousands of descriptors, fragment descriptors occupy a special place. Fragment descriptors represent selected subgraphs of a 2D molecular graph; structure–property approaches use their occurrences in molecules or binary values (0, 1) to indicate their presence or absence in the given graph.

The unique properties of fragment descriptors are related to the fact that (i) any molecular graph invariant (*i.e.*, any molecular descriptor or property)

Chemoinformatics Approaches to Virtual Screening
Edited by Alexandre Varnek and Alex Tropsha

Published by the Royal Society of Chemistry, www.rsc.org

can be uniquely represented as a linear combination of fragment descriptors;[7–9] (ii) any symmetric similarity measure can be uniquely expressed in terms of fragment descriptors;[10,11] and (iii) any regression or classification structure–property model can be represented as a linear equation involving fragment descriptors.[12,13]

An important advantage of fragment descriptors is related to the simplicity of their calculation, storage and interpretation (see review articles[14–18]). They belong to information-based descriptors,[19] which tend to code the information stored in molecular structures. This contrasts with knowledge-based (or semi-empirical) descriptors derived from consideration of the mechanism of action. Owing to their versatility, fragment descriptors can efficiently be used to build structure–property models, perform similarity search, virtual screening and *in silico* design of chemical compounds with desired properties.

This chapter reviews fragment descriptors with respect to their use in structure–property studies, similarity search and virtual screening. After a short historical survey, different types of fragment descriptors are considered thoroughly. This is followed by a brief review of the application of fragment descriptors in virtual screening, focusing mostly on filtering, similarity search and direct activity/property assessment using quantitative structure–property models.

1.2 Historical Survey

Among a multitude of descriptors currently used in Structure–Activity Relationships/Quantitative Structure–Activity Relationships/Quantitative Structure–Property Relationships (SAR/QSAR/QSPR) studies,[20] fragment descriptors occupy a special place. Their application as atoms and bonds increments in the framework of *additive schemes* can be traced back to the 1930–1950s; Vogel,[21] Zahn,[22] Souders,[23,24] Franklin,[25,26] Tatevskii,[27,28] Bernstein,[29] Laidler,[30] Benson and Buss[31] and Allen[32] pioneered this field. Smolenskii was one of the first, in 1964, to apply graph theory to tackle the problem of predictions of the physico-chemical properties of organic compounds.[33] Later on, these first additive schemes approaches have gradually evolved into *group contribution methods*. The latter are closely linked with thermodynamic approaches and, therefore, they are applicable only to a limited number of properties.

The epoch of QSAR (Quantitative Structure–Activity Relationships) studies began in 1963–1964 with two seminal approaches: the σ-ρ-π analysis of Hansch and Fujita[34,35] and the Free–Wilson method.[36] The former approach involves three types of descriptors related to electronic, steric and hydrophobic characteristics of substituents, whereas the latter considers the substituents themselves as descriptors. Both approaches are confined to strictly congeneric series of compounds. The Free–Wilson method additionally requires all types of substituents to be sufficiently present in the training set. A combination of these two approaches has led to QSAR models involving *indicator variables*, which indicate the presence of some structural fragments in molecules.

The non-quantitative SAR (Structure–Activity Relationships) models developed in the 1970s by Hiller,[37,38] Golender and Rosenblit,[39,40] Piruzyan, Avidon *et al.*,[41] Cramer,[42] Brugger, Stuper and Jurs,[43,44] and Hodes *et al.*[45] were inspired by the, at that time, popular artificial intelligence, expert systems, machine learning and pattern recognition paradigms. In those approaches, chemical structures were described by means of indicators of the presence of structural fragments interpreted as topological (or 2D) pharmacophores (biophores, toxophores, *etc.*) or topological pharmacophobes (biophobes, toxophobes, *etc.*). Chemical compounds were then classified as active or inactive with respect to certain types of biological activity.

Methodologies based on fragment descriptors in QSAR/QSPR studies are not strictly confined to particular types of properties or compounds. In the 1970s Adamson and coworkers[46,47] were the first to apply fragment descriptors in multiple linear regression analysis to find correlations with some biological activities,[48,49] physicochemical properties,[50] and reactivity.[51]

An important class of fragment descriptors, the so-called *screens* (or structural *keys*, *fingerprints*), were also developed in 1970s.[52–56] As a rule, they represent the bit strings that can effectively be stored and processed by computers. Although their primary role is to provide efficient substructure searching in large chemical structure databases, they can be efficiently used also for similarity searching,[57,58] clustering large chemical databases,[59,60] assessing their diversity,[61] as well as for SAR[62] and QSAR[63] modeling.

Another important contribution was made in 1980 by Cramer who invented BC(DEF) parameters obtained by means of factor analysis of the physical properties of 114 organic liquids. These parameters correlate strongly with various physical properties of diverse liquid organic compounds.[64] On the other hand, they could be estimated by linear additive-constitutive models involving fragment descriptors.[65] Thus, a set of QSPR models encompassing numerous physical properties of diverse organic compounds has been developed using only fragment descriptors.

One of the most important developments of the 1980s was the CASE (Computer-Automated Structure Evaluation) program by Klopman *et al.*[66–69] This ‘‘self-learning artificial intelligent system’’[69] can recognize activating and deactivating fragments (biophores and biophobes) with respect to the given biological activity and to use this information to determine the probability that a test chemical is active. This methodology has been successfully applied to predict various types of biological activity: mutagenicity,[67,70,71] carcinogenicity,[66,69,71–73] hallucinogenic activity,[74] anticonvulsant activity,[75] inhibitory activity with respect to sparteine monooxygenase,[76] β-adrenergic activity,[77] μ-receptor binding (opiate) activity,[78] antibacterial activity,[79] antileukemic activity,[80] *etc.* Using the multivariate regression technique, CASE can also build quantitative models involving fragment descriptors.[72,77]

Starting in the early 1990s, various approaches and related software tools based on fragment descriptors have been developed and are listed in several conceptual and mini-review papers.[14–18] Because of the wide scope and large variety of different approaches and applications in this field, many important ideas were reinvented many times and continue to be reinvented. In this review we try to present a clear state-of-the-art picture in this area.

1.3 Main Characteristics of Fragment Descriptors

In this section different types of fragments are classified with respect to their topology and the level of abstraction of molecular graphs.

1.3.1 Types of Fragments

A tremendous number of various fragments are used in structure–property studies: atoms, bonds, "topological torsions", chains, cycles, atom- and bond-centered fragments, maximum common substructures, line notation (WLN and SMILES) fragments, atom pairs and topological multiplets, substituents and molecular frameworks, basic subgraphs, *etc*. Their detailed description is given below.

Depending on the application area, two types of values taken by fragment descriptors are considered: binary and integer. Binary values indicate the presence (*true*, *yes*, 1) or the absence (*false*, *no*, 0) of a given fragment in a structure. They are usually used as screens and elements of fingerprints for chemical database management and virtual screening using similarity-based approaches as well as in SAR studies. Integer values corresponding to the occurrences of fragments in structures are used in QSAR/QSPR modeling.

1.3.1.1 Simple Fixed Types

Disconnected atoms represent the simplest type of fragments. They are used to assess a chemical or biological property P in the framework of an additive scheme based on atomic contributions:

$$P \approx \sum_{i=1}^{N} n_i \cdot A_i \tag{1.1}$$

where n_i is the number of atoms of i-type, A_i is corresponding atomic contributions. Usually, the atom types account for not only the type of chemical element but also hybridization, the number of attached hydrogen atoms (for heavy elements), occurrence in some groups or aromatic systems, *etc*. Nowadays, atom-based methods are used to predict some physicochemical properties and biological activities. Thus, several works have been devoted to assess the octanol–water partition coefficient log P: the ALOGP method by Ghose-Crippen,[81–83] later modified by Ghose and co-workers,[84,85] and by Wildman and Crippen,[86] the CHEMICALC-2 method by Suzuki and Kudo,[87] the SMILOGP program by Convard and co-authors,[88] and the XLOGP method by Wang and co-authors.[89,90] Hou and co-authors[91] used Equation (1.1) to calculate aqueous solubility. The ability of this approach to assess biological activities was demonstrated by Winkler *et al.*[92]

Chemical bonds are another type of simple fragment. The first bond-based additive schemes, such as those of Zahn,[22] Bernstein[29,93] and Allen,[32,94] appeared almost simultaneously with the atom-based ones and dealt, presumably, with predictions of some thermodynamic properties.

"Topological torsions" invented Nilakantan *et al.*[95] are defined as a linear sequence of four consecutively bonded non-hydrogen atoms. Each atom there is described by the type of corresponding chemical element, the number of attached non-hydrogen atoms and the number of π-electron pairs. Molecular descriptors indicating the presence or absence of topological torsions in chemical structures have been used to perform qualitative predictions of biological activity in structure–activity (SAR) studies.[95] Later on, Kearsley *et al.*[96] recognized that characterizing atoms by element types can be too specific for similarity searching and, therefore, it does not provide sufficient flexibility for large-scaled virtual screening. To solve this problem, they suggested assigning atoms in the Carhart's atom pairs and Nilakantan's topological torsions to one of seven classes: cations, anions, neutral hydrogen bond donors, neutral hydrogen bond acceptors, polar atoms, hydrophobic atoms and other.

The above-mentioned structural fragments – atoms, bonds and topological torsions – can be regarded as *chains* of different lengths. Smolenskii[33] suggested using the occurrences of chains in an additive scheme to predict the formation enthalpy of alkanes. For the last four decades, chain fragments have proved to be one of the most popular and useful type of fragment descriptors in QSPR/QSAR/SAR studies. Fragment descriptors based on enumerating chains in molecular graphs are efficiently used in many popular structure–property and structure–activity programs: CASE[66–69] and MULTICASE (MultiCASE, MCASE) by Klopman[97,98] NASAWIN[99] by Baskin *et al.*, BIBIGON[100] by Kumskov, TRAIL[101,102] and ISIDA[18] by Solov'ev and Varnek. "Molecular pathways" by Gakh and co-authors,[103] and "molecular walks" by Rücker,[104] represent chains of atoms.

In contrast to chains, cyclic and polycyclic fragments are relatively rarely applied as descriptors in QSAR/QSPR studies. Nevertheless, *implicitly* cyclicity is accounted for by means of: (i) introducing special "cyclic" and "aromatic" types of atoms and bonds, (ii) "collapsing" the whole cycles and even polycyclic systems into "pharmacophoric" pseudo-atoms and (iii) generating cyclic fragments as a part of large fragments [Maximum Common Substructure (MCS), molecular framework, substituents]. Besides, the cyclic fragments are widely used as screens for chemical database processing.[105,106]

1.3.1.2 WLN and SMILES Fragments

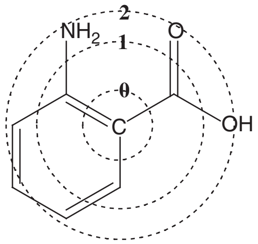
WLN and SMILES fragments correspond respectively to substrings of the Wiswesser Line Notation[107] or Simplified Molecular Input Line Entry System[108,109] strings used for encoding the chemical structures. Since simple string operations are much faster than processing of information in connection tables, the use of WLN descriptors was justified in the 1970s when computers were still very slow. At that time Adamson and Bawden published some linear QSAR models based on WLN fragments.[48,50,51,110,111] They have also applied this kind of descriptor for hierarchical cluster analysis and automatic classification of chemical structures.[112] Qu *et al.*[113,114] have developed AES (Advanced Encoding System), a new WLN-based notation encoding chemical information for group contribution methods. Interest in line notation descriptors has not disappeared completely with the advent of powerful computers. Thus, SMILES fragment descriptors are used in the SMILOGP program to predict log *P*,[88] whereas the recently developed LINGO system for assessing some biophysical properties and intermolecular similarities uses holographic representations of canonical SMILES strings.[115]

1.3.1.3 Atom-centered Fragments

Atom-Centered Fragments (ACF) consist of a single central atom surrounded by one or several shells of atoms separated from the central one by the same topological distance. This type of structural fragments was introduced in the early 1950s by Tatevskii,[27,28,116–119] and then by Benson[31] to predict some physicochemical properties of organic compounds in the framework of additive schemes.

ACF fragments containing only one shell of atoms around the central one (*i.e.*, atom-centered neighborhoods of radius 1) were introduced into chemoinformatics practice in 1971 under the names "atom-centered fragments" and "augmented atoms" by Adamson,[120,121] who studied their distribution in large chemical databases with the intention of using them as screens in chemical database searching. Hodes used, in SAR studies, both "augmented atoms"[45] and "ganglia augmented atoms"[325] representing ACF fragments with radius 2 and generalized second-shell atoms. Subsequently, ACF fragments with radius 1 were implemented in NASAWIN,[122–124] TRAIL[101,102,125] and ISIDA[18] programs. ACF fragments with arbitrary radius were implemented by Filimonov, Poroikov and co-authors in the PASS[126] program under the name Multilevel Neighborhoods of Atoms (MNA),[127] by Xing and Glen as "tree structured fingerprints",[128] by Bender and Glen as "atom environments"[129,130] and "circular fingerprints"[131–133] (Figure 1.1), and by Faulon as "molecular signatures".[134–136]

Several types of ACF fragments were designed to store local spectral parameters (chemical shifts) in spectroscopy data bases. Thus, Bremser has developed Hierarchically Ordered Spherical Environment (HOSE), a system of substructure codes aimed at characterizing the spherical environment of single atoms and complete ring systems.[137] The codes are generated automatically from 2D graphs and describe structural entities corresponding to chemical shifts. A very similar idea has also been implemented by Dubois *et al.* in the DARC system based on FREL (Fragment Réduit à un Environment Limité) fragments.[138,139] Xiao *et al.* have applied Atom-Centered Multilayer Code



Layers:	0	1	2
	C.ar (sp^2)	C.ar (sp^2)	C.ar (sp^2)
		C.ar (sp^2)	C.ar (sp^2)
		C (sp^2)	N (sp^3)
			O (sp^2)
			O (sp^3)

Figure 1.1 Circular fingerprints with Sybyl mol2 atom typing. An individual fingerprint is calculated for each atom in the molecule, considering those atoms up to two bonds from the central atom (level 2). The molecular fingerprint consists of the individual atom fingerprints of all the heavy atoms in the structure. (Adapted from ref. 132.)

