CHAPTER 5

VIRTUAL SCREENING AND MOLECULAR DESIGN BASED ON HIERARCHICAL QSAR TECHNOLOGY

VICTOR E. KUZ'MIN, A.G. ARTEMENKO, EUGENE N. MURATOV,
P.G. POLISCHUK, L.N. OGNICHENKO, A.V. LIAHOVSKY,
A.I. HROMOV, AND E.V. VARLAMOVA

*A.V. Bogatsky Physical-Chemical Institute NAS of Ukraine, Lustdorfskaya Doroga 86,
Odessa 65080, Ukraine, e-mail: victor@ccmsi.us*

Abstract: This chapter is devoted to the hierarchical QSAR technology (HiT QSAR) based on simplex representation of molecular structure (SiRMS) and its application to different QSAR/QSPR tasks. The essence of this technology is a sequential solution (with the use of the information obtained on the previous steps) of the QSAR paradigm by a series of enhanced models based on molecular structure description (in a specific order from 1D to 4D). Actually, it's a system of permanently improved solutions. Different approaches for domain applicability estimation are implemented in HiT QSAR. In the SiRMS approach every molecule is represented as a system of different simplexes (tetratomic fragments with fixed composition, structure, chirality, and symmetry). The level of simplex descriptors detailed increases consecutively from the 1D to 4D representation of the molecular structure. The advantages of the approach presented are an ability to solve QSAR/QSPR tasks for mixtures of compounds, the absence of the "molecular alignment" problem, consideration of different physical–chemical properties of atoms (e.g., charge, lipophilicity), and the high adequacy and good interpretability of obtained models and clear ways for molecular design. The efficiency of HiT QSAR was demonstrated by its comparison with the most popular modern QSAR approaches on two representative examination sets. The examples of successful application of the HiT QSAR for various QSAR/QSPR investigations on the different levels (1D–4D) of the molecular structure description are also highlighted. The reliability of developed QSAR models as the predictive virtual screening tools and their ability to serve as the basis of directed drug design was validated by subsequent synthetic, biological, etc. experiments. The HiT QSAR is realized as the suite of computer programs termed the "HiT QSAR" software that so includes powerful statistical capabilities and a number of useful utilities.

Keywords: HiT QSAR, Simplex representation, SiRMS

Abbreviations

A/I/EVS	Automatic/Interactive/Evolutionary Variables Selection
ACE	Angiotensin Converting Enzyme

T. Puzyn et al. (eds.), Recent Advances in QSAR Studies, 127–176.
DOI 10.1007/978-1-4020-9783-6_5, © Springer Science+Business Media B.V. 2010

AchE	Acetylcholinesterase
CoMFA	Comparative Molecular Fields Analysis QSAR approach
CoMSIA	Comparative Molecular Similarity Indexes Analysis QSAR approach
DA	Applicability Domain
DSTP	dispirotripiperazine
EVA	Eigenvalue Analysis QSAR approach
GA	Genetic Algorithm
HiT QSAR	Hierarchical QSAR Technology
HQSAR	Hologram QSAR approach
HRV	Human Rhinovirus
HSV	Herpes Simplex Virus
MLR	Multiple Linear Regression statistical method
PLS	Partial Least Squares or Projection on Latent Structures statistical method
Q^2	cross-validation determination coefficient
QSAR/QSPR	Quantitative Structure-Activity/Property Relationship
R^2	determination coefficient for training set
R^2_{test}	determination coefficient for test set
SD	Simplex Descriptor
SI	Selectivity Index
SiRMS	Simplex Representation of Molecular Structure QSAR approach
TV	Trend-Vector statistical method

5.1. INTRODUCTION

Nowadays the creation of a new medicine costs more than one billion dollars and the price of this process is growing steadily day by day [1]. During recent decades different theoretical approaches have been used to facilitate and accelerate the process of new drugs creation that is not only very expensive, but also is a multistep and long-term activity [2]. The choice of approaches depends on a presence or absence of information regarding a biological target and the substances interacting with it. A situation, when we have a set of biologically active compounds (ligands) and have no information about a biological target (e.g., receptor) is the most common. Different quantitative structure–activity relationship (QSAR) approaches are used in this case. For many years, QSAR has been used successfully for the analysis of huge variety of endpoints, e.g., antiviral and anticancer activity, toxicity [3–14]. Its staying power may be attributed to the strength of its initial postulate that activity is a function of structure and the rapid and extensive development of the methodology and computational techniques. The overall goals of QSAR retain their original essence and remain focused on the predictive ability of the approach and its receptiveness to mechanistic interpretation [15].

Many different QSAR methods [16–20] have been developed since the second half of the last century and new techniques and improvements are still being created

[21]. These approaches differ mainly by the principles and levels of representation and description of molecular structure. The degree of adequacy of molecular structure models varies from 1D to 4D level:

- 1D models consider only the gross formula of a molecule (for example, glycine – $C_2H_4NO_2$). Actually, such models reflect only a composition of a molecule. Obviously, it is quite impossible to solve adequately the "structure–activity" tasks using such approaches. So, usually these models have an auxiliary role only, but sometimes they can be used as independent virtual screening tools [22].
- 2D models contain information regarding the structure of a compound and are based on its structural formula [20]. Such models reflect only the topology of the molecule. These models are very popular [3, 23]. The capacity of such approaches is due to the fact that the topological models of molecular structure, in an implicit form, contain information about possible conformations of the compound. Our operational experience shows that 2D level of representation of the molecular structure is enough for the solution of more than 90% of existing QSAR/QSPR tasks.
- 3D-QSAR models [16, 17, 19, 20] give full structural information taking into account composition, topology, and spatial shape of molecule for one conformer only. These models are widespread. However, the choice of the conformer of the molecule analyzed is mostly accidental.

The description of the molecular structure is realized more adequately by 4D-QSAR models [10, 24]. These models are similar to 3D models, but compared to them the structural information is considered for a set of conformers (conditionally the fourth dimension), instead of one fixed conformation (also see Chapter 3).

The description of compounds from 1D to 4D models reflects the hierarchy of molecular structure representation. However, it's only one of the principles of HiT QSAR. In this work the hierarchic strategy related to all the aspects of the QSAR models development has been considered.

The developed strategy has been realized as a complex of computer programs known as the "HiT QSAR" software. Innovative aspect and main advantages of HiT QSAR involve

- Simplex representation of molecular structure that provides universality, diversity, and flexibility of the description of compounds related to different structural types.
- HiT QSAR that, depending on the concrete aims of research, allows for the construction of the optimal strategy for QSAR model generation, avoiding at the same time superfluous complications that do not result in an increase in the adequacy of the model.
- HiT QSAR does not have the restrictions of such well-known and widely used approaches as CoMFA, CoMSIA, and HASL. Usage of such methods is limited by the requirement for a structurally homogeneous set of molecules and the use of only one conformer.

- HiT QSAR does not have the HQSAR restrictions that are related to the ambiguity of descriptor system formation.
- At every stage of HiT QSAR use, we can determine the molecular structural features that are important for the studied activity and exclude the rest. It shows unambiguously the limits of QSAR models' complication and ensures that resources are not wasted on needless calculations.

The efficiency of the HiT QSAR has been demonstrated through the example of various QSAR tasks, e.g., given in [3, 10–12, 22, 25–37].

5.2. MULTI-HIERARCHICAL STRATEGY OF QSAR INVESTIGATION

5.2.1. HiT QSAR Concept

In this chapter, the hierarchic QSAR technology (HiT QSAR) [31, 32, 36, 37] based on the simplex representation of molecular structure (SiRMS) has been considered. This method has proved efficient in numerous studies for solving different "structure–activity/property" problems [3, 10–12, 22, 25–37]. The essence of the strategy presented is based on the solution of QSAR problems via the sequence of the permanently improved molecular structure models (from 1D to 4D) (Figure 5-1). Thus, at each stage of the hierarchical system, the QSAR task is not solved ab ovo, but with the use of the information received from a previous stage. In fact, it is proposed to deal with a system of permanently improved solutions. It leads to more effective interpretation of the obtained QSAR models because the approach reveals molecular fragments/models for which the detailed development of structure is important.

The main feature of the strategy presented consists of the multiple-aspect hierarchy (Figure 5-1), related to

- models describing molecular structure (1D → 2D → 3D → 4D);
- scales of activity estimation (binomial → nominal → ordinal → continual);
- mathematical methods used to establish structure–activity relationships [pattern recognition → rank correlation → multivariate regression → partial least squares (PLS)];
- final aims of the solution of the QSAR task (prediction → interpretation → structure optimization → molecular design).

The set of different QSAR models that supplement each other results from the HiT QSAR application. These models altogether, in combination, solve the problems of virtual screening, evaluation of the influence of structural factors on activity, modification of known molecular structures, and the design of new high-potency potential antiviral agents or other compounds with desired properties.

The scheme for HiT QSAR is shown in Figure 5-1. The information from the lowest level QSAR models has been transferred (curved arrow) to the highest

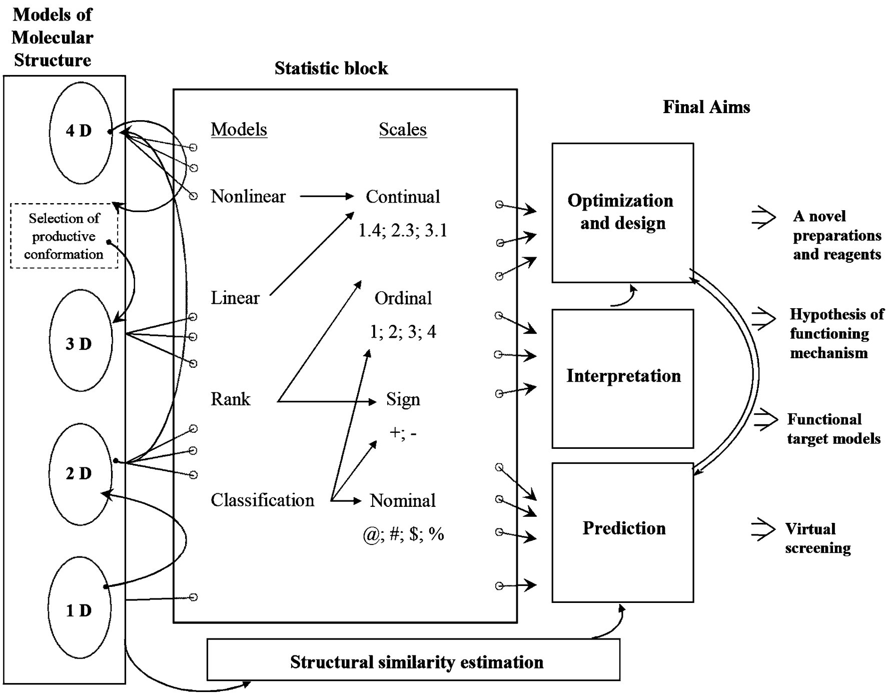


Figure 5-1. Scheme of the hierarchical QSAR technology

level models following corresponding statistical processing ("Statistic block" in Figure 5-1), during which the most significant structural parameters have been chosen. It is necessary to note that after the 2D modeling, the QSAR task is solved at the 4D level, because there is no a priori information available about a "productive" conformation (the conformer that interacts with a biological target most effectively) for 3D-QSAR models. This information comes only after the development of 4D-QSAR models and activity calculation for all conformers considered. Then the information about the "productive" (the most active) conformation is transferred to the 3D-QSAR level. This is the main difference between HiT QSAR and ordinary 3D-QSAR approaches, where the investigated conformers have been chosen through a less vigorous process. When an investigated activity is mainly determined by the interaction of the exact "productive" conformation (not by the set of conformers) with a biological target, it is possible to construct the most adequate "structure–property" models at this stage. In all cases (1D–4D), different statistical methods can be used to obtain the QSAR models (the "Statistic block" in Figure 5-1).

The principal feature of the HiT QSAR is its multi-hierarchy, i.e., not only the hierarchy of different models but also that the hierarchy of the aims has been taken into account (Figure 5-1, unit –"Final Aims"). Evidently, it is very difficult to obtain a model that can solve all the problems related to the influence of the structure of the studied molecules to the property examined. Thus, to solve every definitive task, it is necessary to develop a set of different QSAR models, where some of them are more suitable for the prediction of the studied property, the others for the interpretation of the obtained relationships, and the third for molecular design. These models altogether, in combination, solve the problem of the creation of the new compounds and issue relating to the desired set of properties. The important feature of such an approach is that the general results obtained from a few different independent models always are more relevant. It's also necessary to note that these resulting QSAR models have been chosen in accordance with the QECD principles for the validation of (Q)SARs [38], i.e., they have a defined endpoint, an unambiguous algorithm, a defined domain of applicability (DA), mechanistic interpretation, have good statistical fit, and are robust and predictive. Thus, we assume that the proposed strategy provides a solution to solve all problems dealing with virtual screening, modeling of functional (biological) targets, advancement of hypotheses regarding mechanisms of action, and, finally, the design of the new compounds with desired properties.

5.2.2. Hierarchy of Molecular Models

5.2.2.1. Simplex Representation of Molecular Structure (SiRMS)

In the framework of SiRMS, any molecule can be represented as a system of different simplexes (tetratomic fragments of fixed composition, structure, chirality, and symmetry) [29, 31, 32, 39] (Figure 5-2).

1D models. At the 1D level, a simplex is a combination of four atoms contained in the molecule (Figure 5-2). The simplex descriptor (SD) at this level is the number of quadruples of atoms of the definite composition. For the compound $(A_aB_bC_cD_dE_eF_f\ldots)$, the value of SD $(A_iB_jC_lD_m)$ is $К = f(i)\cdot f(j)\cdot f(l)\cdot f(m)$, where, for example Eq. (5-1),

$$f(i)\frac{a!}{(a-i)!\cdot i!} \tag{5-1}$$

The values of $f(j)$, $f(l)$, $f(m)$ have been calculated analogically. It is possible to define the number of smaller fragments ("pairs," "triples") by the same scheme. In this case some of i, j, l, m parameters are equal to zero.

2D models. At the 2D level, the connectivity of atoms in simplex, atom type, and bond nature (single, double, triple, aromatic) has been considered. Atoms in simplex can be differentiated on the basis of different characteristics, especially

- atom individuality (nature or more detailed type of atom);
- partial atom charge [40] (see Figure 5-2) (reflects electrostatic properties);
- lipophilicity of atom [41] (reflects hydrophobic properties);
- atomic refraction [42] (partially reflects the ability of the atom to dispersion interactions);

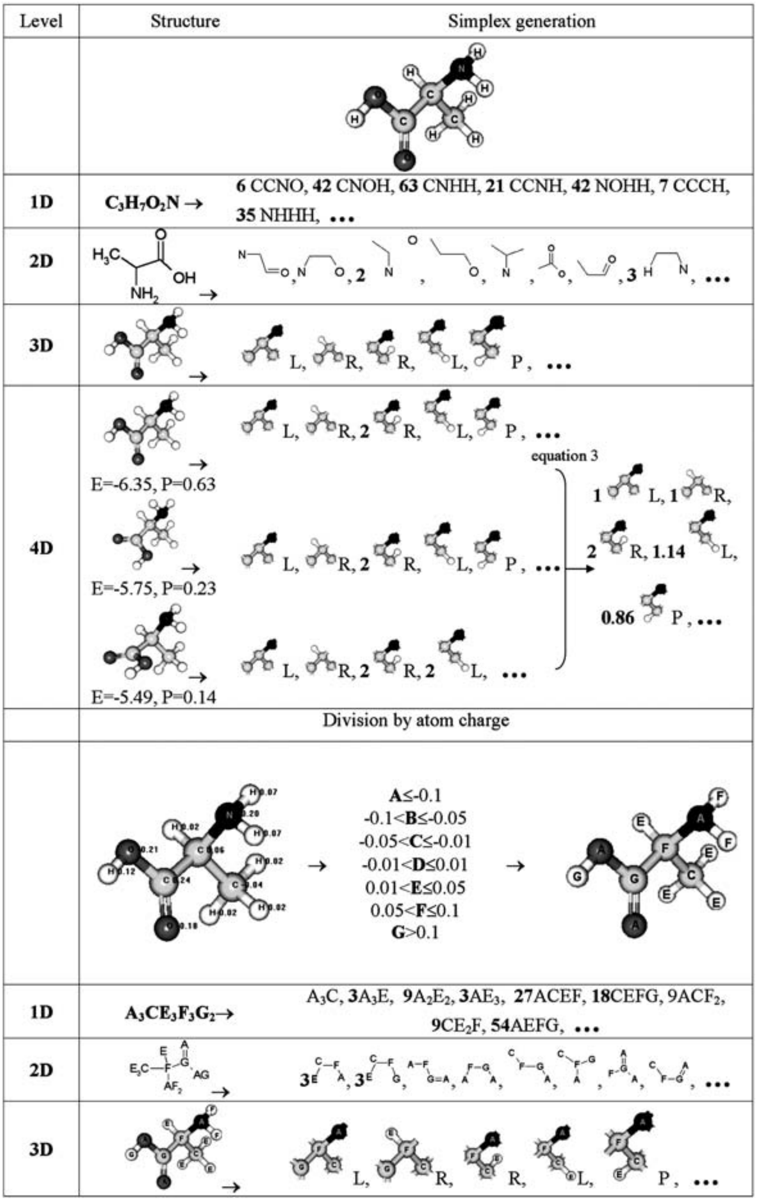


Figure 5-2. Examples of simplex descriptors generation for alanine at the 1D–4D levels

- a mark that characterizes the atom as a possible a Hydrogen donor or acceptor (A – Hydrogen acceptor in H-bond, D – Hydrogen donor in H-bond, I – no bond).

