Chapter 9

Fragment Descriptors in Structure–Property Modeling and Virtual Screening

Alexandre Varnek

Abstract

This chapter reviews the application of fragment descriptors at different stages of virtual screening: filtering, similarity search, and direct activity assessment using QSAR/QSPR models. Several case studies are considered. It is demonstrated that the power of fragment descriptors stems from their universality, very high computational efficiency, simplicity of interpretation, and versatility.

Key words: Fragmental approach, Fragment descriptors, QSAR, QSPR, Filtering, Similarity, Virtual screening, In silico design

1. Introduction

Chemoinformatics aims to discover active and/or selective ligands for biologically related targets by conducting screening, ideally, of all possible compounds against all possible targets, or at least, in practice, available libraries of compounds against main target families [1]. One can hardly imagine to screen experimentally the chemical universe containing from 10^{12} to 10^{180} drug-like compounds [2] against the biological target universe. Nowadays, the number of experimentally screened compounds does not exceed several millions per biological target, whereas a single inexpensive computational study allows one to screen the libraries up to 10^{12} molecules and this number tends to grow up with the evolution of hardware and related software tools. Therefore, this is not surprising that the virtual, or in silico, screening approaches play a key role in chemogenomics.

Virtual screening is usually defined as a process in which large libraries of compounds are automatically evaluated using computational techniques [3]. Its goal is to discover putative hits in large databases of chemical compounds (usually ligands for biological

Jürgen Bajorath (ed.), *Chemoinformatics and Computational Chemical Biology*, Methods in Molecular Biology, vol. 672, DOI 10.1007/978-1-60761-839-3_9, © Springer Science+Business Media, LLC 2011

targets) and remove molecules predicted to be toxic or those possessing unfavorable pharmacodynamic or pharmacokinetic properties. Generally, two types of virtual screening are known: structure-based and ligand-based. The former explicitly uses 3D structure of a biological target at the stage of hit detection, whereas the latter uses only information about structure of small molecules and their properties (activities). Structure-based virtual screening (docking, 3D pharmacophores) has been described in series of review articles, *see* [4–6] and references therein.

In this paper ligand-based virtual screening involving fragment descriptors is discussed. Fragment descriptors represent selected substructures (fragments) of 2D molecular graphs and their occurrences in molecules; they constitute one of the most important types of molecular descriptors [7]. Their main advantage is related to simplicity of their calculation, storage and interpretation (*see* review articles [8–12]). Substructural fragment are information-based descriptors [13] which tend to code the information stored in molecular structures. This contrasts with knowledge-based (or semiempirical) descriptors issued from the mechanistic consideration. Selected descriptors form a "chemical space" in which each molecule is represented as a vector. Due to their versatility, fragment descriptors could be efficiently used to create a chemical space which separates active and non-active compounds.

Historically, molecular fragments were used in first additive schemes developed in the 1950s to estimate physicochemical properties of organic compounds by Tatevskii [14, 15], Bernstein [16], Laidler [17], Benson and Buss [18], and others. The Free–Wilson method [19], one of the first QSAR approaches invented in 1960s, is based on the assumption of the additivity of contributions of structural fragments to the biological activity of the whole molecule. Later on, fragment descriptors were successfully used in expert systems able to classify chemical compounds as active or inactive with respect to certain type of biological activity. Hiller [20, 21], Golender and Rosenblit [22, 23], Piruzyan, Avidon et al. [24, 25], Cramer [26], Brugger, Stuper and Jurs [27, 28], and Hodes et al. [29] pioneered in this field.

An important class of fragmental descriptors, so-called *screens* (structural *keys*, *fingerprints*), has been developed in the 1970s [30–34]. As a rule, they represent bit strings which can effectively be stored and processed by computers. Although their primary role is to provide efficient substructure searching capabilities in large chemical databases, they are efficiently used for similarity searching [35, 36], to cluster large data sets [37, 38], to assess chemical diversity [39], as well as to conduct SAR [40] and QSAR [41] studies. Nowadays, application of modern machine-learning techniques significantly improves predictive performance of structure–property models based on fragment descriptors.

This paper briefly reviews the application of fragment descriptors in virtual screening of large libraries of organic compounds focusing mostly on its three stages: (1) filtering, (2) similarity search, and (3) direct activity/property assessment using QSAR/QSPR models. A particular attention will be paid to new approaches in structure–property modeling (ensemble modeling, applicability domain, inductive learning transfer) and in mining of chemical reactions. Most of examples described here concerns the ISIDA (In SIlico Design and Data Analysis) platform for virtual screening.

2. Types of Fragment Descriptors

Due to their enormous diversity, one could hardly review all types of 2D fragment descriptors used for structural search in chemical database or in SAR/QSAR studies. Here, we focus on some of them which are the most efficiently used in virtual screening and in silico design of organic compounds.

According to Lounkine et al. [42], there exists four major strategies to fragment design: knowledge-based, synthetically oriented, random, and systematic and hierarchical. The knowledge-based methods are based on chemical and pharmaceutical expertise. As examples, one could mention fragments dictionaries for ADME predictions [43] or toxicity alerts [44, 45], and “privileged” substructures recurrent in families of bioactive compounds [46, 47]. In retrosynthetic fragmentation methods, substructures are obtained by breaking bonds in molecules described by cataloged chemical reactions [48, 49]. The main underlying idea is that the resulting fragments can be chemically re-combined in different ways. The most known example of such fragmentation is *Retrosynthetic Combinatorial Analysis Procedure* (RECAP) approach [50] which defines 11 chemical bond types where a cleavage can occur. In random molecular fragmentation methods [42], substructures populations are generated for selected molecules by random deletion of bonds in their connectivity tables followed by sampling of the resulting fragments. Comparing fragments distributions obtained for the set of molecules of given activity class, one can identify class-specific substructures which could be used in virtual screening.

Systematic and hierarchical approaches are based on the predefined rules of fragmentation. Generally, molecular fragments could be classified with respect to their topology (atom-based, chains, cycles, polycycles, etc.), information content of vertices in molecular graphs (atoms, groups of atoms, pharmacophores, descriptor centers) and the level of abstraction when some information concerning atom and bond types is omitted. Some popular fragmentation schemes are discussed below.

Purely structural fragments are used as descriptors in ACD/Labs [51], NASAWIN [52], ISIDA [12], and some other programs. These are 2D subgraphs in which all atoms and/or bonds are represented explicitly and no information about their properties is used. Their typical example is sequences of atoms and/or bonds of variable length, branch fragments, saturated and aromatic cycles (polycycles), and atom-centered fragments (ACF). The latter consist of a single central atom surrounded by one or several shells of atoms with the same topological distance from the central one. The ACF were invented by Tatevskii [14] and Benson and Buss [18] in the 1950s as elements of additive schemes for predicting physicochemical properties of organic compounds. In the early 1970s, Adamson [53] investigated the distribution of one shell ACF in some chemical databases with respect to their possible application as screens. Hodes reinvented one shell ACF as descriptors in SAR studies under the name *augmented atoms* [29], and also suggested *ganglia augmented atoms* [54] representing two shells ACF with generalized second-shell atoms. Later on, one shell ACF were implemented by Baskin et al. in the NASAWIN [52] software and by Solov′ev and Varnek in ISIDA [12] package (*see* Fig. 1). Atom-centered fragments with arbitrary number of shells were implemented by Filimonov and Poroikov in the PASS [55] program as *multilevel neighborhoods of atoms* [56], by Xing and Glen as *tree structured fingerprints* [57], by Bender et al. as *atom environments* [58, 59] and *circular fingerprints* [60–62], and by Faulon as *molecular signatures* [63–65].

It has been found that characterizing atoms only by element types is too specific for similarity searching and therefore does not provide sufficient flexibility needed for large-scaled virtual screening. For that reason, numerous studies were devoted to increase the informational content of fragment descriptors by adding some useful empirical information and/or by representing a part of molecular graph implicitly. The simplest representatives of those descriptors were *atom pairs and topological multiplets* based on the notion of *descriptor center* representing an atom or a group of atoms which could serve as centers of intermolecular interactions. Usually, descriptor centers include heteroatoms,

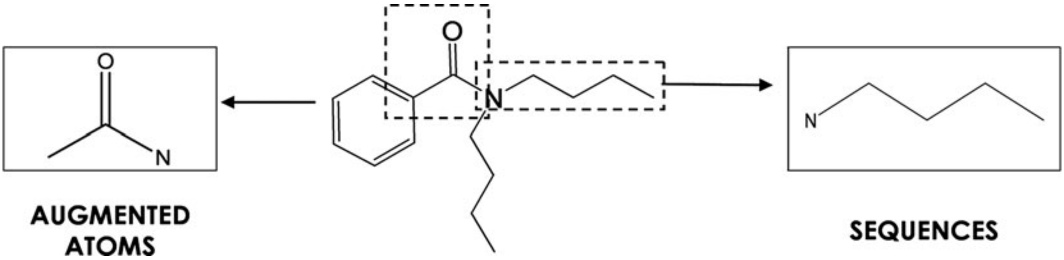


Fig. 1. Decomposition of a chemical structure into fragments. Examples of *sequences* and *augmented atoms* used as descriptors in the ISIDA program [12].

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: $\mathbf{AT}_i$-$\mathbf{AT}_j$-***Dist***, 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 in order to indicate the presence or absence of the corresponding features in studied chemical structures.

The atom pairs were first suggested for SAR studies by Avidon under the name *SSFN* (*Substructure Superposition Fragment Notation*) [25, 66]. Then they were independently reinvented by Carhart and co-authors [67] 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 rather popular in virtual screening. *Topological Fuzzy Bipolar Pharmacophore Autocorrelograms* (*TFBPA*) [68] 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). These descriptors were successfully applied in virtual screening against a panel of 42 biological targets using similarity search based on several fuzzy and non-fuzzy metrics [69], performing only slightly less well than their 3D counterparts [68]. *Fuzzy Pharmacophore Triplets* (*FPT*) by Horvath [70] is an extention of *TFBPA* for three sites pharmacophores. An important innovation in the *FPT* concerns accounting for proteolytic equilibrium as a function of pH [70]. Due to this feature, even small structural modifications leading to a $\mathrm{pK_a}$ shift, may have a profound effect on fuzzy pharmocophore triples. As a result, these descriptors efficiently discriminate structurally similar compounds exhibiting significantly different activities [70] and, therefore, they have been successfully used both in similarity search experiments [68–70] and in structure–property modeling [71].

Some other topological triplets should be mentioned. Thus, *Similog pharmacophoric keys* by Jacoby [72] represent triplets of binary coded types of atoms (pharmacophoric centers) and topological distances between them. 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 [73] represent triplets of MOE atom types separated by binned topological distances. Structure–property models obtained by support vector machine method with these descriptors have been successfully used for virtual screening of COX-2 inhibitors [74] and $\mathrm{D_3}$ dopamine receptor ligands [75].

Topological torsions by Nilakantan et al. [76] is a sequence of four consecutively bonded atoms AT_i-AT_j-AT_k-AT_l, where each atom is characterized by a number of parameters similarly to atoms in Carhart's pairs. In order to enhance efficiency of virtual screening, Kearsley et al. [77] suggested to assign atoms in 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. Kuz'min et al. [78, 79] used both connected and disconnected combinations of 4 atoms ("simplex" fragments) as descriptors in SAR/QSAR studies.

In contrast to QSPR studies based mostly on the use of complete (containing all atoms) or hydrogen-suppressed molecular graphs, handling biological activity at the qualitative level, often demands more abstractions. Namely, it is rather convenient to approximate chemical structures by *reduced graphs*, in which each vertex is an atom or a group of atoms (descriptor or pharmacophoric center), whereas each edge is a topological distance $Dist_{ij}$. Such biology-oriented representation of chemical structures was suggested by Avidon et al. as descriptor center connection graphs [25]. Gillet, Willett, and Bradshaw have proposed the GWB-reduced graphs which use the hierarchical organization of vertex labels. This allows one to control the level of their generalization which may explain their high efficiency in similarity searching.