(ACMC) fragments for structural and substructural searching in large databases of compounds and reactions.[140] An important recent application of ACF fragments concerns target prediction (“target fishing”) in chemogenomic data analysis.[126,141,142]

1.3.1.4 Bond-centered Fragments

Bond-centered fragments (BCF) consist of two atoms linked by the bond and surrounded by one or several shells of atoms separated by the same topological distance from this bond. Although these fragments are rather rarely used in structure–property studies, they can be efficiently used as screens for chemical database processing.[143] BCF have been used as a part of MDL keys[144,145] for substructure search in chemical databases, database clustering[60] and for SAR studies of 17 different types of biological activity.[62] Bond-centered fragments have also been used in the DARC system.[138,139]

1.3.1.5 Maximum Common Substructures

For a set of molecular graphs, a Maximum Common Substructure (MCS) is defined as a largest substructure in all graphs belonging to the given set. In most practical applications, only MCS for graph pairs are considered, *i.e.*, for sets containing only two graphs. MCS can be found by *intersecting* molecular graphs using several different algorithms (for a review see ref. 146), the best known of which involve clique detection in so-called compatibility graphs.

Notably, a pair of graphs can have more than one MCS. The main advantage of MCS fragments is related to the fact that their complexity is not limited and therefore they can be used to detect property-relevant features that could not be detected by fragments (subgraphs) of limited complexity.

MCSs were first applied to SAR studies in the early 1980s by Rozenblit and Golender in the framework of their logical-combinatorial approach.[40,41,147] Since at that time computer power was limited, the authors suggested the use of reduced graphs (Section 1.3.5) built on pharmacophoric centers. The MCS fragments were subsequently applied to perform a similarity search,[148] to cluster chemical databases[149,150] as well to assess biological activities of organic compounds.[99,151,152]

1.3.1.6 Atom Pairs and Topological Multiplets

Characterizing atoms only by element types is too specific for similarity searching and, therefore, does not provide sufficient flexibility for large-scale virtual screening. For that reason, numerous studies have been devoted to increase the informational content of fragment descriptors by adding some useful empirical information and/or by representing a part of the molecular graph implicitly. The simplest representatives of such descriptors were "atom pairs and topological multiplets" based on the notion of a "descriptor center" representing an atom or a group of atoms that could serve as centers of intermolecular interactions. Usually, descriptor centers include heteroatoms, unsaturated bonds and aromatic cycles. An *atom pair* is defined as a pair of atoms (**AT**) or descriptor centers separated by a fixed topological distance: $\mathbf{AT}_i$-$\boldsymbol{Dist}$-$\mathbf{AT}_j$, where $Dist_{ij}$ is the shortest path (the number of bonds) between AT_i and AT_j. Analogously, a topological multiplet is defined as a multiplet (usually triplet) of descriptor centers and topological distances between each pair of them. In most of cases, these descriptors are used in binary form to indicate the presence or absence of the corresponding features in studied chemical structures.

Atom pairs were first suggested for SAR studies by Avidon as Substructure Superposition Fragment Notation (SSFN).[41,153] They were then independently reinvented by Carhart and co-authors[154] for similarity and trend vector analysis. In contrast to SSFN, Carhart's atom pairs are not necessarily composed only of descriptor centers but account for the information about element type, the number of bonded non-hydrogen neighbors and the number of π electrons. Nowadays, Carhart's atom pairs are popular in virtual screening. Topological Fuzzy Bipolar Pharmacophore Autocorrelograms (TFBPA)[155] by Horvath are based on atom pairs, in which real atoms are replaced by pharmacophore sites (hydrophobic, aromatic, hydrogen bond acceptor, hydrogen bond donor, cation, anion), while $Dist_{ij}$ corresponds to different ranges of topological distances between pharmacophores. These descriptors were successfully applied in virtual screening against a panel of 42 biological targets using a similarity search based on several fuzzy and non-fuzzy metrics,[156] performing only slightly less

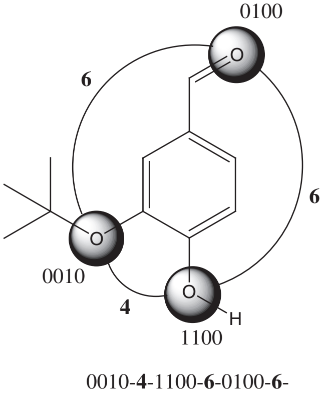


Figure 1.2 Example of a Similog key. (Adapted from ref. 158.)

well than their 3D counterparts.[155] Fuzzy Pharmacophore Triplets (FPT) by Horvath[157] is an extension of FBPF[156] for three-site pharmacophores. An important innovation in the FPT concerns accounting for proteolytic equilibrium as a function of pH.[157] Owing to this feature, even small structural modifications leading to a pK_a shift may have a profound effect on the fuzzy pharmocophore triples. As a result, these descriptors efficiently discriminate structurally similar compounds exhibiting significantly different activities.[157]

Some other topological triplets should be mentioned. Similog pharmacophoric keys by Schuffenhauer *et al.*[158] represent triplets of binary coded types of atoms (pharmacophoric centers) and topological distances between them (Figure 1.2). Atomic types are generalized by four features (represented as four bits per atom): potential hydrogen bond, donor or acceptor, bulkiness and electropositivity. The "topological pharmacophore-point triangles" implemented in the MOE software[159] represent triplets of MOE atom types separated by binned topological distances. Structure–property models obtained by a support vector machine method with these descriptors have been successfully used for virtual screening of COX-2 inhibitors[160] and D_3 dopamine receptor ligands.[161]

1.3.1.7 Substituents and Molecular Frameworks

In organic chemistry, decomposition of molecules into substituents and molecular frameworks is a natural way to characterize molecular structures. In QSAR, both the Hansch–Fujita[34,35] and the Free–Wilson[36] classical approaches are based on this decomposition, but only the second one explicitly accounts for the presence or the absence of substituent(s) attached to molecular framework at a certain position. While the multiple linear regression technique was associated with the Free–Wilson method, recent modifications of this approach involve more sophisticated statistical and machine-learning approaches, such as the principal component analysis[162] and neural networks.[163]

In contrast to substituents, molecular frameworks are rarely used in SAR/QSAR/QSPR studies. In most cases, they are implicitly involved as indicator variables discriminating different types of molecular motifs (see, for example, ref. 164). The distributions of different molecular frameworks and substituents (side chains) in the databases of known drug molecules has been thoroughly studied by Bemis and Murcko.[165,166]

1.3.1.8 Basic Subgraphs

Regarding fragment descriptors, one could imagine a huge number of possibilities to split a molecular graph into constituent fragments. Making a parallel with the decomposition of vectors into a limited number of basis functions, Randič[326] suggested the existence of a small set of *basic subgraphs* representing any structure and which could be used to calculate any molecular property. In particular, for small alkanes a set of disconnected graphs representing paths (chains) of different length has been proposed (Figure 1.3).

However, later it has since been found that this set is not sufficient to differentiate any two structures. Skvortsova *et al.* have extended the set of Randič basic subgraphs by including cyclic fragments and more complex subgraphs consisting of single node attached to a cyclic fragment.[167] This set exhibits good coding uniqueness (*i.e.*, different vectors of descriptors correspond to different structures) and coding completeness (*i.e.*, they can approximate a numerous structure–property functions). Basic fragment descriptors of this kind were used in several QSPR studies.[168]

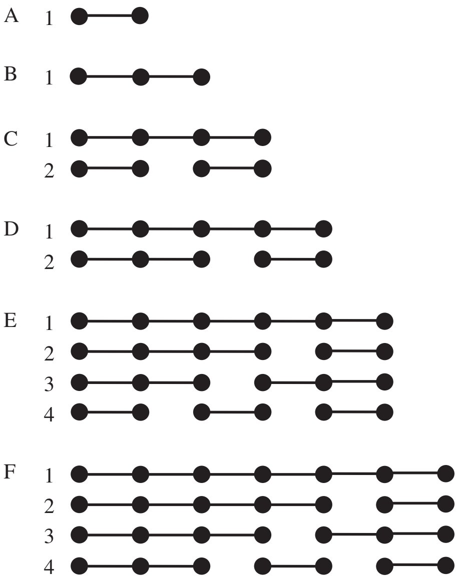


Figure 1.3 Randič basic graphs for a maximum number of nodes of 7.

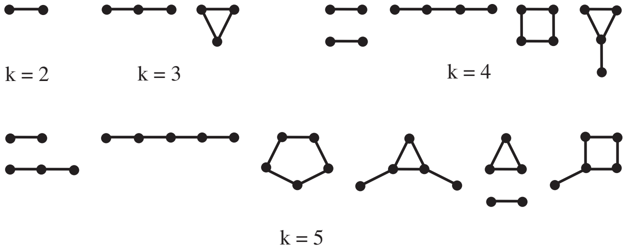


Figure 1.4 Skvortsova's basic graphs for a maximum number of nodes of 5.

In fact, a rigorous solution of the problem of finding a set of basic graph invariants was obtained by Mnukhin[169] for simple graphs and then extended to molecular graphs by Baskin, Skvortsova *et al.*[7–9] (Figure 1.4). It has been shown that the complete set of basic graph invariants could be built on all possible subgraphs, and hence one can not to confine this to any subset of limited size. Nonetheless, for many practical tasks the application of a limited number of basic subgraphs and the corresponding fragment descriptors could be useful.

Another application of basic subgraphs arises from the possibility[8,169] of relating the invariants of molecular graphs to the occurrence numbers of some basic subgraphs. Estrada has developed this methodology for *spectral moments* of the edge-adjacency matrix of molecular graphs – defined as the traces of the different powers of such matrix:[170–172]

$$\mu_k = \mathrm{tr}(\boldsymbol{E}^k) \qquad (1.2)$$

where μ_k is the k-th spectral moment of the edge-adjacency matrix $\boldsymbol{E}$ (which is a symmetric matrix whose elements e_{ij} are 1 only if edge i is adjacent to edge j) and tr is the trace, *i.e.* the sum of the diagonal elements of the matrix. On the other hand, spectral moments can be expressed as linear combinations of the occurrence numbers of certain structural fragments in the molecular graph. These linear combinations for simple molecular graphs not containing heteroatoms have been reported for acyclic[170] and cyclic[172] chemical structures.

To illustrate these notions, consider a correlation between the boiling points of alkanes and their spectral moments reported in ref. 170:

$$\mathrm{bp}(^\circ\mathrm{C}) = - \quad 76.719 + 23.992\mu_0 + 2.506\mu_2 - 2.967\mu_3 + 0.149\mu_5 \qquad (1.3)$$

$$R = 0.9949, \; s = 4.21, \; F = 1650$$

The first six spectral moments of the edge-adjacency matrix $\boldsymbol{E}$ are expressed as linear combinations of the occurrence numbers of fragments listed in Figure 1.5:

$$\mu_0 = |F_1| \qquad (1.4)$$

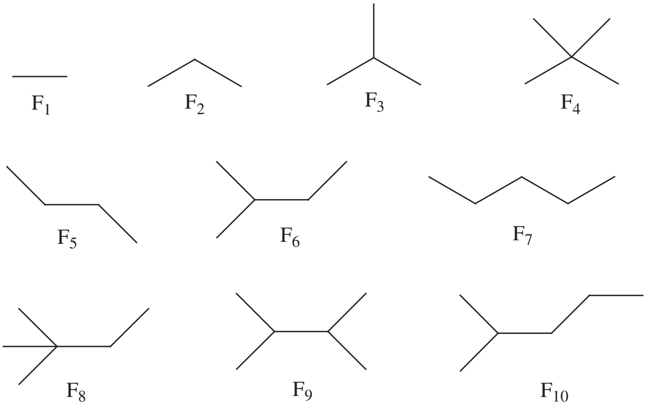


Figure 1.5 First ten structural fragments contained in molecular graphs of alkanes. (Adapted from ref. 170.)

$$\mu_2 = 2 \times |F_2| \tag{1.5}$$

$$\mu_3 = 6 \times |F_3| \tag{1.6}$$

$$\mu_4 = 2 \times |F_2| + 12 \times |F_3| + 24 \times |F_4| + 4 \times |F_5| \tag{1.7}$$

$$\mu_5 = 30 \times |F_3| + 120 \times |F_4| + 10 \times |F_6| \tag{1.8}$$

$$\begin{aligned}\mu_6 =& 2 \times |F_2| + 60 \times |F_3| + 480 \times |F_4| + 12 \times |F_5| + 24 \times |F_6| \\ &+ 6 \times |F_7| + 36 \times |F_8| + 24 \times |F_9|\end{aligned} \tag{1.9}$$

where $|F_i|$ denotes the occurrence number of subgraph F_i in molecular graph.