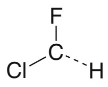
For atomic characteristics, which have real values (charge, lipophilicity, refraction, etc.) the division of values range into definite discrete groups is carried out at the preliminary stage. The number of groups (G) is a tuning parameter and can be varied (as a rule $G = 3–7$).

The usage of sundry variants of simplex vertexes (atoms) differentiation represents an important part of SiRMS. We consider that specification of atoms only by their nature (actually reflects atom identity, for example, C, N, O) realized in many QSAR methods limits the possibilities of pharmacophore fragment selection. For example, if the –NH– group has been selected as the fragment (pharmacophore) determining activity and the ability of H-bond formation is a factor determining its activity, H-bonds donors, for example, the OH-group will be missed. The use of atom differentiation using H-bond marks mentioned above avoids this situation. One can make analogous examples for other atomic properties (lipophilicity, partial charge, refraction, etc.).

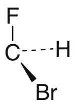
Thus, the SD at the 2D level is a number of simplexes of fixed composition and topology. It is necessary to note that, in addition to the simplex descriptors, other structural parameters, corresponding to molecular fragments of different size, can be used for 1D and 2D-QSAR analysis. The use of 1–4 atomic fragments is preferable because further extension of the fragment length could increase the probability of the model overfitting and decrease its predictivity and DA.

2.5D models. It's well known that the stereochemical moieties of the investigated compounds could affect biological activity to at least at the same level as their topology. Although the most adequate description of stereochemistry of compounds is possible only on 3D and 4D levels of molecular structure modeling, 2D models of molecules can also provide stereochemical information. In the case when a compound contains a chiral center on the atom X (X = C, Si, P, etc.), the special marks X^A, X^R, X^S (A – achiral X atom, R – "right" surrounding of X atom, S – "left" surrounding of X atom) can be used to reflect the stereochemistry information of such a center. In each case, the configuration (R or S) of a chiral center can be determined by the Kahn–Ingold–Prelog rule [43]. For example, in the situation where atom X has been differentiated to three different types depending on its stereochemical surroundings, i.e., X^A, X^R, X^S, the different types are analyzed in the molecular model as separate atoms. Conventionally, such molecular models can be considered as 2.5D because not only topological (molecular graph) but also stereochemical information has been taken into account. If simplex vertexes (X atoms) have been differentiated by some physical–chemical properties (e.g., partial charges, lipophilicity) then the differences between atoms X^A, X^R, X^S will be leveled as in normal 2D models. For subsequent QSAR analysis, the simplexes differentiated by atom individuality have been used separately and in combination with those differentiated by physical–chemical properties.

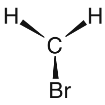
3D models. At the 3D level, not only the topology but also the stereochemistry of molecule is taken into account. It is possible to differentiate all the simplexes as right (R), left (L), symmetrical (S), and plane (P) achiral. For example:



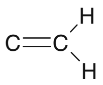
(R)



(L)



(S)



(P)

The stereochemical configuration of simplexes is defined by modified Kahn–Ingold–Prelog rules [39]. A SD at this level is a number of simplexes of fixed composition, topology, chirality, and symmetry.

4D models. For the 4D-QSAR models, each *SD* is calculated by the summation of the products of descriptor values for each conformer (SD_k) and the probability of the realization of the corresponding conformer Eq. (5-2) (P_k).

$$SD = \sum_{k=1}^{N} (SD_k \cdot P_k), \tag{5-2}$$

where *N* is a number of conformers being considered.

As is well known [44], the probability of conformation P_k is defined by its energy equation (5-3):

$$P_k = \left\{ 1 + \sum_{i \neq k} \mathrm{EXP}\left(\frac{-(E_i - E_k)}{RT} \right) \right\}^{-1}, \quad \sum_k P_k = 1, \tag{5-3}$$

where E_i and E_k are the energies of conformations *i* and *k*, respectively.

The conformers are analyzed within an energy band of 5–7 kcal/mol. Thus, the molecular SD at the 4D level takes into account the probability of the realization of the 3D-level SD in the set of conformers. At the 4D level the other 3D whole-molecule parameters, which are efficient for the description of spatial forms of the conformer (e.g., characteristics of inertia ellipsoid, dipole moment), can be used along with SD. An example of the representation of a molecule as sets of simplexes with different levels of structure detailed (1D–4D) is depicted in Figure 5-2.

Double nD models. The interaction of a mixture with a biological target cannot normally be described simply as the average between interactions of its parts, since the last interactions have different reactivity. It is also applicable for mixtures of compounds with synergetic or anti-synergetic action [45]. Because of these issues, the SiRMS approach has been developed and improved in order to make this method suitable for the execution of QSAR analysis for molecular mixtures and ensembles. With this purpose it's necessary to indicate whether the parts of unbound simplexes belong to the same molecule or to a different one. In the latter case, such unbound simplex will reflect the structure not of a single molecule, but will characterize a pair of different molecules. Simplexes of this kind are structural descriptors of the mixtures of compounds (Figure 5-3). Their usage allows for the analysis of synergism, anti-synergism, or competition in the mixture's interaction with the biological target. Obviously, such an approach is suitable for different *n*D-QSAR models, where $n = 1–4$[1]. If in the same task both mixtures and single compounds have been considered, it's necessary to represent individual compounds as the mixture of two similar molecules for the correct description of such systems [46]. Thus, this approach has been named by authors as "double *n*D-QSAR." Although such methodic is suitable only for binary mixtures, it can be easily extended to more complicated tasks. For molecular ensembles (associates), it is necessary to use one more simplex type – simplexes with intermolecular bonds.

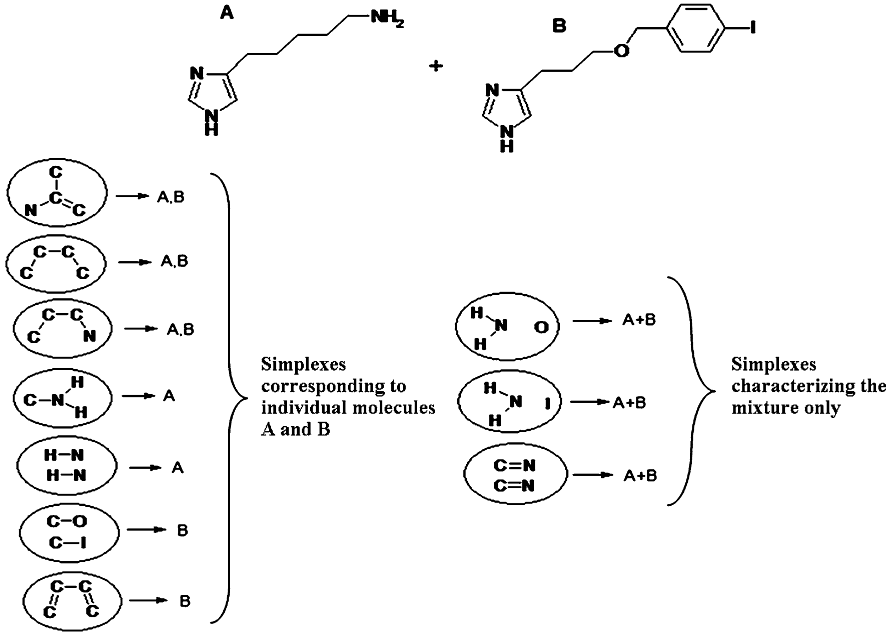

A
NH2
N
N
H
+
B
O
I
N
N
H
A,B
A,B
A,B
A
A
B
B
Simplexes corresponding to individual molecules A and B
A+B
A+B
A+B
Simplexes characterizing the mixture only

Figure 5-3. Example of structure description of the mixture of antagonists of histamine H3 – receptors (A-imphetamine, B-iodoproxiphane)

[1] For 1D-QSAR models unbounded simplexes characterize only the mixtures.

In this chapter, the application of the "double *n*D-QSAR" approach is demonstrated with the example of chiral AChE inhibitors [46] (see Section 5.4.4).

5.2.2.2. Lattice Model

The lattice model (LM) approach has been developed by the authors [19] using similar principles as CoMFA and CoMSIA (see Chapter 4), which utilize a more elaborated description of the molecules and consider parameters reflecting peculiarities of the intermolecular interaction of the compounds analyzed and their spatial structure. However, in addition, molecular properties are described with a variety of complementary parameters. The whole set of parameters generated ranges from the most simple, such as the presence or absence of particular atoms in the molecule, to more sophisticated parameters that could be used for the consideration of the stereochemistry of the analyzed molecule and its interaction with the environment.

The description of compounds includes several steps. In the first, the spatial structure of the analyzed molecules is obtained from experimental data (i.e., X-ray analysis) or from quantum mechanical calculations. In the case of flexible molecules, it is necessary to select one of the stable conformations. This may be achieved using a conformational search [47] or some complementary information regarding the biologically active conformation of the molecule. The conformation of each molecule is placed into a lattice of cubic cells. The size of a cell can be varied, by default it equals 2 Å, that corresponds approximately to the average van der Waals radius of an organogenic atom. The invariant disposition of the molecule in the lattice is achieved by the superposition of the center of mass of the molecules with the origin of the coordinates. In addition, the principal axes of inertia of the molecule are also superimposed with the coordinate axes of the lattice. If the analyzed structures contain a large common structural fragment, their alignment is carried out mainly according to this fragment.

All structural parameters in the LM can be classified as follows:

- Integral parameters describing properties of the whole molecular structure;
- Local parameters describing the separate fragments of the molecule;
- Field parameters describing the influence of the molecule on the enclosing space.

Integral parameters are characteristics of inertia ellipsoid, dipole moment, molecular refraction, lipophilicity, parachor, and average polarizability. If available, some information about the environment and mutual disposition of the pharmacophores can be also included into the analysis [48].

Local parameters were used to describe the properties of cells occupied by atoms. They include parameters corresponding to the presence or absence of some atoms in the cell (i.e., the presence of C or O), average lipophilicity, refraction, polarizability, electrostatic charge, and electronegativity of fragments and atoms. All charge characteristics were calculated using the Jolly-Perry [40, 49] method of smoothing of electronegativity.

Field parameters describe the characteristics of vacant cells. They include

(1) An electrostatic potential in the vacant cell [Eq. (5-4)]:

$$\mathrm{EP}_i = \sum_j^{n_1} \frac{q_j}{r_{ij}} \tag{5-4}$$

where i is the number of the cell, j is the number of the atom, q_j is the charge of the atom j [40, 49], and r_{ij} is the distance between the atom j and the cell i;

(2) A lipophilicity potential [50] in the vacant cell [Eq. (5-5)]:

$$\mathrm{LP}_i = \sum_j \frac{f_j}{(1 + r_{ij})} \tag{5-5}$$

where i is the number of the cell, j is the number of the atom, f_j is the lipophilicity of the atom (group), and r_{ij} is the distance between the atom j and the cell i;

(3) A probability of an occupancy of a vacant cell by different atoms *i, k* ("probe-atoms") or probability of it to be empty [Eq. (5-6)]:

$$P_k = \left\{ 1 + \sum_{i \neq k} \mathrm{EXP}\left(\frac{-(E_i - E_k)}{RT} \right) \right\}^{-1}, \quad \sum_k P_k = 1, \tag{5-6}$$

where E_i or E_k is the energy of interaction between the molecule and the corresponding probe-atom i or k in the analyzed cell.

A set of atoms $\mathrm{C_{sp}}^3$, $\mathrm{N_{sp}}^3$, $\mathrm{O_{sp}}^3$, $\mathrm{C_{sp}}^2$, $\mathrm{N_{sp}}^2$, $\mathrm{O_{sp}}^2$ Cl, H and the absence of any atom ("vacuum") were used as probes. If CoMFA [16] uses energy attributes to characterize the analyzed cells, in LM the probabilities of the occupancy of a cell represents a different approach for the description of interactions between the molecule and the biological target. It might be argued that a probability-based scheme offers improvements over an energy-based method.

(4) A possibility of the presence of hydrogen bond donor or acceptors in the cell. It is assumed that such a hydrogen bond can be formed between this donor or this acceptor and the analyzed molecule.

All structural parameters, i.e., integral, local, and field parameters contain an exhaustive description of the molecular structure. Thousands of descriptors (their exact number depends on the characteristics of the lattice) are generated within the proposed approach for each analyzed molecule. This reduces the probability of missing the most significant parameters required to correlate activity of the analyzed molecules with their structure.

The efficiency of the LM approach has been demonstrated on different tasks, e.g., [19, 48, 51, 52].

5.2.2.3. Whole-Molecule Descriptors and Fourier Transform of Local Parameters

SDs at all levels of differentiation (1D–4D) are the fragmentary parameters which describe not a molecule as a whole, but its different parts. In order to reflect the structural features of a whole molecule, it is necessary to carry out the Fourier transformation [53] for the spectrum of structural parameters. The spectrum of structural parameters is the discrete row of values arranged in a determined order. The mode of ordering is not crucial (frequently descriptors are lexicographically ordered), but it must be the same for all compounds of an investigated task. As a result of the Fourier transformation, the high-frequency harmonics characterize small fragments while the low-frequency harmonics correspond to the global molecule properties. The Fourier transformation of a discrete function of parameters $P(i)$ can be presented as Eq. (5-7):

$$P(i) = \frac{a_0}{2} + \sum_{k=1}^{M-1}\left(a_k \cos\frac{2\pi k(i-1)}{N} + b_k \sin\frac{2\pi k(i-1)}{N}\right) + a_{N/2}\cos\left(\pi(i-1)\right) \tag{5-7}$$

where

$$a_k\frac{2}{N}\cdot\sum_{i=1}^{N} P_i\cdot\cos\left(\frac{2\pi\cdot k\cdot(i-1)}{N}\right),\quad b_k = \frac{2}{N}\cdot\sum_{i=1}^{N} P_i\cdot\sin\left(\frac{2\pi\cdot k\cdot(i-1)}{N}\right) \tag{5-8}$$

or in an alternative form [Eq. (5-9)]

$$p(i) = \frac{q_0}{2} + \sum_{k=1}^{M-1}\left(q_k \sin\left[\frac{2\pi k(i-1)}{N} + \psi_k\right]\right) + q_{n/2}\cos\left[\pi(i-1)\right], \tag{5-9}$$

The amplitudes and phase angle in Eq. 5-9 are defined as follows:

$$\text{Amplitudes: } q_k = \sqrt{a_k^2 + b_k^2}, \text{ Phase angle: } \psi_k = \arctan\left((a_k)/b_k\right). \tag{5-10}$$

where k is the number of harmonics, N is the total number of simplex descriptors, $M = \text{int}(N–1)/2$ is the total number of harmonics, a_k and b_k are the coefficients of expansion procedure, $q_{n/2} = 0$ for even N.

Values of amplitudes (a_k, b_k, q_k) can be used as the parameters for the solution of QSAR tasks [19, 54]. PLS equations containing amplitudes a_k and b_k can be mechanistically interpreted, because they can be represented as a linear combination of source structural parameters (5-7). Amplitudes q_k have poor mechanistic interpretation because of the more complex dependence from the source structural parameters (5-9). However, all the amplitudes (a_k, b_k, q_k) separately or together allow for well-fitted, robust, and predictive models to be obtained; hence, they can be used as an additional (completely different) tool for the virtual screening.