An alternative scheme of reducing molecular graph proposed by Bemis and Murcko [80, 81] involves four-level hierarchical scheme of molecular structure simplification: (1) full molecular structure with all atoms; (2) structure without hydrogen atoms; (3) *scaffolds*, i.e., structures without substituents (which are deleted recursively by means of eliminating the "leaves" of molecular graph); and (4) *molecular frameworks*, i.e., scaffolds, in which all heteroatoms are substituted by carbon atoms, while all multiple bonds are replaced by single bonds. This presentation of molecular graph was found very useful for diversity analysis of large databases [80, 81].

3. Application of Fragment Descriptors in Virtual Screening and In Silico Design

In this chapter, the use of fragment descriptors is considered at different stages of virtual screening: filtering, similarity search, and obtaining and application of SAR/QSAR models.

3.1. Filtering

Filtering is a rule-based approach aimed to perform fast assessment of useful or useless molecules (in the given context). In drug design, this is used to discard toxic compounds as well as those possessing unfavorable pharmacodynamic or pharmacokinetic properties. 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 [82, 83]. Pharmacokinetics concerns mostly 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 [84], mutagenicity [85], and carcinogenicity [86]. Klopman et al. were the first to recognize the potency of using fragmental descriptors for this purpose [87–90]. Their programs CASE [87], MultiCASE [91, 92], as well as more recent MCASE QSAR expert systems [93] proved to be effective tools to assess mutagenicity [88, 92, 93] and carcinogenicity [90, 92] of organic compounds. In these programs, sets of biophores (analogs of structural alerts) were identified and used for activity predictions. A number of more sophisticated fragment-based expert systems of toxicity assessment – DEREK [94], TopKat [95], and Rex [96] – have been developed. DEREK is a knowledge-based system operating with human-coded or automatically generated [97] rules about toxicophores. Fragments in the DEREK knowledge base are defined by means of 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*). To read more information about fragment-based computational assessment of toxicity, including mutagenicity and carcinogenicity, *see* review [98] and references therein.

The most popular filter used in drug design area is based on the Lipinski "rule of five" [99], 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 by Oprea [100], Veber [101], and Hann [102]. Formally, fragment descriptors are not explicitly involved there. However, many computational approaches to assess log *P* are fragment-based [51, 103, 104]; whereas H-donors and acceptor sites are simplest molecular fragments.

3.2. Similarity Search

The similarity-based virtual screening is based on an assumption that all compounds in a chemical database, which are similar to a query compound, could also have similar biological activities. Although this hypothesis is not always valid (*see* discussion in [105]), quite often the set of retrieved compounds is enriched in actives [106].

To achieve high efficacy of similarity-based screening of databases containing millions compounds, molecular structures are usually represented either by *screens* (structural keys) or by fixed-size or variable-size *fingerprints*. Screens and fingerprints may 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 [40], 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, such as Daylight [107], BCI [108], and UNITY 2D [109] fingerprints.

The most popular similarity measure for comparing chemical structures represented by means of fingerprints is the Tanimoto (or Jaccard) coefficient T [110]. Two structures are usually considered similar if $T > 0.85$ [106]. Using this threshold and Daylight fingerprints [107], Martin estimated only 30% of a chance to retrieve actives.

In the CATS (*Chemically Advanced Template Search*) approach by Schneider et al. [111], the chemical structures are described by vectors, each component of which is equal to atom pair occurrence divided by the total number of non-hydrogen atoms. Each atom in these atom pairs is attributed to one of five classes: hydrogen bond donor, hydrogen bond acceptor, positively charged, negatively charged, and lipophilic. Topological distances of up to ten bonds are considered in the atom-pair specification. Similarity search with CATS was shown efficient in virtual screening experiments [111].

Hull et al. have developed the *Latent Semantic Structure Indexing* (LaSSI) approach to perform similarity search in low-dimensional chemical space [112, 113]. 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 [114], whereas the Carhart's atom pairs [67] and the Nilakantan's topological torsions [76] 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" [112].

The issue of "fuzziness" in similarity search was developed by Horvath et al. [68–70] both for the atom pairs and Fuzzy pharmacophore triplets (FPT). The first fuzzy similarity metrics suggested in [68] uses partial similarity scores calculated with respect to the inter-atomic distances distributions for each pharmacophore pair. In that case, the "fuzziness" enables to compare pairs of pharmacophores with different topological or 3D distances. Fuzzy pharmacophore triplets (*see* Subsection 2) can be gradually mapped onto related basis triplets, thus minimizing binary classification artifacts [70]. In the similarity scoring index introduced in reference [70], both simultaneous absence and presence of a pharmacophore triplet in two molecules are taken into account.

Most of similarity search approaches require only a single reference structure. However, in practice several compounds with the same type of biological activity are often available. This motivated Hert et al. [115] 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 [109], BCI [108], and Daylight [107] 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 leads to inability to discriminate between relevant and non-relevant fragment descriptors being used for computing similarity measures. To tackle this problem, Cramer et al. [26] 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. Later on, many similar approaches have been described in the literature [116].

Another way to perform 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 by Godden et al. [117] those features are selected by finding common positions in bit strings for all active compounds. The *potency-scaled DMC* algorithm (POT-DMC) [118] 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 et al. have developed the MolBlaster method [119], in which molecular similarity is assessed by *Differential Shannon Entropy* [120] 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 the smaller values of T the entropy-based index is a more sensitive, since it distinguishes between pairs of molecules having almost identical *T*. To adapt this methodology for large-scale virtual screening, the *Proportional Shannon Entropy* (PSE) metrics was introduced [121]. A key feature of this approach is that class-specific PSE of random fragment distributions enables the identification of the molecules sharing a significant number of signature substructures with known active compounds. Another approach based on random fragments has been developed by Lounkine et al. [42]. Comparison of distributions of random fragments obtained for the set of molecules of given activity class, they have extracted *Activity Class Characteristic Substructures* (ACCS) which only occur in compounds having similar activity. Combinations of ACCS carry compound class-specific information and therefore, they can be encoded as class directed fingerprints and used to search databases for novel active compounds.

Another way to perform potency-related similarity search concerns selection of fragment descriptors selected for QSAR modeling. This strategy of "tailored similarity" [122] has been used by Fourches [123] for similarity search of anticonvulsants in Maybridge and NCI databases. At the first step, QSAR models based on fragment descriptors have been obtained for the training set of 48 compounds. Then, atom/bond sequences involved in the model were used to build chemical space in which a similarity search has been performed.

Sometimes, similarity search based on explicit molecular fragments is not able to explain unexpected large activity difference of the molecules, chemical structures of which look similar ("activity cliffs" [124, 125]). In this case, application of topological pharmacophores (especially those accounting for proteolytic equilibrium effects) and chemically meaningful similarity scores could be particularly useful. Figure 2 illustrates a typical strength of

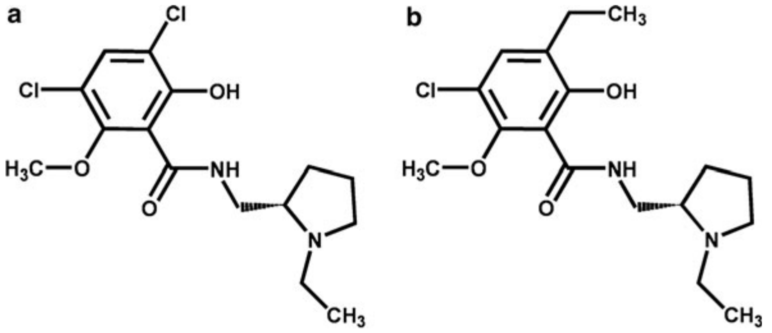


Fig. 2. State-of-the-art similarity evaluations would all agree that the compounds (a) and (b) are virtually identical. However, these molecules actually display significantly different biological activities. Due to its pK_a-sensitive pharmacophore feature flagging scheme, the FPT-based similarity scoring recognizes the difference between them because the compound (a) is an anion, whereas (b) is neutral at pH $= 7$.

fuzzy pharmacophores driven similarity searching, which is able to explain apparent “activity cliffs”. Typically, substitution of an ethyl group by a halogen atom, both flagged as “hydrophobes” by pharmacophore feature flagging routines, would leave the overall pharmacophore pattern unchanged and the two compounds would be virtually undistinguishable (near null dissimilarity score). However, this apparently harmless chemical modification triggers a ionization propensity change of a close proteolytic group and practically toggles the state of an ionic center in the molecule, with important effects on activity. Fuzzy Pharmacophore Triplets [70, 126] successfully take this phenomenon into account and therefore do not overestimate the similarity of the two molecules (Fig. 2).

3.3. SAR/QSAR/QSPR Models

Simplistic and heuristic similarity-based approaches can hardly produce as good predictive models as modern machine-learning methods able to assess biological or physicochemical properties. SAR/QSAR-based virtual screening consists in direct assessment of activity values (qualitative or quantitative) of all compounds in the database followed by selection of hits possessing desirable activity. Generally, approaches of two types – classification and regression – are used in the modeling. The former assesses a probability that a given compound belongs to a particular class (e.g. active or not active) whereas the latter numerically evaluates the activity values. Several examples of SAR/QSAR studies involving fragment descriptors are given below.

Harper et al. [127] have demonstrated a good performance of classification using *binary kernel discrimination* method to screen large databases when Carhart’s atom-pairs [67] and Nilakantan’s topological torsions [76] are used as descriptors.

Aiming to discover new cognition enhancers, Geronikaki et al. [128] applied the PASS program [55], which implements a probabilistic Bayesian-based approach, and the DEREK rule-based system [94] to screen a database of highly diverse chemical compounds. Eight compounds with the highest probability of cognition-enhancing effect were selected. Experimental tests have shown that all of them possessed a pronounced antiamnesic effect.

Bender et al. [58–62] have applied several classification machine-learning methods (naïve Bayesian classifier, inductive logic programming, and support vector inductive learning programming) in combination with circular fingerprints to perform the classification of bioactive chemical compounds and to carry out virtual screening on several biological targets. It has been shown that the performance of support vector inductive learning programming was significantly better than the other two methods [62].

Regression QSAR/QSPR models are used to assess ADME/Tox properties or to detect "hit" molecules capable to bind a certain biological target. Available in the literature fragments based QSAR models for blood–brain barrier [129], skin permeation rate [130], blood–air [131] and tissue–air partition coefficients [131] could be mentioned as examples. Many theoretical approaches of calculation of octanol-water partition coefficient log *P* involve fragment descriptors. The methods by Rekker [132, 133], Leo and Hansch (CLOGP) [103, 134], Ghose-Crippen (ALOGP) [135–137], Wildman and Crippen [138], Suzuki and Kudo (CHEMICALC-2) [139], Convard (SMILOGP) [140], and Wang (XLOGP) [141, 142] represent just a few modern examples. Fragment-based predictive models for estimation solubility in water [143] and DMSO [143] are available.

Benchmarking studies performed in references [129–131, 144] show that QSAR/QSPR models for various biological and physicochemical properties involving fragment descriptors are, at least, as robust as those involving topological, quantum, electrostatic and other types of descriptors.

In fact, classical QSAR has been developed for relatively small congener datasets. Below, we describe some strategies to improve predictive performance of the models developed on relatively large or too small structurally diverse datasets: ensemble modeling, "divide and conquer" technique and inductive learning transfer approach implemented into the ISIDA program package. It should be noted that some ISIDA models for biological activities, ADME properties, aqueous solubility, and stability constants of metals in solution are available for the users via INTERNET interface at http://infochim.u-strasbg.fr.