Thus, by substituting spectral moments in the QSPR Equation (1.4) for their expansions (Equations 1.5–1.10) one can obtain the following QSPR equation with fragment descriptors:

$$\begin{aligned}\mathrm{bp}(^\circ\mathrm{C}) =& -76.719 + 23.992|F_1| + 5.01|F_2| - 13.332|F_3| \\ &+ 17.880|F_4| + 1.492|F_6|\end{aligned} \tag{1.10}$$

Thus, any spectral moment and hence the activities/properties of chemical compounds can be represented by contributions of corresponding fragments. This approach was further extended to molecular graphs containing heteroatoms by weighting the diagonal elements of the bond adjacency matrix.[171]

This methodology has been implemented in TOSS-MODE (TOpological Sub-Structural MOlecular Design) and TOPS-MODE (TOPological Substructural MOlecular DEsign) methods,[173] which were successfully used to assess various physicochemical properties of chemical compounds: retention indices in chromatography,[174] diamagnetic and magnetooptic properties,[175] dipole moments,[176] permeability coefficients through low-density polyethylene,[177] *etc.*), 3D-parameters[178] and a different types of biological activity (sedative/hypnotic activity,[173] anti-cancer activity,[179] anti-HIV activity,[180] skin sensitization,[181] herbicide activity,[182] affinity to A_1 adenosine receptor,[183] inhibition of cyclooxygenase,[184] antibacterial activity,[185] toxicity in *Tetrahymena pyriformis*,[186] mutagenicity,[187–189] *etc.*

1.3.1.9 Mined Subgraphs

The notion of mined subgraphs is closely linked to graph mining (or subgraph mining), a field of searching the graphs (subgraphs) specifically related to some properties or activities.[190–195] The advantage of this approach is that all relevant fragments are available for analysis without the need to consider an almost infinite number of all possible subgraphs, which allows one to select the most "useful" fragments. This methodology[196,197] is based on efficient algorithms for mining the most *frequent fragments* occurring in sets of molecular graphs, such as the AGM (Apriori-based Graph Mining) algorithm by Inokuchi *et al.*,[198] the FSG (Frequent Sub-Graphs) algorithm by Kuramochi and Karypis,[199] the chemical sub-structure discovery algorithm by Borgelt and Berthold,[200] the gSpan (graph-based Substructure pattern mining) algorithm by Yan and Han,[194] the TreeMiner algorithm by Zaki[201] and the HybridTreeMiner and CMTreeMiner algorithms by Chi, Yang and Muntz,[202,203] *etc.* The mined subgraphs approach was originally used to classify chemical structures.[204,205] "Weighted substructure mining, in conjunction with linear programming boosting,[206] allows one to build QSAR regression models involving mined fragment descriptors.[195]

1.3.1.10 Random Subgraphs

The success of different fragmentation schemes in SAR/QSAR studies strongly depends on the initial choice of relevant fragment types. Since it is unrealistic to consider all possible fragments because of their enormous number, one should always select their small subsets. However, any attempt to apply a limited subtype of them (*e.g.*, to use only chains with the user specified length) risks being inefficient because of missing of important fragments. One possible solution is to generate substructural fragments using stochastic techniques. Such an approach has been used by Graham *et al.*, who generated "tape recordings" of chemical structures from atom-bond-atom fragments extracted from molecular graphs by random walks.[207] In the MolBlaster method by Batista, Godden and Bajorath, for each molecule the program generates a "random fragment profile" representing a population of fragments generated by randomly deleting bonds in hydrogen-suppressed molecular graph.[208] This method was successfully applied in similarity-based virtual screening.[209]

1.3.1.11 Library Subgraphs

Many studies employ fixed sets of fragments taken from some libraries containing preliminary selected fragments. Thus, most additive schemes and group contribution methods have been derived using fixed sets of fragments. Some SAR/QSAR/QSPR expert systems also employ fixed sets of selected fragments and often apply an internal language specifically designed for handling the descriptors lists. For example, to describe fragments, the DEREK expert system for assessing toxicity uses the PATRAN language,[210] whereas the ALogP method[86] for predicting the octanol–water partition coefficient log P is based on the SMARTS line notation [as implemented in the MOE (Molecular Operating Environment) software suite[159]].

1.3.2 Fragments Describing Supramolecular Systems and Chemical Reactions

Using "special" bond types, molecular graphs can represent not only individual molecules but also more complex species: supramolecular systems, chemical reactions and polymers with periodic structure. For example, the ISIDA program can recognize a "coordination bond" between central metal atom and donor atoms of the ligand in the metal complexes and "hydrogen bond" in supramolecular assemblies.[32] Varnek *et al.* used fragment descriptors derived from "supramolecular" graphs in QSPR modeling of free energy and enthalpy of formation of 1 : 1 hydrogen bonded complexes.[18]

The concept of molecular graphs can also be expanded to describe chemical reactions by introducing special types of "dynamical" bonds corresponding to formation, modification and breaking of chemical bonds (for a review see ref. 211). The resulting reaction graph contains all necessary information to reconstruct both reactants and products in the corresponding reaction equation. Partial reaction graphs containing only "dynamical" bonds were used to classify and enumerate organic reactions in the framework of Ugi–Dugundji matrix formalism[212] and the Zefirov–Tratch formal-logical approach.[213,214] Vladutz condensed reactants and products of a chemical reaction into a single Superimposed Reaction Skeleton Graph (SRSG)[215] containing both dynamical and conventional (not modified in the reaction) bonds. Similar reaction graphs under the name "imaginary transition state" were also suggested by Fujita[216,217] for classification and enumeration of organic reactions. This approach has been extended recently by Varnek *et al.*[18] in Condensed Graphs of Reactions (CGRs) containing both "dynamical" and conventional bonds (Figure 1.6). Fragment descriptors derived from CGRs were used in similarity search of reactions, in reaction classification and in the development of QSPR models of the rate constant of S_N2 reactions in water.[218]

To encode reaction transformations Borodina *et al.* have developed Reacting Multilevel Neighborhood of Atom (RMNA)[219] descriptors representing an extended version of the MNA descriptors. Unlike CGRs, where reaction information is condensed, in the RMNA approach the information about modified, created or broken bonds is added to the list of the MNA descriptors generated for all products and reactants. The RMNA descriptors were applied to predict metabolic P450-mediated aromatic hydroxylation.[219]

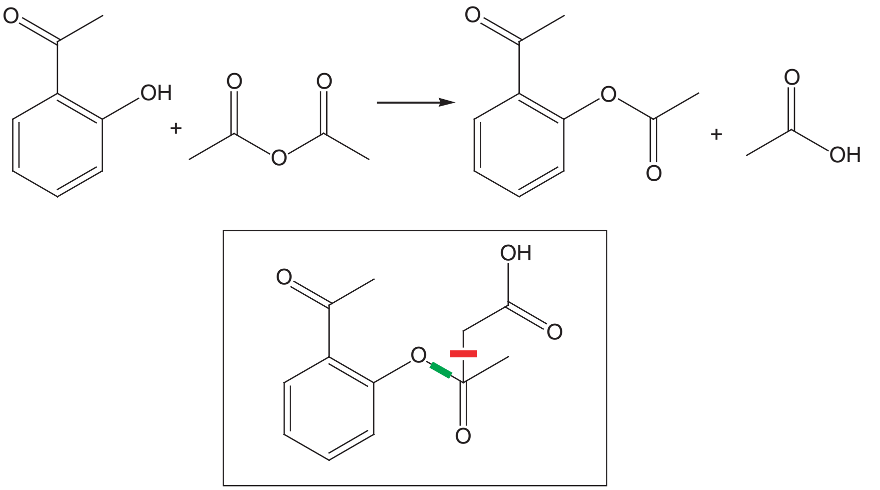


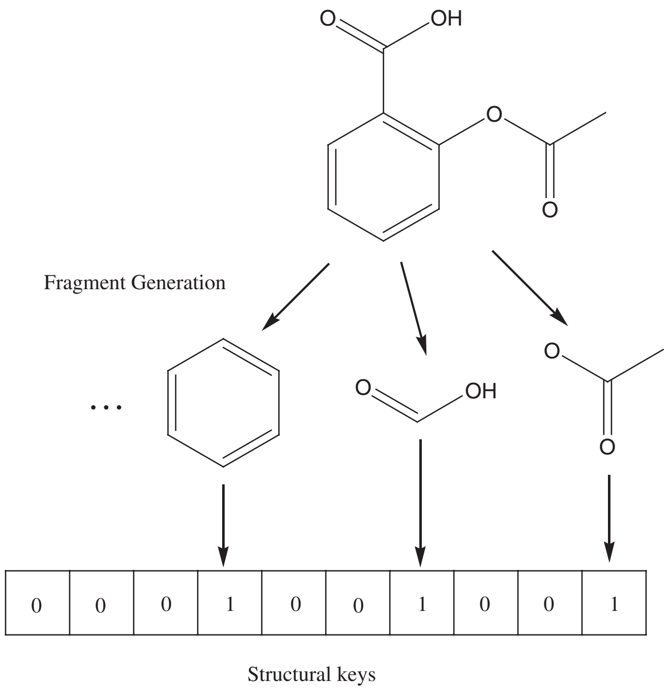
Figure 1.6 Phenol acetylation and related Condensed Graph of Reaction. "Dynamical" bonds marked with green and red correspond, respectively, to formation and breaking a single bond.

1.3.3 Storage of Fragment Information

This section discusses different techniques to store the information about molecular fragments. The most common way is present a given chemical structure as a fixed-size array (vector), in which each element corresponds to the occurrence of a given molecular fragment. Structural keys are descriptor vectors containing binary values indicating presence of absence of fragments. Since structural keys can be kept in computer memory as bit strings they are processed very rapidly, which explains their popularity in chemical database management, similarity search, SAR/QSAR studies and in virtual screening (Figure 1.7).

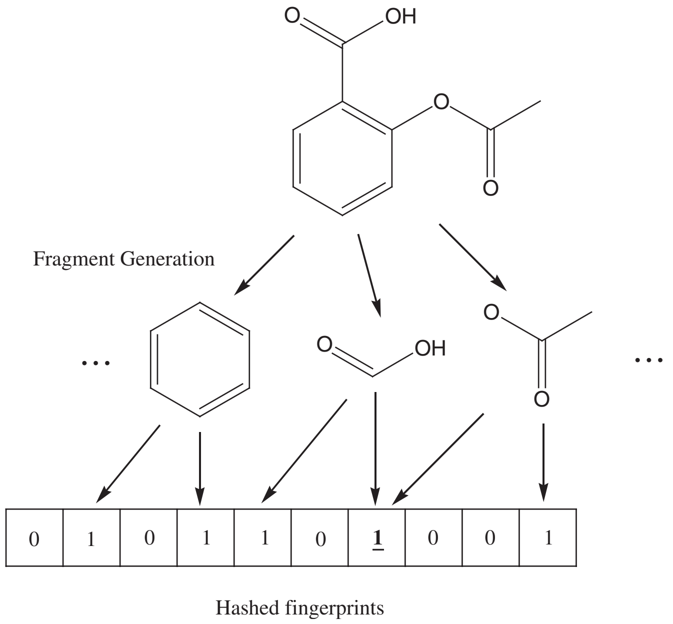
The composition and length of structural keys always depend on the choice of constituent fragments. Often, structural keys become very sparse, *i.e.*, they contain very few non-zero values. Such highly imbalanced data presentation is rather inefficient for computer processing. As a partial solution to this problem, fragment descriptors can be stored in a list containing the codes (names) of fragments "ON". Although application of lists reduces the storage's size, it is still time consuming to be used for a substructural search in large databases.

Search efficiency can be improved significantly by using hash tables, allowing one to link directly the name of descriptor and location of the descriptor's value. This technology is used in *hashed molecular fingerprints* operating with binary values (Figure 1.8). In contrast to structural keys, in molecular fingerprints each fragment is mapped onto several cells, positions of which are computed from the fragment code. The advantage of hashed fingerprints is a



0	0	0	1	0	0	1	0	0	1

Figure 1.7 Generation of structural keys for a molecule of aspirin.



0	1	0	1	1	0	**1**	0	0	1

Figure 1.8 Generation of hashed fingerprints. Each fragment leads to "switching on" of several bits. A bit with collisions is underlined and shown in bold.

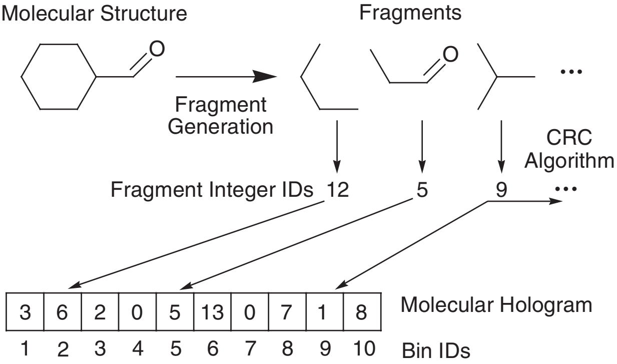


Figure 1.9 Generation of a molecular hologram. A molecule is broken into several structural fragments that are assigned fragment integer identifications (IDs) using the CRC algorithm. Each fragment is then placed in a particular bin based on its fragment integer ID corresponding to the bin ID. The bin occupancy numbers are the molecular hologram descriptors that count structural fragments in each bin. (Adapted from ref. 63.)

possibility to include a big number of fragments in a bit string of reasonable length. Their drawback is related to the existence of collisions when two or more fragments are mapped in the same bit. Nonetheless, this problem could be solved by trade-off between the length of bit string, the number of fragments types and the number of bits allocated for each fragment.

An interesting way of encoding structural information is realized in molecular holograms, which represent an integer array of bins of predetermined length (hologram length) that contains information about the occurrences of fragments. In the course of generating a molecular hologram, each fragment is coded using the SLN (SYBYL Line Notation).[220] Using the cyclic redundancy check (CRC) algorithm,[221] this code is transformed into a fragment integer ID, indicating the location of the particular bin in the molecular hologram (Figure 1.9). The occupancy of bins is then incremented by one as soon as the corresponding fragments occur. Since the hologram length *I* always smaller than the number of fragments, several different fragments map to the same bin in the molecular hologram. The resulting bin occupancy is equal to the sum of occurrence numbers of all these fragments. Molecular holograms were specially designed to be used in the Holographic QSAR (HQSAR) approach.[63]

1.3.4 Fragment Connectivity

Fragments used for building fragment descriptors can be *connected* and *disconnected*. Most applications are based on connected fragments. The point is that the indicators of presence or occurrences of disconnected fragments can always be expressed through the corresponding values obtained for connected fragments.[8] Hence, descriptors based on disconnected fragments are redundant, since they do not carry any additional information compared to their connected counterparts.

Nonetheless, in some cases disconnected fragments descriptors could simplify QSAR/QSPR equations. In particular, nonlinear models involving connected fragments can be replaced with linear models built on disconnected fragments, because the occurrences of disconnected and connected fragments are nonlinearly related. Thus, the use of disconnected fragments may be viewed as an implicit way of introducing nonlinearity into QSARs/QSPRs. If binary descriptor values are used, disconnected fragments implicitly introduce conjunctions (logical .AND.) into logical expressions instead of nonlinear terms for connected fragments. Tarasov *et al.*[222] have shown that the *compound structural descriptors* defined as combinations of unrelated fragments improve significantly the efficiency of mutagenicity predictions. Implicitly, disconnected fragments, as conjugations of binary (logical) connected fragment descriptors, were used to build probabilistic SAR models for some biological activities (see ref. 223 and references therein).

1.3.5 Generic Graphs

In contrast to QSPR studies based on complete (containing all atoms) or hydrogen-suppressed molecular graphs, assessment of biological activity, especially at the qualitative level, often requires greater generalization. In that case, it is convenient to describe chemical structures by *reduced graphs*, in which each vertex – descriptor center or pharmacophoric center – represents an atom or a group of atoms capable of interacting with biological targets, whereas each edge measures the number of bonds between them. Such a biology-oriented representation of chemical structures was invented in 1982 by Avidon *et al.* under the name Descriptor Center Connection Graphs (DCCG)[41] as a generalization of SSFN descriptors (Section 1.3.1.6).