Such whole-molecule parameters, such as characteristics of inertia ellipsoid (moments of inertia I_X, I_Y, I_Z and its ratio I_X/I_Y, I_Y/I_Z, I_X/I_Z), dipole moment, molecular refraction, lipophilicity, also can be used for different levels of representation of the molecular structure.

All mentioned integral parameters can be united with SD which usually leads to the most adequate model that unites the advantages of molecular descriptors of every mentioned type.

5.2.3. Hierarchy of Statistical Methods

As was mentioned above, different statistical methods have been used in HiT QSAR to establish the structure–activity relationship depending on the scale of the investigated property (binomial → nominal → ordinal → continual).

5.2.3.1. Classification Trees

The classification tree (CT) approach is a non-parametric statistical method of analysis [55]. It allows for the analysis of data sets regardless of the number of investigated compounds and the number of their characteristics (descriptors). In the CT approach, the models obtained represent the hierarchical sets of rules based on descriptors selected for the description of the investigated property. The rule represents "IF-THEN" logical construction. For example, a simple rule can be "IF lipophilicity > 3 THEN compound is active." In fact, such model is presented by a set of consecutive nodes, and each of them contains certain sets of compounds which correspond to this node rule. The CT method has several advantages: obtaining of intuitively understandable models using natural language, quick learning and predicting processes, non-linearity of obtained models, and the ability to develop models using ranked values of the activity (it allows for the analysis of sets of compounds with heterogeneous experimental activity values).

The usage of CT methods for QSAR analysis is limited due to the poor mechanistic interpretation of the models. It is difficult to make quantitative estimation of the influence of descriptors used in the model and to determine structural fragments interfering or promoting activity.

A new approach for the interpretation of CT models, based on a trend-vector procedure (see Section 5.2.3.3), has been proposed to solve this problem. It allows for the determination of the quantitative influence of descriptors used in the model built on the investigated property [Eq. (5-11)]:

$$T_j = \frac{1}{m} \sum_{i=1}^{m} [(A_i - A_{\text{mean}})] \tag{5-11}$$

where T_j is the relative influence of jth descriptor on investigated property, m is the number of compounds in the certain node, A_i is the activity rank of ith compound, and A_{mean} is the mean value of activity rank for the whole set of compounds.

The relative influence (T_j) of each descriptor used in the CT model are calculated by applying Eq. (5-11) to each node of the model (excepting the root node). Furthermore, each calculated influence has a corresponding range of descriptor values (D) according to node rule, within which this influence has been implemented. As a result of such analyses, ranges of descriptor values and corresponding relative influences can be determined. When descriptor has several overlapping ranges of values then the relative influence values should be summarized in the overlapping interval.

The approach described is valid only for models with classification scale of activity. It can be considered as a restriction of the method. However, estimation of activity level is an appropriate result in many cases relating to the investigation of biological activity. In the case of the usage of simplex (fragmentary) descriptors for the representation of molecular structure, T_j values obtained in this manner are the cumulative influences of all simplexes of a certain type in the molecule. It allows for the calculation of the relative atomic influences for each investigated compound according to Eq. (5-12).

$$T_a = \frac{T_j}{4N_j} \tag{5-12}$$

where T_a is the relative influence of each atom included in the jth simplex of certain molecule, T_j is the relative influence of the jth simplex, $4N_j$ is the number of jth simplexes (value of jth descriptor) in certain molecules multiplied by four (number of atoms in a simplex).

Calculated relative atom influences can be visualized on the investigated compounds. They allow for the determination of the relative influences of separate molecular fragments by summarizing the influences of individual atoms included in certain fragments.

5.2.3.2. Trend-Vector

The trend-vector (TV) procedure [19, 56, 57] does not depend on the form of corresponding dependence and can use many structural parameters. This method can predict the properties of analyzed molecules only in a rank scale and can be used if biological data are represented in an ordinal scale (see Figure 5-1). Similar to a dipole moment vector, TV characterizes a division of "conventional charges" (corresponding to active and inactive classes) in the multi-dimensional space of structural parameters S_{ij} ($i = \overline{1,n}$ – number of molecules, $j = \overline{1,m}$ – number of structural parameters). Each component of a TV is determined by Eq. (5-13)

$$T_j = \frac{1}{n} \cdot \sum_{i=1}^{n} (A_i - \bar{A}) \cdot S_{ij}, \tag{5-13}$$

and reflects a degree and direction of influence of the jth structural parameter on the magnitude of a property A. The prediction of activity is obtained using the following relation:

$$\mathrm{rank}(A_i) = \mathrm{rank}\left(\sum_{j=1}^{m} T_j S_{ij}\right) \tag{5-14}$$

It is important to note that each component of the TV is calculated independently from the others and its contribution to a model is not adjusted. Thus, the influence on the reliability of the model of the number of structural parameters used is not so critical, as in the case of the regression methods. The quality of the structure–property relationship can be estimated by the Spearman rank correlation coefficient calculated between ranks of the experimental and calculated activities A_i.

The search for models using the TV method in HiT QSAR is achieved by the methods of exhaustive or partial search after the removal of mutual correlations. It was discovered by the authors [10, 32] that descriptors involved in the best TV models (several decades of models with approximately identical quality) form a good subset for the subsequent usage in PLS. Noise elimination can be one of the probable explanations of the success of the TV procedure.

5.2.3.3. Multiple Linear Regression

The greatest number of QSAR/QSPR investigations has been made using linear statistic methods [58]. In such approaches, the investigated property is represented as a linear function of calculated descriptors [Eq. (5-15)]:

$$y' = a_0 + \sum_{i=1}^{n} a_i x_i \tag{5-15}$$

where y' is the calculated values of investigated property (y), x_i is the structural descriptors (independent variables), a_i is the regression coefficients determined during the analysis by the least squares method, n is the number of variables in the regression equation.

The use of linear approaches is very convenient for investigations because the theory of selection of the most important attributes and obtaining of the final equations is well developed for such methods. The quality of the obtained model is estimated by the correlation coefficient *R* between the observed values of the investigated property (y) and those predicted by Eq. (5-15) (y'). The R^2 value is explained by regression measure of the part of common scatter relative to average y. The term of adequacy of the obtained regression model with the chosen level of risk α will be *F* [Eq. (5-16)] [58]:

$$F = \frac{R^2(m-n-1)}{(1-R^2)\cdot n} \geq F_{xp}\ (n-1, m-n, \alpha), \tag{5-16}$$

where m is the number of molecules in the training set and F_{xp} (n–1, m–n, 1–α) is the percent points of the F-distribution for given level of significance 1–α.

The relative simplicity of regression approaches is also their shortcoming; they show poor results during the extrapolation of complicated structure–activity relationships. Their usage is further hampered in the case of large numbers of descriptors, since the total number of descriptors in a MLR equation must be at least ten times fewer than the number of training set compounds [59].

5.2.3.4. Partial Least Squares or Projection to Latent Structures (PLS)

A great number of simplex descriptors have been generated in HiT QSAR. The PLS-method has proved efficient for working with a great number of variables [60–62]. The PLS regression model may be written as Eq. (5-17) [62]:

$$Y = b_0 + \sum_{i=1}^{N} b_i x_i, \tag{5-17}$$

where Y is an appropriate activity, b_i are the PLS regression coefficients, x_i is the ith descriptor value, and N is the total number of descriptors.

This is not apparently different from MLR (see Section 5.2.3.3), except that the values of the coefficients b are calculated using PLS. However, the assumptions underlying PLS are radically different from those of MLR. In PLS one assumes the x-variables to be collinear and PLS estimates the covariance structure in terms of a limited number of weights and loadings. In this way, PLS can analyze any number of x-variables (K) relating to the number of objects (N) [62].

5.2.4. Data Cleaning and Mining

The removal of highly correlated and constant descriptors, the use of genetic algorithms (GA) [63], trend-vector methods [56, 57], and automatic variable selection (AVS) strategies that are similar to interactive variable selection (IVS) [61] and evolutionary variable selection (EVS) [60] have been used for selection of descriptors in PLS. The removal of highly correlated descriptors is not necessary for PLS analysis, since descriptors are reduced to series of uncorrelated latent variables. However, this procedure frequently helps to obtain more adequate models and reduce a number of used variables up to five times. During this procedure one descriptor from each pair having a pair correlation coefficient r satisfying $|r| > 0.90$ has been eliminated.

5.2.4.1. Automatic Variable Selection (AVS) Strategy in PLS

The AVS strategy in PLS is used to obtain highly adequate models by removing the "noise" data, i.e., systematic variations in X (descriptors space) that are orthogonal to Y (investigated property). This strategy is similar to IVS [61], EVS [60], OSC [64], and O-PLS [65] and has the same objective but uses different means.

The essence of AVS consists of the following: at the first step of the AVS the model containing all descriptors is obtained. Then variables with the smallest normalized regression coefficients (b_i, Eq. (5-17)) are excluded from the X-matrix and in the next step the PLS model is obtained. This procedure has been repeated stepwise until the amount of variables equals 1. The AVS strategy can be used either for all structural parameters or after different variable selection procedures (e.g., removal of highly correlated descriptors, TV procedure, GA). An application of the AVS procedure resulted in the decreasing of the model complexity (number of descriptors and latent variables) and an increase in model predictivity and robustness.

5.2.4.2. Genetic Algorithms

GA imitates such properties of living nature as natural selection, adaptability, heredity. The use of the heuristic organized operations of "reproduction," "crossing," and "mutation" from casual or user-selected starting "populations" generates the new "chromosomes" – or models. The utility of the GA is its flexibility. With adjustment of the small set of algorithm parameters (number of generations, crossover and mutation type, crossover and mutation probability, and type of selection), it is possible to find a balance between the time for search and the quality of decision. In the HiT QSAR, GA is used as a tool for the selection of adequate PLS, MLR, and TV models. Descriptors from the best model obtained by the preliminary AVS procedure have usually been used as the starting "population." GA is not a tool for the elucidation of the global maximum or minimum, and very often a subsequent AVS procedure and different enumerative techniques allow one to increase the quality of the obtained PLS models.

5.2.4.3. Enumerative Techniques

As mentioned above, the usage of the methods of exhaustive or partial searching (depending on the number of selected descriptors) after AVS or GA very often allow one to increase the quality of the obtained models (PLS, MLR, and TV). After the statistical processing model or models with the best combinations of statistic characteristics (R^2, Q^2) have been selected from the obtained resulting list, and they may be submitted for subsequent validation using an external test set. The general scheme

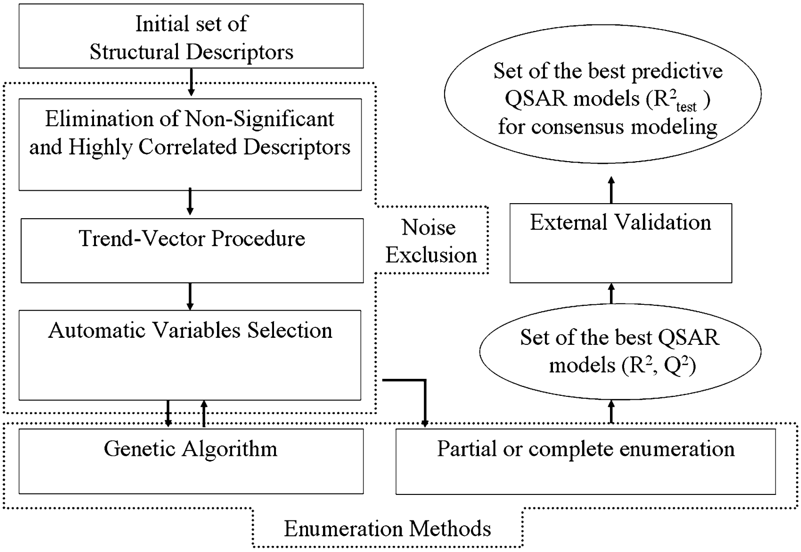

Initial set of Structural Descriptors
Elimination of Non-Significant and Highly Correlated Descriptors
Trend-Vector Procedure
Automatic Variables Selection
Noise Exclusion
Genetic Algorithm
Partial or complete enumeration
Enumeration Methods
Set of the best QSAR models (R^2, Q^2)
External Validation
Set of the best predictive QSAR models (R^2_{test}) for consensus modeling

Figure 5-4. General scheme of the PLS models generation and selection applied in the HiT QSAR

of the PLS model generation and selection applied in the HiT QSAR is presented in Figure 5-4. This procedure can be repeated several times using as input an initial set of SD of different levels of molecular structure representation (usually 2D–4D) and/or with various kinds of atom differentiation (see above) with the purpose to develop several resulting "predictive" QSAR models for consensus modeling. This approach is believed to yield more accurate predictions.

5.2.5. Validation of QSAR Models

To have any practical utility, up-to-date QSAR investigations must be used to make predictions [66]. The statistical fit of a QSAR can be assessed in many easily available statistical terms (e.g., correlation coefficient R^2, cross-validation correlation coefficient Q^2, standard error of prediction S).

Cross-validation is the statistical practice of partitioning a sample of data into subsets such that the analysis is initially performed on a single subset, while the other subset(s) are retained for subsequent use to confirm and validate the initial analysis. The initial subset of data is called the training set; the other subset(s) are called validation sets. In QSAR analysis, only two types of cross-validation are used:

(1) *K-fold cross-validation*. In K-fold cross-validation, the original sample is partitioned into K subsamples. Of the K subsamples, a single subsample is retained as the validation data for testing the model and the remaining K – 1 subsamples are used as training data. The cross-validation process is then repeated K times (the folds), with each of the K subsamples used exactly once as the validation data.

(2) *Leave-one-out cross-validation.* As the name suggests, leave-one-out cross-validation (LOOCV) involves using a single observation from the original sample as the validation datum and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data. This is the same as a K-fold cross-validation with K being equal to the number of observations in the original sample.

The determination coefficient (Q^2) calculated in cross-validation terms is the main characteristic of model robustness. Q^2 is calculated by the following formula:

$$Q^2 = 1 - \frac{\sum_Y (Y_{\text{pred}} - Y_{\text{actual}})^2}{\sum_Y (Y_{\text{actual}} - Y_{\text{mean}})^2} \tag{5.18}$$

where Y_{pred} is a predicted value of activity, Y_{actual} is an actual or experimental value of activity, and Y_{mean} is the mean activity value.

The shortfalls of cross-validation are the following:

(1) The training task must be solved N times leading to substantial calculative expenses in time and resources.
(2) The estimation of cross-validation assumes that the training algorithm is already given. It has no idea how to obtain "good" algorithms and which properties must be inherent to them.
(3) An attempt to use cross-validation for training as an optimizable criterion leads to loss of its unbiasedness property and there is a risk of overfitting.

At the same time statistical fit should not be confused with the ability of a model to make predictions. The only method to obtain a meaningful assessment of statistical fit is to utilize the so-called "test set". During this procedure a certain proportion of the data set molecules (10–85%) are removed to form the test set before the modeling process begins (remaining molecules form the training set). Once a model has been developed, predictions can be made for the test set. This is the only method by which the validity of a QSAR can be more or less truly assessed. However, one must understand that sometimes it means only the model ability to predict the certain test set. It is important that both training and test sets cover the structural space of the complete data set as much as possible.

In the HiT QSAR, the following procedure has been used for the formation of the test set: a dissimilarity matrix for all initial training set molecules has been developed on the basis of relevant structural descriptors. Such a descriptor set can be obtained using different procedures for descriptor selection (for example, see Chapter 4) or directly from the model generated for all investigated compounds. In our opinion the use of the whole set of descriptors generated at the very beginning is not completely correct, because during QSAR research we are interested not in structural similarity by itself, but from the point of view of the investigated activity and the descriptor selection will help the avoidance of some distortions caused by the insignificance of structural parameters from the initial set for this task.

A dissimilarity matrix is based on the estimation of structural dissimilarity between all investigated molecules. A measure of the structural dissimilarity for molecules M, M' can be calculated using the Euclidean distance in the multidimensional space of structural parameters S [Eq. (5-19)]:

$$\mathrm{SD}(M, M') = \sqrt{\sum_{i=1}^{n} (S_i - S'_i)^2}, \tag{5.19}$$

where n is the a number of molecules in data set.