3.3.1. Ensemble Modeling

Relationships between chemical structures of compounds and their properties may have a very complex nature. As a consequence, a single QSAR approach may be insufficient to reflect the structure–activity relationships accurately enough. Therefore, in order to improve performance of predictions, one could use Consensus Model (CM) approach which combines several individual models [144–146]. There exist several possible ways to generate ensembles of models. The ISIDA program builds many individual models on one same training set issued from different initial pools of descriptors, each of which corresponds to a given fragmentation type. Only models for which leave-one out cross-validation correlation coefficient Q^2 is larger than a user-defined threshold are selected. Then, for each query compound, the program calculates the predicted property as an arithmetic mean of values obtained with the selected models (excluding, according to Grubbs's test [147], those leading to outlying values). This approach has been successfully used to obtain predictive MLR models for various ADME related properties (Skin Permeation

Rate [129], Blood–Air and Tissue–Air Partition Coefficients [131], Blood–Brain Barrier Permeation [129]), some biological activities [148], thermodynamic parameters of metal complexation and extraction [149, 150], free energies of hydrogen-bond complexes [12], and melting points of ionic liquids [144]. The Stochastic QSAR Sampler (SQS) program builds individual models issued from different descriptor subsets selected in genetic algorithm-based variable selection procedure. SQS has been successfully used to build predictive consensus models involving FPT descriptors for various biological activities [71, 151].

One can also build consensus model combining different machine-learning approaches. Recently, QSAR models for aquatic toxicity ($pIGC_{50}$) of organic molecules against *Tetrahymena pyriformis* have been obtained in the framework of collaborative project between six research teams [152]. Initial dataset was randomly split into a training set (644 compounds) and a test set (339 compounds) to afford an external validation of training set models. Incidentally, a second test set (110 compounds) has become available after the model building was completed. Each group used their favorite machine-learning approaches and descriptors (Dragon [153], MolconnZ [153] and ISIDA fragments), as well as their definition of models applicability domains. Totally, from 9 to 15 individual models have been obtained and applied to predict $pIGC_{50}$ for compounds in both test sets. Applicability domain assessment leads to significant improvement of the prediction accuracy for both test sets but dramatically reduces the chemical space coverage of the models. To increase the model coverage, different types of consensus models were developed by averaging predicted toxicity values computed from all models, with or without taking into account their respective applicability domains. In all cases, consensus models leads to better prediction accuracy for the external test sets as well as to the largest space coverage, as compared to any individual constituent model. Thus, on the Regression Error Curves plot, the curve corresponding to the consensus model lays higher than the curves of individual models (Fig. 3).

It has been shown [151, 154] that the variance of predicted value for a query molecule could be used as criterion of the prediction performance of the consensus models. More individual models converge toward one number, more reliable the predicted value is.

3.3.2. "Divide and Conquer" Technique

For large structurally diverse datasets, "*Divide and Conquer*" (DC) strategy could be particularly useful. It consists in split of the initial dataset into smaller congener subsets followed by obtaining the *local* QSAR models on each subset. The local models together with *global* ones obtained for the whole initial set are then used for consensus model calculations on the external test

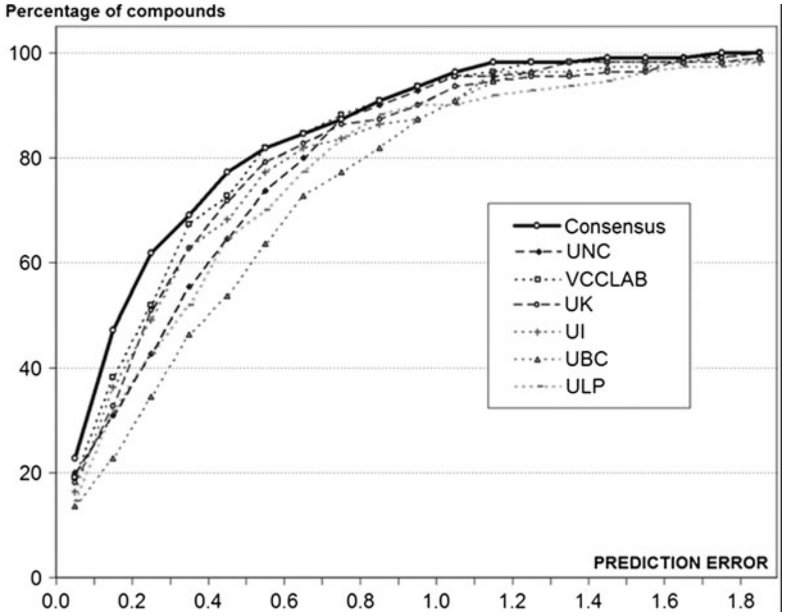


Fig. 3. Percentage of compounds for the test set two (containing 110 compounds) versus prediction errors [152]. Individual models were prepared by six teams participating in the project: *UNC* University of North Carolina at Chapel Hill in USA, *ULP* Louis Pasteur University in France, *UI* University of Insubria in Italy, *UK* University of Kalmar in Sweden, *VCCLAB* Virtual Computational Chemistry Laboratory in Germany, and *UBC* University of British Columbia in Canada.

set. In that case, the applicability domain must be respected in order to avoid application the model to the query compounds dissimilar to the related training (sub)set.

The DC approach [145] has been applied to develop QSAR models for intrinsic aqueous solubility (log *S*) for the set of 1,630 compounds compiled from the references [155–158]. The initial set has been split onto four subsets using an algorithm combing both hierarchical and non-hierarchical clustering approaches [12, 159]. Both for the initial set and for each of subsets, several individual linear models have been selected according to leave-one out cross-validation correlation coefficient. The prediction performance of the models has been tested on the external test set of 412 compounds. Two consensus models (CM) were used: conventional CM involving only global models, and DC–CM involving both global and local models. For the linear correlation log *S* (predicted) versus. log *S* (experimental), root-mean squared error of DC–CM (0.86 log *S* units) is significantly smaller than that of the conventional CM (1.13 log *S* units). Thus, these calculations demonstrate that predictive performance of the DC models significantly outperforms that of linear conventional models [145].

3.3.3. Inductive Learning Transfer Approach

QSAR modeling of pharmacokinetics properties represent of real challenge because in many cases experimental data are available only for relatively small and structurally diverse data sets. In conventional calculations (Single Task Learning, STL), the models are developed for a given property "from the scratch" without any involvements of available information for other related properties. For small initial data sets, they may fail because of lack of experimental information. In that case, Multi-Task Learning (MTL) and Feature Net (FN) [160] approaches integrating the knowledge extracted from different data sets could become a reasonable solution. In MTL, the knowledge is cumulated when the models are simultaneously trained for several related properties. In FN, estimated values of related properties are used as descriptors.

Higher performance of MTL and FN approaches over STL has been demonstrated in QSAR modeling of tissue–air partition coefficients (log *K*) [145, 161]. The initial dataset contained 11 different log *K* types obtained for a diverse set containing 199 organic compounds for human (*H*) and rat (*R*) species [131]. For each molecule, experimental data were not systematically available for all tissues and species. Thus, individual datasets included, respectively, 138, 35, 42, 30, 34 and 38 compounds for *H*-blood, *H*-brain, *H*-fat, *H*-liver, *H*-kidney and *H*-muscle, and 59, 99, 100, 27 and 97 compounds for *R*-brain, *R*-fat, *R*-liver, *R*-kidney and *R*-muscle partition coefficients. Associative Neural Networks (ASNN) approach [162] and fragment descriptors were used to build the models. In three layers neural networks, each neuron in the initial layer corresponded to one molecular descriptor. Hidden layer contained from three to six neurons, whereas the output layer contained one (for STL and FN) or 11 (MTL) neurons, corresponding to the number of simultaneously treated properties. In STL and MTL calculations, only fragment descriptors were used as an input. In FN calculations, the models were built only for one target property, whereas other ten properties served as complementary descriptors. Each model was validated using external fivefold cross-validation procedure [163]. The model was accepted if squared determination coefficient (R^2) for the linear correlation between predicted and experimental property values exceeds a threshold of $R^2 > 0.5$. Figure 4 shows that conventional STL modeling results to predictive models only for four properties corresponding to relatively large (about 100 compounds and more) data sets: *H*-blood, *R*-fat, *R*-liver, and *R*-muscle. Application of MTL and FN approaches allowed us to significantly improve the reliability of the calculations: predictive models were obtained for nine types of partition coefficients tissue–air (Fig. 4), *see* details in reference [161].

3.3.4. Applicability Domain

The question arises whether QSAR models built on the training set of limited size (usually, several hundred molecules) could be

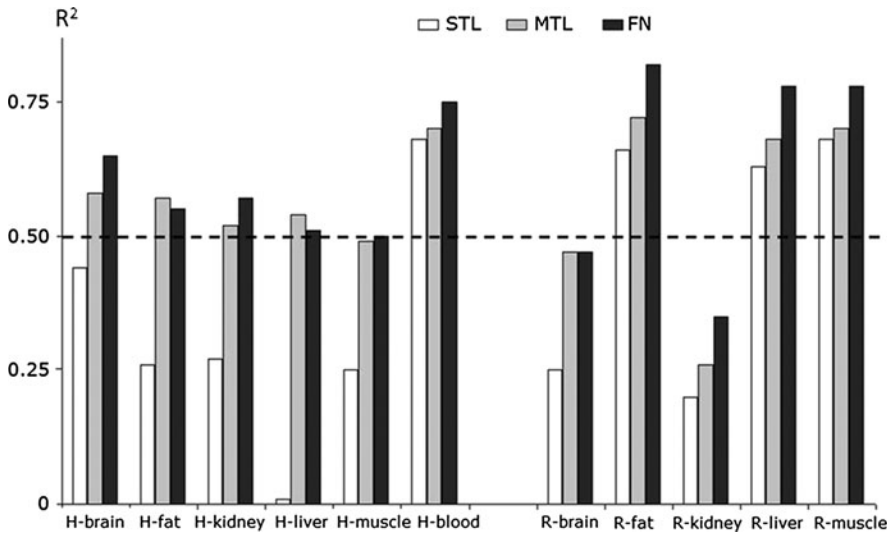


Fig. 4. Performance of different learning strategies to predict Human or Rat air tissue partition coefficient. MTL and FN calculations involved all 11 studied properties. The horizontal line at $R^2 > 0.5$ corresponds to model acceptance threshold (*see* details in [161]).

successfully applied to predict properties of the molecules different from training examples? The question is not trivial because defining an applicability domain (AD) amounts to the calibration of a meta-model based on its own specific attributes and equations. For a given query molecule, the AD is supposed to assess a “predictability score” allowing user to take a decision concerning the application of QSAR model associated to this AD.

AD definition research is nowadays a hot topic in the QSAR field. Typically, state-of-the-art AD models can be roughly classified into range-based [164–166], distance-based [167–169] and density-based [167–171] approaches which could be applied to the models involving any type of descriptors. On the other hand, there exist some AD approaches specifically related to fragment descriptors [145, 149, 172]. Thus, the *Fragment Control* (FC) algorithm [145, 149, 172] prevents to apply a given model to query molecule containing molecular fragments which do not occur in the initial pool of descriptors. This is a very simple but rather efficient approach which may significantly improve the quality of predictions. Figure 5 shows that statistical parameters of the models accepted by FC are rather close to those obtained in external cross-validation for the training set, which is not the case for the rejected models.