Figure 1.10(b) shows the DCCG for phenothiazine. In this case, the reduced graph consists of 16 edges and 10 vertices corresponding to descriptor centers shown in Figure 1.10(a). Descriptor centers involve four heteroatoms (1–4; see numbering in Figure 1.10a), which can take part in donor–acceptor interaction with biomolecules and in the formation of hydrogen bonds, three methyl groups (5–7), which can take part in hydrophobic interaction with biomolecules, two benzene rings (8, 9) and one heterocycle (10), which can take part in π–π and π–cation interactions with biomolecules. Eleven edges in the DCCG labeled with positive numbers indicate the topological distances (counted as the number of bonds) between the atoms included in the corresponding descriptor centers, while the negative labels denote relations between rings within a polycyclic system. Such graphs are very useful not only as a

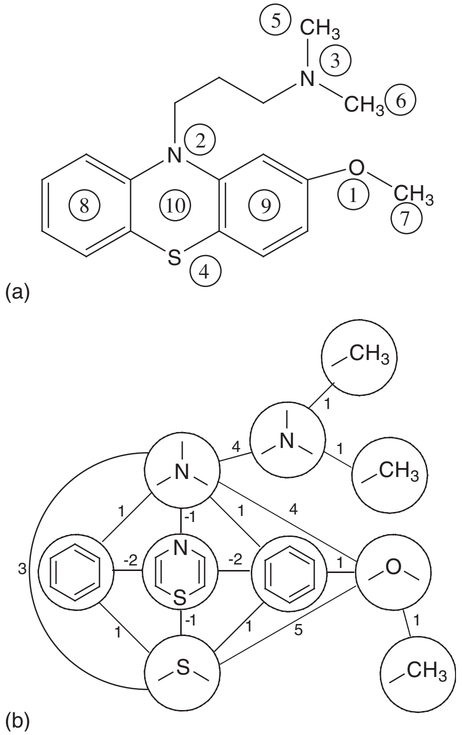


Figure 1.10 (a) Structure of phenothiazine with descriptor centers marked on it. (Adapted from ref. 41.) (b) Descriptor center connection graph for phenothiazine. (Adapted from ref. 41.)

source of biology-oriented fragment descriptors but also for pharmacophore based virtual screening.

The atom-pairs proposed by Carhart *et al.*[154] are rather similar to the SSFN descriptors. They can be considered as two-vertex connected fragments of reduced graphs, in which edges correspond to paths between certain atoms. Modifications introduced to the atom-pairs descriptors by Kearsley *et al.*[96] through encoding physicochemical properties of atoms render these fragments even more generic. In 2003 Gillet, Willett and Bradshaw (GWB) introduced another type reduced graphs and proved their high efficiency in a similarity search.[224] A GWB reduced graph consisting of six vertices and five edges is shown in Figure 1.11. Its three vertices R correspond to rings, its two vertices L to linkers, while the vertex F corresponds to a feature – an oxygen atom in this case, which can form hydrogen bonds. In contrast to DCCG, the edges of GWB reduced graphs are not labeled and correspond to ordinary chemical bonds.

An important feature of the GWB reduced graphs is a hierarchical organization of vertex labels. For example, the label Ar_n (non-hydrogen-bonding aromatic cycle) is less general than the label Ar (any aromatic cycle), which, in turn, is less general than R (any ring). Due to this feature, GWB reduced graphs

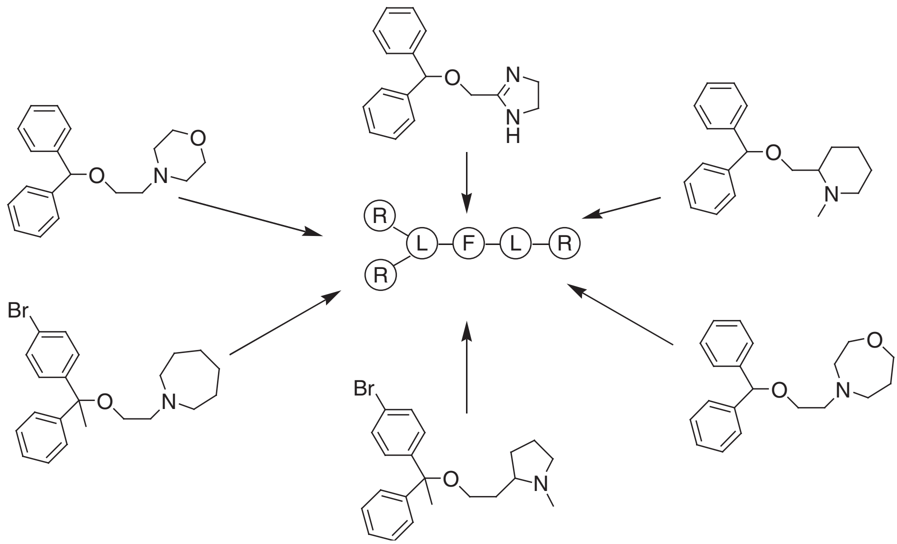


Figure 1.11 Examples of chemical structures corresponding to the same GWB reduced graph of type R/F (shown in center). (Adapted from ref. 224.)

can also be organized hierarchically, and the level of their generalization can be controlled (Figure 1.12). Besides similarity searching, fragment descriptors based on GWB reduced graphs have been applied to derive SAR models using decision trees.[225]

1.3.6 Labeling Atoms

In some cases selected atoms in molecules could be marked with special labels, indicating their particular role in a modeled property. Some examples are (i) local properties, such as atomic charges or NMR chemical shifts, which should always be attributed to a given atom(s), (ii) anchor atoms in the given scaffold to which substituents are attached (Figure 1.13), (iii) atoms forming a main chain in polymers and (iv) reaction centers in a set of reactions. Zefirov *et al.* have applied labeling in QSPR studies of pK_a[226,227] chemical NMR shifts and reaction rate constant for the acid hydrolysis of esters.[226,228] Varnek *et al.*[18] labeled hydrogen bond donor and acceptor centers to model free energies and enthalpies of formation of the 1 : 1 hydrogen-bond complexes.

1.4 Application in Virtual Screening and *In Silico* Design

This section considers the application of fragment descriptors at different stages of virtual screening and *in silico* design.

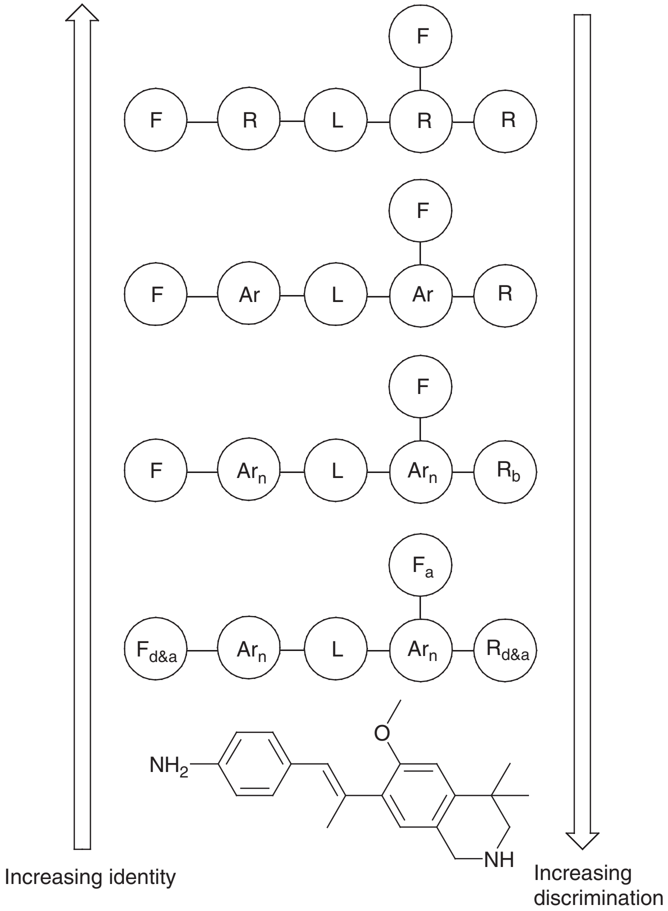


Figure 1.12 A hierarchy of GWB reduced graphs. (Adapted from ref. 224.)

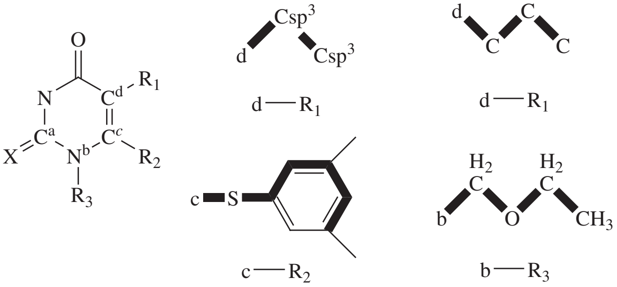


Figure 1.13 Examples of fragments with marked atoms used for modeling inhibitor activity against HIV-I reverse transcriptase for a congeneric set of HEPT derivatives.

1.4.1 Filtering

Filtering is a rule-based approach aimed to perform fast assessment of usefulness of molecules in the given context. In terms of drug design, the filtering is used to eliminate compounds with unfavorable pharmacodynamic or pharmacokinetic properties as well as toxic compounds. Pharmacodynamics considers binding drug-like organic molecules (ligands) to chosen biological target. Since the efficiency of ligand–target interactions depends on spatial complementarity of their binding sites, the filtering is usually performed with 3D-pharmacophores, representing "optimal" spatial arrangements of steric and electronic features of ligands.[229,230] Pharmacokinetics is mostly related to absorption, distribution, metabolism and excretion (ADME) related properties: octanol–water partition coefficients (log P), solubility in water (log S), blood–brain coefficient (log BB), partition coefficient between different tissues, skin penetration coefficient, *etc.*

Fragment descriptors are widely used for early ADME/Tox prediction both explicitly and implicitly. The easiest way to filter large databases concerns detecting undesirable molecular fragments (structural alerts). Appropriate lists of structural alerts are published for toxicity,[231] mutagenicity,[232] and carcinogenicity.[233] Klopman *et al.* were the first to recognize the potency of fragment descriptors for this purpose.[66,67,69] Their programs CASE,[66] MultiCASE,[97,234] as well as more recent MCASE QSAR expert systems,[235] proved to be effective tools to assess the mutagenicity[67,234,235] and carcinogenicity[69,234] of organic compounds. In these programs, sets of biophores (analogs of structural alerts) were identified and used for activity predictions. Several more sophisticated fragment-based expert systems of toxicity assessment – DEREK,[210] TopKat[236] and Rex[237] – have been developed. DEREK is a knowledge-based system operating with human-coded or automatically generated[238] rules concerning toxicophores. Fragments in the DEREK knowledge base are defined by means of the linear notation language PATRAN, which codes the information about atom, bonds and stereochemistry. TopKat uses a large predefined set of fragment descriptors, whereas Rex implements a special kind of atom-pairs descriptors (links). For more information about fragment-based computational assessment of toxicity, including mutagenicity and carcinogenicity, see ref. 239 and references therein.

The most popular filter used in drug design area is the Lipinski "rule of five",[240] which takes into account the molecular weight, the number of hydrogen bond donors and acceptors, along with the octanol–water partition coefficient log P, to assess the bioavailability of oral drugs. Similar rules of "drug-likeness" or "lead-likeness" were later proposed by Oprea,[241] Veber[242] and Hann.[243] Formally, fragment descriptors are not explicitly involved there. However, most computational approaches that assess log P are fragment-based;[244–246] whereas H-donors and acceptor sites are the simplest molecular fragments.

1.4.2 Similarity Search

The notion of molecular similarity (or chemical similarity) is one of the most useful and at the same time one of the most contradictory concepts in chemoinformatics.[247,248] The concept of molecular similarity plays an important role in many modern approaches to predicting the properties of chemical compounds, designing chemicals with a predefined set of properties and, especially, in conducting drug design studies by screening large databases containing structures of available (or potentially available) chemicals. These studies are based on the similar property principle of Johnson and Maggiora, which states: similar compounds have similar properties.[247] The similarity-based virtual screening assumes that all compounds in a database that are similar to a query compound have similar biological activity. Although this hypothesis is not always valid (see discussion in ref. 249), quite often the set of retrieved compounds is considerably enriched with actives.[250]

To achieve high efficacy of similarity-based screening of databases containing millions compounds, molecular structures are usually represented by *screens* (structural keys) or fixed-size or variable-size *fingerprints*. Screens and fingerprints can contain both 2D- and 3D-information. However, the 2D-fingerprints, which are a kind of binary fragment descriptors, dominate in this area. Fragment-based structural keys, like MDL keys,[62] are sufficiently good for handling small and medium-sized chemical databases, whereas processing of large databases is performed with fingerprints having much higher information density. Fragment-based Daylight,[251] BCI,[252] and UNITY 2D[253] fingerprints are the best known examples.

The most popular similarity measure for comparing chemical structures represented by means of fingerprints is the Tanimoto (or Jaccard) coefficient T.[254] Two structures are usually considered similar if $T > 0.85$[250] (for Daylight fingerprints[251]). Using this threshold, Taylor estimated a probability to retrieve actives as 0.012–0.50,[255] whereas according to Delaney this probability is even higher, *i.e.*, 0.40–0.60 (ref. 256) (using Daylight fingerprints[251]). These computer experiments confirm the usefulness of the similarity approach as an instrument of virtual screening.

Schneider *et al.* have developed a special technique for performing virtual screening referred to as Chemically Advanced Template Search (CATS).[257] Within its framework, chemical structures are described by means of so-called correlation vectors, each component of which is equal to the occurrence of a given atom pair divided by the total number of non-hydrogen atoms in it. Each atom in the atom pair is specified as belonging to one of five classes (hydrogen-bond donor, hydrogen-bond acceptor, positively charged, negatively charged, and lipophilic), while topological distances of up to ten bonds are also considered in the atom-pair specification. In ref. 257, the similarity is assessed by Euclidean distance between the corresponding correlation vectors. CATS has been shown to outperform the MERLIN program with Daylight fingerprints[251] for retrieving thrombin inhibitors in a virtual screening experiment.[257]

Hull *et al.* have developed the Latent Semantic Structure Indexing (LaSSI) approach to perform similarity search in low-dimensional chemical space.[258,259] To reduce the dimension of initial chemical space, the singular value decomposition method is applied for the descriptor-molecule matrix. Ranking molecules by similarity to a query molecule was performed in the reduced space using the cosine similarity measure,[260] whereas the Carhart's atom pairs[154] and the Nilakantan's topological torsions[95] were used as descriptors. The authors claim that this approach "has several advantages over analogous ranking in the original descriptor space: matching latent structures is more robust than matching discrete descriptors, choosing the number of singular values provides a rational way to vary the 'fuzziness' of the search".[258]

The issue of "fuzzification" of similarity search has been addressed by Horvath *et al.*[155–157] The first fuzzy similarity metric suggested[155] relies on partial similarity scores calculated with respect to the inter-atomic distances distributions for each pharmacophore pair. In this case the "fuzziness" enables comparison of pairs of pharmacophores with different topological or 3D distances. Similar results[156] were achieved using fuzzy and weighted modified Dice similarity metric.[260] Fuzzy pharmacophore triplets (FPT, see Section 1.3.1.6) can be gradually mapped onto related basis triplets, thus minimizing binary classification artifacts.[157] In a new similarity scoring index introduced in ref. 157, the simultaneous absence of a pharmacophore triplet in two molecules is taken into account. However, this is a less-constraining indicator of similarity than simultaneous presence of triplets.