Thus, total structural dissimilarity toward the rest of initial training set compounds could be calculated for every molecule from a sum of the corresponding Euclidean distances. In the meanwhile, all the compounds were divided into groups depending on their activity, where the number of groups equals the number of molecules that one wants to include into test set. Then one compound from each group has been chosen to go to the test set according to its maximal (or minimal) total Euclidean distance from the other molecules in this group, or by random choice. Most likely, the use of several (three is the enough minimum) test sets constructed by different principles and subsequent comparison and averaging of the obtained results is more preferable than the use of only one set for the model validation. In that way, the first test set has been constructed to maximize its diversity from the training set, i.e., the compounds with maximal dissimilarity were chosen. This is the most rigorous estimation, sometimes it can lead to the elimination of all of the dissimilar compounds from the training set, i.e., such splitting of the training set when the test set structures would not be predicted correctly by the developed model and would be situated outside of DA. The second test set is created in order to minimize its diversity from the training set, i.e., less dissimilar compounds from each group were removed. The last test set has been chosen in random manner taking into account activity variation only.

5.2.6. Hierarchy of Aims of QSAR Investigation

HiT QSAR provides not only hierarchy of molecular models, systems of descriptors, and statistical models, but also the hierarchy of the aims of QSAR investigation (Figure 5-1). Targets of the first level are activity prediction or virtual screening. Any descriptors could be used here, even those that are only poorly interpretable or noninterpretable, e.g., different topological indices, informational-topological indices, eigenvalues of various structural matrices. In other words, at this level descriptors which are not expected to be used for subsequent analysis of structural factors promoting or interfering with activity can be used.

The aims of the second level must include the interpretability of obtained QSAR models. Only descriptors which have clear physico-chemical meaning, e.g., reflecting such parameters of the molecule such as dipole moment, lipophilicity, polarizability, van der Waals volume, can be used at this level. Analysis of

QSAR models corresponding to this level allows one to reveal structural factors promoting or interfering with the investigated property. Such information can be useful for the generation of hypotheses about mechanisms of biological action and assumptions about the structure of biological target. Finally, the presence of information useful for molecular design is expected from QSAR models corresponding to the third level of purposes. As a rule, fragmentary descriptors have been used in such models. In this case, the analysis of the degree and direction of influence of such descriptors on activity can give immediate information for the optimization of known structures and design of novel substances with desired properties.

5.2.6.1. Virtual Screening (Including Consensus Modeling and DA)

As mentioned above, QSAR investigations must be used to make predictions for compounds with unknown activity values (so-called "virtual screening"). In order to increase the quality of predictions, these authors recently started to apply consensus QSAR modeling which has become more and more popular [67]. It also represents one of the crucial concepts of HiT QSAR [31, 36] and can be briefly described by the statement "More models that are good and different." The efficiency of this technique can be easily explained by the fact that nearly the same predictions obtained by different and independent methods (either statistical or descriptors generation) are more reliable than single prediction made by even the best fitted and predictable model.

From another aspect, in order to analyze the predictivity of PLS models and according to the OECD QSAR principles [38], different DA procedures have been included in the HiT QSAR. The first procedure is an integral DA called "ellipsoid" developed by the authors [11]. It represents a line at the 1D level; an ellipse at the 2D level; an ellipsoid at the 3D level; and multidimensional ellipsoids in more complicated *n*-dimensional spaces. Its essence consists of the following: the distribution of training set molecules in a space of latent variables T_1–T_A (axes of coordinates) can be obtained from PLS. For each coordinate axis (T_1 and T_2 in our case) the root-mean-square deviations S_{T1} and S_{T2} have been determined. DA represents an ellipsoid that is built from the molecules of the training set distribution center ($T_1 = 0$; $T_2 = 0$) with the semi-axes length $3S_{T1}$ and $3S_{T2}$, respectively [11] (Figure 5-5). Further, the correct positions in relation to this center have been calculated for every molecule (including molecules from prediction set). If a work set molecule does not correspond to the DA criteria, it is termed "influential," i.e., it has unique (for given training set) structural features that distinguish it from the other compounds. If a new molecule from the prediction set is situated out of the DA (region outside ellipsoid), its prognosis from the corresponding QSAR model is less reliable (model extrapolation). And, naturally, the prognoses for molecules nearest to the center of the DA are most reliable.

The second approach – the integral DA rectangle has been also developed by the authors [11]. Two extreme points (so-called virtual activity and inactivity etalons)

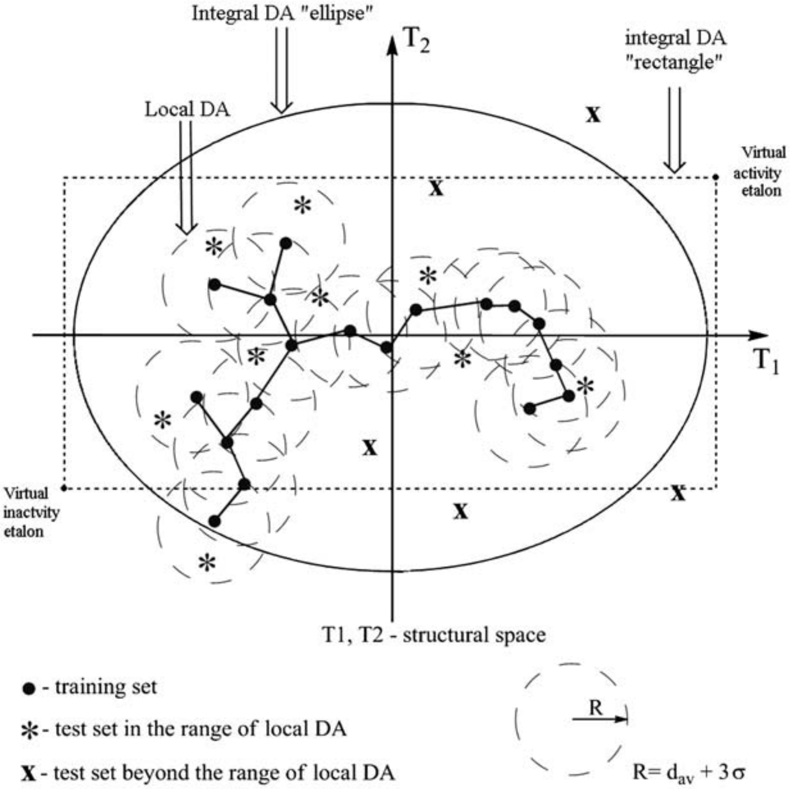


Figure 5-5. Different domain applicability procedures in the HiT QSAR: integral (*ellipsoid, rectangle*) and local

are determined in a space of structural features. The first one has maximal values of descriptors (training set data) promoting activity and minimal interfering. This point corresponds to a hypothetic molecule – the peculiar activity etalon. The second point, analogically, is an inactivity etalon, i.e. contains maximal values of descriptors interfering activity and minimal promoting. Vectors that unite these points (directed from inactive to active) depict the tendency of activity change in the variable space. This vector is a diagonal for the rectangle that determines DA [11] (Figure 5-5). All the mentioned trends concern the "influential" points from the training set and model extrapolation for new molecules from the prediction set remain and for the DA rectangle approach.

The third method is based on the estimation of leverage value h_i [68]. It has been visualized as a Williams plot [69] and is described in detail in [70]. For leverage, a value of 3 is commonly used as a cut-off value for accepting predictions, because points that lie ±3 standard deviations from the mean cover 99% of the normally distributed data. For training set molecules high leverage values do not always indicate outliers from the model, i.e., points that are outside the model domain. If high leverage points ($h_i > h_{cr}$, separated by vertical bold line) fit the model well (i.e., have small residuals), they are called "good high leverage points" or good influence points. Such points stabilize the model and make it more precise. High leverage points, which do not fit the model (i.e., have large residuals) are called "bad high leverage points" or bad influence points. They destabilize the model [70]. A new molecule is situated out of the DA (model extrapolation) if it has $h_i > h_{cr} = 3(A+1)/M$, where A – number of the PLS latent variables and M – number of molecules in a work set.

Recently, a local (Tree) approach for DA estimation has been developed by authors in order to avoid the inclusion of hollow space into the DA that is the lack of integral DA methods. The following are required for its realization:

(1) Obtaining of a distance matrix between the training set molecules in the structural space of descriptors of the QSAR model. The molecules in the given approach have been analyzed in the coordinates of the latent variables of the PLS model considered.
(2) Detection of the shortest distances between molecules using the above-mentioned matrix. Building of an extreme short distance tree for all training set molecules.
(3) Finding of average distance (d_{av}) and its root-mean-square deviation (σ) for inclusion in the tree average values. Such a distance is the characteristic of average density of molecules distribution in the structural space.

Following this procedure, all the points corresponding to test set molecules have been taken into account in the structural space. If any of test set molecules have been situated on the distance bigger than $d_{\mathrm{av}}+3\sigma$ from the nearest training set point, it means that this test set molecule is situated outside DA. Respectively, molecules belonging to the DA are situated on the distance less than $d_{\mathrm{av}}+3\sigma$ from the training set points. The scheme of DA estimation has been depicted in Figure 5-5.

Such an approach for DA estimation is similar, to some extent, to methods described in [70]. As opposed to integral approaches, e.g. [11], where the convex region (polyhedron, ellipsoid) which could contain vast cavities has been determined in the structural space, the approach presented here is local. The space of the structural parameters has been analyzed locally, i.e., regions around every training set point are analyzed. The presence of cavities in the structural space which correspond to DA is undesirable and it has been eliminated in the given approach.

Summarizing, it's necessary to note that if a new structure is lying inside the DA, it is not a final argument for a correct prediction; rather, it is an indication of the reduced uncertainty of a prediction. In exactly the same way, the situation of the compound outside the DA does not lead to the rejection of the prediction; it is just an indication of the increased uncertainty of the subsequent virtual screening prediction. Naturally, such compounds could be predicted (by model extrapolation) with great accuracy, but it will be more by co-incidence than design. Unfortunately, there is currently no unbiased estimation of prognosis reliability, and the relative character of any DA procedure was reflected in [11, 70]. Thus, it should be remembered that the DA is not a guide to action but only a probable recommendation.

All of the mentioned DA procedures together, or separately, are applied to selected single models before being averaged in consensus model. The accuracy of the DA consensus model has been compared with the adequacy of consensus models without DA consideration. The authors recommend the use of consensus DA models for subsequent virtual screening excepting the case of substantial loss of coverage of training and prediction sets with only a limited benefit in predictivity.

5.2.6.2. *Inverse Task Solution and Interpretation of QSAR Models*

Using Eq. (5-15) it is not difficult to make the inverse analysis (interpretation of QSAR models) in the frameworks of the SiRMS approach. The contribution of each *j*-atom (C_j) in the molecule can be defined as the ratio of the sum of the PLS regression coefficients (b_i) of all simplexes this atom contains (M) to a number of atoms (n) in the simplex (or fragment) [Eq. (5-20)]:

$$C_j = \frac{1}{n}\sum_{i=1}^{M} b_i, \quad (\text{for simplex}\, n = 4) \qquad (5\text{-}20)$$

According to this formula, the atom contribution depends on the number of simplexes which include this atom. This value (number of simplexes) is not constant; it varies in different molecules and depends on other constituents (surroundings), and hence, this contribution is non-additive. Atoms that have a positive or negative influence on the studied biological activity of compounds can be colored. It helps to present the results and to determine visually (additionally to the automate search) the groups of atoms affecting the activity in different directions and with varying strength. The example of the representation of the obtained results on the molecule using color-coding according to the contribution of atoms into antirhinoviral activity [11] is represented in Figure 5-6. Atoms and structural fragments reducing antiviral activity are colored in *red* (*dark gray* in printed version) and that enhance antiviral

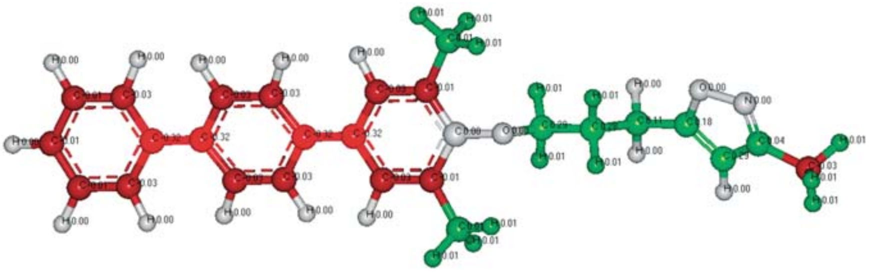
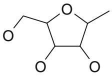
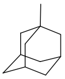
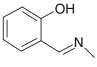
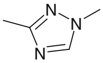
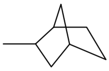
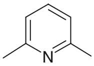


Figure 5-6. Color-coded structure according to atoms contributions to activity against HRV-2 [11]. Atoms and structural fragments reducing antiviral activity are colored in *dark gray* and that enhancing antiviral activity in *light gray* and *white*

activity in *green* (*light gray and white* in printed version). Atoms and fragments with no effect are colored in *gray*.

The automatic search procedure for pre-defined fragments from the data set and their relative effect on activity has been realized in HiT QSAR. The procedure of the fragment searching in molecule is based on a fast algorithm for solving the maximum clique problem [71]. Some molecular fragments promoting and interfering anti-influenza activity [12, 29, 34] are represented in Table 5-1 as well as their average relative influence on it.

Table 5-1. Molecular fragments governing the anti-influenza activity change (Δ lgTID$_{50}$) and their average relative influence on it [12, 29, 34]

Enhance the activity		
		
3.0	2.4	1.9
	—(CH$_2$)$_2$—O—	
1.7	1.4	0.8
Decrease the activity		
—(CH$_2$)$_n$—NH— $n = 2$–3		—CO—NH$_2$
–0.3 to –0.4	–0.2	–0.2

5.2.6.3. Molecular Design

It is possible to design compounds with a desired activity level from the SiRMS via the generation of allowed combinations of simplexes determining the investigated property. The simplest way is soft drug design [72] that consists of replacing of undesired substituents by more active ones, or by the insertion of fragments, promoting the activity instead of non-active parts of molecule or hydrogen atoms. The use of this technique allows one to retain newly designed compounds in the same region of structural space as the training set compounds. The accuracy of prognosis can be estimated using the DA techniques (see below). However, the use of soft drug design keeps within the limits of the initial chemical class of training set compounds. More drastic drug design is, certainly, more risky, but it allows for much more dramatic results. Almost certainly, new structures would lie outside the DA region. That, however, does not mean uncertainty of prediction, but extrapolation of the model predictivity and a certain lack of any DA procedure. However, at the same time, we can receive compounds of completely different (from initial training set) chemical classes as the output of such design. It was demonstrated in [12, 28, 29], where, in searching for a new antiviral and anticancer agents, we started our investigations from macrocyclic pyridinophanes and through several convolutions of QSAR analyses came out with nitrogen analogues of crown ethers in the first and acyclic aromatic structures with the azomethine fragment in the second case.

5.2.7. HiT QSAR Software

The HiT QSAR software for Windows has been designed and developed as an instrument for high-value QSAR investigations including the solution of the following tasks:

- Creation of QSAR projects;
- Calculation of lipophilicities and partial atom charges;
- Molecules superposition in the lattice approaches;
- Generation of different integral, simplex, lattice (local and field), and harmonic descriptors;
- Data mining (see Section 5.2.4);
- Obtaining of statistical models by PLS, MLR, and TV approaches with the usage of total and partial enumeration methods, GA, AVS strategy, etc.
- Inverse task solution – interpretation of the equations developed as color-coded diagrams for the molecules or their fields;
- Determination of the contributions (increments) of the fragments in the property investigated;
- Consensus modeling of the property investigated taking into account the DA of the model.

Graphic visualization of molecules, the atoms’ influence on the investigated properties, lattice models, different fields, etc. was implemented using the open graphic language (OpenGL) library from Silicon Graphics©. HiT QSAR software is accessible on your request. Please contact the authors if you have any questions about its usage. Summarizing the information above, the HiT QSAR workflow (Figure 5-7) has recently been developed and used by authors for the solution of different QSAR/QSPR tasks.