Unlike FC, *Model Fragment Control* (MFC) technique [172] is model-dependent and deals with only fragment descriptors involved in the model. Let us suggest that a given model involves N_{tot} descriptors. Each individual molecule in the training set contains $N_i \leq N_{\text{tot}}$ descriptors, where N_i varies from $N_{\min}$ to

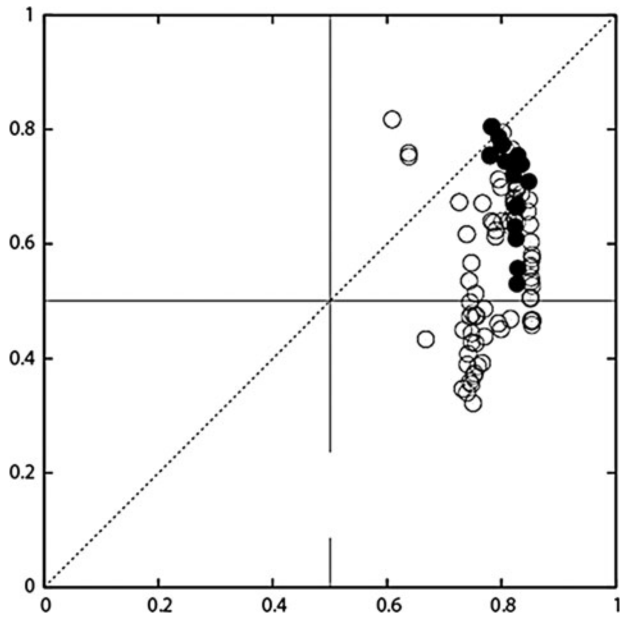


Fig. 5. Predictive performance of individual classification models developed for antibiotic activity for the compounds from the French National Library with Naïve Bayes method. ROC AUC of cross-validation for the training set (4,563 compounds, 62 actives) versus that obtained for the external test set (122 molecules). Each point represents a model issued from different initial descriptors pool with (*black cycles*) or without (*empty cycles*) accounting for the model's applicability domain. The *horizontal* and *vertical lines* at ROC AUC = 0.5 correspond to the prediction performances of a random selection ("no model").

$N_{\max}$ ($N_{\min}$, $N_{\max} \leq N_{\text{tot}}$). The query compound is discarded if the related N_{query} value is outside of this range. MCF is a robust procedure for discarding meaningless queries [172]. Thus, MCF discarded alkans (which are definitely not metal binders) from the set of molecules where QSPR models for metal complexation have been applied. Surprisingly, the range-based, distance-based and FC methods do not recognize alkans as being outside of AD.

Applying an AD, one should always look for a trade-off between accuracy of prediction and test set coverage – the number of the query molecules accepted by AD. More restrictive AD as a better prediction performance can be achieved for smaller fraction of the test set. Application of consensus models may significantly increase the test set coverage. The point is that AD discards some individual models involved in the consensus model, but usually not all of them. This has been demonstrated by Zhu et al. [152, 154] in the combinatorial QSAR modeling of chemical toxicants tested against *T. pyriformis*. The consensus model involving nine individual models was applied to two external test sets. Each individual model was associated with its applicability domain. The Mean Average Error (MAE) of predictions with the consensus model (0.27 and 0.34 $pIGC_{50}$ units for test sets one and two, respectively) was equal or lower than that for the individual models (0.27–0.44 and 0.33–0.43, respectively). On the other hand, the consensus calculation covered all molecules in the test sets one and two, whereas the coverage of the individual models varied from 80 to 97%, and from 43 to 98%; respectively. Thus, a joint application of the ensemble modeling and applicability domain approaches leads to reasonable balance between test set coverage and prediction accuracy [152, 154].

3.4. In Silico Design

In this section, we consider examples of virtual screening performed on a database containing only virtual (still non-synthesized or unavailable) compounds. Generation of virtual libraries is usually performed using combinatorial chemistry approaches [173–175]. One of simplest ways is to attach systematically user-defined substituents R_1, R_2, …, 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 $N = \prod_i n_i$, although taking symmetry into account could reduce the library's size. The number n_i of substituents R_i should be carefully selected in order to avoid a generation of too large set of structures (combinatorial explosion). The "optimal" substituents could be prepared using fragments selected at the QSAR stage, since their contributions into 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 into ISIDA and applied to computer-aided design of new uranyl binders belonging to two different families of organic molecules: phosphoryl containing podands [176] and monoamides [146]. QSAR models have been developed using different machine-learning methods (multi-linear regression analysis, associative neural networks [177] and support vector machines [178]) and fragment descriptors (atom/bond sequences and augmented atoms). Then, these models were used to screen virtual combinatorial libraries containing up to 11,000 compounds. Selected hits were synthesized and tested experimentally. Experimental data correspond well to predicted uranyl binding affinity. Thus, initial data sets were significantly enriched with new efficient uranyl binders, and one of hits was found more efficient than previously studied compounds. A similar study was conducted for development of new 1-[2-(hydroxyethoxy)methyl]-6-(phenylthio) thymine (HEPT) derivatives potentially possessing high anti-HIV activity [148]. This demonstrates universality of fragment descriptors and broad perspectives of their use in virtual screening and in silico design.

4. Mining Chemical Reactions Data Using Condensed Reaction Graphs Approach

Compared to the huge number of reported QSAR and similarity search applications to datasets of individual molecules, very few articles are devoted to chemical reactions. Indeed, chemical reactions are difficult objects because they involve several species of two different types: reactants and products. The "Condensed Graph of Reaction" (CGR) approach [179–181] opens new perspectives in the mining of reaction databases since it allows one to transform several 2D molecular graphs describing a chemical reaction into one single graph. Besides conventional chemical bonds (simple, double, aromatic, etc.), a CGR contains dynamical bonds corresponding to created, broken or transformed bonds. Thus, a chemical reactions database can be transformed into a set of "pseudo-compounds" to which most chemoinformatics methods developed for individual molecules can be applied. Here, we briefly discuss application of CGR approach for the reactions classification, similarity search and quantitative structure–reactivity modeling.

Reactions Classification. A possibility to use CGRs for the analysis of the content of reaction databases has been described in [182]. A sample containing 3,983 Diels-Alder (DA) and 736 metathesis (MT) reactions has been selected from the *ChemInform* and *Reflib* databases using queries given on Fig. 6. All selected reactions were then transformed into CGRs followed by their fragmentation into atom/bond sequences containing at least one dynamical bond. Then, the hierarchical clustering has been performed using Tanimoto similarity coefficient as a metrics. This resulted in four distinct clusters two of which contained exclusively MT reactions, one cluster included exclusively DA reactions,

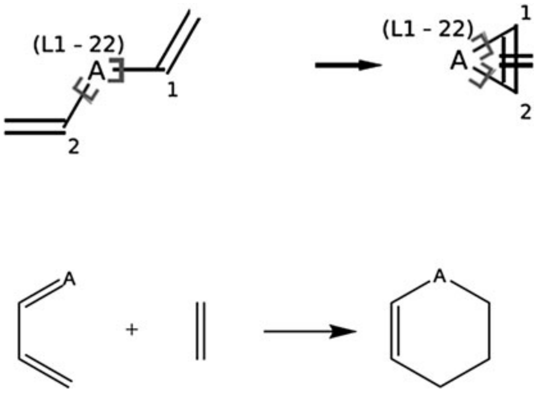


Fig. 6. Queries used for substructural search of metathesis (*top*) and Diels–Alder (*bottom*) reactions in the *ChemInform* and *Reflib* databases. "A" corresponds to any non-hydrogen atom.

and one cluster contained a mixture of DA and MT reactions. Detailed analysis of this mixed cluster shows that 28 reactions initially attributed to MT, in fact, represent *Domino Heck-Diels-Alder* (DHDA) reactions proceeding in two steps: a single bond formation between carbon atoms 1 and 2 followed by the cyclization step according to DA mechanism. Analysis of the clusters contents resulted in preparation of "reaction signatures" for each type of reactions (Fig. 7). One may see that the CGRs of DHDA and DA are very similar: the only difference concerns one dynamical bond which is "created double" for DHDA and "created single" for DA. Substructural search using CGRs on Fig. 7 as queries perfectly separates MT and DHDA reactions which is not always possible using canonical representation of both reactions themselves and reactions queries.

Reaction Similarity Search. As any molecular graphs, CGRs can be fragmented and related fingerprints could be then used for the similarity search. The pertinence of this search depends on the fragments type. Thus, a similarity search using as query the CGR for the domino Heck–Diels–Alder reaction on Fig. 8 and Tanimoto coefficient (*T*) as metrics has been performed using (1) fragments containing, at least, one dynamical bond, and, (2) fragments containing only dynamical bond(s) conjugated with canonical double bonds. In search (1), for $T > 0.8$, only 9 of 28 DHDA reactions have been retrieved, whereas the all 28 reactions have been found in the search (2).

Structure–Reactivity Relationships. Conventional QSAR modeling of the thermodynamic, kinetic or any other parameters of chemical reactions involving many species is a big problem because

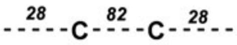
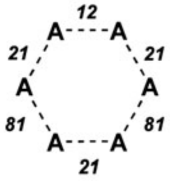
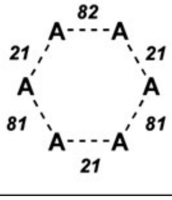
Metathesis	
Diels-Alder	 12 A A 21 21 A A 81 81 A A 21
Domino-Heck-Diels-Alder	 82 A A 21 21 A A 81 81 A A 21

Fig. 7. Examples of CGRs recommended as queries for substructural search for metathesis, Diels–Alder, and Domino Heck–Diels–Alder reactions. The numbers correspond to the types of dynamical bonds (*see* Fig. 9.8). Type "12" corresponds to single bond transformed to double bond.

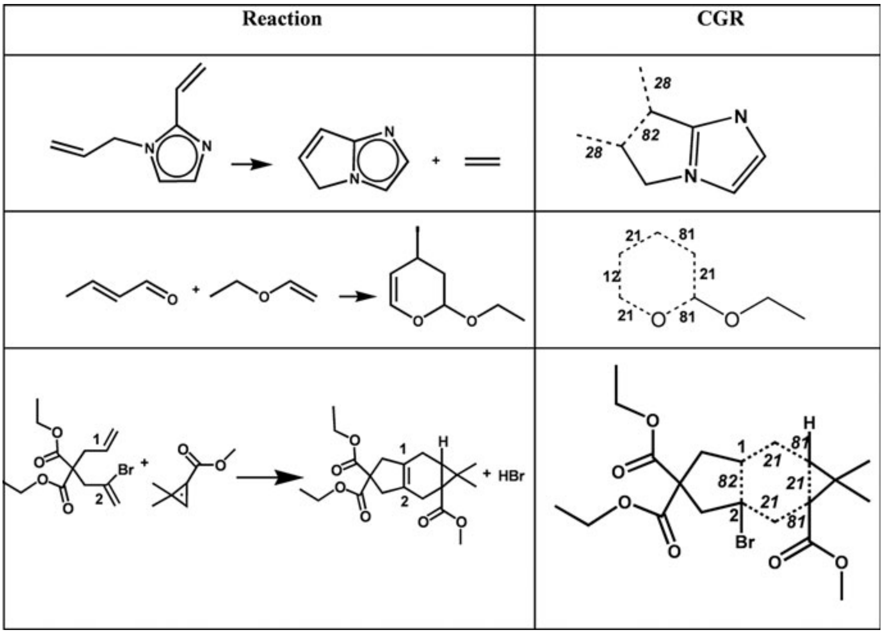


Fig. 8. Typical examples of metathesis (*top*), Diels–Alder (*middle*) and Domino Heck–Diels–Alder (*bottom*) reactions and related Condensed Reaction Graphs (CGR). The following labels are used for the dynamical bonds in GCR: 21 (double bond transformed to single bond), 28 (broken double bond), 82 (created double bond), 81 (created single bond).

it is not clear for which species the descriptors should be calculated. In this situation, CGR could become a reasonable solution since it represents a simple chemical graph for which fragment descriptors can be easily calculated. Recently [183], the CGR approach has been used to build predictive models for reaction rate (log *k*). The training set contained 463 structurally diverse S_N2 reactions in water at different temperature (totally 1,014 data). ISIDA fragments and reverse temperature have been used as descriptors in SVM calculations. The models have been validated in external tenfold cross-validation procedure repeated ten times after randomization of the data set. Figure 9 shows that log *k* is reasonably well predicted: squared determination coefficient $R^2 = 0.6$ and root-mean squared error is 1.14. The latter is rather close to the experimental error estimated as 1 log *k* unit.