Most similarity search approaches require only a single reference structure. However, in practice several lead compounds are often available. This motivated Hert *et al.*[261] to develop the data fusion method, which allows one to screen a database using all available reference structures. Then, the similarity scores are combined for all retrieved structures using selected fusion rules. Searches conducted on the MDL Drug Data Report database using fragment-based UNITY 2D,[253] BCI,[252] and Daylight[251] fingerprints have proved the effectiveness of this approach.

The main drawback of the conventional similarity search concerns an inability to use experimental information on biological activity to adjust similarity measures. This results in an inability to discriminate relevant and non-relevant fragment descriptors used for computing similarity measures. To tackle this problem, Cramer *et al.* [42] developed substructural analysis, in which each fragment (represented as a bit in a fingerprint) is weighted by taking into account its occurrence in active and in inactive compounds. Subsequently, many similar approaches have been described in the literature.[262]

One more way to conduct a similarity-based virtual screening is to retrieve the structures containing a user-defined set of "pharmacophoric" features. In the Dynamic Mapping of Consensus positions (DMC) algorithm[263] those features are selected by finding common positions in bit strings for all active compounds. The potency-scaled DMC algorithm (POT-DMC)[264] is a modification of DMC in which compounds activities are taken into account. The latter two methods may be considered as intermediate between conventional similarity search and probabilistic SAR approaches.

Batista, Godden and Bajorath have developed the MolBlaster method,[208] in which molecular similarity is assessed by Differential Shannon Entropy[265] computed from populations of randomly generated fragments. For the range $0.64 < T < 0.99$, this similarity measure provides with the same ranking as the

Tanimoto index *T*. However, for smaller values of *T* the entropy-based index is more sensitive, since it distinguishes between pairs of molecules having almost identical *T*. To adapt this methodology for large-scale virtual screening, Proportional Shannon Entropy (PSE) metrics were introduced.[209] A key feature of this approach is that class-specific PSE of random fragment distributions enables the identification of the molecules sharing with known active compounds a significant number of signature substructures.

Similarity search methods developed for individual compounds are difficult to apply directly for chemical reactions involving many species subdivided by two types: reactants and products. To overcome this problem, Varnek *et al.*[18] suggested condensing all participating reaction species in one molecular graph [Condensed Graphs of Reactions (CGR),[18] see Section 1.3.2] followed by its fragmentation and application of developed fingerprints in ''classical'' similarity search. Besides conventional chemical bonds (simple, double, aromatic, *etc.*), a CGR contains dynamical bonds corresponding to created, broken or transformed bonds. This approach could be efficiently used for screening of large reaction databases.

1.4.3 SAR Classification (Probabilistic) Models

Simplistic and heuristic similarity-based approaches can hardly produce as good predictive models as modern statistical and machine learning methods that are able to assess quantitatively biological or physicochemical properties. QSAR-based virtual screening consists of direct assessment of activity values (numerical or binary) of all compounds in the database followed by selection of hits possessing desirable activity. Mathematical methods used for models preparation can be subdivided into classification and regression approaches. The former decide whether a given compound is active, whereas the latter numerically evaluate the activity values. Classification approaches that assess probability of decisions are called probabilistic.

Various classification approaches have been reported to be used successfully in conjunction with fragment descriptors for building classification SAR models: the Linear Discriminant Analysis (LDA),[266,267] the Partial Least Square Discriminant Analysis (PLS-DA),[268] Soft Independent Modeling by Class Analogy (SIMCA),[269] Artificial Neural Networks (ANN),[270] Support Vector Machines (SVM),[271] Decision Trees (DT), [269,272,273] Spline Fitting with Genetic Algorithm (SFGA),[269] *etc.* Probabilistic methods usually used with fragment descriptors are: Naïve Bayes (NB)[142] and its modification implemented in PASS,[126] Binary Kernel Discrimination,[6] Inductive Logic Programming (ILP),[274] Support Vector Inductive Logic Programming (SVILP),[133] *etc.*

Numerous studies have been devoted to classification (probabilistic) approaches used in conjunction with fragment descriptors for virtual screening. Here we present several examples.

Harper *et al.* [6] have demonstrated a much better performance of probabilistic ''binary kernel discrimination'' method to screen large databases compared to backpropagation neural networks or conventional similarity search. The Carhart's atom-pairs[154] and Nilakantan's topological torsions[95] were used as descriptors.

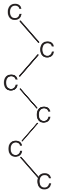
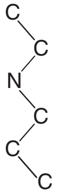
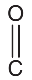
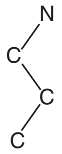
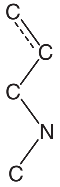
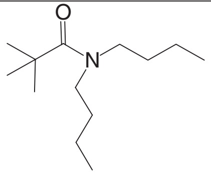
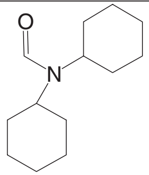
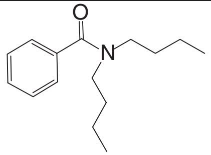
Aiming to discover new cognition enhancers, Geronikaki *et al.*[275] applied the PASS program,[126] which implements a probabilistic Bayesian-based approach, and the DEREK rule-based system[210] to screen a database of highly diverse chemical compounds. Eight compounds with the highest probability of cognition-enhancing effect were selected. Experimental tests showed that all of them possess a pronounced antiamnesic effect.

Bender, Glen *et al.* have applied[129–133] several probabilistic machine learning methods (naïve Bayesian classifier, inductive logic programming, and support vector inductive learning programming) in conjunction with circular fingerprints for making classification of bioactive chemical compounds and performing virtual screening on several biological targets. The latter of these three methods (*i.e.*, support vector inductive learning programming) performed significantly better than the other two methods.[133] The advantages of using circular fingerprints were pointed out.[131]

1.4.4 QSAR/QSPR Regression Models

The Multiple Linear Regression (MLR) method was historically the first and to date the most popular method used to develop QSAR/QSPR models with fragment descriptors (Figure 1.14). Linear models involving fragments are built in several program packages: CASE,[66–69] MULTICASE,[97,98] TRAIL,[101,102] ISIDA,[18] EMMA,[276] QSAR Builder from Pharma Algorithms[277] and some others. The Partial Least Squares (PLS) regression,[278,279] an alternative technique for building linear quantitative models, has also been successfully coupled with fragment descriptors.[63,128,280–282] This approach is efficiently used the Holographic QSAR (HQSAR)[63] (implemented in the Sybyl software[253]) and the "Generalized Fragment-Substructure Based Property Prediction Method".[282] The success of treating the fragment descriptors in PLS is explained by efficient handling of multicollinearity, which is a typical problem of fragment descriptors. Two other methods, the Group Method of Data Handling (GMDH)[283] and the more recent Maximal Margin Linear Programming Method (MMLPM),[284,285] also displayed their efficiency in building the linear models from an initial pool of highly correlated fragment descriptors.

Among nonlinear regression methods used in conjunction with fragment descriptors, the Back-Propagation Neural Networks (BPNN)[286–289] occupy a special place. It has been proved[7,8] that any molecular graph invariant can be approximated by an output of a BPNN using fragment descriptors as an input. Indeed, numerous studies have shown that the BPNN models based on fragment descriptors efficiently predict various physicochemical properties[16,290–294] and some biological activities[16,163,295] of organic compounds. A popular ASNN (Associative Neural Networks) approach consists of an ensemble of BPNN coupled with kNN correction in the space of models.[296] This technique,

Dataset	Matrix of Fragment Descriptors					Property Values
						
	0	10	1	5	0	-0.222
	0	8	1	4	0	0.973
	0	4	1	2	4	-0.066

QSAR/QSPR MODEL	$Y_{CALC} = -0.36 * N_{C-C-C=N-C-C} + 0.27 * N_{C=O} + 0.12 * N_{C-N-C*C} + ..$

Figure 1.14 General scheme of constructing linear QSAR/QSPR models based on fragment descriptors.

together with fragment descriptors, has been successfully used to model the thermodynamic parameters of metal complexation[285] and melting point of ionic liquids.[297] Besides, the Radial Basis Function Neural Networks[298] (RBFNNs) have also been used with fragment descriptors for predicting the properties of organic compounds.[285,299] The Support Vector Regression (SVR) technique[300–303] is a serious "competitor" of neural networks, as has been demonstrated in QSAR/QSPR studies[285,304] involving fragment descriptors.

In drug design, regression QSAR/QSPR models are often used to assess ADME/Tox properties or to detect "hit" molecules capable of binding a certain biological target. Thus, one could mention fragments based QSAR models for blood–brain barrier,[305] skin permeation rate,[306] blood–air[307] and tissue-air partition coefficients.[307] Many theoretical approaches to calculating the octanol–water partition coefficient log P involve fragment descriptors. In particular, it concerns the methods by Rekker,[308,309] Leo and Hansch (CLOGP),[245,310] Ghose-Crippen (ALOGP),[81–83] Wildman and Crippen,[86] Suzuki and Kudo (CHEMICALC-2),[87] Convard (SMILOGP)[88] and by Wang (XLOGP).[89,90] Fragment-based predictive models for estimation of solubility in water[311] and DMSO[311] are also available.

Benchmarking studies on various biological and physicochemical properties[305–307,312] show that QSAR/QSPR models for involving fragment descriptors in many cases outperform those built on topological, quantum, electrostatic and other types of descriptors.

1.4.5 *In Silico* Design

In this section we consider several examples of virtual screening performed on a database containing only virtual (still non-synthesized or unavailable) compounds. Virtual libraries are usually generated using combinatorial chemistry approaches.[313–315] One of simplest ways is to attach systematically user-defined substituents $R_1, R_2, \ldots, R_N$ to a given scaffold. If the list for the substituent R_i contains n_i candidates, the total number of generated structures is:

$$N = \prod_i n_i \tag{1.11}$$

although taking symmetry into account could reduce the library's size. The number of substituents R_i (n_i) should be carefully selected to avoid generation of too large a set of structures (combinatorial explosion). The "optimal" substituents could be prepared using fragments selected at the QSAR stage, since their contributions to activity (for linear models) allow one to estimate an impact of combining the fragment into larger species (R_i). In such a way, a focused combinatorial library could be generated.

The technology based on combining QSAR, generation of virtual libraries and screening stages has been implemented in the ISIDA program and applied to computer-aided design of new uranyl binders belonging to two different families of organic molecules: phosphoryl containing podands[316] and monoamides.[317] QSAR models have been developed using different machine-learning methods (multi-linear regression analysis, associative neural networks[296] and support vector machines[301]) and fragment descriptors (atom/bond sequences and augmented atoms). These models were then used to screen virtual combinatorial libraries containing up to 11000 compounds. Selected hits were synthesized and tested experimentally. Predicted uranyl binding affinity was shown to agree well with the experimental data. Thus, initial data sets were significantly enriched with new efficient uranyl binders, and one of new molecules was found to be more efficient than previously studied compounds. A similar study was conducted for the development of new 1-(2-hydroxyethoxy)methyl)-6-(phenylthio)thymine (HEPT) derivatives potentially possessing high anti-HIV activity.[318] This demonstrates the universality of fragment descriptors and the broad perspectives of their use in virtual screening and *in silico* design.

1.5 Limitations of Fragment Descriptors

Despite the many advantages of fragment descriptors they are not devoid of certain drawbacks, which deserve serious attention. Two main problems should be mentioned: (i) "missing fragments";[319] and (ii) modeling of stereochemically dependent properties.

The term "missing fragments" concerns comparison of the lists of fragments generated for the training and test sets. A test set molecule may contain fragments that, on one hand, belong to the same family of descriptors used for the modeling, and, on the other hand, are different from those in the initial pool calculated for the training set. The question arises whether the model built from that initial pool can be applied to those test set molecules? This is a difficult problem because *a priori* it is not clear if the "missing fragments" are important for the property being predicted. Several possible strategies to treat this problem have been reported. The ALOGPS program,[320] predicting lipophilicity and aqueous solubility of chemical compounds, flags calculations as unreliable if the analyzed molecule contains one or more E-state atom or bond types missed in the training set. In such a way, the program detects about 90% of large prediction errors.[319] The ISIDA program[18] calculates a consensus model as an average over the "best" models developed with different sets of fragment descriptors. Each model corresponds to its "own" initial pool of descriptors. If a new molecule contains fragments different from those in that pool, the corresponding model is ignored. As demonstrated by benchmarking studies,[285] this improves the predictive performance of the method. For each model, the NASAWIN software[99] creates a list of "important" fragments including cycles and all one-atom fragments. The test molecule is rejected if its list of "important" fragments contains those absent in the training set.[321] The LOGP program for lipophilicity predictions[322] uses a set of empirical rules to calculate the contribution of missed fragments.

The second problem of using fragment descriptors deals with accounting for stereochemical information. In fact, its adequate treatment is not possible at the graph-theoretical level and requires explicit consideration of hypergraphs.[323] However, in practice, it is sufficient to introduce special labels indicating stereochemical configuration of chiral centers or (*E*/*Z*)-isomers around a double bond, and then to use them in the specification of molecular fragments. Such an approach has been used in hologram fragment descriptors[324] as well as in the PARTAN language.[238]

1.6 Conclusion

Fragment descriptors constitute one of the most universal types of molecular descriptors. The scope of their application encompasses almost all existing areas of SAR/QSAR/QSPR studies. Their universality stems from the basic character of structural theory in chemistry as well as from the fundamental possibility of molecular graph invariants being expressed in terms of subgraph occurrence numbers.[8] The main advantages of fragment descriptors lie in the simplicity of their computation, the easiness of their interpretation as well as in efficiency of their applications in similarity searches and SAR/QSAR/QSPR modeling. Progress of their use in virtual screening could be related to the development of new types of fragments and of new mathematical approaches of their processing.