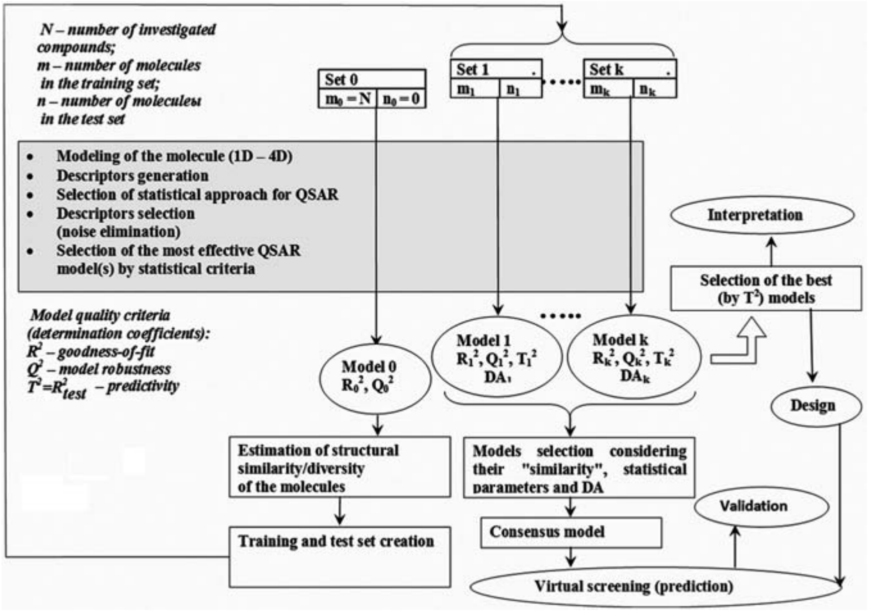

N – number of investigated compounds;
m – number of molecules in the training set;
n – number of moleculesн in the test set
Set 0
m₀ = N
n₀ = 0
Set 1
m₁
n₁
Set k
mₖ
nₖ
Modeling of the molecule (1D – 4D)
Descriptors generation
Selection of statistical approach for QSAR
Descriptors selection (noise elimination)
Selection of the most effective QSAR model(s) by statistical criteria
Model quality criteria (determination coefficients):
R² – goodness-of-fit
Q² – model robustness
T²=R²test – predictivity
Model 0
R₀², Q₀²
Model 1
R₁², Q₁², T₁²
DA₁
Model k
Rₖ², Qₖ², Tₖ²
DAₖ
Interpretation
Selection of the best (by T²) models
Design
Estimation of structural similarity/diversity of the molecules
Models selection considering their "similarity", statistical parameters and DA
Training and test set creation
Consensus model
Validation
Virtual screening (prediction)

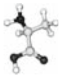
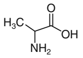
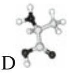
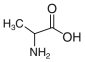
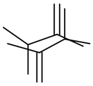
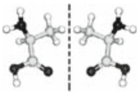
Figure 5-7. HiT QSAR workflow

As was mentioned above, the proposed technology operates on a set of different models. At the preliminary stage "Model 0" (Figure 5-7) is generated for the initial division of investigated molecules into training and test sets. Subsequent generation of sets 1–K is required for the development of consensus QSAR models. In all cases, such statistical characteristics as R^2, Q^2, $R^2{}_{\text{test}}$ have been taken into account as well as the model DA.

5.3. COMPARATIVE ANALYSIS OF HiT QSAR EFFICIENCY

The HiT QSAR based on SiRMS has proved efficient in numerous studies to solve different "structure–activity/property" problems [3, 10–12, 25–30, 33, 35] and it has been interesting to compare it with the other successful QSAR approaches and software. The results of a comparative analysis are shown in Table 5-2. Obviously, HiT QSAR does not have the problem of the optimal alignment of the set of molecules considered that is inherent to CoMFA and its analogues [16–19]. The SiRMS approach is similar to HQSAR [20] in certain ways, but has none of its restrictions (only topological representation of molecular structure an ambiguity of descriptor formation during the molecular hologram hashing). In addition, contrary to HQSAR, different physical and chemical properties of atoms (charge, lipophilicity, etc.) can be taken into account in SiRMS (Table 5-2).

Table 5-2. Comparison of different QSAR methods

Criterion		HiT QSAR	CoMFA CoMSIA HASL GRID	CODESSA DRAGON	HQSAR
Adequacy of representation of molecular structure	1D-4D	1D - 4D	3D 	2D  3D 	2D 
Absence of "molecular alignment" problem		Yes	No	Yes	Yes
Explicit con sideration of stereochem istry and chirality		Yes	Partly	No	No
Consideration of physical-chemical properties of atoms	charge, lipophilicity, polarizability etc.	Yes	Partly	Partly	No
Possibility of molecular design		Yes	Partly	No	Partly

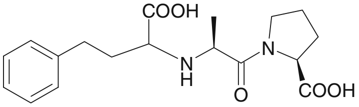
Thus, main advantages of the HiT QSAR are the following:

- The use of different (1D–4D) levels of molecular modeling;
- The absence of the “molecular alignment” problem;
- Explicit consideration of stereochemical features of molecules;
- Consideration of different physical and chemical properties of atoms;
- Clear methods (rules) for molecular design.

5.3.1. Angiotensin Converting Enzyme (ACE) Inhibitors

After such a theoretical comparative analysis, it was logical to test the efficiency of the proposed HiT QSAR on real representative sets of compounds. All such sets only contain structurally similar compounds to avoid the “molecular alignment” problem and, therefore, to facilitate the usage of the “lattice” approaches (CoMFA and CoMSIA). One hundred and fourteen angiotensin converting enzyme (ACE) inhibitors [73] represent the first set. Different statistic models obtained by

HiT QSAR have been compared with those published in [73]. The structure of enalaprat – a representative compound from the ACE data set is displayed below:



The ability of ACE inhibition (pIC_{50}) has been investigated. The training set consists of 76 compounds and 38 structures were used in a test set [73]. In the given work, we have compared the resulting PLS-models built with the use of descriptors generated from the following QSAR approaches:

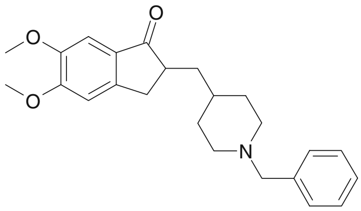
(a) CoMFA – comparative molecular fields analysis [16];
(b) CoMSIA – comparative molecular similarity indexes analysis [18];
(c) EVA – eigenvalue analysis [74];
(d) HQSAR – hologram QSAR [20];
(e) the Cerius 2 program (Accelrys, Inc., San Diego, CA) – method of traditional integral (whole-molecule) 2D and 2.5D[2] descriptors generation;
(f) HiT QSAR based on SiRMS [3, 11, 32].

Because all the mentioned approaches compare parameters generated at 2D or 3D levels of molecular structure representation, the corresponding SD, the Fourier parameters, and united models with mixed (simplex + Fourier) parameters were taken for comparison. The advantage of HiT QSAR over other methods is revealed by the comparison of such statistical descriptions of the QSAR models, as the determination coefficient for training (R^2) and test ($R^2{}_{\text{test}}$) sets; the determination coefficient calculated in the cross-validation terms (Q^2) as well as the standard errors of prediction for both sets (see Table 5-3). For example, for SiRMS Q^2= 0.81–0.87, for the Fourier models Q^2= 0.73–0.80, and for the other methods $Q^2 = 0.65$–0.72. It is necessary to note that the transition to 3D level allows for the improvement of the quality of the QSAR models obtained. At the same time, the usage of the Fourier parameters does not lead to good predictive models ($R^2{}_{\text{test}} = 0.37$–0.51) for this task. United models (simplex + Fourier) have the same predictive power as the simplex ones, but, because of the presence of integral parameters, they are sufficiently different to provide another aspect of the property.

5.3.2. Acetylcholinesterase (AChE) Inhibitors

The second set used for comparative analysis consisted of 111 acetylcholinesterase (AChE) inhibitors. The structure of E2020 – a representative compound from the AChE data set is displayed below:

[2] This classification is offered by the authors of Cerius2.



The ability to model AChE inhibition (pIC_{50}) has been investigated. The training set consists of 74 compounds and 37 structures were used as a test set [73]. The methods compared and the principles of comparison are similar to the ones described above. The main trends revealed for the ACE set were also the same for the AChE inhibitors. The advantage of HiT QSAR over other methods have been observed with all statistical parameters (Table 5-3), but especially on predictivity of the models: for SiRMS $R^2_{\text{test}} = 0.74$–0.82, for the Fourier models $R^2_{\text{test}} = 0.59$–0.61, and for the other methods $R^2_{\text{test}} = 0.16$–0.47. As in the previous case, consideration of the spatial structure of investigated compounds improved the quality of the models obtained.

Table 5-3. Statistical characteristics of the QSAR models obtained for ACE and AChE data sets by different methods

QSAR method	R^2		Q^2		R^2_{test}		S_{ws}		S_{test}		A	
	ACE	AChE	ACE	AChE	ACE	AChE	ACE	AChE	ACE	AChE	ACE	AChE
CoMFA*	0.80	0.88	0.68	0.52	0.49	0.47	1.04	0.41	1.54	0.95	3	5
CoMSIA(basic)*	0.76	0.86	0.65	0.45	0.52	0.44	1.15	0.45	1.48	0.98	3	6
CoMSIA(extra)*	0.73	0.86	0.66	0.46	0.49	0.44	1.22	0.45	1.53	0.98	2	4
EVA*	0.84	0.96	0.70	0.41	0.36	0.28	0.93	0.23	1.72	1.11	4	4
HQSAR*	0.84	0.72	0.72	0.33	0.30	0.37	0.95	0.64	1.80	1.01	4	5
Cerius 2*	0.82	0.38	0.72	0.3	0.51	0.16	1.00	0.95	1.50	1.2	4	1
Simplex 2D	0.87	0.81	0.81	0.65	0.73	0.74	0.86	0.53	1.13	0.67	2	2
Simplex 3D	0.92	0.89	0.87	0.84	0.85	0.82	0.68	0.41	0.85	0.56	2	2
Fourier 2D	0.83	0.71	0.80	0.61	0.37	0.61	0.96	0.66	1.7	0.82	5	4
Fourier 3D	0.78	0.81	0.73	0.71	0.51	0.59	1.1	0.53	1.5	0.84	4	4
Mix** 2D	0.86	0.81	0.80	0.69	0.75	0.74	0.9	0.53	1.07	0.67	2	2
Mix** 3D	0.90	0.89	0.88	0.84	0.85	0.82	0.74	0.4	0.83	0.56	2	2

where
R^2 – correlation coefficient
Q^2 – cross-validation correlation coefficient (10-fold, see Chapter 5)
R^2_{test} – correlation coefficient for test set
S_{ws} – standard error of a prediction for training set
S_{test} – standard error of a prediction for test set
A – number of PLS latent variables
*Statistic characteristics from [73] were shown
**Mix = Simplex + Fourier descriptors

Summarizing, it is necessary to note that we understand that the advantage of simplex descriptors generated in HiT QSAR may be partially a result of some of the differences in the statistical approaches applied (e.g., in addition to GA, TV and AVS procedures have been used). However, these mathematical differences are not responsible for all the improvements in the investigated approaches. Thus, it is obvious from the results obtained that HiT QSAR simplex models are well-fitted, robust and, in the main, they are much more predictive than QSAR models developed by other approaches.

5.4. HiT QSAR APPLICATIONS

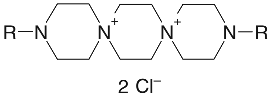
The application of HiT QSAR for the solution of different QSAR/QSPR tasks on different levels of representation of molecular structure is highlighted briefly below. The PLS method has been used for the development of QSAR models in all the cases described below.

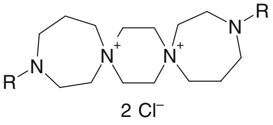
5.4.1. Antiviral Activity

Because a lot of different viral serotypes and strains exist, vaccine development for prevention of a wide variety of viral infections is considered to be impracticable. The present treatment options for such infections are unsatisfactory [75–77]. However, there are ongoing attempts to develop antiviral drugs [78–84]. That is why computational approaches, which can distinguish highly active inhibitors from less useful compounds and predict more potent substances, have been used for the analysis of antiviral activity for many years [4, 6, 7, 12, 13, 29].

5.4.1.1. Antiherpetic Activity of N,N′*-(bis-5-nitropyrimidyl) Dispirotripiperazine Derivatives*[3] *(2D)*

HiT QSAR was applied to evaluate the influence of the structure of 48 *N,N*′-(bis-5-nitropyrimidyl)dispirotripiperazines (see structures below) on their antiherpetic activity, selectivity, and cytotoxicity with the purpose to understand the chemico-biological interactions governing their activities, and to design new compounds with strong antiviral activity [3].





[3] The authors express sincere gratitude to Dr. M. Schmidtke, Prof. P. Wutzler, Dr. V. Makarov, Dr. O. Riabova, Mr. N. Kovdienko and Mr. A. Hromov for fruitful cooperation that made the development of this task possible.

The common logarithms of 50% cytotoxic concentration (CC_{50}) in GMK cells, 50% inhibitory concentration (IC_{50}) against HSV-1, and the selectivity index (SI = CC_{50}/IC_{50}) were used to develop 2D-QSAR models. Spirobromine – a medicine with a nitrogen-containing dispiro structure possessing anti-HSV-1 activity was included in the training set. The statistic characteristics of QSAR models obtained are quite high ($R^2 = 0.84$–0.91; $Q^2 = 0.61$–0.68; $R^2_{\text{test}} = 0.68$–0.71) and allow for the prediction of antiherpetic activity, cytotoxicity, and selectivity of new compounds. Electrostatic factors (38%) and hydrophobicity (25%) were the most important determinants of antiherpetic activity (Figure 5-8). The results of the QSAR analysis demonstrate a high impact of individual structural fragments for antiviral activity. Molecular fragments that promote and interfere with antiviral activity were defined on the basis of the models obtained. Thus, for example, the insertion of non-cationic linkers such as *N*-(2-aminoethyl)ethane-1,2-diamine, ethylenediamine, or piperazine instead of dispirotripiperazine leads to a complete loss of activity while the presence of methyloxirane leads to a strong increase. Using the established results and observations, several new dispirotripiperazine derivatives – potential antiviral agents – were computationally designed. Two of these new compounds (**1** and **2**, Table 5-4) were synthesized. The results of biological tests confirm the predicted high values of antiviral activity and selectivity (they are about two logarithmic units more active and one order more selective than spirobromine) as well as low toxicity of these compounds.

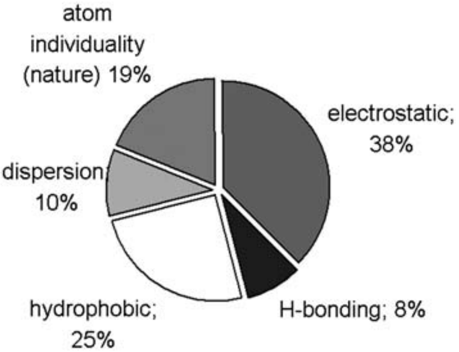


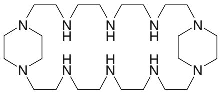
Figure 5-8. Relative influence of some physico-chemical factors on variation of anti-HSV-1 activity estimated on the basis of QSAR models

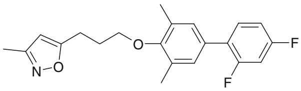
5.4.1.2. *Antiherpetic Activity of Macrocyclic Pyridinophanes*[4]

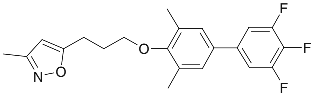
The antiherpetic data set was similar to that for the anti-influenza study and was also characterized by essential structural variety: different macrocyclic pyridinophanes and their acyclic analogues plus well-known antiviral agents including acyclovir as a reference compound:

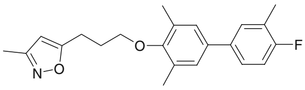
[4] Anti-influenza and antiherpetic investigations described below were carried out as a result of fruitful cooperation with Dr. V.P. Lozitsky, Dr. R.N. Lozytska, Dr. A.S. Fedtchouk, Dr. T.L. Gridina, Dr. S. Basok, Dr. D. Chikhichin, Mr. V. Chelombitko and Dr. J.-J. Vanden Eynde. The authors express sincere gratitude for all mentioned above colleagues.