5. Limitations of Fragment Descriptors

Despite many advantages of fragment descriptors, they are not devoid of certain drawbacks, which deserve serious attention. Two main problems should be mentioned: (1) “missing fragments” [184] and (2) modeling of stereochemically dependent properties.

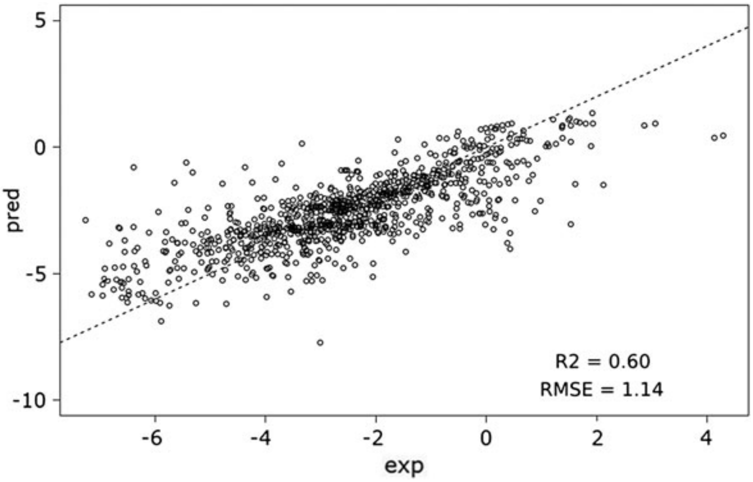


Fig. 9. Predicted versus experimental values of rate constant of S_N2 reactions in water (log *k*). The models were built on the set of 1,014 reactions using SVM method and fragment descriptors extracted from Condensed Graphs of Reactions.

The term “missing fragments” concerns comparison of the lists of fragments generated for the training and test sets. Test set molecules may contain fragments 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 [162] 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 [184]. The ISIDA program [12] applies Fragment Control and Model Fragment Control AD to consensus models, which improve the accuracy of prediction at reasonably high coverage of test sets. The NASAWIN program [185], for each model, 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 [186]. The LOGP program for lipophilicity predictions [187] 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. However, in practice, it is sufficient to introduce special labels indicating stereochemical configuration of chiral centers or E/Z isomers around double bond, then to use them in specification of molecular fragments. Such approach has been used in hologram fragment descriptors [188] as well as in the PARTAN language [97].

6. Conclusion

Fragment descriptors constitute one of the most universal type 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 the structural theory in chemistry as well as from the fundamental possibility of molecular graph invariants to be expressed in terms of subgraph occurrence numbers [189]. The main advantages of fragment descriptors lie in the ease of their computation as well as in the natural character of structural interpretation of SAR/QSAR/QSPR models. Due to all these factors, fragment descriptors play very important role in structure–property studies and ligand-based virtual screening.

Acknowledgment

The author thanks I. Baskin, D. Horvath, N. Lachiche, F. Hoonakker, A. Wagner, O. Klimchuk and G. Marcou for fruitful discussion and help for preparation of the manuscript.

References

1. Kubinyi, H., and Muler, G. (2004) *Chemogenomics in Drug Discovery*, Wiley-VCH Publishers, Weinheim.
2. Gorse, A. D. (2006) Diversity in Medicinal Chemistry Space. *Curr. Top. Med. Chem.* **6**, 3–18.
3. Walters, W. P., Stahl, M. T., and Murcko, M. A. (1998) Virtual Screening – An Overview. *Drug Discov. Today* **3**, 160–178.
4. Seifert, M. H., Kraus, J., and Kramer, B. (2007) Virtual High-Throughput Screening of Molecular Databases. *Curr. Opin. Drug. Discov. Dev.* **10**, 298–307.
5. Cavasotto, C. N., and Orry, A. J. (2007) Ligand Docking and Structure-Based Virtual Screening in Drug Discovery. *Curr. Top. Med. Chem.* **7**, 1006–1014.
6. Ghosh, S., Nie, A., An, J., and Huang, Z. (2006) Structure-Based Virtual Screening of Chemical Libraries for Drug Discovery. *Curr. Opin. Chem. Biol.* **10**, 194–202.
7. Todeschini, R., and Consonni, V. (2000) *Handbook of Molecular Descriptors*. Wiley-VCH Publishers, Weinheim.
8. Zefirov, N. S., and Palyulin, V. A. (2002) Fragmental Approach in QSPR. *J. Chem. Inf. Comput. Sci.* **42**, 1112–1122.
9. Japertas, P., Didziapetris, R., and Petrauskas, A. (2002) Fragmental Methods in the Design of New Compounds. Applications of The Advanced Algorithm Builder. *Quant. Struct. Act. Relat.* **21**, 23–37.
10. Artemenko, N. V., Baskin, I. I., Palyulin, V. A., and Zefirov, N. S. (2003) Artificial Neural Network and Fragmental Approach in Prediction of Physicochemical Properties of Organic Compounds. *Russ. Chem. Bull.* **52**, 20–29.

11. Merlot, C., Domine, D., and Church, D. J. (2002) Fragment Analysis in Small Molecule Discovery. *Curr. Opin. Drug Discov. Dev.* **5**, 391–399.
12. Varnek, A., Fourches, D., Hoonakker, F., and Solov'ev, V. P. (2005) Substructural Fragments: An Universal Language to Encode Reactions, Molecular and Supramolecular Structures. *J. Comput. Aided Mol. Des.* **19**, 693–703.
13. Jelfs, S., Ertl, P., and Selzer, P. (2007) Estimation of pKa for Drug Like Compounds Using Semiempirical and Information-Based Descriptors. *J. Chem. Inf. Model.* **47**, 450–459.
14. Tatevskii, V. M. (1950) Chemical Structure of Hydrocarbons and Their Heats of Formation. *Dokl. Akad. Nauk SSSR* **75**, 819–822.
15. Tatevskii, V. M., Mendzheritskii, E. A., and Korobov, V. (1951) The Additive Scheme of the Heat of Formation of Hydrocarbons and the Problem of the Heat of Sublimation of Graphite. *Vestn. Mosk. Univ.* **6**, 83–86.
16. Bernstein, H. J. (1952) The Physical Properties of Molecules in Relation to Their Structure. I: Relations Between Additive Molecular Properties in Several Homologous Series. *J. Chem. Phys.* **20**, 263–269.
17. Laidler, K. J. (1956) System of Molecular Thermochemistry for Organic Gases and Liquids. *Can. J. Chem.* **34**, 626–648.
18. Benson, S. W., and Buss, J. H. (1958) Additivity Rules for the Estimation of Molecular Properties: Thermodynamic Properties. *J. Chem. Phys.* **29**, 546–572.
19. Free, S. M., Jr., and Wilson, J. W. (1964) A Mathematical Contribution to Structure-Activity Studies. *J. Med. Chem.* **7**, 395–399.
20. Hiller, S. A., Golender, V. E., Rosenblit, A. B., Rastrigin, L. A., and Glaz, A. B. (1973) Cybernetic Methods of Drug Design. I: Statement of the Problem – The Perceptron Approach. *Comput. Biomed. Res.* **6**, 411–421.
21. Hiller, S. A., Glaz, A. B., Rastrigin, L. A., and Rosenblit, A. B. (1971) Recognition of Phisiological Activity of Chemical Compounds on Perceptron with Random Adaptation of Structure. *Dokl. Akad. Nauk SSSR* **199**, 851–853.
22. Golender, V. E., and Rozenblit, A. B. (1974) Interactive System for Recognition of Biological Activity Features in Complex Chemical Compounds. *Avtomatika i Telemekhanika* 99–105.
23. Golender, V. E., and Rozenblit, A. B. (1980) Logico-Structural Approach to Computer-Assisted Drug Design. *Med. Chem.* **11**, 299–337.
24. Piruzyan, L. A., Avidon, V. V., Rozenblit, A. B., Arolovich, V. S., Golender, V. E., Kozlova, S. P., Mikhailovskii, E. M., and Gavrishchuk, E. G. (1977) Statistical Study of an Information File on Biologically Active Compounds: Data Bank of the Structure and Activity of Chemical Compounds. *Khimiko-Farmatsevticheskii Zhurnal* **11**, 35–40.
25. Avidon, V. V., Pomerantsev, I. A., Golender, V. E., and Rozenblit, A. B. (1982) Structure-Activity Relationship Oriented Languages for Chemical Structure Representation. *J. Chem. Inf. Comput. Sci.* **22**, 207–214.
26. Cramer, R. D., III, Redl, G., and Berkoff, C. E. (1974) Substructural analysis: A Novel Approach to the Problem of Drug Design. *J. Med. Chem.* **17**, 533–535.
27. Stuper, A. J., and Jurs, P. C. (1976) ADAPT: A Computer System for Automated Data Analysis Using Pattern Recognition Techniques. *J. Chem. Inf. Model.* **16**, 99–105.
28. Brugger, W. E., Stuper, A. J., and Jurs, P. C. (1976) Generation of Descriptors from Molecular Structures. *J. Chem. Inf. Model.* **16**, 105–110.
29. Hodes, L., Hazard, G. F., Geran, R. I., and Richman, S. (1977) A Statistical-Heuristic Methods for Automated Selection of Drugs for Screening. *J. Med. Chem.* **20**, 469–475.
30. Milne, M., Lefkovitz, D., Hill, H., and Powers, R. (1972) Search of CA Registry (1.25 Million Compounds) with the Topological Screens System. *J. Chem. Doc.* **12**, 183–189.
31. Adamson, G. W., Cowell, J., Lynch, M. F., McLure, A. H. W., Town, W. G., and Yapp, A. M. (1973) Strategic Considerations in the Design of a Screening System for Substructure Searches of Chemical Structure Files. *J. Chem. Doc.* **13**, 153–157.
32. Feldman, A., and Hodes, L. (1975) An Efficient Design for Chemical Structure Searching. I: The Screens. *J. Chem. Inf. Model.* **15**, 147–152.
33. Willett, P. (1979) A Screen Set Generation Algorithm. *J. Chem. Inf. Model.* **19**, 159–162.
34. Willett, P. (1979) The Effect of Screen Set Size on Retrieval from Chemical Substructure Search Systems. *J. Chem. Inf. Model.* **19**, 253–255.
35. Willett, P., Winterman, V., and Bawden, D. (1986) Implementation of Nearest-Neighbor Searching in an Online Chemical Structure Search System. *J. Chem. Inf. Model.* **26**, 36–41.
36. Fisanick, W., Lipkus, A. H., and Rusinko, A. (1994) Similarity Searching on CAS Registry Substances. 2: 2D Structural Similarity. *J. Chem. Inf. Model.* **34**, 130–140.