Acknowledgements

The authors thank GDRE SupraChem and ARCUS "Alsace –Russia/Ukraine" project for support and also Dr V. Solov'ev for fruitful discussions.

References

1. J. Gasteiger and T. Engel, eds., *Chemoinformatics: A Textbook*, Wiley-VCH, Weinheim, 2003.
2. J. Gasteiger, ed., *Handbook of Chemoinformatics: From Data to Knowledge.*, Wiley-VCH, Weinheim, 2003.
3. T. Engel, *J. Chem. Inf. Model.*, 2006, **46**, 2267–2277.
4. W. L. Chen, *J. Chem. Inf. Model.*, 2006, **46**, 2230–2255.
5. N. Brown, *Computing Surveys*, 2006.
6. G. Harper, J. Bradshaw, J. C. Gittins, D. V. S. Green and A. R. Leach, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 1295–1300.
7. I. I. Baskin, M. I. Skvortsova, I. V. Stankevich and N. S. Zefirov, *Dokl. Chem.*, 1994, **339**, 231–234.
8. I. I. Baskin, M. I. Skvortsova, I. V. Stankevich and N. S. Zefirov, *J. Chem. Inf. Comput. Sci.*, 1995, **35**, 527–531.
9. M. I. Skvortsova, I. I. Baskin, L. A. Skvortsov, V. A. Palyulin, N. S. Zefirov and I. V. Stankevich, *Theochem.*, 1999, **466**, 211–217.
10. M. I. Skvortsova, I. V. Stankevich, I. I. Baskin, V. A. Palyulin and N. A. Zefirov, *Doklady Akademii Nauk*, 1996, **350**, 786–788.
11. M. I. Skvortsova, I. I. Baskin, I. V. Stankevich, V. A. Palyulin and N. S. Zefirov, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 785–790.
12. M. I. Skvortsova, I. I. Baskin, O. L. Slovokhotova and N. S. Zefirov, *Doklady Akademii Nauk*, 1994, **336**, 496–499.
13. M. I. Skvortsova, I. I. Baskin, I. V. Stankevich and N. S. Zefirov, *Doklady Akademii Nauk*, 1996, **351**, 78–80.
14. N. S. Zefirov and V. A. Palyulin, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 1112–1122.

15. P. Japertas, R. Didziapetris and A. Petrauskas, *Quant. Struct.-Act. Relat.*, 2002, **21**, 23–37.
16. N. V. Artemenko, I. I. Baskin, V. A. Palyulin and N. S. Zefirov, *Russ. Chem. Bull.*, 2003, **52**, 20–29.
17. C. Merlot, D. Domine and D. J. Church, *Curr. Opin. Drug Discov. Devel.*, 2002, **5**, 391–399.
18. A. Varnek, D. Fourches, F. Hoonakker and V. P. Solov'ev, *J. Comput. Aided Mol. Des.*, 2005, **19**, 693–703.
19. S. Jelfs, P. Ertl and P. Selzer, *J. Chem. Inf. Model.*, 2007, **47**, 450–459.
20. R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors.*, Wiley-VCH Publishers, Weinheim, 2000.
21. A. I. Vogel, *Chemistry & Industry*, 1934, 85.
22. C. T. Zahn, *J. Chem. Phys.*, 1934, **2**, 671–680.
23. M. Souders, C. S. Matthews and C. O. Hurd, *Ind. Eng. Chem.*, 1949, **41**, 1037–1048.
24. M. Souders, C. S. Matthews and C. O. Hurd, *Ind. Eng. Chem.*, 1949, **41**, 1048–1056.
25. J. L. Franklin, *Ind. Eng. Chem.*, 1949, **41**, 1070–1076.
26. J. L. Franklin, *J. Chem. Phys.*, 1953, **21**, 2029–2033.
27. V. M. Tatevskii, *Doklady Akademii Nauk SSSR*, 1950, **75**, 819–822.
28. V. M. Tatevskii, E. A. Mendzheritskii and V. Korobov, *Vestnik Moskovskogo Universiteta*, 1951, **6**, 83–86.
29. H. J. Bernstein, *J. Chem. Phys.*, 1952, **20**, 263–269.
30. K. J. Laidler, *Canadian J. Chem.*, 1956, **34**, 626–648.
31. S. W. Benson and J. H. Buss, *J. Chem. Phys.*, 1958, **29**, 546–572.
32. T. L. Allen, *J. Chem. Phys.*, 1959, **31**, 1039–1049.
33. E. A. Smolenskii, *Zhurnal Fizicheskoi Khimii*, 1964, **38**, 1288–1291.
34. C. Hansch, R. M. Muir, T. Fujita, P. P. Maloney, F. Geiger and M. Streich, *J. Am. Chem. Soc.*, 1963, **85**, 2817–2824.
35. C. Hansch and T. Fujita, *J. Am. Chem. Soc.*, 1964, **86**, 1616–1626.
36. S. M. Free Jr. and J. W. Wilson, *J. Med. Chem.*, 1964, **7**, 395–399.
37. S. A. Hiller, A. B. Glaz, L. A. Rastrigin and A. B. Rosenblit, *Doklady Akademii Nauk SSSR.*, 1971, **199**, 851–853.
38. S. A. Hiller, V. E. Golender, A. B. Rosenblit, L. A. Rastrigin and A. B. Glaz, *Comput. Biomed. Res.*, 1973, **6**, 411–421.
39. V. E. Golender and A. B. Rozenblit, *Avtomatika i Telemekhanika*, 1974, 99–105.
40. V. E. Golender and A. B. Rozenblit, *Med. Chem. (Academic Press)*, 1980, **11**, 299–337.
41. V. V. Avidon, I. A. Pomerantsev, V. E. Golender and A. B. Rozenblit, *J. Chem. Inf. Comput. Sci.*, 1982, **22**, 207–214.
42. R. D. Cramer 3rd, G. Redl and C. E. Berkoff, *J. Med. Chem.*, 1974, **17**, 533–535.
43. W. E. Brugger, A. J. Stuper and P. C. Jurs, *J. Chem. Inf. Model.*, 1976, **16**, 105–110.
44. A. J. Stuper and P. C. Jurs, *J. Chem. Inf. Model.*, 1976, **16**, 99–105.

45. L. Hodes, G. F. Hazard, R. I. Geran and S. Richman, *J. Med. Chem.*, 1977, **20**, 469–475.
46. G. W. Adamson, *Proceedings of the Analytical Division of the Chemical Society*, 1977, **14**, 26–28.
47. G. W. Adamson and J. A. Bush, *Nature*, 1974, **248**, 406–407.
48. G. W. Adamson and D. Bawden, *J. Chem. Inf. Comput. Sci.*, 1975, **15**, 215–220.
49. G. W. Adamson and J. A. Bush, *Journal of the Chemical Society, Perkin Transactions 1*, 1976, 168–172.
50. G. W. Adamson and D. Bawden, *J. Chem. Inf. Comput. Sci.*, 1977, **17**, 164–171.
51. G. W. Adamson and D. Bawden, *J. Chem. Inf. Comput. Sci.*, 1976, **16**, 161–165.
52. M. Milne, D. Lefkovitz, H. Hill and R. Powers, *J. Chem. Doc.*, 1972, **12**, 183–189.
53. G. W. Adamson, J. Cowell, M. F. Lynch, A. H. W. McLure, W. G. Town and A. M. Yapp, *J. Chem. Doc.*, 1973, **13**, 153–157.
54. A. Feldman and L. Hodes, *J. Chem. Inf. Model.*, 1975, **15**, 147–152.
55. P. Willett, *J. Chem. Inf. Model.*, 1979, **19**, 159–162.
56. P. Willett, *J. Chem. Inf. Model.*, 1979, **19**, 253–255.
57. P. Willett, V. Winterman and D. Bawden, *J. Chem. Inf. Model.*, 1986, **26**, 36–41.
58. W. Fisanick, A. H. Lipkus and A. Rusinko, *J. Chem. Inf. Model.*, 1994, **34**, 130–140.
59. L. Hodes, *J. Chem. Inf. Model.*, 1989, **29**, 66–71.
60. M. J. McGregor and P. V. Pallai, *J. Chem. Inf. Model.*, 1997, **37**, 443–448.
61. D. B. Turner, S. M. Tyrrell and P. Willett, *J. Chem. Inf. Model.*, 1997, **37**, 18–22.
62. J. L. Durant, B. A. Leland, D. R. Henry and J. G. Nourse, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 1273–1280.
63. W. Tong, D. R. Lowis, R. Perkins, Y. Chen, W. J. Welsh, D. W. Goddette, T. W. Heritage and D. M. Sheehan, *J. Chem. Inf. Model.*, 1998, **38**, 669–677.
64. R. D. Cramer, *J. Am. Chem. Soc.*, 1980, **102**, 1837–1849.
65. R. D. Cramer, *J. Am. Chem. Soc.*, 1980, **102**, 1849–1859.
66. G. Klopman, *J. Am. Chem. Soc.*, 1984, **106**, 7315–7321.
67. G. Klopman and H. S. Rosenkranz, *Mutat. Res.*, 1984, **126**, 227–238.
68. G. Klopman and A. N. Kalos, *J. Comput. Chem.*, 1985, **6**, 492–506.
69. H. S. Rosenkranz, C. S. Mitchell and G. Klopman, *Mutat. Res.*, 1985, **150**, 1–11.
70. G. Klopman, M. R. Frierson and H. S. Rosenkranz, *Environmental Mutagenesis*, 1985, **7**, 625–644.
71. H. S. Rosenkranz and G. Klopman, *Progress in Clinical and Biological Research*, 1986, **209A**, 71–104.
72. G. Klopman, K. Namboodiri and A. N. Kalos, *Progress in Clinical and Biological Research*, 1985, **172**, 287–298.

73. G. Klopman, *Environmental Health Perspectives*, 1985, **61**, 269–274.
74. G. Klopman and O. T. Macina, *J. Theor. Biol.*, 1985, **113**, 637–648.
75. G. Klopman and R. Contreras, *Mol. Pharmacol.*, 1985, **27**, 86–93.
76. G. Klopman and R. E. Venegas, *Acta Pharmaceutica Jugoslavica*, 1986, **36**, 189–209.
77. G. Klopman and A. N. Kalos, *J. Theor. Biol.*, 1986, **118**, 199–214.
78. G. Klopman, O. T. Macina, E. J. Simon and J. M. Hiller, *Theochem*, 1986, **27**, 299–308.
79. G. Klopman, O. T. Macina, M. E. Levinson and H. S. Rosenkranz, *Antimicrobial Agents and Chemotherapy*, 1987, **31**, 1831–1840.
80. G. Klopman and O. T. Macina, *Mol. Pharmacol.*, 1987, **31**, 457–476.
81. A. K. Ghose and G. M. Crippen, *J. Comput. Chem.*, 1986, **7**, 565–577.
82. A. K. Ghose and G. M. Crippen, *J. Chem. Inf. Comput. Sci.*, 1987, **27**, 21–35.
83. A. K. Ghose, A. Pritchett and G. M. Crippen, *J. Comput. Chem.*, 1988, **9**, 80–90.
84. V. N. Viswanadhan, A. K. Ghose, G. R. Revankar and R. K. Robins, *J. Chem. Inf. Comput. Sci.*, 1989, **29**, 163–172.
85. A. K. Ghose, V. N. Viswanadhan and J. J. Wendoloski, *Journal of Physical Chemistry A*, 1998, **102**, 3762–3772.
86. S. A. Wildman and G. M. Crippen, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 868–873.
87. T. Suzuki and Y. Kudo, *J. Comput. Aided. Mol. Des.*, 1990, **4**, 155–198.
88. T. Convard, J.-P. Dubost, H. Le Solleu and E. Kummer, *Quant. Struct.-Act. Relat.*, 1994, **13**, 34–37.
89. R. Wang, Y. Fu and L. Lai, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 615–621.
90. R. Wang, Y. Gao and L. Lai, *Persp. Drug Discov. Design*, 2000, **19**, 47–66.
91. T. J. Hou, K. Xia, W. Zhang and X. J. Xu, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 266–275.
92. D. A. Winkler, F. R. Burden and A. J. R. Watkins, *Quantitative Structure-Activity Relationships*, 1998, **17**, 14–19.
93. H. J. Bernstein, *Trans. Faraday Soc.*, 1962, **58**, 2285–2306.
94. A. J. Kalb, A. L. H. Chung and T. L. Allen, *J. Am. Chem. Soc.*, 1966, **88**, 2938–2942.
95. R. Nilakantan, N. Bauman, J. S. Dixon and R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.*, 1987, **27**, 82–85.
96. S. K. Kearsley, S. Sallamack, E. M. Fluder, J. D. Andose, R. T. Mosley and R. P. Sheridan, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 118–127.
97. G. Klopman, *Quant. Struct.-Act. Relat.*, 1992, **11**, 176–184.
98. G. Klopman, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 78–81.
99. I. I. Baskin, N. M. Halberstam, N. V. Artemenko, V. A. Palyulin and N. S. Zefirov, in: *EuroQSAR 2002 Designing Drugs and Crop Protectants: processes, problems and solutions.*, M. Ford ed., Blackwell Publishing, 2003, pp. 260–263.
100. M. I. Kumskov, *Zhurnal Organicheskoi Khimii*, 1995, **31**, 1495–1498.