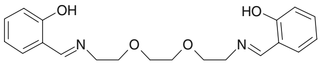
Table 5-4. Perspective potent compounds – results of computer-assisted molecular design

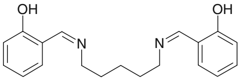

1

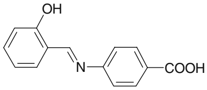

2

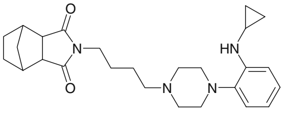

3

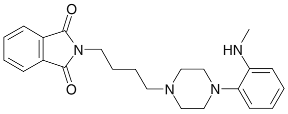

4

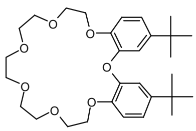

5

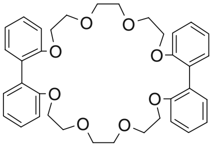

6

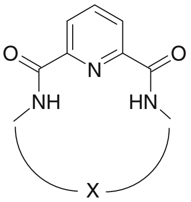

7

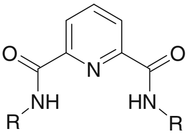

8

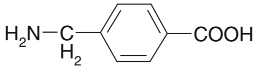

9

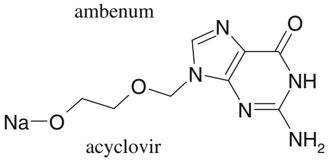

10


11






ambenum

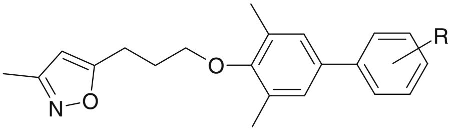

acyclovir

The antiherpetic activity against HSV-1 strain US was expressed as a percentage of the inhibition of HSV reproduction in treated cell cultures (Hep-2) in comparison with untreated ones. As in previous cases, the antiherpetic study has a multistep cyclic character: synthesis – biological tests – QSAR analysis – virtual screening and computer-assisted drug design – synthesis –, etc. [25, 28, 29, 34]. Initially, 14 compounds (mostly macrocyclic pyridinophanes and their acyclic analogues) have been investigated for antiherpetic activity [29]. At the present stage [25], after the several QSAR convolutions, 37 compounds were divided between training and test sets (26 and 11 compounds respectively) and the set of QSAR models with different adequacy levels (2D, 4D, and 3D) has been obtained as a result of the investigations. All the obtained QSAR models were well fitted, robust, predictive ($R^2 = 0.82$–0.90, $Q^2 = 0.60$–0.65, $R^2{}_{\text{test}} = 0.70$–$0.78$), and have a defined DA and clear mechanistic interpretation. For the 3D-QSAR investigations the set of "productive" conformers has been used. They were determined as the most active from the results of 4D-QSAR modeling.

All the models developed (2D–4D) indicate the impact of hydrophobic (~50%) and electrostatic (~20%) factors on the variation of antiherpetic activity. The strong promotion of antiherpetic activity by aminoethylene fragments was revealed. It was also discovered that an important factor for the HSV inhibition is the presence of an amino group connected to aliphatic fragment. A tendency of antiviral activity increasing with the strengthening of acceptor properties of compound's aromatic rings was revealed. This information was used for the design of potent antiherpetic agent **1** (Table 5-4). The use of SiRMS allows to progress in searching for new antiherpetic agents starting from macrocyclic pyridinophanes [29] and finishing in symmetric piperazine containing macroheterocycle 1,4,7,10,13,16,19,22,25,28-Decaaza-tricyclo[26.2.2.2*13,16*]tetratriacontane (**1**).

5.4.1.3. [(Biphenyloxy)propyl]isoxazole Derivatives – Human Rhinovirus 2 Replication Inhibitors[5] *(2D)*

QSAR analysis of antiviral activity of [(biphenyloxy)propyl]isoxazole derivatives



was developed using HiT QSAR based on SiRMS to reveal chemico-biological interactions governing their activities as well as their probable mode of action, and to design new compounds with a strong antiviral activity [11]. The common logarithms

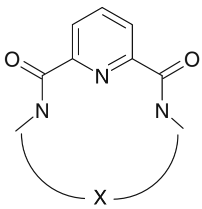
[5] The authors express sincere gratitude to Dr. M. Schmidtke, Prof. P. Wutzler, Dr. V. Makarov, Dr. O. Riabova and Ms. Volineckaya for fruitful cooperation that made possible the development of this task.

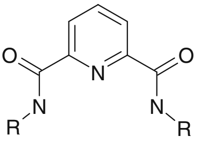
of 50% cytotoxic concentration (CC_{50}) in HeLa cells, the 50% inhibitory concentration (IC_{50}) against human rhinovirus 2 (HRV-2), and the selectivity index (SI = CC_{50}/IC_{50}) of [(biphenyloxy)propyl]isoxazole derivatives were used as cytotoxicity, antiviral activity, and selectivity assessments, respectively. The set of molecules consists of 18 compounds including pleconaril as a reference compound. They have not been divided into training and test sets because of the low number of compounds (i.e., the structural information contained in each molecule in this case is unique and useful). The statistic characteristics of the resulting 2D-QSAR models are quite satisfactory ($R^2 = 0.84$–0.92; $Q^2 = 0.70$–0.87) for the prediction of CC_{50}, IC_{50}, and SI values and permit the virtual screening and molecular design of new compounds with high anti-HRV-2 activity. The results indicate the high influence of atom′s individuality on all the investigated properties (~40%), electrostatic factors on selectivity (~50%), where these factors along with atom individuality play the determining role, and hydrophobic interactions on the antiviral activity (~40%). The presence of terminal 5-trifluoromethyl-1,2,4-oxadiazole and *p*-fluorophenyl fragments in a molecule leads to strong enhancement of its useful properties, i.e., increase of activity toward HRV-2 as well as selectivity and decrease of cytotoxicity. An additional terminal aromatic ring – naphthalene or phenyl – strongly reduces activity toward HRV-2 and, to a lesser degree, SI. The virtual screening and molecular design of new well-tolerated compounds with strong anti-HRV-2 activity has been performed on the basis of QSAR results. Three different DA approaches (DA rectangle and ellipsoid as well as leverage) give nearly the same results for each QSAR model and additionally allow for the estimation of the quality of the prediction for all designed compounds. A hypothesis to the effect that external benzene substituent must have negative electrostatic potential and definite length L (approximately 5.5–5.6 A) to possess strong antiviral activity has been suggested. Most probably, the fluorine atom in the *para*-position of terminal aromatic ring (compounds **2-4**, Table 5-4) is quite complementary ($L = 5.59$ A) to the receptor cavity for such an interaction. It is necessary to note that pleconaril ($L = 5.54$ A) completely satisfies the indicated criteria. In the case of nitroaromatics, the accumulation of nitro groups in the region of receptor cavity will lead to strengthening of electrostatic interactions with the biological target and, therefore, to an increase in activity.

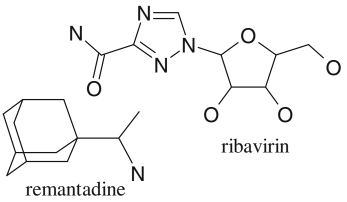
Several new compounds have been designed computationally and predicted as having high activity and selectivity. Three of them (**2–4**, Table 5-4) were synthesized. Subsequent experimental testing revealed a strong coincidence between experimental and predicted anti-HRV-2 activity and SI. Compounds **2–4** are similar in their cytotoxicity level to plecanoril, but they are more active and selective.

5.4.1.4. Anti-influenza Activity of Macrocyclic Pyridinophanes[4] (2D–4D)

All the advantages of HiT QSAR were demonstrated during the investigation of anti-influenza activity on the data set possessing structural variety: different macrocyclic pyridinophanes, their acyclic analogues, and well-known antiviral agents (deiteforin, remantadine, ribavirin, ambenum, and others) [12, 29]:







remantadine

ribavirin

Anti-influenza activity (virus A/Hong Kong/1/68 (H3N2)) was expressed in lgTID$_{50}$ and reflected the suppression of viral replication in "experimental" samples in comparison with "controls." The structures investigated were divided between training and test sets (25 and 6 compounds, respectively).

In accordance with the hierarchical principles of the approach offered, the QSAR analysis was solved sequentially on the 2D, 4D, and 3D levels.[6] The set of QSAR models with different adequacy levels (2D, 4D, and 3D) was obtained as a result of the investigations. All the obtained QSAR models were well fitted, robust, predictive ($R^2 = 0.94$–0.98, $Q^2 = 0.85$–0.95, and $R^2{}_{\text{test}} = 0.98$–$0.99$)[7], and have defined DA and clear mechanistic interpretation. For 3D-QSAR investigations the set of "productive" conformers has been used. They were determined as the most active from the results of 4D-QSAR modeling. The results indicate the great impact of atom individuality on the variation of anti-influenza activity (37–50%). Hydrophobic/hydrophilic and electrostatic interactions also played an important role (15–22%). The shape of molecules (4D and 3D models) also effects anti-influenza activity but has the smallest influence (11 and 16%, respectively). The cylindrical form of molecules ($I_X/I_Y \rightarrow 1$) with small diameters ($I_Y \rightarrow \min$) promotes anti-influenza activity. The molecular fragments governing the change of anti-influenza activity and their average relative influence (Table 5-1) were determined. For example, the presence of oxyethylene or 2-iminomethylphenol fragments promotes antiviral activity and aminoethylene fragments decreases it.

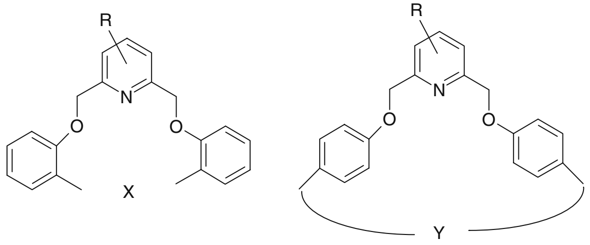
The purposeful design of new molecules **5–7** (Table 5-4) with adjusted activity level was developed by obtained results. The high level of all predicted (all the resulting 2D–4D models show the strong coincidence of predictions) values of anti-influenza activity was confirmed experimentally. Thus, during the QSAR investigations [12, 29] the search for active compounds began from macrocyclic pyridinophanes and finally results in benzene derivatives containing the 2-iminomethyl-phenol fragment (**5–7**, Table 5-4).

[6] In this and antiherpetic research 1D modeling were not performed.

[7] We are aware that these models can approximate not only variation of activity but also variation of experimental errors. The high values of $R^2{}_{\text{test}}$ can be explained by the fact that test compounds are very similar to those in the training set, that there are only few compounds in test set, by high quality of obtained models, by simple good luck or by combination of all mentioned factors.

5.4.2. Anticancer Activity of MacroCyclic Schiff Bases[8] (2D and 4D)

The investigation of influence of the molecular structure of macrocyclic Schiff bases (see structures below) on their anticancer activity has been carried out by means of the 4D-QSAR SiRMS approach [10]. The panel of investigated human malignant tumors includes 60 lines of the following nine cell cultures: leukemia, CNS cancer, prostate cancer, breast cancer, melanoma, non-small cell lung cancer, colon cancer, ovarian cancer, and renal cancer. Anticancer activity was expressed as the percent of the corresponding cell growth. The training set is very structurally dissimilar and consists of 30 macrocyclic pyridinophanes, their analogues, and some other compounds.



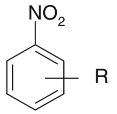
The use of simple topological models generated by EMMA [85] allows the description of the anticancer activity of macrocyclic pyridinophanes (MCP) for only five cell cultures [86]. These studies show that even within the simple topological model it is possible to detect some patterns of the relationship between the structure of MCP and their activity. The consideration of spatial structure improves the situation, but only at the 4D level reliable QSAR models (R^2 = 0.74–0.98; Q^2 = 0.54–0.84) were obtained for all of the investigated cells (except leukemia, where $Q^2 < 0.5$; however, even in this case the designed compound was predicted correctly) and averaged activity (most of lines and cells are highly correlated) that indicate the importance of not the most active or favorable single conformer but the set of interacting conformers within the limits of energy gap of 3 kcal/mol. It was discovered that the presence of the N^1,N^3-dimethylenepropane-1,3-diamine fragment strongly promotes anticancer activity. This fragment was used as a linker between two naphthalen-2-oles that leads to the creation of universal anticancer agent active against all mentioned tumors except prostate cancer. It is necessary to note that the use of SiRMS allow one starting from 12 macrocyclic pyridinophanes [86] in the search for anticancer agents to finally result in symmetric open-chained aromatic compounds connected by above-mentioned linker [10].

[7] The authors express sincere gratitude to Dr. V.P. Lozitsky, Dr. R.N. Lozytska and Dr. A.S. Fedtchouk for fruitful cooperation during the development of this task.

5.4.3. Acute Toxicity of Nitroaromatics

5.4.3.1. Toxicity to Rats[8] *(1D–2D)*

HiT QSAR based on 1D and 2D simplex models and some other approaches for the description of molecular structure have been applied for (i) evaluation of the influence of the characteristics (constitutional and structural) on the toxicity of 28 nitroaromatic compounds (some of them belonging to a widely known class of explosives, see structures below); (ii) prediction of the toxicity of new nitroaromatic derivatives; (iii) analysis of the effects of substitution in nitroaromatic compounds on in vivo toxicity



R = H, F, Cl, OH, NO_2, COOH, CH_3, CH_2Cl

The 50% lethal dose to rats (LD_{50}) has been used to develop the QSAR models based on simplex representation of molecular structure. The preliminary 1D-QSAR results show that even the information on the composition of molecules reveals the main characteristics for the variations in toxicity [87].

A novel 1D-QSAR approach that allows for the analysis of the non-additive effects of molecular fragments on toxicity has been proposed [87]. The necessity of the consideration of substituents' impact for the development of adequate QSAR models of nitroaromatics' toxicity was demonstrated.

The statistic characteristics for all the 1D-QSAR models developed, with the exception of the additive models, were quite satisfactory (R^2 = 0.81–0.92; Q^2 = 0.64–0.83; $R^2{}_{\text{test}}$ = 0.84–0.87). Successful performance of such models is due to their non-additivity, i.e., the possibility of taking into account the mutual influence of substituents in a benzene ring which governs variations in toxicity and could be mediated through the different C–H fragments of the ring.

The passage to 2D level, i.e., consideration of topology, allows for the improvement of the quality of the obtained QSAR models (R^2 = 0.96–0.98; Q^2 = 0.91–0.93; $R^2{}_{\text{test}}$ = 0.89–0.92) to predict the activity for 41 novel compounds designed by the application of new combinations of substituents represented in the training set [37]. The comprehensive analysis of variations in toxicity as a function of the position and nature of the substituent was performed. Among the contributions analyzed in this work are the electrostatic, hydrophobic, and van der Waals interactions of toxicants to biological targets. Molecular fragments that promote and interfere with toxicity were defined on the basis of models obtained. In particular, it was found that in most cases, insertion of fluorine and hydroxyl groups into nitroaromatics increases toxicity, whereas insertion of a methyl group has the opposite effect. The influence

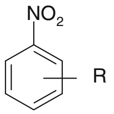
[8] The authors express sincere gratitude to Prof. J. Leszczynski, Dr. L. Gorb and Dr. M. Quasim for fruitful cooperation during the development of this task.

of chlorine on toxicity is ambiguous. Insertion of chlorine at the *ortho*-position to the nitro group leads to substantial increase in toxicity, whereas the second chlorine atom (at the *para*-position to the first) results in a considerable decrease in toxicity. The mutual influence of substituents in the benzene ring is substantially non-additive and plays a crucial role regarding toxicity. The influence of different substituents on toxicity can be mediated via different C–H fragments of the aromatic ring.

The correspondence between observed and predicted toxicity obtained by the 1D and 2D models was good. The single models obtained were summarized in the most adequate consensus model that allows for an improved accuracy of toxicity prediction and demonstrate its ability to be used as a virtual screening tool.

5.4.3.2. Toxicity to Tetrahymena Pyriformis[9] *(2D)*

The present study applies HiT QSAR to evaluate the influence of the structure of 95 various nitroaromatic compounds (including some widely known explosives, see structures below) to the toxicity to the ciliate *T. pyriformis* (QSTR – quantitative structure–toxicity relationship); for the virtual screening of toxicity of new nitroaromatic derivatives; analysis of the characteristics of the substituents in nitroaromatic compounds as to their influence on toxicity.



R = H, F, Cl, OH, NO_2,
COOH, CH_3, OAlc,
CHO, CN, NH2, etc.

The negative logarithm of the 50% inhibition growth concentration (IGC_{50}) was used to develop 2D simplex QSTR models.

During the first part of the work the whole initial set of compounds was divided into three overlapping sets depending on the possible mechanism of action [88]. The 2D-QSTR PLS models obtained were quite satisfactory ($R^2 = 0.84$–0.95; $Q^2 = 0.68$–0.86). The predictive ability of the QSTR models was confirmed through the use of three different test sets (maximal similarity with training set, also minimal one and random choice, taking into account toxicity range only) for any obtained model ($R^2_{\text{test}} = 0.57$–0.85).