37. Hodes, L. (1989) Clustering a Large Number of Compounds. 1: Establishing the Method on an Initial Sample. *J. Chem. Inf. Model.* **29**, 66–71.
38. McGregor, M. J., and Pallai, P. V. (1997) Clustering of Large Databases of Compounds: Using the MDL "Keys" as Structural Descriptors. *J. Chem. Inf. Model.* **37**, 443–448.
39. Turner, D. B., Tyrrell, S. M., and Willett, P. (1997) Rapid Quantification of Molecular Diversity for Selective Database Acquisition. *J. Chem. Inf. Model.* **37**, 18–22.
40. Durant, J. L., Leland, B. A., Henry, D. R., and Nourse, J. G. (2002) Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **42**, 1273–1280.
41. Tong, W., Lowis, D. R., Perkins, R., Chen, Y., Welsh, W. J., Goddette, D. W., Heritage, T. W., and Sheehan, D. M. (1998) Evaluation of Quantitative Structure-Activity Relationship Methods for Large-Scale Prediction of Chemicals Binding to the Estrogen Receptor. *J. Chem. Inf. Model.* **38**, 669–677.
42. Lounkine, E., Batista, J., and Bajorath, J. (2008) Random Molecular Fragment Methods in Computational Medicinal Chemistry. *Curr. Med. Chem.* **15**, 2108–2121.
43. Clark, M. (2005) Generalized Fragment-Substructure Based Property Prediction Method. *J. Chem. Inf. Model.* **45**, 30–38.
44. Matter, H., Baringhaus, K. H., Naumann, T., Klabunde, T., and Pirard, B. (2001) Computational Approaches Towards the Rational Design of Drug-Like Compound Libraries. *Comb. Chem. High Throughput Screen.* **4**, 453–475.
45. Oprea, T., Davis, A., Teague, S., and Leeson, P. (2001) Is There a Difference Between Leads and Drugs? A Historical Perspective. *J. Chem. Inf. Comput. Sci.* **41**, 1308–1315.
46. Patchett, A. A. N., and Nargund, R. P. (2000) Privileged Structures: An Update. *Annu. Rep. Med. Chem.* **35**, 289–298.
47. Aronov, A. M., McClain, B., Moody, C. S., and Murcko, M. A. (2008) Kinase-Likeness and Kinase-Priviledged Fragments: Toward Virtual Pharmacology. *J. Med. Chem.* **51**, 1214–1222.
48. Gillet, V. M., Myatt G., Zsoldos, Z., and Johnson, P. (1995) SPROUT, HIPPO and CAESA: Tools for De Novo Structure Generation and Estimation of Synthetic Accessibility. *Perspect. Drug Discov. Des.* **3**, 34–50.
49. Schneider, G. F., and Fechner, U. (2005) Computer-Based De Novo Design of Drug-Like Molecules. *Nat. Rev. Drug. Discov.* **4**, 649–663.
50. Lewell, X. Q., Judd D. B., Watson, S. P., and Hann, M. M. (1998) RECAP – Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **38**, 511–522.
51. Petrauskas, A. A., and Kolovanov, E. A. (2000) ACD/Log P Method Description. *Perspect. Drug Discov. Des.* **19**, 99–116.
52. Artemenko, N. V., Baskin, I. I., Palyulin, V. A., and Zefirov, N. S. (2001) Prediction of Physical Properties of Organic Compounds Using Artificial Neural Networks Within the Substructure Approach. *Dokl. Chem.* **381**, 317–320.
53. Adamson, G. W., Lynch, M. F., and Town, W. G. (1971) Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. Part II: Atom-Centered Fragments. *J. Chem. Soc. C*, 3702–3706.
54. Hodes, L. (1981) Selection of Molecular Fragment Features for Structure-Activity Studies in Antitumor Screening. *J. Chem. Inf. Comput. Sci.* **21**, 132–136.
55. Poroikov, V. V., Filimonov, D. A., Borodina, Y. V., Lagunin, A. A., and Kos, A. (2000) Robustness of Biological Activity Spectra Predicting by Computer Program Pass for Noncongeneric Sets of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **40**, 1349–1355.
56. Filimonov, D., Poroikov, V., Borodina, Y., and Gloriozova, T. (1999) Chemical Similarity Assessment through Multilevel Neighborhoods of Atoms: Definition and Comparison with the Other Descriptors. *J. Chem. Inf. Comput. Sci.* **39**, 666–670.
57. Xing, L., and Glen, R. C. (2002) Novel Methods for the Prediction of logP, pKa, and logD. *J. Chem. Inf. Comput. Sci.* **42**, 796–805.
58. Bender, A., Mussa, H. Y., Glen, R. C., and Reiling, S. (2004) Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naive Bayesian Classifier. *J. Chem. Inf. Comput. Sci.* **44**, 170–178.
59. Bender, A., Mussa, H. Y., Glen, R. C., and Reiling, S. (2004) Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. *J. Chem. Inf. Comput. Sci.* **44**, 1708–1718.
60. Glen, R. C., Bender, A., Arnby, C. H., Carlsson, L., Boyer, S., and Smith, J. (2006) Circular Fingerprints: Flexible Molecular

Descriptors with Applications from Physical Chemistry to ADME. *IDrugs* **9**, 199–204.

61. Rodgers, S., Glen, R. C., and Bender, A. (2006) Characterizing Bitterness: Identification of Key Structural Features and Development of a Classification Model. *J. Chem. Inf. Model.* **46**, 569–576.
62. Cannon, E. O., Amini, A., Bender, A., Sternberg, M. J. E., Muggleton, S. H., Glen, R. C., and Mitchell, J. B. O. (2007) Support Vector Inductive Logic Programming Outperforms the Naive Bayes Classifier and Inductive Logic Programming for the Classification of Bioactive Chemical Compounds. *J. Comput. Aided Mol. Des.* **21**, 269–280.
63. Faulon, J.-L., Visco, D. P., Jr., and Pophale, R. S. (2003) The Signature Molecular Descriptor. 1: Using Extended Valence Sequences in QSAR and QSPR Studies. *J. Chem. Inf. Comput. Sci.* **43**, 707–720.
64. Faulon, J.-L., Churchwell, C. J., and Visco, D. P., Jr. (2003) The Signature Molecular Descriptor. 2: Enumerating Molecules from Their Extended Valence Sequences. *J. Chem. Inf. Comput. Sci.* **43**, 721–734.
65. Churchwell, C. J., Rintoul, M. D., Martin, S., Visco, D. P., Jr., Kotu, A., Larson, R. S., Sillerud, L. O., Brown, D. C., and Faulon, J. L. (2004) The Signature Molecular Descriptor. 3: Inverse-Quantitative Structure-Activity Relationship of ICAM-1 Inhibitory Peptides. *J. Mol. Graph. Model.* **22**, 263–273.
66. Avidon, V. V., and Leksina, L. A. (1974) Descriptor Language for the Analysis of Structural Similarity of Organic Compounds. *Nauchno. Tekhn. Inf., Ser. 2*, 22–25.
67. Carhart, R. E., Smith, D. H., and Venkataraghavan, R. (1985) Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **25**, 64–73.
68. Horvath, D. (2001) High Throughput Conformational Sampling & Fuzzy Similarity Metrics: A Novel Approach to Similarity Searching and Focused Combinatorial Library Design and Its Role in the Drug Discovery Laboratory. in *Combinatorial Library Design and Evaluation: Principles, Software Tools and Applications* (Ghose, A., and Viswanadhan, V., Eds.), 429–472, Marcel Dekker, New York.
69. Horvath, D., and Jeandenans, C. (2003) Neighborhood Behavior of In Silico Structural Spaces with Respect to In Vitro Activity Spaces: A Novel Understanding of the Molecular Similarity Principle in the Context of Multiple Receptor Binding Profiles. *J. Chem. Inf. Comput. Sci.* **43**, 680–690.
70. Bonachera, F., Parent, B., Barbosa, F., Froloff, N., and Horvath, D. (2006) Fuzzy Tricentric Pharmacophore Fingerprints. 1: Topological Fuzzy Pharmacophore Triplets and Adapted Molecular Similarity Scoring Schemes. *J. Chem. Inf. Model.* **46**, 2457–2477.
71. Horvath, D., Bonachera, F., Solov'ev, V., Gaudin, C., and Varnek, A. (2007) Stochastic versus Stepwise Strategies for Quantitative Structure-Activity Relationship Generation – How Much Effort May the Mining for Successful QSAR Models Take? *J. Chem. Inf. Mod.* **47**, 927–939.
72. Schuffenhauer, A., Floersheim, P., Acklin, P., and Jacoby, E. (2003) Similarity Metrics for Ligands Reflecting the Similarity of the Target Proteins. *J. Chem. Inf. Comput. Sci.* **43**, 391–405.
73. MOE, Molecular Operating Environment, Chemical Computing Group Inc., Montreal, Canada, *MOE, Molecular Operating Environment, Chemical Computing Group Inc., Montreal, Canada*. www.chemcomp.com.
74. Franke, L., Byvatov, E., Werz, O., Steinhilber, D., Schneider, P., and Schneider, G. (2005) Extraction and Visualization of Potential Pharmacophore Points Using Support Vector Machines: Application to Ligand-Based Virtual Screening for COX-2 Inhibitors. *J. Med. Chem.* **48**, 6997–7004.
75. Byvatov, E., Sasse, B. C., Stark, H., and Schneider, G. (2005) From Virtual to Real Screening for D3 Dopamine Receptor Ligands. *Chembiochem.* **6**, 997–999.
76. Nilakantan, R., Bauman, N., Dixon, J. S., and Venkataraghavan, R. (1987) Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.* **27**, 82–85.
77. Kearsley, S. K., Sallamack, S., Fluder, E. M., Andose, J. D., Mosley, R. T., and Sheridan, R. P. (1996) Chemical Similarity Using Physiochemical Property Descriptors. *J. Chem. Inf. Comput. Sci.* **36**, 118–127.
78. Kuz'min, V. E., Muratov, E. N., Artemenko, A. G., Gorb, L. G., Qasim, M., and Leszczynski, J. (2008) The Effects of Characteristics of Substituents on Toxicity of the Nitroaromatics: HiT QSAR Study. *J. Comput. Aid. Mol. Des.* **22**, 747–759.
79. Kuz'min, V. E., Artemenko, A. G., Muratov, E. N., Lozitsky, V. P., Fedchuk, A. S., Lozitska, R. N., Boschenko, Y. A., and Gridina., T. L. (2005) The Hierarchical QSAR Technology for Effective Virtual Screening and Molecular Design of the

Promising Antiviral Compounds. *Antivir. Res.* **65**, A70–A71.