101. V. P. Solov'ev, A. Varnek and G. Wipff, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 847–858.
102. A. Varnek, G. Wipff and V. P. Solovev, *Solvent Extraction and Ion Exchange*, 2001, **19**, 791–837.
103. A. A. Gakh, E. G. Gakh, B. G. Sumpter and D. W. Noid, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 832–839.
104. G. Rucker and C. Rucker, *J. Chem. Inf. Comput. Sci.*, 1993, **33**, 683–695.
105. G. W. Adamson, J. Cowell, M. F. Lynch, W. G. Town and A. M. Yapp, *J. Chem. Soc., Perkin Trans. 1*, 1973, 863–865.
106. G. W. Adamson, S. E. Creasey, J. P. Eakins and M. F. Lynch, *J. Chem. Soc., Perkin Trans. 1*, 1973, **1**, 2071–2076.
107. W. J. Wiswesser, *J. Chem. Inf. Comput. Sci.*, 1982, **22**, 88–93.
108. D. Weininger, *J. Chem. Inf. Comput. Sci.*, 1988, **28**, 31–36.
109. D. Weininger, A. Weininger and J. L. Weininger, *J. Chem. Inf. Comput. Sci.*, 1989, **29**, 97–101.
110. G. W. Adamson and D. Bawden, *J. Chem. Inf. Model.*, 1980, **20**, 97–100.
111. G. W. Adamson and D. Bawden, *J. Chem. Inf. Model.*, 1980, **20**, 242–246.
112. G. W. Adamson and D. Bawden, *J. Chem. Inf. Comput. Sci.*, 1981, **21**, 204–209.
113. D. Qu, B. Fu, M. Muraki and T. Hayakawa, *J. Chem. Inf. Model.*, 1992, **32**, 443–447.
114. D. Qu, J. Su, M. Muraki and T. Hayakawa, *J. Chem. Inf. Model.*, 1992, **32**, 448–452.
115. D. Vidal, M. Thormann and M. Pons, *J. Chem. Inf. Model.*, 2005, **45**, 386–393.
116. V. M. Tatevskii, *The Classical Theory of the Structure of Molecules and Quantum Mechanics*, Khimiya, M., 1973.
117. V. M. Tatevskii, *The Theory of Physicochemical Properties of Molecules and Substances.*, MSU Publishing House, 1987.
118. V. M. Tatevskii, *Chemical Structure of Hydrocarbons and Regularities in Their Physicochemical Properties.*, MSU Publishing House, 1953.
119. V. M. Tatevskii, V. A. Benderskii and S. S. Yarovoi, *Methods for Calculating Physicochemical Properties of Paraffin Hydrocarbons.*, MSU Publishing House, M., 1960.
120. G. W. Adamson, M. F. Lynch and W. G. Town, *J. Chem. Soc. C*, 1971, 3702–3706.
121. G. W. Adamson, D. R. Lambourne and M. F. Lynch, *J. Chem. Soc., Perkin Trans. 1*, 1972, 2428–2433.
122. N. V. Artemenko, I. I. Baskin, V. A. Palyulin and N. S. Zefirov, *Dokl. Chem.*, 2001, **381**, 317–320.
123. I. I. Baskin, V. A. Palyulin and N. S. Zefirov, Molecular Graphs in Chemistry Studies, Kalinin, 1990.
124. I. I. Baskin, V. A. Palyulin and N. S. Zefirov, 1st All-Union Conference on Theoretical Organic Chemistry, Volgograd, 1991.
125. O. A. Rayevsky, A. M. Sapegin, V. V. Chistiakov, V. P. Solov'ev and N. S. Zefirov, *Koordinatsionnaya Khimiya*, 1990, **16**, 1175–1184.

126. V. V. Poroikov, D. A. Filimonov, Y. V. Borodina, A. A. Lagunin and A. Kos, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 1349–1355.
127. D. Filimonov, V. Poroikov, Y. Borodina and T. Gloriozova, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 666–670.
128. L. Xing and R. C. Glen, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 796–805.
129. A. Bender, H. Y. Mussa, R. C. Glen and S. Reiling, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 170–178.
130. A. Bender, H. Y. Mussa, R. C. Glen and S. Reiling, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1708–1718.
131. R. C. Glen, A. Bender, C. H. Arnby, L. Carlsson, S. Boyer and J. Smith, *IDrugs*, 2006, **9**, 199–204.
132. S. Rodgers, R. C. Glen and A. Bender, *J. Chem. Inf. Model.*, 2006, **46**, 569–576.
133. E. O. Cannon, A. Amini, A. Bender, M. J. E. Sternberg, S. H. Muggleton, R. C. Glen and J. B. O. Mitchell, *Journal of Computer-Aided Molecular Design*, 2007, **21**, 269–280.
134. J.-L. Faulon, D. P. Visco Jr. and R. S. Pophale, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 707–720.
135. J.-L. Faulon, C. J. Churchwell and D. P. Visco Jr., *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 721–734.
136. C. J. Churchwell, M. D. Rintoul, S. Martin, D. P. Visco Jr., A. Kotu, R. S. Larson, L. O. Sillerud, D. C. Brown and J. L. Faulon, *J. Mol. Graph. Model.*, 2004, **22**, 263–273.
137. W. Bremser, *Analytica Chimica Acta*, 1978, **103**, 355–365.
138. J.-E. Dubois, A. Panaye and R. Attias, *J. Chem. Inf. Comput. Sci.*, 1987, **27**, 74–82.
139. J. E. Dubois, J. P. Doucet, A. Panaye and B. T. Fan, in *Topological Indices and Related Descriptors in QSAR and QSPR*, eds. J. Devillers and A. T. Balaban, Gordon and Breach Sciences Publishers, Amsterdam, 1999, pp. 613–673.
140. Y. Xiao, Y. Qiao, J. Zhang, S. Lin and W. Zhang, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 701–704.
141. A. Bender, D. W. Young, J. L. Jenkins, M. Serrano, D. Mikhailov, P. A. Clemons and J. W. Davies, *Comb. Chem. High Throughput Screen.*, 2007, **10**, 719–731.
142. M. G. Nidhi, J. W. Davies and J. L. Jenkins, *J. Chem. Inf. Model.*, 2006, **46**, 1124–1133.
143. G. W. Adamson, J. A. Bush, A. H. W. McLure and M. F. Lynch, *J. Chem. Doc.*, 1974, **14**, 44–48.
144. MDL Information Systems, Inc., www.mdli.com.
145. E. K. F. Ahrensin *Chemical Structures*, ed. W. A. Warr, Springer, London, UK, 1988, pp. 97–111.
146. J. W. Raymond and P. Willett, *J. Comput. Aided Mol. Des.*, 2002, **16**, 521–533.
147. A. B. Rozenblit and V. E. Golender, *Logical-Combinatorial Methods in the Development of Drugs*, Zinatne, Riga, 1983.

148. T. R. Hagadone, *J. Chem. Inf. Model.*, 1992, **32**, 515–521.
149. I. L. Ruiz, C. G. Garcia and M. A. Gomez-Nieto, *J. Chem. Inf. Model.*, 2005, **45**, 1178–1194.
150. M. Stahl and H. Mauser, *J. Chem. Inf. Model.*, 2005, **45**, 542–548.
151. P. A. Bacha, H. S. Gruver, B. K. Den Hartog, S. Y. Tamura and R. F. Nutt, *J. Chem. Inf. Model.*, 2002, **42**, 1104–1111.
152. R. P. Sheridan, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1037–1050.
153. V. V. Avidon and L. A. Leksina, *Nauchno.-Tekhn. Inf., Ser. 2*, 1974, 22–25.
154. R. E. Carhart, D. H. Smith and R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.*, 1985, **25**, 64–73.
155. D. Horvath, in *Combinatorial Library Design and Evaluation: Principles, Software Tools and Applications*, A. Ghose and V. Viswanadhan eds., Marcel Dekker, New York, 2001, pp. 429–472.
156. D. Horvath and C. Jeandenans, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 680–690.
157. F. Bonachera, B. Parent, F. Barbosa, N. Froloff and D. Horvath, *J. Chem. Inf. Model.*, 2006, **46**, 2457–2477.
158. A. Schuffenhauer, P. Floersheim, P. Acklin and E. Jacoby, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 391–405.
159. MOE, Molecular Operating Environment, Chemical Computing Group Inc., Montreal, www.chemcomp.com.
160. L. Franke, E. Byvatov, O. Werz, D. Steinhilber, P. Schneider and G. Schneider, *J. Med. Chem.*, 2005, **48**, 6997–7004.
161. E. Byvatov, B. C. Sasse, H. Stark and G. Schneider, *ChemBioChem.*, 2005, **6**, 997–999.
162. R. Fleischer, P. Frohberg, A. Büge, P. Nuhn and M. Wiese, *Quant. Struct.-Act. Relat.*, 2000, **19**, 162–172.
163. S. Hatrik and P. Zahradnik, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 992–995.
164. I. I. Baskin, A. O. Ait, N. M. Halberstam, V. A. Palyulin, M. V. Alfimov and N. S. Zefirov, *Dokl. Akad. Nauk.*, 1997, **357**, 57–59.
165. G. W. Bemis and M. A. Murcko, *J. Med. Chem.*, 1996, **39**, 2887–2893.
166. G. W. Bemis and M. A. Murcko, *J. Med. Chem.*, 1999, **42**, 5095–5099.
167. M. I. Skvortsova, K. S. Fedyaev, I. I. Baskin, V. A. Palyulin and N. S. Zefirov, *Dokl. Chem.*, 2002, **382**, 33–36.
168. M. I. Skvortsova, K. S. Fedyaev, V. A. Palyulin and N. S. Zefirov, *Russian Chemical Bulletin*, 2004, **53**, 1587–1595.
169. V. B. Mnukhin, *in Mathemutical Analysis and its Applications*, Rostov-na-Donu, 1983, pp. 55–60.
170. E. Estrada, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 844–849.
171. E. Estrada, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 320–328.
172. E. Estrada, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 23–27.
173. E. Estrada, A. Pena and R. Garcia-Domenech, *J. Comput. Aided Mol. Des.*, 1998, **12**, 583–595.
174. E. Estrada and Y. Gutierrez, *Journal of Chromatography A*, 1999, **858**, 187–199.

175. E. Estrada, Y. Gutierrez and H. Gonzalez, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 1386–1399.
176. E. Estrada and H. Gonzalez, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 75–84.
177. M. P. Gonzalez, A. M. Helguera and H. G. Diaz, *Polymer*, 2004, **45**, 2073–2079.
178. E. Estrada, E. Molina and I. Perdomo-Lopez, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 1015–1021.
179. E. Estrada, E. Uriarte, A. Montero, M. Teijeira, L. Santana and E. De Clercq, *J. Med. Chem.*, 2000, **43**, 1975–1985.
180. E. Estrada, S. Vilar, E. Uriarte and Y. Gutierrez, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 1194–1203.
181. E. Estrada, G. Patlewicz and Y. Gutierrez, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 688–698.
182. M. P. Gonzalez, H. G. Diaz, R. M. Ruiz, M. A. Cabrera and R. R. de Armas, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1192–1199.
183. M. P. Gonzalez and M. D. T. Moldes, *Bull. Math. Biol.*, 2004, **66**, 907–920.
184. M. P. Gonzalez, L. C. Dias, A. M. Helguera, Y. M. Rodriguez, L. G. de Oliveira, L. T. Gomez and H. G. Diaz, *Bioorganic & Medicinal Chemistry*, 2004, **12**, 4467–4475.
185. E. Molina, H. Gonzales Diaz, M. P. Gonzalez, E. Rodriguez and E. Uriarte, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 515–521.
186. M. P. Gonzalez, H. G. Diaz, M. A. Cabrera and R. M. Ruiz, *Bioorganic & medicinal chemistry*, 2004, **12**, 735–744.
187. A. M. Helguera, M. P. Gonzalez and J. R. Briones, *Polymer*, 2004, **45**, 2045–2050.
188. M. P. Gonzalez, L. C. Dias and A. M. Helguera, *Polymer*, 2004, **45**, 5353–5359.
189. M. P. Gonzalez, M. d. C. T. Moldes, Y. Fall, L. C. Dias and A. M. Helguera, *Polymer*, 2005, **46**, 2783–2790.
190. S. Kramer, L. De Raedt and C. Helma, Seventh ACM SIGKDD international conference on Knowledge discovery and data mining, San Francisco, California, August 26–29, 2001, 2001.
191. L. De Raedt and S. Kramer, The Seventeenth International Joint Conference on Articial Intelligence, 2001.
192. S. Kramer and L. De Raedt, The eighteenth International Conference on Machine Learning, 2001.
193. A. Inokuchi, in *Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM'04)* IEEE Computer Society, 2004, pp. 415–418.
194. X. Yan and J. Han, in *Proceedings of the 2002 IEEE International Conference on Data Mining*, IEEE Computer Society, Washington DC, USA, 2002, pp. 721–724.
195. H. Saigo, T. Kadowaki and K. Tsuda, International Workshop on Mining and Learning with Graphs 2006, 2006.
196. T. Asai, K. Abe, S. Kawasoe, H. Arimura, H. Satamoto and S. Arikawa, in *SIAM SDM'02*, 2002.

197. Y. Chi, R. R. Muntz, S. Nijssen and J. N. Kok, *Fundamenta Informaticae*, 2005, **66**, 161–198.
198. A. Inokuchi, T. Washio and H. Motoda, 4th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD), Lyon, France, September 2000, 2000.
199. M. Kuramochi and G. Karypis, 1st IEEE Conference on Data Mining, 2001.
200. C. Borgelt, T. Meinl and M. Berthold, in *Proceedings of the 1st international Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations* ACM Press, New York, NY, Chicago, Illinois, August 21–21, 2005, 2005, pp. 6–15.
201. M. J. Zaki, in*Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM Press, Edmonton, Alberta, 2002, pp. 71–80.
202. Y. Chi, Y. Yang and R. R. Muntz, in *The 16th International Conference on Scientific and Statistical Database Management (SSDBM'04), June 2004*, Editon edn., 2004.
203. Y. Chi, Y. Yang, Y. Xia and R. R. Muntz, in *The Eighth Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD'04)*, May 2004 Editon, Springer, London, UK, 2004.
204. L. Dehaspe, H. Toivonen and R. D. King, in *4th International Conference on Knowledge Discovery and Data Mining*, R. Agrawal, P. Stolorz and G. Piatetsky-Shapiro eds., AAAI Press, 1998, pp. 30–36.
205. M. Deshpande, M. Kuramochi and G. Karypis, in *Proceedings of the Third IEEE international Conference on Data Mining (November 19–22, 2003). ICDM.*, IEEE Computer Society, Washington, DC, 2003, pp. 35–49.
206. A. Demiriz, K. P. Bennett and J. Shawe-Taylor, *Mach. Learn.*, 2002, **46**, 225–254.
207. D. J. Graham, C. Malarkey and M. V. Schulmerich, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1601–1611.
208. J. Batista, J. W. Godden and J. Bajorath, *J. Chem. Inf. Model.*, 2006, **46**, 1937–1944.
209. J. Batista and J. Bajorath, *J. Chem. Inf. Model*, 2007, **47**, 59–68.
210. D. M. Sanderson and C. G. Earnshaw, *Hum. Exp. Toxicol.*, 1991, **10**, 261–273.
211. L. Chen, in *Handbook of Chemoinformatics*, ed. J. Gasteiger, Wiley-VCH, Weinheim, 2003, vol. 1, pp. 348–388.
212. J. Dugundji and I. Ugi, *Topics Curr. Chem.*, 1973, **39**, 19–64.
213. N. S. Zefirov and S. S. Trach, *Chemica Scripta*, 1980, **15**, 4–12.
214. N. S. Zefirov, *Accounts of Chemical Research*, 1987, **20**, 237–243.
215. G. Vladutz, in *Approaches to Chemical Reaction Searching*, ed. P. Willett, Gower, London, 1986, pp. 202–220.
216. S. Fujita, *J. Chem. Inf. Comput. Sci.*, 1986, **26**, 205–212.
217. S. Fujita, *J. Chem. Inf. Comput. Sci.*, 1987, **27**, 120–126.
218. F. Hoonakker, PhD Thesis, ULP, Strasbourg, 2008.