The initial division into different sets was confirmed by the QSTR analysis, i.e., the models developed for structures with one mechanism (e.g., redox cyclers) cannot satisfactorily predict the others (e.g., those participating in nucleophilic attack). However, the reliable predictive model can be obtained for all the compounds, regardless of mechanism, when structures of different modes of action are sufficiently represented in the training set.

In addition, the classification and regression trees (CRT) algorithm has been used to obtain models that can predict possible mechanism of action. The quality of the

[9] The authors express sincere gratitude to Prof. J. Leszczynski, Dr. L. Gorb, Dr. M. Quasim and Prof. A. Tropsha for fruitful cooperation during the development of this task.

CRT models obtained is also quite good. The final models had only 15–20% misclassification errors. The obtained models have correctly predicted mechanism of action for compounds of the test set (76–81%).

The comparative analysis of similarity/difference of all nine selected QSAR models has been carried out using the correlation coefficient and Euclidean distance between the sets of toxicity predicted values. It has been shown that all of them are quite close between themselves and the vector of observed activity values. Hence, *T. pyriformis* toxicity by nitroaromatic compounds is complicated and multifactorial process where, most probably, factors determining penetration and delivery of toxicant to biological target play the most important role. Reactivity of nitroaromatics, seemingly, only has an auxiliary role. This was confirmed by the absence of any correlation between toxicity and Hammett constants of substituents. In this regard, the difference in the mechanisms of toxicant interaction with biomolecules (reactions of nucleophylic substitution or radical reduction of nitro group) is important but do not determine for the value of its toxicity.

Molecular fragments that promote and interfere with toxicity were defined using the interpretation of the PLS models obtained. For example, oxibutane and aminophenyl substituents promote the toxicity of nitroaromatics to *T. pyriformis* but carboxyl groups interfere with toxicity. It was also shown that substituent interference in the benzene ring plays the determining role for toxicity. Contributions of the substituents to toxicity are substantially non-additive. Substituents interference effects the activation of aromatic C–H fragments with regard to toxicity.

The structural factors of nitroaromatics which characterize their hydrophobicity and ability to form electrostatic interactions are the most important for the toxic action of the compounds investigated; local structural characteristics (presence of one or other fragments) are more important than integral (whole-molecule) ones.

All the nine selected models were used for consensus predictions of toxicity of an external test set which consists of 63 nitroaromatics. PLS models based on compounds from one mechanism of action were used for consensus predictions only in the case when the CRT model was able to predict such a mechanism. Thus, the predictivity of the consensus model on the external test set was quite satisfactory ($R^2{}_{\text{test}} = 0.64$).

5.4.4. AChE Inhibition[10] (2.5D, Double 2.5D, and 3D)

HiT QSAR has been used for the consensus QSAR analysis of AChE inhibition by various organophosphate compounds. SiRMS and LM QSAR approaches have been used for descriptor generation. Different chiral organophosphates represented by their (R)- and (S)-isomers, racemic mixtures, and achiral structures (totally 42 points) have been investigated. A successful consensus model ($R^2 = 0.978$) based on 14 best QSAR models ($R^2 = 0.91$–0.99; $Q^2 = 0.86$–0.98; $R^2{}_{\text{test}} = 0.82$–$0.97$),

[10] The authors express sincere gratitude to Prof. J. Leszczynski, Dr. L. Gorb and Dr. J. Wang for fruitful cooperation during the development of this task.

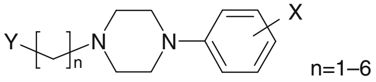
obtained using different QSAR approaches and training sets for several levels and methods of molecular structure representation (2.5D, double 2.5D, and 3D), was used for the prediction of AChE inhibition of new compounds. The trend established on the training set compounds [(S)-isomers are more active than (R)-ones] applies to all new predicted structures.

Atom individuality (including stereochemistry of the chiral surroundings of the asymmetric phosphorus atom) plays the determining role in the variation of activity and is followed by the dispersion and electrostatic characteristics of the OPs. The molecular fragments promoting or interfering with the activities investigated were determined. Identical fragments in the achiral compounds have smaller contributions to activity in comparison with their role in chiral molecules. The influence of phosphorus on the AChE inhibition has a wide range of variation and is very dependent on its surroundings. The substitution of oxygen in $\geq P = O$ by sulfur leads to decreasing AChE inhibition. The presence of the 2-sulphanylpropane fragment facilitates a decrease in activity. Oxyme-containing fragments are actively promoting with activity. The most active predicted compound (2-[(E)-({[cyano(cyclopentyloxy)phosphoryl]oxy}imino)methyl]-1-methylpyridinium) contains oxyme and cyclopentyl parts and is more toxic than oxyme-containing OPs from the training set.

It was also shown in the given work that the topological models of molecular structure (2.5D and double 2.5D) with the identification of stereochemical center of investigated compounds allow for the description of the OPs' ability to inhibit AChE.

5.4.5. 5-HT$_{1A}$ Affinity (1D–4D)[11]

This work was devoted to the analysis of the influence of the structure of *N*-alkyl-*N′*-arylpiperazine derivatives (see structures below) on their affinity for the 5-HT$_{1A}$ receptors (5-HT$_{1A}$R).



Several PLS and MLR models have been obtained for the training set containing 42 ligands of 5-HT$_{1A}$R represented on the 1D–4D levels by SiRMS [32]. All the models obtained have acceptable statistical characteristics ($R^2 = 0.71$–0.96, $Q^2 = 0.66$–0.88). There is improvement in the models from 1D → 2D → 4D → 3D. Molecular fragments which have an influence on the affinity for 5-HT$_{1A}$R have been identified. Analysis of the spatial structure of "productive" conformers determined according to 4D-QSAR model shows considerable similarity to the existing pharmacophore models [89–91] and has allowed for improvement.

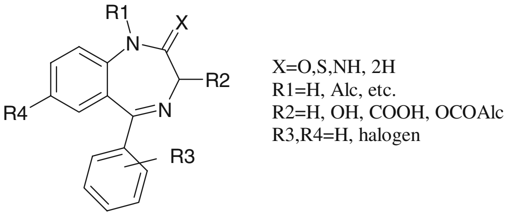
[11] The authors express sincere gratitude to Academician S.A. Andronati and Dr. S.Yu. Makan for fruitful cooperation during the development of this task.

The 2D-QSAR classification task has been solved using the PLS and CRT methods for the set of 364 ligands of 5-HT$_{1A}$R (284 in the training set and 62 in the test set) [92]. The PLS model showed a 65% accuracy for the prediction of test set compounds and the CRT model – 74%. The results of these models have a considerable correspondence between each other that additionally confirmed their validity. It has been shown that, in general, a polymethylene chain comprising three or fewer CH_2 groups has a negative influence on affinity for 5-HT$_{1A}$R and a chain comprising four or more CH_2 groups has a positive influence. Electron-donating substituents (*o*–OCH_3, *o*–OH, *o*–Cl) at the *ortho*-position of phenyl ring strongly promoted affinity. A 2,3-dihydrobenzodioxin-5-yl residue has a similar influence on affinity. Electron-accepting substituents (*m*-CF_3) in phenyl have high affinity. Electron-accepting substituents at the *para*-position of the phenyl ring (*p*-NO_2, *p*-F) have a stronger negative influence on affinity to 5-HT$_{1A}$R than electron-donating ones (*p*–OCH_3). The following conclusions have been made about the influence of the terminal fragments (substituents of *N*-alkyl group) on affinity. Saturated polycyclic fragments and small aromatic residues demonstrated positive influence on affinity and larger aromatic fragments show a negative effect. According to the following analysis, the optimal van der Waals volume for the terminal moiety must be approximately 500 Å^3 or less.

Molecular design and virtual screening of new potential ligands of 5-HT$_{1A}$R has been developed on the basis of the obtained results. Several most promising compounds have been chosen for subsequent investigations, two of them are represented in Table 5-4 (**8** and **9**).

5.4.6. Pharmacokinetic Properties of Substituted Benzodiazepines (2D)

The influence of the structure of substituted benzodiazepines (27 compounds, see below)[12] on the variation of their pharmacokinetic properties including bioavailability, semi-elimination period, clearance, and volume of distribution in the organism of man has been studied [94].



Simplex descriptors in addition to some integral parameters generated by the Dragon software [93] were used for the development of statistic models.

[12] The authors express sincere gratitude to Dr I.Yu. Borisyuk and Acad. N.Ya. Golovenko for a fruitful collaboration.

Reasonably adequate quantitative "structure-pharmacokinetic properties" relationships were obtained using the PLS and MLR statistical approaches ($R^2 = 0.91–0.95$, $Q^2 = 0.81–0.94$) [94]. Structural factors affecting the change of pharmacokinetic properties of substituted benzodiazepines were revealed on the basis of the obtained models.

Bioavailability. Although there is no correlation between absolute bioavailability (F) and lipophilicity ($R \approx 0$), the trend of increasing of molecular fragments' contribution to common bioavailability alongside with increasing of its lipophilicity is observed quite clearly. This trend is the most evident in case of aromatic fragments. Pentamerous aromatic heterocycles have the greatest influence on bioavailability.

Thus, the presence of benzene rings in a molecule increases its bioavailability in a series of substituted benzodiazepines and substitution on the aromatic rings leads to a decrease in bioavailability. Also one can note that the more oxygen atoms in a molecule, the lower the bioavailability. It has been determined that the oxygen atoms are hydrogen bond acceptors. This is in agreement with Lipinski's "rule of five" [24], whereby good bioavailability is observed when the drug corresponds to the following physico-chemical characteristics: molecular weight < 500; $\log P \leq 5$; number of groups – proton donors ≤ 5; number of groups – proton acceptors ≤ 10.

Clearance. For clearance (Cl) of the investigated series, the trend is opposite to that for bioavailability. Thus, the presence of H-donors in a molecule, substitution in aromatic rings as well as an increase of molecule saturation leads to an increase in clearance.

Time of semi-elimination. The influence of structural fragments on the variation of the time of semi-elimination is similar to that described for bioavailability. Thus, all lipophilic aromatic fragments have high values for increasing semi-elimination time.

Volume of distribution. During the analysis of the influence of structure of benzodiazepines on their volume of distribution, the same trends as for clearance were revealed. Thus, refraction (electronic polarizability) increases the volume of distribution and high aromaticity and hydrophilicity decrease it.

The resulting PLS models have been used for the development of virtual screening of pharmacokinetic properties of novel compounds belonging to bezdiazepines family [94].

5.4.7. Catalytic Activity of Crown Ethers[13] (3D)

HiT QSAR was applied to develop the QSPR analysis of the phase-transfer catalytic properties of crown ethers in the reaction of benzyl alcohol oxidation by potassium chlorochromate:

$$3PhCH_2OH + 2KCrO_3Cl \xrightarrow[CH_2Cl_2]{\text{Crown Ether}} 3PhCHO + 2KCl + Cr_2O_3 + 3H_2O \quad (5\text{-}21)$$

[13] The authors express sincere gratitude to Prof. G.L. Kamalov, Dr. S.A. Kotlyar and Dr. G.N. Chuprin for fruitful cooperation during the development of this task.

The objects of the investigation were 66 structurally dissimilar crown ethers, their acyclic analogues and related compounds. The compounds were not divided into training and test sets. Catalytic activity was expressed as the percentage of conversion acceleration.

The distinctive feature of this study is the absence of any reliable relationship between topololgical (2D) structure of crown ethers and their catalytic properties. At the 4D level a not very robust ($Q^2 = 0.46$) relationship was obtained and, only at the 3D level, after the selection of the conformations with the most acceptable formation of complexes with potassium, was a reliable model formed ($R^2 = 0.87$; $Q^2 = 0.66$). Alongside the positive effect of biphenyl and diphenyloxide fragments on catalytic activity of the investigated compounds, the slight preference of "transoid" on *cis*-conformations of crown ethers containing mentioned fragments was shown. The undesirability of the cyclohexyl fragment was determined as well as the certain limits of crown ether dentacy (4–8). These findings, as well as the predominant role of electrostatic factors in investigated process (~50%), correspond to the known mechanisms of catalytic action of the crown ethers. Two potent catalysts **10** and **11** (Table 5-4) were designed and introduced as a result of the QSPR analysis.

5.4.8. Aqueous Solubility[14] (2D)

This work was devoted to the development of new QSPR equations which will accurately predict S_w for compounds of interest to the US Army (explosives and their metabolites) using the SiRMS approach with subsequent validation of the obtained results using a broad spectrum of available experimentally determined data.

The series of the different QSPR models that supplement each other excludes the application of additive schemes and provides a solution to the problems of virtual screening, the evaluation of influence of the structural factors on solubility, etc., have been developed and used with the consensus part of hierarchical QSAR technology.

The training set consists of 135 compounds and the test set includes 156 compounds. Two-dimensional simplex and derived from them Fourier integral descriptors have been used to obtain the set of well-fitted, robust, and predictive (internally and externally) QSPR models ($R^2 = 0.90$–0.95; $Q^2 = 0.85$–0.91; $R^2_{\text{test}} = 0.78$–0.87). External validation using four different test sets also reflects a high level of predictivity ($R^2_{\text{test1}} = 0.7$–0.87; $R^2_{\text{test2}} = 0.82$–$0.88$; $R^2_{\text{test3}} = 0.66$–$0.76$; $R^2_{\text{test4}} = 0.86$–$0.91$). Here test_1 – mixed set of 27 compounds from different chemical classes; test_2 – set of 100 pesticides; test_3 – McFarland set of 18 drugs and pesticides; and test_4 – Arthursson set of 11 drugs. When all 156 compounds have been united in one external set, $R^2_{\text{testU}} = 0.87$ has been reached. The application of DA estimated by the two different approaches (Ellipsoid DA and Williams Plot) leads to a loss of coverage but does not improve the quality of the prediction ($R^2_{\text{testU}} = 0.87$).

[14] The authors express sincere gratitude to Prof. J. Leszczynski, Dr. L. Gorb and Dr. M. Quasim for fruitful cooperation during the development of this task.

Special attention was paid to the accurate prediction of the solubility of polynitro military compounds, e.g., HMX, RDX, CL-20. Comparison of the solubility values for such compounds predicted by our QSPR results and EPI SuiteTM and SPARC techniques indicates that both DoD and Environmental Protection Agency will have considerable advantage using the SiRMS models developed here.

5.5. CONCLUSIONS

In summary, it can be concluded that the QSAR technology considered is a universal instrument for the development of effective QSAR models which provide reliable enough virtual screening and targeted molecular design of various compounds with desired properties. This is a result of its hierarchical structure and wide descriptor system.

The comparative analysis of HiT QSAR with the most popular modern QSAR approaches reflects its advantage, especially in predictivity. The efficiency of HiT QSAR was demonstrated on various QSAR/QSPR tasks at different (1D–4D) levels of molecular modeling. HiT QSAR is under permanent development and improvement. Currently the system of descriptors devoted to adequate description of structure of nanomaterials on the basis of carbon polyhedrons (fullerenes, nanotubes, etc.), algorithms of consensus modeling, and procedures for QSAR analysis of complex mixtures are under development. The technology developed has been realized as a complex of computer programs "HiT QSAR." The trial version is available on request for everyone who is interested in it.