80. Bemis, G. W., and Murcko, M. A. (1996) The Properties of Known Drugs. 1: Molecular Frameworks. *J. Med. Chem.* **39**, 2887–2893.
81. Bemis, G. W., and Murcko, M. A. (1999) Properties of Known Drugs. 2: Side Chains. *J. Med. Chem.* **42**, 5095–5099.
82. Guener, O. F. (2000) *Pharmacophore Perception, Development, and Use in Drug Design*, Wiley-VCH Publishers, Weinheim.
83. Langer, T., and Hoffman, R. D. (2000) *Pharmacophores and Pharmacophore Searches*, Wiley-VCH Publishers, Weinheim.
84. Wang, J., Lai, L., and Tang, Y. (1999) Structural Features of Toxic Chemicals for Specific Toxicity. *J. Chem. Inf. Comput. Sci.* **39**, 1173–1189.
85. Kazius, J., McGuire, R., and Bursi, R. (2005) Derivation and Validation of Toxicophores for Mutagenicity Prediction. *J. Med. Chem.* **48**, 312–320.
86. Cunningham, A. R., Rosenkranz, H. S., Zhang, Y. P., and Klopman, G. (1998) Identification of 'Genotoxic' and 'Non-Genotoxic' Alerts for Cancer in Mice: The Carcinogenic Potency Database. *Mutat. Res.* **398**, 1–17.
87. Klopman, G. (1984) Artificial Intelligence Approach to Structure-Activity Studies. Computer Automated Structure Evaluation of Biological Activity of Organic Molecules. *J. Am. Chem. Soc.* **106**, 7315–7321.
88. Klopman, G., and Rosenkranz, H. S. (1984) Structural Requirements for the Mutagenicity of Environmental Nitroarenes. *Mutat. Res.* **126**, 227–238.
89. Klopman, G. (1985) Predicting Toxicity Through a Computer Automated Structure Evaluation Program. *Environ. Health Perspect.* **61**, 269–274.
90. Rosenkranz, H. S., Mitchell, C. S., and Klopman, G. (1985) Artificial Intelligence and Bayesian Decision Theory in the Prediction of Chemical Carcinogens. *Mutat. Res.* **150**, 1–11.
91. Klopman, G. (1992) MULTICASE. 1: A Hierarchical Computer Automated Structure Evaluation Program. *Quant. Struct. Act. Relat.* **11**, 176–184.
92. Klopman, G., and Rosenkranz, H. S. (1994) Approaches to SAR in Carcinogenesis and Mutagenesis: Prediction of Carcinogenicity/Mutagenicity Using MULTI-CASE. *Mutat. Res.* **305**, 33–46.
93. Klopman, G., Chakravarti, S. K., Harris, N., Ivanov, J., and Saiakhov, R. D. (2003) In-Silico Screening of High Production Volume Chemicals for Mutagenicity Using the MCASE QSAR Expert System. *SAR QSAR Environ. Res.* **14**, 165–180.
94. Sanderson, D. M., and Earnshaw, C. G. (1991) Computer Prediction of Possible Toxic Action from Chemical Structure: The DEREK System. *Hum. Exp. Toxicol.* **10**, 261–273.
95. Gombar, V. K., Enslein, K., Hart, J. B., Blake, B. W., and Borgstedt, H. H. (1991) Estimation of Maximum Tolerated Dose for Long-Term Bioassays from Acute Lethal Dose and Structure by QSAR. *Risk Anal.* **11**, 509–517.
96. Judson, P. N. (1992) QSAR and Expert Systems in the Prediction of Biological Activity. *Pestic. Sci.* **36**, 155–160.
97. Judson, P. N. (1994) Rule Induction for Systems Predicting Biological Activity. *J. Chem. Inf. Comput. Sci.* **34**, 148–153.
98. Barratt, M. D., and Rodford, R. A. (2001) The Computational Prediction of Toxicity. *Curr. Opin. Chem. Biol.* **5**, 383–388.
99. Lipinski, C. A., Lombardo, F., Dominy, B. W., and Feeney, P. J. (2001) Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.* **46**, 3–26.
100. Oprea, T. I. (2000) Property Distribution of Drug-Related Chemical Databases. *J. Comput. Aided Mol. Des.* **14**, 251–264.
101. Veber, D. F., Johnson, S. R., Cheng, H. Y., Smith, B. R., Ward, K. W., and Kopple, K. D. (2002) Molecular Properties that Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **45**, 2615–2623.
102. Hann, M. M., and Oprea, T. I. (2004) Pursuing the Leadlikeness Concept in Pharmaceutical Research. *Curr. Opin. Chem. Biol.* **8**, 255–263.
103. Leo, A. J. (1993) Calculating log Poct from Structures. *Chem. Rev.* **93**, 1281–1306.
104. Tetko, I. V., and Livingstone, D. J. (2006) Rule-Based Systems to Predict Lipophilicity. in *Comprehensive Medicinal Chemistry II: In Silico Tools in ADMET* (Testa, B., and van de Waterbeemd, H., Eds.), 649–668, Elsevier, Oxford, UK.
105. Kubinyi, H. (1998) Similarity and Dissimilarity: A Medicinal Chemist's View. *Perspect. Drug Discov. Des.* **9–11**, 225–252.
106. Martin, Y. C., Kofron, J. L., and Traphagen, L. M. (2002) Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **45**, 4350–4358.
107. *Daylight Chemical Information Systems Inc.* http://www.daylight.com.
108. *Barnard Chemical Information Ltd.* http://www.bci.gb.com/.

109. *Tripos Inc.* http://www.tripos.com.
110. Jaccard, P. (1901) Distribution de la flore alpine dans le Bassin des Dranses et dans quelques regions voisines. *Bull. Soc. Vaud. Sci. Nat.* **37**, 241–272.
111. Schneider, G., Neidhart, W., Giller, T., and Schmid, G. (1999) "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem. Int. Ed.* **38**, 2894–2896.
112. Hull, R. D., Singh, S. B., Nachbar, R. B., Sheridan, R. P., Kearsley, S. K., and Fluder, E. M. (2001) Latent Semantic Structure Indexing (LaSSI) for Defining Chemical Similarity. *J. Med. Chem.* **44**, 1177–1184.
113. Hull, R. D., Fluder, E. M., Singh, S. B., Nachbar, R. B., Kearsley, S. K., and Sheridan, R. P. (2001) Chemical Similarity Searches Using Latent Semantic Structural Indexing (LaSSI) and Comparison to TOPOSIM. *J. Med. Chem.* **44**, 1185–1191.
114. Willett, P., Barnard, J. M., and Downs, G. M. (1998) Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **38**, 983–996.
115. Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., and Schuffenhauer, A. (2004) Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Comput. Sci.* **44**, 1177–1185.
116. Ormerod, A., Willett, P., and Bawden, D. (1989) Comparison of Fragment Weighting Schemes for Substructural Analysis. *Quant. Struct. Act. Relat.* **8**, 115–129.
117. Godden, J. W., Furr, J. R., Xue, L., Stahura, F. L., and Bajorath, J. (2004) Molecular Similarity Analysis and Virtual Screening by Mapping of Consensus Positions in Binary-Transformed Chemical Descriptor Spaces with Variable Dimensionality. *J. Chem. Inf. Comput. Sci.* **44**, 21–29.
118. Godden, J. W., Stahura, F. L., and Bajorath, J. (2004) POT-DMC: A Virtual Screening Method for the Identification of Potent Hits. *J. Med. Chem.* **47**, 5608–5611.
119. Batista, J., Godden, J. W., and Bajorath, J. (2006) Assessment of Molecular Similarity from the Analysis of Randomly Generated Structural Fragment Populations. *J. Chem. Inf. Model.* **46**, 1937–1944.
120. Godden, J. W., and Bajorath, J. (2001) Differential Shannon Entropy as a Sensitive Measure of Differences in Database Variability of Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **41**, 1060–1066.
121. Batista, J., and Bajorath, J. (2007) Chemical Database Mining Through Entropy-Based Molecular Similarity Assessment of Randomly Generated Structural Fragment Populations. *J. Chem. Inf. Model.* **47**, 59–68.
122. Gute B. D., Basak S. C., Mills, D. and Hawkins, D. M. (2002) Tailored Similarity Spaces for the Prediction of Physicochemical Properties. *Internet Electron. J. Mol. Des.* **1**, 374–387.
123. Fourches, D. (2007) Modèles multiples en QSAR/QSPR: développement de nouvelles approches et leurs applications au design «in silico» de nouveaux extractants de métaux, aux propriétés ADMETox ainsi qu'à différentes activités biologiques de molécules organiques. Louis Pasteur University of Strasbourg, Strasbourg.
124. Guha, R., and VanDrie, J. H. (2008) Structure-Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model.* **48**, 646–658.
125. Peltason, L., and Bajorath, J. (2007) SAR Index: Quantifying the Nature of Structure-Activity Relationships. *J. Med. Chem.* **50**, 5571–5578.
126. Bonachera, F., and Horvath, D. (2008) Fuzzy Tricentric Pharmacophore Fingerprints. 2: Application of Topological Fuzzy Pharmacophore Triplets in Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model.* **48**, 409–425.
127. Harper, G., Bradshaw, J., Gittins, J. C., Green, D. V. S., and Leach, A. R. (2001) The Prediction of Biological Activity for High-Throughput Screening Using Binary Kernel Discrimination. *J. Chem. Inf. Comput. Sci.* **41**, 1295–1300.
128. Geronikaki, A. A., Dearden, J. C., Filimonov, D., Galaeva, I., Garibova, T. L., Gloriozova, T., Krajneva, V., Lagunin, A., Macaev, F. Z., Molodavkin, G., Poroikov, V. V., Pogrebnoi, S. I., Shepeli, F., Voronina, T. A., Tsitlakidou, M., and Vlad, L. (2004) Design of New Cognition Enhancers: From Computer Prediction to Synthesis and Biological Evaluation. *J. Med. Chem.* **47**, 2870–2876.
129. Katritzky, A. R., Kuanar, M., Slavov, S., Dobchev, D. A., Fara, D. C., Karelson, M., Acree, W. E., Jr., Solov'ev, V. P., and Varnek, A. (2006) Correlation of Blood-Brain Penetration Using Structural Descriptors. *Bioorg. Med. Chem.* **14**, 4888–4917.
130. Katritzky, A. R., Dobchev, D. A., Fara, D. C., Hur, E., Tamm, K., Kurunczi, L., Karelson, M., Varnek, A., and Solov'ev, V. P. (2006) Skin Permeation Rate as a Function of Chemical Structure. *J. Med. Chem.* **49**, 3305–3314.
131. Katritzky, A. R, Kuanar, M., Fara, D. C., Karelson, M., Acree, W. E., Jr., Solov'ev, V. P., and Varnek, A. (2005) QSAR Modeling of

Blood: Air and Tissue: Air Partition Coefficients Using Theoretical Descriptors. *Bioorg. Med. Chem.* **13**, 6450–6463.