219. Y. Borodina, A. Rudik, D. Filimonov, N. Kharchevnikova, A. Dmitriev, V. Blinova and V. Poroikov, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1998–2009.
220. S. Ash, M. A. Cline, R. W. Homer, T. Hurst and G. B. Smith, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 71–79.
221. D. E. Knuth, *Sorting and searching*, Addison-Wesley, Reading, MA, 1988.
222. V. A. Tarasov, O. N. Mustafaev, S. K. Abilev and V. A. Mel'nik, *Russian Journal of Genetics*, 2005, **41**, 814–821.
223. C. S. Kadyrov, L. A. Tjurina, V. D. Simonov and V. A. Semenov, *Machine Search for Chemicals with Specified Properies*, FAN, Tashkent, 1989.
224. V. J. Gillet, P. Willett and J. Bradshaw, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 338–345.
225. E. J. Barker, E. J. Gardiner, V. J. Gillet, P. Kitts and J. Morris, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 346–356.
226. I. I. Baskin, N. I. Zhokhova, V. A. Palyulin, A. A. Ivanova, A. N. Zefirov and N. S. Zefirov, in *Book of Abstracts of the XVI European Symposium on Quantitative Structure-Activity Relationships and Molecular Modelling, 10–17 September 2006, Mediterranean Sea, Italy*, 2006, p. 206.
227. A. A. Ivanona, I. I. Baskin, V. A. Palyulin and N. S. Zefirov, *Dokl. Chem.*, 2007, **413**, 90–94.
228. N. I. Zhokhova, I. I. Baskin, V. A. Palyulin, A. N. Zefirov and N. S. Zefirov, *Dokl. Chem.*, 2007, **417**, 282–284.
229. O. F. Guener, *Pharmacophore Perception, Development, and Use in Drug Design.*, Wiley-VCH Publishers, Weinheim, 2000.
230. T. Langer and R. D. Hoffman, *Pharmacophores and Pharmacophore Searches.*, Wiley-VCH Publishers, Weinheim, 2000.
231. J. Wang, L. Lai and Y. Tang, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 1173–1189.
232. J. Kazius, R. McGuire and R. Bursi, *J. Med. Chem.*, 2005, **48**, 312–320.
233. A. R. Cunningham, H. S. Rosenkranz, Y. P. Zhang and G. Klopman, *Mutat. Res.*, 1998, **398**, 1–17.
234. G. Klopman and H. S. Rosenkranz, *Mutat. Res.*, 1994, **305**, 33–46.
235. G. Klopman, S. K. Chakravarti, N. Harris, J. Ivanov and R. D. Saiakhov, *SAR QSAR Environ. Res.*, 2003, **14**, 165–180.
236. V. K. Gombar, K. Enslein, J. B. Hart, B. W. Blake and H. H. Borgstedt, *Risk Anal.*, 1991, **11**, 509–517.
237. P. N. Judson, *Pestic. Sci.*, 1992, **36**, 155–160.
238. P. N. Judson, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 148–153.
239. M. D. Barratt and R. A. Rodford, *Curr. Opin. Chem. Biol.*, 2001, **5**, 383–388.
240. C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Adv. Drug Deliv. Rev.*, 2001, **46**, 3–26.
241. T. I. Oprea, *J. Comput. Aided Mol. Des.*, 2000, **14**, 251–264.
242. D. F. Veber, S. R. Johnson, H. Y. Cheng, B. R. Smith, K. W. Ward and K. D. Kopple, *J. Med. Chem.*, 2002, **45**, 2615–2623.

243. M. M. Hann and T. I. Oprea, *Curr. Opin. Chem. Biol.*, 2004, **8**, 255–263.
244. A. A. Petrauskas and E. A. Kolovanov, *Perspectives in Drug Discovery and Design*, 2000, **19**, 99–116.
245. A. J. Leo, *Chem. Rev.*, 1993, **93**, 1281–1306.
246. I. V. Tetko and D. J. Livingstone, in *Comprehensive Medicinal Chemistry II: In silico tools in ADMET*, B. Testa and H. van de Waterbeemd eds., Elsevier, 2006, vol. 5, pp. 649–668.
247. A. M. Johnson and G. M. Maggiora eds., *Concepts and Applications of Molecular Similarity*, John Willey & Sons, New York, 1990.
248. N. Nikolova and J. Jaworska, *QSAR & Combinatorial Science*, 2003, **22**, 1006–1026.
249. H. Kubinyi, *Persp. Drug Discov. Design*, 1998, **9–11**, 225–252.
250. Y. C. Martin, J. L. Kofron and L. M. Traphagen, *J. Med. Chem.*, 2002, **45**, 4350–4358.
251. Daylight Chemical Information Systems Inc., http://www.daylight.com (accessed May 2008).
252. Barnard Chemical Information Ltd., http://www.bci.gb.com/ (accessed May 2008).
253. Tripos Inc., http://www.tripos.com (accessed May 2008).
254. P. Jaccard, *Bull. Soc. Vaud. Sci. Nat.*, 1901, **37**, 241–272.
255. R. Taylor, *J. Chem. Inf. Comput. Sci.*, 1995, **35**, 59–67.
256. J. S. Delaney, *Mol. Divers.*, 1996, **1**, 217–222.
257. G. Schneider, W. Neidhart, T. Giller and G. Schmid, *Angew. Chem. Int. Ed.*, 1999, **38**, 2894–2896.
258. R. D. Hull, S. B. Singh, R. B. Nachbar, R. P. Sheridan, S. K. Kearsley and E. M. Fluder, *J. Med. Chem.*, 2001, **44**, 1177–1184.
259. R. D. Hull, E. M. Fluder, S. B. Singh, R. B. Nachbar, S. K. Kearsley and R. P. Sheridan, *J. Med. Chem.*, 2001, **44**, 1185–1191.
260. P. Willett, J. M. Barnard and G. M. Downs, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 983–996.
261. J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby and A. Schuffenhauer, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1177–1185.
262. A. Ormerod, P. Willett and D. Bawden, *Quant. Struct.-Act. Relat.*, 1989, **8**, 115–129.
263. J. W. Godden, J. R. Furr, L. Xue, F. L. Stahura and J. Bajorath, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 21–29.
264. J. W. Godden, F. L. Stahura and J. Bajorath, *J. Med. Chem.*, 2004, **47**, 5608–5611.
265. J. W. Godden and J. Bajorath, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 1060–1066.
266. P. C. Jurs, T. R. Stouch, M. Czerwinski and J. N. Narvaez, *J. Chem. Inf. Comput. Sci.*, 1985, **25**, 296–308.
267. R. I. Zalewski and J. Jasiczak, *J. Chem. Inf. Model.*, 1994, **34**, 179–183.
268. H. Sun, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1506–1514.
269. J. J. Sutherland, L. A. O'Brien and D. F. Weaver, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1906–1915.

270. M. Brinn, P. T. Walsh, M. P. Payne and B. Bott, *SAR and QSAR in Environmental Research*, 1993, **1**, 169–210.
271. C. Helma, T. Cramer, S. Kramer and L. De Raedt, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1402–1411.
272. A. Rusinko III, M. W. Farmen, C. G. Lambert, P. L. Brown and S. S. Young, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 1017–1026.
273. M. Wagener and V. J. Van Geerestein, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 280–292.
274. R. D. King, A. Srinivasan and L. Dehaspe, *Journal of Computer-Aided Molecular Design*, 2001, **15**, 173–181.
275. A. A. Geronikaki, J. C. Dearden, D. Filimonov, I. Galaeva, T. L. Garibova, T. Gloriozova, V. Krajneva, A. Lagunin, F. Z. Macaev, G. Molodavkin, V. V. Poroikov, S. I. Pogrebnoi, F. Shepeli, T. A. Voronina, M. Tsitlakidou and L. Vlad, *J. Med. Chem.*, 2004, **47**, 2870–2876.
276. D. E. Petelin, D. V. Sukhachev, I. I. Baskin, V. A. Palyulin and N. S. Zefirov, 9th All-Union Conference on Chemical Informatics, 1992.
277. Pharma Algorithms, http://pharma-algorithms.com/qsar_builder.htm (accessed May 2008).
278. H. Martens and T. Naes, *Multivariate Calibration*, Wiley, Chichester, etc, 1989.
279. A. Höskuldsson, *J. Chemometrics*, 1988, **2**, 211–228.
280. L. Xing, R. C. Glen and R. D. Clark, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 870–879.
281. D. Butina and J. M. R. Gola, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 837–841.
282. M. Clark, *J. Chem. Inf. Comput. Sci.*, 2005, **45**, 30–38.
283. H. R. Madala and A. G. Ivakhnenko, *Inductive Learning Algorithms for Complex System Modeling*, CRC Press, Boca Raton, Ann Arbor, London, Tokyo, 1994.
284. A. V. Antonov, I. V. Tetko, M. T. Mader, J. Budczies and H. W. Mewes, *Bioinformatics*, 2004, **20**, 644–652.
285. I. V. Tetko, V. P. Solov'ev, A. V. Antonov, X. Yao, J. P. Doucet, B. Fan, F. Hoonakker, D. Fourches, P. Jost, N. Lachiche and A. Varnek, *J. Chem. Inf. Model.*, 2006, **46**, 808–819.
286. D. E. Rumelhart, G. E. Hinton and R. J. Williams, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations.*, D. E. Rumelhart and J. L. McClelland eds., MIT Press, Cambridge, MA, 1986, pp. 318–362.
287. J. Zupan and J. Gasteiger, *Neural Networks in Chemistry*, Wiley-VCH, Weinheim, 1999.
288. D. A. Winkler and F. R. Burden, *Methods Mol. Biol.*, 2002, **201**, 325–367.
289. N. M. Halberstam, I. I. Baskin, A. Palyulin Vladimir and N. S. Zefirof, *Russian Chemical Reviews*, 2003, **72**, 629–649.
290. I. I. Baskin, V. A. Palyulin and N. S. Zefirov, *Doklady Akademii Nauk*, 1993, **332**, 713–716.

291. N. V. Artemenko, V. A. Palyulin and N. S. Zefirov, *Doklady Chemistry (Translation of the chemistry section of Doklady Akademii Nauk)*, 2002, **383**, 114–116.
292. N. I. Zhokhova, I. I. Baskin, V. A. Palyulin, A. N. Zefirov and N. S. Zefirov, *Russian Chemical Bulletin (Translation of Izvestiya Akademii Nauk, Seriya Khimicheskaya)*, 2003, **52**, 1885–1892.
293. N. I. Zhokhova, I. I. Baskin, V. A. Palyulin, A. N. Zefirov and N. S. Zefirov, *Russian Journal of Applied Chemistry (Translation of Zhurnal Prikladnoi Khimii)*, 2003, **76**, 1914–1919.
294. N. I. Zhokhova, I. I. Baskin, V. A. Palyulin, A. N. Zefirov and N. S. Zefirov, *Journal of Structural Chemistry*, 2004, **45**, 626–635.
295. T. M. Martin and D. M. Young, *Chem Res Toxicol*, 2001, **14**, 1378–1385.
296. I. V. Tetko, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 717–728.
297. A. Varnek, N. Kireeva, I. V. Tetko, I. I. Baskin and V. P. Solov'ev, *J. Chem. Inf. Model.*, 2007, **47**, 1111–1122.
298. E. Hartman, D. Keeler and J. Kawalski, *Neural Computation*, 1990, **2**, 210–215.
299. J. Tetteh, T. Suzuki, E. Metcalfe and S. Howells, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 491–507.
300. B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, MA, London, England, 2002.
301. V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995.
302. N. Christianini and J. Shawe-Taylor, *An introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, 2000.
303. R. Herbrich, *Learning Kernel Classifiers: Theory and Algorithms*, MIT Press, 2002.
304. P. Lind and T. Maltseva, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1855–1859.
305. A. R. Katritzky, M. Kuanar, S. Slavov, D. A. Dobchev, D. C. Fara, M. Karelson, W. E. Acree Jr., V. P. Solov'ev and A. Varnek, *Bioorg. Med. Chem.*, 2006, **14**, 4888–4917.
306. A. R. Katritzky, D. A. Dobchev, D. C. Fara, E. Hur, K. Tamm, L. Kurunczi, M. Karelson, A. Varnek and V. P. Solov'ev, *J. Med. Chem.*, 2006, **49**, 3305–3314.
307. A. R. Katritzky, M. Kuanar, D. C. Fara, M. Karelson, W. E. Acree Jr, V. P. Solov'ev and A. Varnek, *Bioorg. Med. Chem.*, 2005, **13**, 6450–6463.
308. R. Mannhold, R. F. Rekker, C. Sonntag, A. M. ter Laak, K. Dross and E. E. Polymeropoulos, *J. Pharm. Sci.*, 1995, **84**, 1410–1419.
309. G. G. Nys and R. F. Rekker, *Eur. J. Med. Chem.*, 1973, **8**, 521–535.
310. A. Leo, P. Y. C. Jow, C. Silipo and C. Hansch, *J. Med. Chem.*, 1975, **18**, 865–868.
311. K. V. Balakin, N. P. Savchuk and I. V. Tetko, *Curr. Med. Chem.*, 2006, **13**, 223–241.
312. A. Varnek, N. Kireeva, I. V. Tetko, Baskin II and V. P. Solov'ev, *J. Chem. Inf. Model.*, 2007, **47**, 1111–1122.