REFERENCES

1. Ooms F (2000) Molecular modeling and computer aided drug design. Examples of their applications in medicinal chemistry. Curr Med Chem 7:141–158
2. Thomas G (2008) Medicinal chemistry: An introduction, 2nd edn John Wiley & Sons Inc, New York
3. Artemenko AG, Muratov EN, Kuz'min VE et al. (2007) Identification of individual structural fragments of N,N′-(bis-5-nitropyrimidyl)dispirotripiperazine derivatives for cytotoxicity and antiherpetic activity allows the prediction of new highly active compounds. J Antimicrob Chemother 60:68–77
4. Bailey TR, Diana GD, Kowalczyk PJ et al. (1992) Antirhinoviral activity of heterocyclic analogs of win 54954. J Med Chem 35:4628–4633
5. Butina D, Gola JMR (2004) Modeling aqueous solubility. J Chem Inf Comp Sci 43:837–841
6. de Jonge MR, Koymans LM, Vinkers HM et al. (2005) Structure based activity prediction of HIV-1 reverse transcriptase inhibitors. J Med Chem 48:2176–2183
7. Jenssen H, Gutteberg TJ, Lejon T (2005) Modelling of anti-HSV activity of lactoferricin analogues using amino acid descriptors. J Pept Sci 11:97–103
8. Kovatcheva A, Golbraikh A, Oloff S et al. (2004) Combinatorial QSAR of ambergris fragrance compounds. J Chem Inf Comp Sci 44:582–595
9. Kubinyi H (1990) Quantitative structure–activity relationships (QSAR) and molecular modeling in cancer research. J Cancer Res Clin Oncol 116:529–537
10. Kuz'min VE, Artemenko AG, Lozitska RN et al. (2005) Investigation of anticancer activity of macrocyclic Schiff bases by means of 4D-QSAR based on simplex representation of molecular structure. SAR QSAR Environ Res 16:219–230

11. Kuz'min VE, Artemenko AG, Muratov EN et al. (2007) Quantitative structure–activity relationship studies of [(biphenyloxy)propyl]isoxazole derivatives – human rhinovirus 2 replication inhibitors. J Med Chem 50:4205–4213
12. Muratov EN, Artemenko AG, Kuz'min VE et al. (2005) Investigation of anti-influenza activity using hierarchic QSAR technology on the base of simplex representation of molecular structure. Antivir Res 65:A62–A63
13. Verma RP, Hansch C (2006) Chemical toxicity on HeLa cells. Curr Med Chem 13:423–448
14. Zhang S, Golbraikh A, Tropsha A (2006) The development of quantitative structure–binding affinity relationship (QSBR) models based on novel geometrical chemical descriptors of the protein–ligand interfaces. J Med Chem 49:2713–2724
15. Selassie CD (2003) History of QSAR. In: Abraham DJ (ed) Burger's medicinal chemistry and drug discovery. Wiley, New York, p 960
16. Cramer RD, Patterson DI, Bunce JD (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape binding to carrier proteins. J Am Chem Soc 110:5959–5967
17. Doweyko AM (1988) The hypothetical active site lattice. An approach to modeling active sites from data on inhibitor molecules. J Math Chem 31:1396–1406
18. Klebe G, Abraham U, Mietzner T (1994) Molecular similarity indeces in comparative anaysis (CoMSIA) of molecules to correlate and predict their biological activity. J Med Chem 37:4130–4146
19. Kuz'min VE, Artemenko AG, Kovdienko NA et al. (2000) Lattice model for QSAR studies. J Mol Model 6:517–526
20. Seel M, Turner DB, Wilett P (1999) HQSAR – a highly predictive QSAR technique based on molecular holograms. QSAR 18:245–252
21. Pavan M, Consonni V, Gramatica P et al. (2006) New QSAR modelling approach based on ranking models by genetic algorithms – variable subset selection (GA-VSS). In: Brüggeman R, Carlsen L (eds) Partial order in environmental sciences and chemistry. Springer Berlin Heidelberg, Berlin, pp 181–217
22. Kuz'min VE, Muratov EN, Artemenko AG et al. (2008) The effect of nitroaromatics composition on theirs toxicity in vivo. 1D QSAR research. Chemosphere 72:1373–1380
23. Baurin N, Mozziconacci JC, Arnoult E et al. (2004) 2D QSAR consensus prediction for high-throughput virtual screening. An application to COX-2 inhibition modeling and screening of the NCI database. J Chem Inf Model 44:276–285
24. Vedani A, Dobler M (2000) Multi-dimensional QSAR in drug design. Progress in Drug Res 55: 107–135
25. Artemenko A, Kuz'min V, Muratov E et al. (2007) Molecular design of active antiherpetic compounds using hierarchic QSAR technology. Antivir Res 74:A76
26. Artemenko A, Muratov E, Kuz'min V et al. (2006) Molecular design of novel antimicrobial agents on the base of 4-thiazolidone derivatives. Clin Microbiol Infec 12:1557
27. Artemenko A, Muratov E, Kuz'min V et al. (2006) Influence of artifical ribonucleases structure on their anti-HIV activity. Antivir Res 70:A43
28. Artemenko AG, Kuz'min VE, Muratov EN et al. (2005) Investigation of antiherpetic activity using hierarchic QSAR technology on the base of simplex representation of molecular structure. Antivir Res 65:A77
29. Kuz'min VE, Artemenko AG, Lozitsky VP et al. (2002) The analysis of structure-anticancer and antiviral activity relationships for macrocyclic pyridinophanes and their analogues on the basis of 4D QSAR models (simplex representation of molecular structure). Acta Biochim Polon 49: 157–168
30. Kuz'min VE, Artemenko AG, Muratov EN et al. (2007) QSAR analysis of anti-coxsackievirus B3 nancy activity of 2-amino-3-nitropyrazole[1,5-α]pyrimidines by means of simplex approach. Antivir Res 74:A49–A50

31. Kuz'min VE, Artemenko AG, Muratov EN et al. (2005) The hierarchical QSAR technology for effective virtual screening and molecular design of the promising antiviral compounds. Antivir Res 65:A70–A71
32. Kuz'min VE, Artemenko AG, Polischuk PG et al. (2005) Hierarchic system of QSAR models (1D-4D) on the base of simplex representation of molecular structure. J Mol Model 11:457–467
33. Muratov E, Artemenko A, Kuz'min V et al. (2006) Computational design of the new antimicrobials based on the substituted crown ethers. Clin Microbiol Infec 12:1558
34. Muratov EN (2004) Quantitative evaluation of the structural factors influence on the properties of nitrogen-, oxygen- and sulfur-containing macroheterocycles. National Academy of Sciences of Ukraine, A.V. Bogatsky Physical-Chemical Institute, Odessa, p 202
35. Muratov EN, Kuz'min VE, Artemenko AG et al. (2006) QSAR studies demonstrate the influence of structure of [(biphenyloxy)propyl]isoxazole derivatives on inhibition of coxsackievirus B3 (CVB3) replication. Antivir Res 70:A77
36. Kuz'min VE, Artemenko AG, Muratov EN (2008) Hierarchical QSAR technology on the base of simplex representation of molecular structure. J Comp Aid Mol Des 22:403–421
37. Kuz'min VE, Muratov EN, Artemenko AG et al. (2008) The effects of characteristics of substituents on toxicity of the nitroaromatics: HiT QSAR study. J Comp Aid Mol Des 22:747–759. doi:10.1007/s10822-10008-19211-x
38. QSAR, Expert, Group (2004) The report from the expert group on (quantitative) structure–activity relationships [(Q)SARs] on the principles for the validation of (Q)SARs. In: OECD series on testing and assessment. Organisation for Economic Co-operation and Development, Paris, p 206
39. Kuz'min VE (1995) About homo- and heterochirality of dissymetrical tetrahedrons (chiral simplexes). Stereochemical tunneling. Zh Strucur Khim (in Russ) 36:873–878
40. Jolly WL, Perry WB (1973) Estimation of atomic charges by an electronegativity equalization procedure calibration with core binding energies. J Am Chem Soc 95:5442–5450
41. Wang R, Fu Y, Lai L (1997) A new atom-additive method for calculating partition coefficients. J Chem Inf Comp Sci 37:615–621
42. Ioffe BV (1983) Chemistry refractometric methods, 3 ed. Himiya, Leningrad
43. Cahn RS, Ingold CK, Prelog V (1966) Specification of molecular chirality. Angew Chem Int Ed 5:385–415
44. Burkert U, Allinger N (1982) Molecular mechanics. ACS Publication, Washington, DC
45. Hodges G, Roberts DW, Marshall SJ et al. (2006) Defining the toxic mode of action of ester sulphonates using the joint toxicity of mixtures. Chemosphere 64:17–25
46. Kuz'min VE, Muratov EN, Artemenko AG et al. (2009) Consensus QSAR modeling of phosphor-containing chiral AChE inhibitors. J Comp Aid Mol Des 28:664–677
47. Hyperchem 7.5 software. Hypercube, Inc. 1115 NW 4th Street, Gainesville, FL 32601, USA
48. Kuz'min VE, Artemenko AG, Kovdienko NA et al. (1999) Lattice models of molecules for solution of QSAR tasks. Khim-Pharm Zhurn (in Russ) 9:14–20
49. Kuz'min VE, Beresteckaja EL (1983) The program for calculation of atom charges using the method of orbital electronegativities equalization. Zh Struct Khimii (in Russ) 24:187–188
50. Croizet F, Langlois MH, Dubost JP et al. (1990) Lipophilicity force field profile: An expressive visualization of the lipophilicity molecular potential gradient. J Mol Graphics 8:53
51. Artemenko AG, Kovdienko NA, Kuzmin VE et al. (2002) The analysis of "structure-anticancer activity" relationship in a set of macrocyclic pyridinophanes and their acyclic analogues on the basis of lattice model of molecule using fractal parameters. Exp Oncol 24:123–127
52. Lozitsky VP, Kuz'min VE, Artemenko AG et al. (2000) The analysis of structure–anti-influenza relationship on the basis molecular lattice model for macrocyclic piridino-phanes and their analogs. Antivir Res 50:A85

53. Marple SL Jr (1987) Digital spectral analysis with applications. Prentice-Hall Inc., Englewood Cliffs, NJ
54. Kuz'min VE, Trigub LP, Shapiro YE et al. (1995) The parameters of shape of peptide molecules as a descriptors in the QSAR tasks. Zh Struct Khimii (in Russ) 36:509–517
55. Breiman L, Friedman JH, Olshen RA et al. (1984) Classification and regression trees. Wadsworth, Belmont
56. Carhart RE, Smith DH, Venkataraghavan R (1985) Atom pairs as molecular features in structure–activity studies. Definition and application. J Chem Inf Comput Sci 25:64–73
57. Vitiuk NV, Kuz'min VE (1994) Mechanistic models in chemometrics for the analysis of multidimensional data of researches. Analogue of dipole-moments method in the structure(composition)–property relationships analysis. ZhAnalKhimii 49:165–167
58. Ferster E, Renz B (1979) Methoden der Korrelations und Regressionanalyse. Verlag Die Wirtschaft, Berlin
59. Topliss JG, Costello RJ (1972) Chance correlations in structure–activity studies using multiple regression analysis. J Med Chem 15:1066–1068
60. Kubinyi H (1996) Evolutionary variable selection in regression and PLS analyses. J Chemometr 10:119–133
61. Lindgren F, Geladi P, Rannar S et al. (1994) Interactive variable selection (IVS) for PLS. Part 1: Theory and algorithms. J Chemometr 8:349–363
62. Rannar S, Lindgren F, Geladi P et al. (1994) A PLS kernel algorithm for data sets with many variables and fewer objects. Part 1: Theory and algorithm. J Chemometr 8:111–125
63. Rogers D, Hopfinger AJ (1994) Application of genetic function approximation to quantitative structure–activity relationships and quantitative structure–property relationships. J Chem Inf Comp Sci 34:854–866
64. Wold S, Antti H, Lindgren F et al. (1998) Orthogonal signal correction of nearinfrared spectra. Chemometrics Intell Lab Syst 44:175–185
65. Trygg J, Wold S (2002) Orthogonal projections to latent structures (O-PLS). J Chemometr 16: 119–128
66. Cronin MTD, Schultz TW (2003) Pitfalls in QSAR. J Mol Struct (Theochem) 622:39–51
67. Zhang S, Golbraikh A, Oloff S et al. (2006) A novel automated lazy learning QSAR (ALL-QSAR) approach: Method development, applications, and virtual screening of chemical databases using validated ALL-QSAR models. J Chem Inf Model 46:1984–1995
68. Neter J, Kutner MH, Wasseman W et al. (1996) Applied linear statistical models. McGraw-Hill, New York
69. Meloun M, Militku J, Hill M et al. (2002) Crucial problems in regression modelling and their solutions. Analyst 127:433–450
70. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T (2005) QSAR applicability domain estimation by projection of the training set in descriptor space: A review. Altern Lab Anim 33:445–459
71. Östergard PRJ (2002) A fast algorithm for the maximum clique problem. Discrete Appl Math 120:195–205
72. Bodor N, Buchwald P (2000) Soft drug design: General principles and recent applications. Med Res Rev 20:58–101
73. Sutherland JJ, O'Brien LA, Weaver DF (2004) A comparison of methods for modeling quantitative structure–activity relationships. J Med Chem 47:5541–5554
74. Heritage TV, Ferguson AM, Turner DB et al. (1998) EVA: A novel theoretical descriptor for QSAR studies. Persp Drug Disc Des 11:381–398
75. Barnard DL (2006) Current status of anti-picornavirus therapies. Curr Pharm Des 12:1379–1390
76. Patick AK (2006) Rhinovirus chemotherapy. Antivir Res 71:391–396
77. Rotbart HA (2002) Treatment of picornavirus infections. Antivir Res 53:83–98

78. Binford SL, Maldonado F, Brothers MA et al. (2005) Conservation of amino acids in human rhinovirus 3C protease correlates with broad-spectrum antiviral activity of rupintrivir, a novel human rhinovirus 3C protease inhibitor. Antimicrob Agents Chemother 49:619–626
79. Conti C, Mastromarino P, Goldoni P et al. (2005) Synthesis and anti-rhinovirus properties of fluoro-substituted flavonoids. Antivir Chem Chemother 16:267–276
80. Cutri CC, Garozzo A, Siracusa MA et al. (2002) Synthesis of new 3-methylthio-5-aryl-4-isothiazolecarbonitriles with broad antiviral spectrum. Antiviral Res 55:357–368
81. Diana GD, Cutcliffe D, Oglesby RC et al. (1989) Synthesis and structure–activity studies of some disubstituted phenylisoxazoles against human picornavirus. J Med Chem 32:450–455
82. Dragovich PS, Prins TJ, Zhou R et al. (2002) Structure-based design, synthesis, and biological evaluation of irreversible human rhinovirus 3C protease inhibitors. 6. Structure-activity studies of orally bioavailable, 2-pyridone-containing peptidomimetics. J Med Chem 45:1607–1623
83. Gaudernak E, Seipelt J, Triendl A et al. (2002) Antiviral effects of pyrrolidine dithiocarbamate on human rhinoviruses. J Virol 76:6004–6015
84. Kaiser L, Crump CE, Hayden FG (2000) In vitro activity of pleconaril and AG7088 against selected serotypes and clinical isolates of human rhinoviruses. Antiviral Res 47:215–220
85. Suchachev DV, Pivina TS, Shliapochnikov VA et al. (1993) Investigation of quantitative "structure-shock-sensitivity" relationships for organic polynitrous compounds. Dokl RAN (in Russ) 328:50–57
86. Kuz'min VE, Lozitsky VP, Kamalov GL et al. (2000) The analysis of "structure–anticancer activity" relationship in a set of macrocyclic 2,6-bis (2- and 4-formylaryloxymethyl) pyridines Schiff bases. Acta Biochim Polon 47:867–875
87. Kuz'min VE, Muratov EN, Artemenko AG et al. (2008) The effect of nitroaromatics' composition on their toxicity in vivo: Novel, efficient non-additive 1D QSAR analysis. Chemosphere 72(9):1373–1380. doi:10.1016/j.chemosphere.2008.1004.1045
88. Katritzky AR, Oliferenko P, Oliferenko A et al. (2003) Nitrobenzene toxicity: QSAR correlations and mechanistic interpretations. J Phys Org Chem 16:811–817
89. Chilmonczyk Z, Szelejewska-Wozniakowska A, Cybulski J et al. (1997) Conformational flexibility of serotonin$_{1A}$ receptor ligands from crystallographic data. Updated model of the receptor pharmacophore. Archiv der Pharmazie 330:146–160
90. Hibert MF, Gittos MW, Middlemiss DN et al. (1988) Graphics computer-aided receptor mapping as a predictive tool for drug design: Development of potent, selective, and stereospecific ligands for the 5-HTlA receptor. J Med Chem 31:1087–1093
91. Hibert MF, Mcdermott I, Middlemiss DN et al. (1989) Radioligand binding study of a series of 5-HT1A receptor agonists and definition of a steric model of this site. Eur J Med Chem 24:31–37
92. Kuz'min VE, Polischuk PG, Artemenko AG et al. (2008) Quantitative structure–affinity relationship of 5 HT1A receptor ligands by the classification tree method. SAR & QSAR in Env Res 19:213–244
93. Todeschini R, Consonni V (2000) Handbook of molecular descriptors, 1st ed. Wiley-VCH, Weinheim
94. Artemenko AG, Kuz'min VE, Muratov EN et al. (2009) The analysis of influence of benzodiazepine derivatives structure on its pharmacocinetic properties. Khim-Pharm Zhurn 43:36–45 (in Russ)