132. Mannhold, R., Rekker, R. F., Sonntag, C., ter Laak, A. M., Dross, K., and Polymeropoulos, E. E. (1995) Comparative Evaluation of the Predictive Power of Calculation Procedures for Molecular Lipophilicity. *J. Pharm. Sci.* **84**, 1410–1419.
133. Nys, G. G., and Rekker, R. F. (1973) Statistical Analysis of a Series of Partition Coefficients with Special Reference to the Predictability of Folding of Drug Molecules: Introduction of Hydrophobic Fragmental Constants (f-Values). *Eur. J. Med. Chem.* **8**, 521–535.
134. Leo, A., Jow, P. Y. C., Silipo, C., and Hansch, C. (1975) Calculation of Hydrophobic Constant (log P) from pi and f Constants. *J. Med. Chem.* **18**, 865–868.
135. Ghose, A. K., and Crippen, G. M. (1987) Atomic Physicochemical Parameters for Three-Dimensional-Structure-Directed Quantitative Structure-Activity Relationships. 2: Modeling Dispersive and Hydrophobic Interactions. *J. Chem. Inf. Comput. Sci.* **27**, 21–35.
136. Ghose, A. K., and Crippen, G. M. (1986) Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships I: Partition Coefficients as a Measure of Hydrophobicity. *J. Comput. Chem.* **7**, 565–577.
137. Ghose, A. K., Pritchett, A., and Crippen, G. M. (1988) Atomic Physicochemical Parameters for Three Dimensional Structure Directed Quantitative Structure-Activity Relationships III: Modeling Hydrophobic Interactions. *J. Comput. Chem.* **9**, 80–90.
138. Wildman, S. A., and Crippen, G. M. (1999) Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **39**, 868–873.
139. Suzuki, T., and Kudo, Y. (1990) Automatic log P Estimation Based on Combined Additive Modeling Methods. *J. Comput. Aided. Mol. Des.* **4**, 155–198.
140. Convard, T., Dubost, J.-P., Le Solleu, H., and Kummer, E. (1994) SMILOGP: A Program for a Fast Evaluation of Theoretical log-p from the Smiles Code of a Molecule. *Quant. Struct. Act. Relat.* **13**, 34–37.
141. Wang, R., Gao, Y., and Lai, L. (2000) Calculating Partition Coefficient by Atom-Additive Method. *Perspect. Drug Discov. Des.* **19**, 47–66.
142. Wang, R., Fu, Y., and Lai, L. (1997) A New Atom-Additive Method for Calculating Partition Coefficients. *J. Chem. Inf. Comput. Sci.* **37**, 615–621.
143. Balakin, K. V., Savchuk, N. P., and Tetko, I. V. (2006) In Silico Approaches to Prediction of Aqueous and DMSO Solubility of Drug-Like Compounds: Trends, Problems and Solutions. *Curr. Med. Chem.* **13**, 223–241.
144. Varnek, A., Kireeva, N., Tetko, I. V., Baskin, I. I., and Solov'ev, V. P. (2007) Exhaustive QSPR Studies of a Large Diverse Set of Ionic Liquids: How Accurately Can We Predict Melting Points? *J. Chem. Inf. Model.* **47**, 1111–1122.
145. Varnek, A., Fourches, D., Horvath, D., Klimchuk, O., Gaudin, O., Vayer, P., Solov'ev, V., Hoonakker, F., Tetko, I. V. and Marcou, G. (2008) ISIDA: Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *Curr. Comput. Aided Drug Des.* **4**, 191–198.
146. Varnek, A., Fourches, D., Solov'ev, V., Klimchuk, O., Ouadi, A., and Billard, I. (2007) Successful "In Silico" Design of New Efficient Uranyl Binders. *Solvent Extr. Ion Exch.* **25**, 433–462.
147. Grubbs, F. E. (1969) Procedures for Detecting Outlying Observations in Samples. *Technometrics* **11**, 1–21.
148. Solov'ev, V. P., and Varnek, A. (2003) Anti-HIV Activity of HEPT, TIBO, and Cyclic Urea Derivatives: Structure-Property Studies, Focused Combinatorial Library Generation, and Hits Selection Using Substructural Molecular Fragments Method *J. Chem. Inf. Comput. Sci.* **43**, 1703–1719.
149. Fourches, D., Kireeva, N., Klimchuk, O., Marcou, G., Solov'ev, V., and Varnek, A. (2008) Computer-Aided Design of New Metal Binders. *Radiochim. Acta* **96**, 505–511.
150. Varnek, A., and Solov'ev, V. (2008) Quantitative Structure-Property Relationships in Solvent Extraction and Complexation of Metals. in *Ion Exchange and Solvent Extraction* (Sengupta, A. K., and Moyer, B.A., Eds.), Taylor and Francis, Philadelphia.
151. Horvath, D., Marcou, G., and Varnek A. (2009) Predicting the Predictability: A Unified Approach to the Applicability Domain Problem. *J. Chem. Inf. Model.* **49**, 1762–1776.
152. Hao Zhu, D. F., Varnek, A., Papa, E., Gramatica, P., Tetko, I.V., Öberg, T., Cherkasov, A., and Tropsha, A. (2008) Combinational QSAR Modeling of Chemical Toxicants Tested Against *Tetrahymena pyriformis. J. Chem. Inf. Model.* **48**, 766–784.
153. *MolConnZ*, version 4.05; eduSoft LC: Ashland, VA, 2003.
154. Tetko, I. V., Sushko, I., Pandey, A. K., Zhu, H., Tropsha, A., Papa, E., Öberg, T.,

Todeschini, R., Fourches, D., and Varnek, A. (2008) Critical Assessment of QSAR Models of Environmental Toxicity Against *Tetrahymena pyriformis*: Focusing On Applicability Domain and Overfitting by Variable Selection. *J. Chem. Inf. Model.* **48**, 1733–1746.

155. Huuskonen, J. (2000) Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **40**, 773–777.
156. McElroy, N., and Jurs, P. (2001) Prediction of Aqueous Solubility of Heteroatom-Containing Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **41**, 1237–1247.
157. Ran, Y., Jain, N., and Yalkowsky, S. (2001) Prediction of Aqueous Solubility of Organic Compounds by the General Solubility Equation (GSE). *J. Chem. Inf. Comput. Sci.* **41**, 1208–1217.
158. Yaffe, D., Cohen, Y., Espinosa, G., Arenas, A., and Giralt, F. (2001) A Fuzzy ARTMAP Based on Quantitative Structure-Property Relationships (QSPRs) for Predicting Aqueous Solubility of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **41**, 1177–1207.
159. Downs, G., and Barnard, J. (2002) Clustering Methods and Their Uses in Computational Chemistry. *Rev. Comput. Chem.* **18**, 1–40.
160. Caruana, R. (1997) Multitask Learning. *Mach. Learn.* **28**, 41–75.
161. Varnek, A., Gaudin, C., Marcou, G.; Baskin, I., Pandey, A. K., and Tetko, I. V. (2009) Inductive Transfer of Knowledge: Application of Multi-Task Learning and Feature Net Approaches to Model Tissue-Air Partition Coefficients. *J. Chem. Inf. Model.* **49**, 133–144.
162. Tetko, I. V., Tanchuk, V. Y., and Villa, A. E. P. (2001) Prediction of n-Octanol/Water Partition Coefficients from PHYSPROP Database Using Artificial Neural Networks and E-State Indices. *J. Chem. Inf. Comput. Sci.* **41**, 1407–1421.
163. Efron, B. (1983) Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *J. Am. Stat. Assoc.* **78**, 316–331.
164. Netzeva, T. I., Worth, A. P., Aldenberg, T., Benigni, R., Cronin, M. T. D., Gramatica, P., Jaworska, J. S., Kahn, S., Klopman, G., Marchant, C. A., Myatt, G., Nikolova-Jeliazkova, N., Patlewicz, G. Y., Perkins, R., Roberts, D. W., Schultz, T. W., Stanton, D. T., van de Sandt, J. J. M., Tong, W., Veith, G., and Yang, C. (2005) Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure-Activity Relationships. The Report and Recommendations of ECVAM Workshop 52. *Altern. Lab. Anim.* **33**, 155–173.
165. Jaworska, J., Nikolova-Jeliazkova, N., and Aldenberg, T. (2005) QSAR Applicability Domain Estimation by Projection of the Training Set in Descriptor Space: A Review. *Altern. Lab. Anim.* **33**, 445–459.
166. Sheridan, R. P., Feuston, B. P., Maiorov, V. N., and Kearsley, S. K. (2004) Similarity to Molecules in the Training Set is a Good Discriminator for Prediction Accuracy in QSAR. *Altern. Lab. Anim.* **44**, 1912–1928.
167. Fukumizu, K., and Watanabe, S. (1993) Probabililty Density Estimation by Regularization Method, in *Proceed. of the International Joint Conf. on Neural Networks*, pp 1727–1730.
168. Parzen, E. (1962) On Estimation of a Probability Density Function and Mode. *Ann. Math. Stat.* **33**, 1065–1076.
169. Schioler, H., and Hartmann, U. (1992) Mapping Neural Network Derived from the Parzen Window Estimator. *Neural Netw.* **5**, 903–909.
170. Duda, R., and Hart, P. (1973) *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York.
171. van der Eijkel, G. C., Jan, van der Lubbe, J., and Backer, E. (1997) A Modulated Parzen-Windows Approach for Probability Density Estimation, in *Proceedings of the Second International Symposium on Advances in Intelligent Data Analysis, Reasoning about Data*, Springer-Verlag.
172. Kireeva, N. (2009) QSPR Ensemble Modeling of Stabilities of Metal-Ligand Complexes and Melting Point of Ionic Liquids. PhD thesis. Louis Pasteur University, Strasbourg.
173. Feuston, B. P., Chakravorty, S. J., Conway, J. F., Culberson, J. C., Forbes, J., Kraker, B., Lennon, P. A., Lindsley, C., McGaughey, G. B., Mosley, R., Sheridan, R. P., Valenciano, M., and Kearsley, S. K. (2005) Web Enabling Technology for the Design, Enumeration, Optimization and Tracking of Compound Libraries. *Curr. Top. Med. Chem.* **5**, 773–783.
174. Green, D. V., and Pickett, S. D. (2004) Methods for Library Design and Optimisation. *Mini Rev. Med. Chem.* **4**, 1067–1076.
175. Green, D. V. (2003) Virtual Screening of Virtual Libraries. *Prog. Med. Chem.* **41**, 61–97.
176. Varnek, A., Fourches, D., Solov'ev, V. P., Baulin, V. E., Turanov, A. N., Karandashev, V. K., Fara, D., and Katritzky, A. R. (2004) “In Silico” Design of New Uranyl Extractants Based on Phosphoryl-Containing

Podands: QSPR Studies, Generation and Screening of Virtual Combinatorial Library and Experimental Tests. *J. Chem. Inf. Comput. Sci.* **44**, 1365–1382.

177. Tetko, I. V. (2002) Neural Network Studies. 4: Introduction to Associative Neural Networks. *J. Chem. Inf. Comput. Sci.* **42**, 717–728.
178. Vapnik, V. N. (1999) An Overview of Statistical Learning Theory. *IEEE Trans. Neural Netw.* **10**, 988–999.
179. Fujita, S. (1986) Description of Organic Reactions Based on Imaginary Transition Structures. 1: Introduction of New Concepts. *J. Chem. Inf. Comput. Sci.* **26**, 205–212.
180. Jauffret, P., Tonnelier, C., Hanser, T., Kaufmann, G., and Wolff, R. (1990) Machine Learning of Generic Reactions: Toward an Advanced Comp uter Representation of Chemical Reactions. *Tetrahedron Comput. Methodol.* **3**, 335–349.
181. Vladutz, G. (1986) Modern Approaches to Chemical Reaction Searching, in *Approaches to Chemical Reaction Searching* (Willett, P., Ed.), 202–220, Gower, London.
182. Hoonakker, F. (2007) Graphes condensés de réeactions, applications à la recherche par similarité, la classification et la modélisation. Louis Pasteur University, Strasbourg.
183. Hoonakker, F., Lachiche, N., Varnek, A., and Wagner, A. (2009) Condensed Graph of Reaction: Considering a Chemical Reaction As one Single Pseudo Molecule. *The 19th International Conference on Inductive Logic Programming.* http://lsiit.u-strasbg.fr/Publications/2009/HLVW09.
184. Tetko, I. V., Bruneau, P., Mewes, H. -W., Rohrer, D. C., and Poda, G. I. (2006) Can We Estimate the Accuracy of ADMET Predictions? *Drug Discov. Today* **11**, 700–707.
185. Baskin, I. I., Halberstam, N. M., Artemenko, N. V., Palyulin, V. A., and Zefirov, N. S. (2003) NASAWIN – A Universal Software for QSPR/QSAR Studies. in *EuroQSAR 2002 Designing Drugs and Crop Protectants: Processes, Problems and Solutions.* (Ford, M., Ed.), 260–263, Blackwell Publishing, Oxford, UK.
186. Halberstam, N. M. (2001) *Modeling Properties and Reactivity of Organic Compounds Using Artificial Neural Networks. Department of Chemistry,* Moscow State University, Moscow.
187. Leo, A. J., and Hoekman, D. (2000) Calculating log P (oct) with No Missing Fragments: The Problem of Estimating New Interaction Parameters. *Perspect. Drug. Discov. Des.* **18**, 19–38.
188. Honorio, K. M., Garratt, R. C., and Andricopulo, A. D. (2005) Hologram Quantitative Structure-Activity Relationships For A Series of Farnesoid X Receptor Activators. *Bioorg. Med. Chem. Lett.* **15**, 3119–3125.
189. Baskin, I. I., Skvortsova, M. I., Stankevich, I. V., and Zefirov, N. S. (1995) On the Basis of Invariants of Labeled Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **35**, 527–531.